DISS. ETH NO. 25883

AN RKHS APPROACH TO MODELLING AND INFERENCE FOR SPATIOTEMPORAL GEODETIC DATA WITH APPLICATIONS TO TERRESTRIAL RADAR INTERFEROMETRY

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCES OF ETH ZURICH
(DR. SC. ETH ZURICH)

presented by

JEMIL AVERS BUTT
M.SC. CLAUSTHAL UNIVERSITY OF TECHNOLOGY

BORN ON 29.04.1988
CITIZEN OF GERMANY

accepted on the recommendation of

Prof. Dr. A. Wieser, examiner
Prof. Dr. R. Hanssen, co-examiner
Prof. Dr. J. Teichmann, co-examiner

2019

Abstract

Reproducing kernel Hilbert spaces (RKHS) are interpretable as normed spaces of functions furnished with a probability distribution and as such are especially suitable for modelling and inference of spatial, temporal, or spatiotemporal phenomena exhibiting a large degree of randomness. Phenomena of this type occur frequently in geodesy in the form of for example noise, instrument drifts, or meteorologically induced systematic deviations whose superimposition on the desired information yields the measurements. To facilitate RKHS-based signal separation and thereby solutions to the above mentioned geodetic problems of splitting data into components of different origin, the correspondence between elements of an RKHS and the random variables comprising stochastic processes is investigated. RKHS can be operated on by abstract algebraic operations like linear transformations, direct sums, and tensor products. Due to the dualistic nature of RKHS as isomorphic to spaces of random variables, these operations have probabilistic analogues that amount to different assumptions and constructive guarantees concerning the decomposability of the involved stochastic processes that are both interesting from a theoretical perspective and have practically relevant implications in terms of numerical stability and decreased computational loads.

The set of RKHS is in one-to-one correspondence with the set of positive definite kernels that in turn completely determine a unique RKHS and can be interpreted as a comprehensive description of the associated stochastic processes correlation structure. The choice of an appropriate kernel is therefore essential to ensure the performance of estimators derived within the RKHS framework. Spectrally decomposing positive definite kernels and their associated compact kernel operators leads not only to insights linking functional calculus and signal separation but also suggests a new stochastic model for reproducing kernels that enables their inference from observational data. Solving the inference problem requires the development of numerical methods that parallel the ones employed in semidefinite programming. The resulting algorithm proves to be extendable to allow for the inclusion of affine constraints making it applicable for problems like variance components estimation. Further geodetically motivated examples of RKHS-based processing are given throughout.

The separation of signal and noise in data originating from terrestrial radar interferometry is a particularly challenging problem due to the spatiotemporal nature of the measurements and highly intricate meteorological effects that exhibit an instationary and topographically affected correlation structure. Following a detailed analysis of the kernels most suitable to model the stochastics of the atmospheric artifacts to be removed, several optimization problems in RKHS are formulated whose solutions are best estimators of the deformations whose quantification is the original purpose of terrestrial radar interferometry. Based on data gathered during a monitoring campaign carried out in the Swiss Alps, the performance of the RKHS-based estimators is evaluated and compared to filtering schemes presented elsewhere in the literature; the latter estimators are shown to be special cases of the RKHS approach.

Zusammenfassung

Hilberträume mit reproduzierendem Kern (RKHS) sind interpretierbar als mit einer Wahrscheinlichkeitsverteilung ausgestattete normierte Räume von Funktionen und als solche insbesondere geeignet für die Modellierung und statistische Inferenz von räumlichen, zeitlichen, oder raumzeitlichen Phänomenen, die einen hohen Grad an Zufälligkeit aufweisen. Phänomene diesen Types treten häufig in der Geodäsie auf; etwa in Form von Rauschen, Instrumentengängen, oder meteorologisch induzierten systematischen Abweichungen, die den zu erhebenden Informationen überlagert sind. Um RKHS-basierte Signaltrennung und dadurch auch eine Lösung der obengenannten geodätischen Probleme betreffend Dekomposition von Daten in Komponenten unterschiedlicher Herkunft zu ermöglichen, wird der Zusammenhang zwischen Elementen eines RKHS und den einen stochastischen Prozess formierenden Zufallsvariablen untersucht. Abstrakte algebraische Operationen wie lineare Transformationen, direkte Summen und Tensorprodukte können auf RKHS angewendet werden und haben aufgrund der dualistischen Natur von RKHS als isomorph zu Räumen von Zufallsvariablen probabilistische Entsprechungen, die sich umwandeln lassen in verschiedene Annahmen und Konstruktionsgarantien betreffend Zerlegbarkeit der involvierten stochastischen Prozesse.

Die Menge der RKHS befindet sich in bijektiver Korrespondenz mit der Menge aller positiv definiter Kerne. Diese wiederum determinieren einen RKHS vollständig und können interpretiert werden als umfassende Beschreibung der Korrelationsstruktur des mit dem RKHS assoziierten stochastischen Prozesses. Die korrekte Wahl des Kernes ist daher essentiell, um die Leistungsfähigkeit der Schätzer zu garantieren, die im Rahmen der RKHS Methode abgeleitet werden. Die Spektralzerlegung positiv definiter Kerne und der mit ihnen assoziierten kompakten Integraloperatoren legt auch ein neues nichtparametrisches stochastisches Modell für reproduzierende Kerne nahe, welches deren Inferenz aus Messdaten erlaubt. Dieses Inferenzproblem zu lösen, erfordert die Entwicklung numerischer Methoden ähnlich derer eingesetzt in der semidefiniten Programmierung. Der sich aus ihnen ergebende Algorithmus erlaubt die Berücksichtigung affiner Beschränkungen, was ihn anwendbar macht für etwa die Varianzkomponentenschätzung. Weitere geodätisch motivierte Beispiele RKHS-basierter Prozessierungsschemata werden ebenfalls präsentiert.

Die Trennung von Signal und Rauschen in Daten stammend aus terrestrischer Radarinterferometrie ist ein herausforderndes Problem insbesondere aufgrund der komplexen meteorologischen Effekte, welche Instationaritäten und topographisch beeinflusste Korrelationsstrukturen aufweisen. Folgend einer Analyse der am besten zur Modellierung der aus den Daten zu filternden atmospärischen Artefakte geeigneten Kerne, werden verschiedene Optimierungsprobleme in RKHS vorgestellt. Deren Lösungen sind beste Schätzer für die Deformationen, deren Quantifizierung der originäre Zweck der Messungen mit terrestrischem Radar ist. Basierend auf Daten, die während einer Monitoringkampagne in den Schweizer Alpen gesammelt wurden, wird die Leistungsfähigkeit der RKHS-basierten Schätzer evaluiert und mit anderen aus der Literatur bekannten Filteransätzen verglichen.

Contents

Preface

As is the case with all carefully chosen titles, the author hopes that ”An RKHS approach to modelling and inference for spatiotemporal geodetic data with applications to terrestrial radar interferometry” gives the reader already an almost complete account of what he or she can expect to encounter during the reading of this monograph. While the expression ”geodetic data” obviously names the object in need of processing strategies and hints at dynamical phenomena possessing a spatial dimension, ”modelling and inference” constitutes the goal of our efforts enabling us to describe stochastically such phenomena and derive information regarding their current states and further progression despite possibly incomplete and only indirect observations. The reproducing kernel Hilbert spaces (RKHS), interpretable as the closure of certain vector spaces of functions with respect to the topology induced by a norm quantifying in some sense the likelihood of its inhabitants, then finally provides us with a framework in which to analyse problems of the aforementioned kind and subsequently also with a means for their systematic treatment.

These functional spaces turn out to be populated by the solutions to differential equations and norm minimization problems that also arise in geodesy when model parameters are to be optimally estimated in the least-squares sense. Consequently the mathematical procedures we employ will initially parallel those found in the set of methods known as adjustment theory. However, as the monograph progresses and due to the different flavour of tasks we face — the objects to be estimated are random themselves and not fixed parameters in a model — the reader can expect a significant departure in mathematics as well as in spirit from what one could consider canonical knowledge in geodesy. As the spotlight is shifted away from parametric models and towards nonparametric representations of functions, questions of a more topological nature related to closure, convergence, continuity, differentiability and spectral decomposability of compact linear operators emerge. They are only sensibly posable in infinite dimensional settings and do not enter the fray in the finite dimensional analysis even though they offer nontrivial conclusions also in that case.

Ultimately, this monograph is an outgrowth of the author’s attempt to deal with the very practical problem of separating pure noise and atmospheric influences from deformations in terrestrial radar interferometry. As such, the underlying motivations for several seemingly abstract constructions have often been concretely founded in the desire to make concise the structure found in real spatiotemporal data and to generalize the methods available in the literature to be applicable to estimation tasks imbued with the complexity and uncertainty inherent in real-world problems. It is the author’s intended goal to not only systematically motivate the theory of optimal estimation in the framework of RKHS but also to demonstrate its usefulness to practically relevant problems. Several examples are given throughout the monograph to help the reader potentially interested in the topic (but lacking the time for a rigorous study) to furnish an intuitive understanding of what some of the theorems actually mean and how the whole theory is to be applied in practice. It is in this spirit, that the monograph at times diverges from the classical Definition-Theorem-Proof format and sacrifices self-containedness for compactness, accessibility and hopefully joy on parts of the reader, who is only assumed to be acquainted with basic linear algebra and a willingness to dive into the worked examples taken from a geodesist's repertoire of routine tasks.

When one's writing is concerned primarily with the application of mathematics, one can hardly claim sole ownership or originality of what is presented. Much is owed to the classical expositions of Helmberg [93] and Riesz and Nagy [164] where functional analysis results are concerned, whereas the texts authored by Bezhaev and Vasilenko [21] and Berlinet and Thomas-Agnan [20] are invaluable RKHS-related resources, whose study is heartily recommended. I want to express my gratitude to my coworkers for interesting discussions and new perspectives. Specifically I want to mention Ephraim Friedli, Zan Gojcic, David Salido, Eugenio Serantoni, Lorenz Schmid, Valens Frangez, and Caifa Zhou who proofread this manuscript and helpfully suggested many improvements that increased correctness and readability. Special thanks go to my direct supervisor Andreas Wieser for his trust, his seemingly bottomless enthusiasm and his guidance during difficult times and to Ramon Hanssen and Josef Teichmann taking time out of their busy schedule to co-supervise this thesis. It is needless to say that some people have greatly influenced the genesis of this text even though they themselves are not scientifically inclined. To my friends and family, and to my parents for their unconditional support, I dedicate this monograph.

Zürich, Spring 2019 Jemil Avers Butt

List of notation

Acronyms and conventions

Acronyms

APS	Atmospheric phase screen	RKHS	Reproducing kernel Hilbert space
BLUE	Best linear unbiased estimator	SAR	Synthetic aperture radar
GNSS	Global navigation satellite systems	SDP	Semidefinite program
KLE	Karhunen-Loewe expansion	SNR	Signal-to-noise ratio
LCA	Locally compact abelian group	TLS	Terrestrial laser scanning
p.d.	Positive definite		

Notational conventions

The following correspondences between letter fonts and classes of mathematical entities are to be understood as a rough guideline that is not to be considered definitive for all special cases.

Lower case letters	indicate functions and elements of Hilbert spaces
Capital letters	indicate operators and sets
Greek lower case letters	indicate scalars and basis functions
Greek upper case letters	indicate matrices containing basis functions
Blackboard-type letters	indicate fields of numbers or subsets thereof
Caligraphy-type letters	indicate Hilbert spaces and linear manifolds
Fraktur-type letters	indicate convex cones

Symbol	Description / Explanation

Special symbols and functions

Symbol	Description / Explanation
$\delta(\cdot,\cdot)$	The Kronecker delta function taking values of 1 if the two input arguments are equal and 0 otherwise. Also often denoted by δ_{kl} when the input variables k, l are discrete.
$\lambda_j, \lambda_{\max}, \lambda_{\min}$	The j-th, maximal, or minimal eigenvalue of an operator C based on its eigenvalue decomposition $C = \sum_{j=1}^{\infty} \lambda_j \varphi_j \otimes \varphi_j^*$.
$\operatorname{supp} f$	The support of f, i.e. those elements x of the domain of f leading to nonzero $f(x)$.
$\operatorname{sign} f$	The signum function $\operatorname{sign} f : T \to \{-1, 1\}$ assigns to a function $f : T \to \mathbb{R}$ and an element $t \in T$ the sign of $f(t)$.
$\min(\cdot,\cdot)$	The function $\min(\cdot,\cdot) : T \times T \to T$ assigns to each pair (t_1, t_2) of elements of an ordered set T the one that is smaller.
$\mathfrak{Re}, \mathfrak{Im}$	The real or imaginary parts of a complex number.
$\{e_j\}_{j=1}^n$	A sequence of elements forming the (canonical Euclidean) basis of a vector space.
$\lim_{k\to\infty} f_k$	The limit $f \in \mathcal{H}$ of the sequence $\{f_k\}_{k=1}^{\infty} \subset \mathcal{H}$ in some Hilbert space $\mathcal{H}$.
$B_\epsilon^V(f)$	The open ϵ-ball in V around $f \in V$.

Hilbert spaces

Symbol	Description / Explanation
$\mathcal{H}, \mathcal{H}_K$	A Hilbert space; if it has a reproducing kernel K this is denoted by attaching the subscript K.
$\{f_k\}_{k\in J}$	Sequences of elements f_k indexed by $k \in J$ for some set $\mathcal{H} \ni f_k$ (typically a Hilbert space).
ℓ^2, L^2	The Hilbert spaces of square summable sequences and of equivalence classes of square integrable functions, respectively.
$\Vert\cdot\Vert_x$	The x-norm where $x = p \in [1,\infty]$ (p-norm) , $x = \mathrm{op}$ (operatornorm) , $x = F$ (Frobenius norm), or $x = HS$ (Hilbert-Schmidt norm).
$\vert\cdot\vert_P$	A seminorm defined with the help of a subspace P annihilated by $\vert\cdot\vert_P$, i.e. $\vert f\vert_P = 0 \Leftrightarrow f \in P$.
$\langle\cdot,\cdot\rangle_{\mathcal{H}}$	The bilinear and positive definite inner product mapping from $\mathcal{H} \times \mathcal{H} \to \mathbb{C}$ for some pre-Hilbert space $\mathcal{H}$.
$\mathcal{M}, \overline{\mathcal{M}}$	A linear manifold $\mathcal{M}$ or its completion.
$\mathcal{H}_1 \,\overline{\boxslash}\mathcal{H}_2$	The Hilbert space $\mathcal{H}_1$ is a subspace of the Hilbert space $\mathcal{H}_2$.
$\overline{\operatorname{span}}$	The closure of the linear span generated by a set of elements of a vector space.
$[f]$	The equivalence class of f under some equivalence relation.
$\dim(\mathcal{H})$	The dimension of a vector space $\mathcal{H}$ as given by the cardinality of any of its bases.

$\mathcal{H}_1 \oplus_{(e,i)} \mathcal{H}_2$	The external (e) or internal (i) direct sum of two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$.
$\mathcal{H}_1 \otimes \mathcal{H}_2$	The tensor product of two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$.
$\mathcal{H}_1^\perp$	The orthogonal complement of $\mathcal{H}_1$ in a larger Hilbert spaces $\mathcal{H}_2$ that includes $\mathcal{H}_1$. Sometimes the alternative notation $\mathcal{H}_2 \ominus \mathcal{H}_1$ is used.
$\mathcal{H}_2/\mathcal{H}_1$	The quotient space of equivalence classes corresponding to elements in $\mathcal{H}_2$ that differ only by an element in $\mathcal{H}_1$.

Operators and operator algebras

$P_{\mathcal{H}_1}$	The orthogonal projection operator mapping elements of a Hilbert space onto its subspace $\mathcal{H}_1$.
$\mathcal{B}(\mathcal{H})$	The Banach algebra of linear operators on a Hilbert space $\mathcal{H}$, as a generalization $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ denotes the space of linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$.
$\mathcal{F}f$	The Fourier transform of $f : T \to \mathbb{C}$ where T is an LCA and $(\mathcal{F}f)(\omega)$ is its value at $\omega \in \hat{T}$.
Δ	The Laplace operator mapping a suitably often differentiable function $f : T \to \mathbb{C}$ to its second derivative. Δ is also the generic symbol for any kind of increment.
tr	The function $\mathrm{tr} : \mathcal{B}(\mathcal{H}) \to \mathbb{C}$ that assigns to a linear operator on a Hilbert space its trace.
$\log$	The natural logarithm of a real number or an operator as defined via functional calculus.
$\det, \lvert \cdot \rvert$	The determinant of a matrix.
A^+	The pseudoinverse of a linear operator A. Coincides with the inverse of A iff it exists.
A^T, A^H, A^*	The transposes, Hermitians, adjoints of the linear operator A.
$I_{\mathcal{H}}$	The identity operator on $\mathcal{H}$ mapping functions $f \in \mathcal{H}$ onto themselves. If no opportunity for confusion arises, the subscript will be suppressed.

Probability and stochastic processes

$E[\cdot]$	The expectation operator mapping random variables to their expected value.
$f \sim \mathcal{Q}$	The quantity f is distributed according to the probability distribution $\mathcal{Q}$, where typically $\mathcal{Q} = \mathcal{N}$ (Normal distribution) or $\mathcal{Q} = \mathcal{W}$ (Wishart distribution).
$\sigma_f(\cdot)$	The abstract spline estimator for the function $f(\cdot)$.
$\mathcal{N}(\mu, \Sigma)$	Multivariate normal distribution with expected vector μ and covariance matrix Σ.
$\mathcal{W}_m(q, \Lambda)$	Wishart distribution of positive semidefinite matrices with scale matrix Λ, dimension m and q degrees of freedom.
$A \coprod B \vert C$	The event or random variable A is conditionally independent of B given that C is known.
$X^\cdot_\cdot$	A set of random variables $X^\cdot_t : \Omega \ni \omega \mapsto X^\omega_t \in \mathbb{C}$ indexed by elements $t \in T$. Can be equivalently be interpreted as a function $T \to \mathbb{C}$.

Spectral theory and kernel inference

T	The index set forming the domain of functions. Interpreted as time in one-dimensional cases.
$K(\cdot, \cdot)$	A positive definite kernel of two variables (typically) defined on an index set T.
C_K	The covariance operator as given by integration against the kernel $K(\cdot, \cdot)$.
$S = U\Lambda U^*$	The spectral decomposition of a selfadjoint operator. Λ is diagonal and U is unitary.
$\sigma(X)$	The spectrum of X where X may either be a single operator A or a whole operatoralgebra $\mathcal{A}$.
$\hat{T}$	The dual group of the LCA group T.
$f * g$	The convolution of two functions f and g.
$\chi_\omega(\cdot)$	Character of the LCA T where ω is an element of the dual group $\hat{T}$.
$\succeq_{\mathfrak{C}}$	The order relation in a convex cone $\mathfrak{C}$.
$\Psi_A(f) = f(A)$	The functional calculus of the selfadjoint operator A applied to the function f.

Optimization and implicitly defined terms

$\sup, \inf$	The lowest upper bound (supremum) and greatest lower bound (infimum) of a set (typically the image of a function).
argmin	The operator mapping functions on a domain T onto those values $t \in T$ that achieve the minimum value $f(t)$. An appropriately modified statement holds for the operator argmax.
$\nabla_\alpha f$	The gradient of f with respect to parameters α.

Chapter 1

Introduction

This chapter surveys scope and structure of the monograph and provides an introduction to the mathematics used within the thesis. Special emphasis is placed on motivating the vector space approach to signal analysis by successively introducing, interpreting, and generalizing the classical Euclidean spaces in this context starting with accessible low-dimensional examples. The two dimensional Cartesian plane is one such model example of a vector space in which geometric intuition and the results obtained by linear algebra coincide. As such it presents a suitable starting point for an understanding of the interaction between purely algebraic and geometric quantities whose explanatory and intuitive power will carry over to Euclidean spaces of arbitrary but finite dimension. By reinterpreting those spaces as spaces of functions, some of the abstractness naturally occluding a clear picture in high dimensional settings will be alleviated. This shift in perspective furthermore establishes functions as the objects of prime interest in this thesis. A first approximation of the definition of a Hilbert space is presented and, although omissive of some topological intricacies, motivates the application of geometric operations to functional analytic problems. Before formally introducing Hilbert spaces in chapter 2, a brief enumeration lists what the author would classify as empirical arguments supporting the claim that Hilbert spaces are a natural and useful concept to guide creation and manipulation of deterministic or stochastic models of natural processes in a computationally tractable way. Apart from the many explanatory and motivational remarks linking vector spaces and signal processing, the material covered in this section is completely standard from a mathematical perspective and can be found in almost any text covering linear algebra. The author suggests the books by Strang [190] and Shilov [180] for their depth of treatment and clarity. References are given where proofs are omitted.

§ Motivation, overview, and guide to the reader

Reproducing kernel Hilbert spaces (RKHS) are vector spaces of functions whose special inner product structure makes them suitable settings for a multitude of data-analytic tasks. In their role both as spaces of functions augmented with a probability distribution and as the natural framework for infinite-dimensional feature representations, RKHS have received increased attention in the past decade from statistical and machine learning-oriented communities. They have been used to represent and manipulate stochastic processes and are one of the main tools for analyzing non-linear relations by embedding them into higher-dimensional auxiliary spaces. This dual role, however, in conjunction with the fact that the theory of RKHS is an intrinsically functional analytic one that replaces simple models with finite degrees of freedom by non-parametric, infinite superpositions of basis functions, has lead to RKHS being employed only sparsely in the geodetic literature with its typical focus on explicitly parametric representations and adjustment theory.

Motivated by ill-posed real-world problems for which parametric models are neither justifiable nor practically sufficient, the primary goal of this monograph is to explain, advance, and apply the theory of RKHS in the context of practically relevant geodetic problems and relate it to more traditional processing strategies. One such problem, and in fact the main driving force behind much of the thesis' new developments, is a challenging spatiotemporal signal separation problem encountered in terrestrial radar interferometry (TRI). As TRI seeks to employ electromagnetic waves as a means for all-weather monitoring of deformations over distances of several km, intricate autocorrelated atmospheric artifacts arise, whose correlation structure is dependent on topographical features and is neither stationary nor directly describable by one of the usual available parametric covariance functions. As such the problem can not be successfully tackled by geodetic or geostatistical standard procedures and the need for computationally efficient processing strategies and instationary, potentially non-parametric covariance functions will guide the selection of topics.

This monograph consists of six chapters. Chapters 1 and 2 introduce Hilbert spaces and the spectral theory of linear operators on them before delving into the specific properties of so called reproducing kernel Hilbert spaces (RKHS). Building on those concepts, abstract splines and their relationship to stochastic processes are explored in chapter 3. The computational tools needed for implementation are also presented there together with a collection of applications as diverse as estimation of trajectories and vector fields as well as signal detection and separation. Closed form solutions exist for these problems but include the apriori unknown kernel determining the correlation structure of the process in question. While we will initially sidestep this drawback by ad-hoc arguments and inclusion of prior knowledge, a systematic study of kernel inference making use of theorems developed in abstract harmonic analysis and convex optimization will follow in chapter 4. Chapter 5 is the last chapter with scientific content and deals primarily with the full spatiotemporal signal separation problem for terrestrial radar interferometry. The monograph closes with a brief outlook onto open research questions in chapter 6.

The reader most interested in applications may directly jump to the sections 2.3, 3.1, and 3.3 for an introduction to RKHS and their relation to statistical nomenclature with a compilation of accessible examples and finally section 5.2 for an in-depth description and analysis in the context of TRI. The shortest route to kernel inference includes the sections 2.2 and 3.2 dealing with the spectral properties of kernels and their associated kernel operators as well as most of the fourth chapter in form of the sections 4.2 and 4.3 that present formulation and numerical solutions for the problem of deriving kernels from observations.

While traversing the sequence of topics listed above, tools are developed that are practically useful also in other disciplines of science in which either by nature or necessitated by complexity, a high-dimensional, and therefore often functional, perspective is the prevalent mode of presentation. This includes especially those parts of scientific theories that are concerned with the formation and adaptation of mathematical models and with the collection and extraction of information from data. For them, the theory of RKHS offers a monolithic block of methods for data processing with a clear stochastic interpretation that handles its dependence on prior knowledge in a transparent and economic manner by encoding it into a positive definite kernel. Since, by means of a newly developed procedure presented in the later parts of the monograph, these kernels can be inferred from observations if necessary, this results in algorithms that are almost free of arbitrary, user-specified choices and thereby particularly suited to act in a black-box-fashion if input and output are of a numerically well defined format.

The potential use of RKHS-based signal processing as an intermediate step integrated into a larger algorithmic pipeline is further augmented by the fact that the theory of RKHS and procedures derived from it are relatively self-contained. Therefore, the methods described in the monograph do not only directly improve the quality of displacement maps gathered by TRI and the associated risk assessments; they can also be used in thematically completely unrelated signal separation problems. To emphasize this generality and the accessibility of the approach even to non-experts, many examples of geodetically motivated statistical problems are posed and solved throughout the monograph. To further increase the convenience with which the content can be absorbed, very short preliminary summaries proceed the core material in each subsection. These do neither feature extensive explanations nor references to other sources but should help the reader to re-familiarize him- or herself with material already known or skip technical sections without missing the most important concepts, take-home messages, and the big picture.

Several alternative collections of state-of-the-art machinery like neural network-based deep learning (DL) and independent component analysis (ICA) could have been used to tackle the tasks presented in this monograph. The author has decided against them as, in contrast to the theory of Hilbert spaces, they emulate flexible behavior via nonlinearity. Although the results of DL and ICA are often satisfying, the exact reasoning behind the respective algorithms' decisions can therefore be quite arcane and inaccessible; in conjunction with the lack of stochastic interpretations and the induced absence of measures of the results' reliability, this is problematic in safety-relevant applications. Comparing this to Hilbert spaces, one finds the latter ones to be rather transparent and well surveyed with clear-cut relationships to stochastic processes and random fields. Since, however, these relationships are of an abstract nature, the rest of this introduction will aim to link archetypal low-dimensional Euclidean spaces to signal processing and thereby motivate a perspective centered around Hilbert spaces.

§ **Euclidean 2-space**

Recall the definition of the real numbers $\mathbb{R}$ and the Cartesian product $\mathbb{R} = \mathbb{R} \times \mathbb{R} =: \mathbb{R}^2$ as the set of all pairs of real numbers; $\mathbb{R}^2 := \{(\alpha, \beta) : \alpha \in \mathbb{R}, \beta \in \mathbb{R}\}$. Real numbers $\alpha, \beta \in \mathbb{R}$ can be added, subtracted and multiplied to provide a new real number denoted by $\alpha + \beta, \alpha - \beta$ and $\alpha\beta$ respectively. Addition and subtraction for elements of $\mathbb{R}^2$ are defined pointwise:

$$f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \in \mathbb{R}^2, g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \in \mathbb{R}^2 \Rightarrow$$

$$f +_{\mathbb{R}^2} g := \begin{bmatrix} f_1 + g_1 \\ f_2 + g_2 \end{bmatrix} \qquad f -_{\mathbb{R}^2} g := \begin{bmatrix} f_1 - g_1 \\ f_2 - g_2 \end{bmatrix} \tag{1.1}$$

They are obviously inverse operations in the sense that $f +_{\mathbb{R}^2} g -_{\mathbb{R}^2} g = f$ and commutative as well as associative. The neutral element of addition in $\mathbb{R}^2$ is denoted by $0_{\mathbb{R}^2} = (0,0)$. Multiplication of $f \in \mathbb{R}^2$ by a scalar $\alpha \in \mathbb{R}$ may be defined again pointwise:

$$f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \in \mathbb{R}^2, \alpha \in \mathbb{R} \Rightarrow \quad \alpha f := \alpha \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} \alpha f_1 \\ \alpha f_2 \end{bmatrix} \tag{1.2}$$

and distributivity holds in α as well as in f due to bilinearity, i.e. $(\alpha+\beta)f = \alpha f +_{\mathbb{R}^2} \beta f$ and $\alpha(f +_{\mathbb{R}^2} g) = \alpha f +_{\mathbb{R}^2} \alpha g$. The usual interpretation of this purely algebraic construction is that if $(\mathbb{R}, +)$ may be represented as a line, then $\mathbb{R}^2$ corresponds to a geometrical 2-dimensional Cartesian plane with fixed origin at $0_{\mathbb{R}^2}$, $f \in \mathbb{R}^2$ is a point in the plane and the previously defined operations on $\mathbb{R}^2$ translate to operations in the Cartesian plane as indicated by figure 1.1.

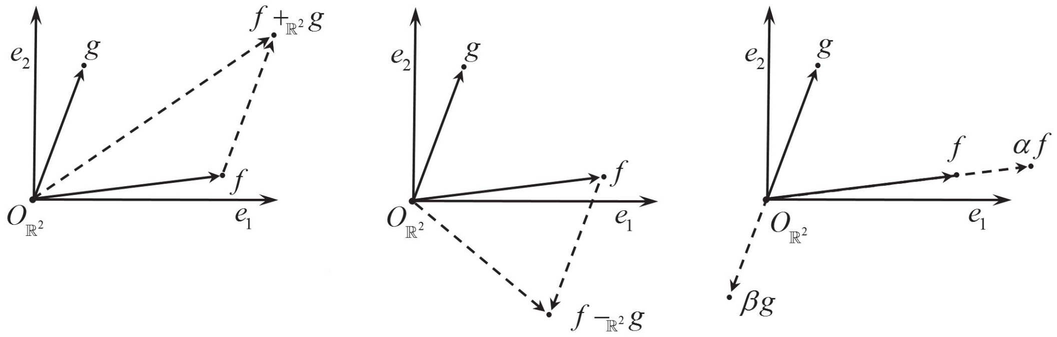


Figure 1.1: Representations of addition $+_{\mathbb{R}^2}$, subtraction $-_{\mathbb{R}^2}$ and scaling by scalars $\alpha, \beta \in \mathbb{R}$ in the Cartesian plane given a fixed choice of origin and direction e_1 and e_2.

Here $e_1 = (1,0)$ and $e_1 = (0,1)$ are chosen to allow any $f \in \mathbb{R}^2$ to be written as $f_1 e_1 + f_2 e_2 = \sum_{k=1}^{2} f_k e_k$. From figure 1.1 it is already clear that $f -_{\mathbb{R}^2} g$ is nothing else but

$$f -_{\mathbb{R}^2} g = \begin{bmatrix} f_1 - g_1 \\ f_2 - g_2 \end{bmatrix} = \begin{bmatrix} f_1 + (-1)g_1 \\ f_2 + (-1)g_2 \end{bmatrix} = f +_{\mathbb{R}^2} (-1g) \tag{1.3}$$

Therefore $-_{\mathbb{R}^2}$ as an operation $-_{\mathbb{R}^2} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ can be replaced by $+_{\mathbb{R}^2} \circ (1, -1) : \mathbb{R}^2 \times \mathbb{R}^2 \ni (f,g) \mapsto 1f +_{\mathbb{R}^2} (-1)g \in \mathbb{R}^2$ where $\circ$ denotes composition of operations.

Remark Note however, that the representation of $\mathbb{R}^2$ as a plane $\mathbb{P}$ is a structure preserving map $\pi : \mathbb{R}^2 \to \mathbb{P}$ rather than an identification. There are objects in $\mathbb{R}^2$ together with a set of operations on $\mathbb{R}^2$ and there are objects in $\mathbb{P}$ together with a set of operations on $\mathbb{P}$ such that objects and operators in $\mathbb{R}^2$ get mapped to objects and operators in $\mathbb{P}$ by π but $\mathbb{R}^2$ and $\mathbb{P}$ are not necessarily identical. Indeed, the choice of reference directions e_1, e_2 in $\mathbb{P}$ as well as the designation of a special point $0_{\mathbb{P}} \in \mathbb{P}$ are arbitrary — $\mathbb{P}$ is a $\mathbb{R}^2$ -torsor not a vector space — and every choice induces a structure preserving map π. More on this topic can be found in [125]. Without delving further into category-theoretical details, the cautionary note is issued that even though $\mathbb{P}$ is a useful model for $\mathbb{R}^2$ it is not its definition; analogue statements hold for $\mathbb{R}^n$.

Apart from addition and scalar multiplication, it is standard to introduce the (obviously symmetric) inner product $\langle \cdot, \cdot \rangle_{\mathbb{R}^2} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ by

$$f, g \in \mathbb{R}^2 \Rightarrow \langle f, g \rangle_{\mathbb{R}^2} := \sum_{k=1}^{2} f_k g_k \tag{1.4}$$

where $f = \sum_{k=1}^{2} f_k e_k$, $g = \sum_{k=1}^{2} g_k e_k$ and $\{e_k\}_{k=1}^{2}$ is the standard basis introduced above. It is then clear that $\langle f, f \rangle_{\mathbb{R}^2} = f_1^2 + f_2^2$ is the squared Euclidean distance between f and the origin $0_{\mathbb{R}^2}$ (see figure 1.2). Its square root corresponds to the length of the arrow pointing from $0_{\mathbb{R}^2}$ to $f \in \mathbb{R}^2$ and is called the norm of f, in symbols $||f||_{\mathbb{R}^2}$.

$$|| \cdot ||_{\mathbb{R}^2} : \mathbb{R}^2 \ni f \mapsto \sqrt{\langle f, f \rangle_{\mathbb{R}^2}} \in \mathbb{R} \tag{1.5}$$

For purposes of intuition, it is helpful to identify f with the arrow connecting the origin with f. For $f, g \in \mathbb{R}^2$ with $||g||_{\mathbb{R}^2} = 1$, $\langle f, g \rangle_{\mathbb{R}^2}$ is the length of f in the direction of g (see figure 1.2). In the future, the subscript $\mathbb{R}^2$ is dropped if no ambiguity exists as to which inner product or norm is meant.

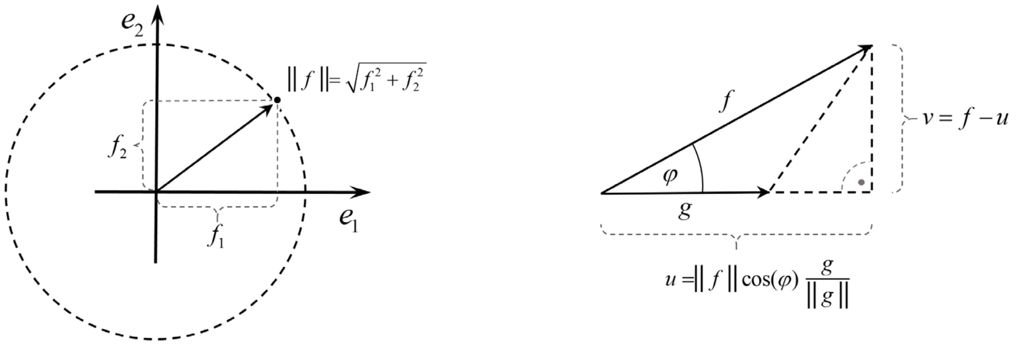

e_2
$||f|| = \sqrt{f_1^2 + f_2^2}$
f_2
f_1
e_1
f
φ
g
$v = f - u$
$u = ||f|| \cos(\varphi) \dfrac{g}{||g||}$

Figure 1.2: The norm of $f \in \mathbb{R}^2$ is its length $||f|| = \sqrt{f_1^2 + f_2^2}$ with f considered as an arrow in the plane and one may talk of the sphere $\mathbb{S}^1 := \{f \in \mathbb{R}^2 : ||f|| = 1\}$ for example. The right image illustrates $\langle f, g \rangle$ as the projection of f onto g for $g \in \mathbb{S}$.

It is furthermore true that the inner product $\langle f, g \rangle$ of f with g is related to the angle

φ between them when considered as arrows. $\forall f, g \in \mathbb{R}^2$, the following holds:

$$\langle f, g\rangle = \|f\|\|g\| \cos(\varphi) \tag{1.6}$$

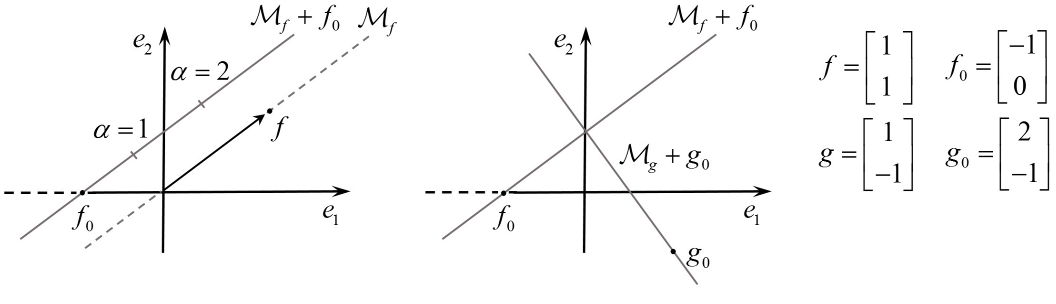
Proof: By the law of cosines $\|f - g\| = \|f\|^2 + \|g\|^2 - 2\|f\|\|g\| \cos(\varphi)$. But $\|f - g\| = \langle f - g, f - g\rangle = \langle f, f\rangle + \langle g, g\rangle - 2\langle f, g\rangle$ and comparing both expansions it follows that $\langle f, g\rangle = \|f\|\|g\| \cos(\varphi)$. □

Remark Note that $\cos\varphi = 0$ if and only if $\varphi = \pm\pi/2$. Consequently for a right angle φ one finds $\langle f, g\rangle = 0$ which implies that that $f \perp g \Leftrightarrow \langle f, g\rangle = 0$.
Remark Every $f \in \mathbb{R}^2$ can be split into two orthogonal components given some $g \in \mathbb{R}^2$ according to the scheme exhibited on the right side of figure 1.2. If $\|g\| = 1$ then $f = f_g + f_g^\perp = \langle f, g\rangle g + (f - \langle f, g\rangle g)$ where $f_g \parallel g$ and $f_g \perp f_g^\perp$ since $\langle f_g, f_g^\perp\rangle = \langle\langle f, g\rangle g, f - \langle f, g\rangle g\rangle = \langle f, g\rangle^2 - \langle f, g\rangle^2\|g\|^2 = 0$.

The set $\mathcal{M}_f := \{\alpha f : \alpha \in \mathbb{R}\}$ for a fixed $f \in \mathbb{R}^2$ is a subset of $\mathbb{R}^2$ and is closed under addition and multiplication by scalars. The set $\mathcal{M}_f + f_0 := \{f_0 + \alpha f : \alpha \in \mathbb{R}\}$ can be identified with the line through f_0 extending in the $f/\|f\|$ unit direction. For two lines $\mathcal{M}_f + f_0$ and $\mathcal{M}_g + g_0$, their intersection is found by solving the following system of linear algebraic equations (SLAE) in the unknowns α_1, α_2:

$$\begin{aligned} f_0 + \alpha_1 f &= g_0 + \alpha_2 g \\ \Leftrightarrow \quad f\alpha_1 + -g\alpha_2 &= g_0 - f_0 \\ \Leftrightarrow \quad \underbrace{\begin{bmatrix} f_1 & -g_1 \\ f_2 & -g_2 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}}_{x} &= \underbrace{\begin{bmatrix} (g_0 - f_0)_1 \\ (g_0 - f_0)_2 \end{bmatrix}}_{y} \end{aligned} \tag{1.7}$$

The SLAE does not necessarily have a unique solution. If $\mathcal{M}_f + f_0 = \mathcal{M}_g + g_0$ then for every choice of α_1 a corresponding α_2 might be found and if $\mathcal{M}_f + f_0 \cap \mathcal{M}_g + g_0 = \emptyset$ then no solution exists at all. Note that both of these special cases are rare if f_1 and f_2 are chosen at random and the system is usually solvable (for an example see figure 1.3) as the set of matrices A such that $\det A = 0$ is a relatively sparse subset of $\mathbb{R}^4$ [65, p. 53].



$\mathcal{M}_f + f_0$
$\mathcal{M}_f$
e_2
$\alpha = 2$
$\alpha = 1$
f
f_0
e_1
$\mathcal{M}_f + f_0$
e_2
$\mathcal{M}_g + g_0$
f_0
e_1
g_0
$f = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
$f_0 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$
$g = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$
$g_0 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$

Figure 1.3: The set $\mathcal{M}_f + f_0$ corresponds to a line parametrized by a real parameter α. The right illustration depicts a specific constellation in which two lines $\mathcal{M}_f + f_0$ and $\mathcal{M}_g + g_0$ intersect in one unique point that can be found by solving a SLAE.

In the example in figure 1.3, the system of linear equations has the form

$$\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix}=\begin{bmatrix}3\\ -1\end{bmatrix} \tag{1.8}$$

which is solved iff $\alpha_1 = 1$ and $\alpha_2 = -2$ corresponding to the point $f_0 + \alpha_1 f = g_0 + \alpha_2 g = (0,1)$. Replace now the symbol α_k by x_k and $x = (x_1, x_2)^T$ to denote its role as an unknown explicitly. Even though nothing interesting seems to have been achieved, writing a system of equations as $Ax = y$ changes the flavor of its interpretation as A represents a mapping $A : \mathbb{R}^2 \to \mathbb{R}^2$ and $Ax = y \Leftrightarrow \sum_k a_k x_k = y, a_k$ the k-th column of A. This then establishes the x_k as coefficients of a suitable zero-error approximation of y by a weighted superposition of basis vectors $a_k \in \mathbb{R}^2$.

Remark The matrix A in the example above can be written as a composition of 3 different matrices

$$A=\begin{bmatrix}1 & -1\\ 1 & 1\end{bmatrix}=\begin{bmatrix}\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}\sqrt{2} & 0\\ 0 & \sqrt{2}\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}=U\Sigma V^T.$$

Therefore Ax can be interpreted as a sequence of actions consisting of scaling the parameters x_1 and x_2 by $\sqrt{2}$ and then multiplying x_1 by basis vector $a_1/\|a_1\| = (1/\sqrt{2}, 1/\sqrt{2})^T$ and x_2 by $a_2/\|a_2\| = (-1/\sqrt{2}, 1/\sqrt{2})^T$ with the total effect then being one of scaling and subsequent rotation. By the proper change of basis $x = U^T\tilde{x}$, $Ax = U\Sigma V^T U^T \tilde{x} = \sqrt{2}UIU^T\tilde{x} = \sqrt{2}\tilde{x}$ where A acts diagonally on $\tilde{x}$ reflecting the fact that f and g form an orthogonal basis. This is indicated by $\langle f, g\rangle = 0$ as can be easily seen in the preceeding figure 1.3 by noting the right angle between $\mathcal{M}_f + f_0$ and $\mathcal{M}_g + g_0$.

Remark The SLAE $Ax = y$ is equivalent to a sequence of one dimensional constraints $\langle a^k, x\rangle = y_k, k = 1, 2$ where a^k is the k-th column of A^T. Every linear constraint for x in $\mathbb{R}^n$ forces x to lie in a plane of dimension $n - 1$. Only in $\mathbb{R}^2$ does this coincide with a line. Letting the j-th entry of a^k be denoted by a^k_j and supposing $a^k_j \neq 0$ for some fixed j, the k-th constraint in equation 1.8 implies

$$x=\begin{bmatrix}\frac{1}{a^k_1}\left(y_k - a^k_2 x_2\right)\\ x_2\end{bmatrix} \qquad \text{or} \qquad x=\begin{bmatrix}x_1\\ \frac{1}{a^k_2}\left(y_k - a^k_1 x_1\right)\end{bmatrix}$$

depending on which of the pair (x_1, x_2) is chosen as the independent variable. Therefore given $y \in \mathbb{R}^2$ and one linear constraint $\langle a^k, x\rangle = y_k$, x only depends on $2-1$ unknowns. Analogous statements hold for n-dimensional settings in which the constraints give rise to feasible sets that form $(n-1)$-dimensional hyperplanes.

Still, even after decoupling $\mathbb{R}^2$ from a strictly geometrical interpretation, two problems persist. The low dimensional setting prohibits useful application and up until now A, x, y were written as collections of real numbers which implies that some orthonormal basis has already been fixed. Both concerns will be addressed in the next paragraphs.

§ Euclidean n-space

When visualizing $f \in \mathbb{R}^2$ as a point or arrow in the plane, much emphasis is given to the ambient space $\mathbb{R}^2$ and little to f as the latter consists of two numbers only. In the general setting with dimension $n > 3$ this approach fails to provide any intuition regarding the ambient space $\mathbb{R}^n$. Considering $f \in \mathbb{R}^2$ as an ordered sequence of numbers with length 2, $f = \{f_k\}_{k=1}^{2}$[1] — or alternatively as a function $f_. : T \ni t \mapsto f_t \in \mathbb{R}$ for $T = \{1, 2\}$ — generalizes to arbitrary finite dimensional Euclidean spaces $\mathbb{R}^n$ without much impediment to the instructivity of the functional interpretation. Figure 1.4 seeks to support this claim.

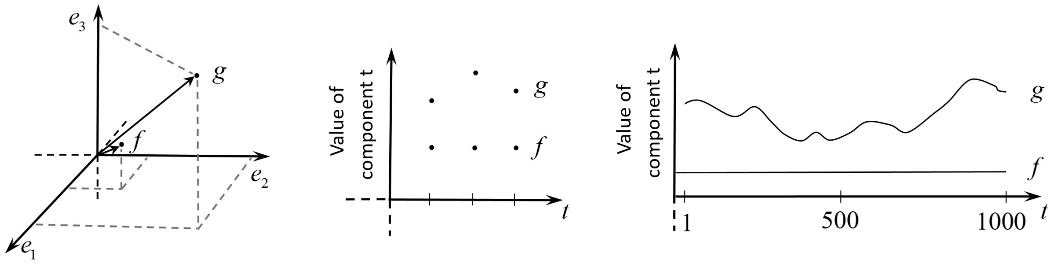


Figure 1.4: In the left and middle plots, two vectors $f, g \in \mathbb{R}^3$ are illustrated. They are interpreted as geometrical objects in three dimensional Euclidean space (left) and as functions $f, g : T \to \mathbb{R}, T = \{1, 2, 3\}$(middle). Their graphs $\{(t, h_t) : t \in T\}, h = f, g$ are plotted. The right side of the figure employs an analogous construction to represent vectors $f, g \in \mathbb{R}^n$ with $n = 1000$.

Definition 1.1.1 For any finite n, the set $\mathbb{R}^n$ of n-tuples of real numbers together with componentwise addition of elements $(f + g)_t = f_t + g_t$, scalar multiplication $(\alpha f)_t = \alpha f_t$ and inner product $\langle f, g\rangle = \sum_{t=1}^{n} f_t g_t \ \forall f, g \in \mathbb{R}^n$, $\forall \alpha \in \mathbb{R}$ is called the Euclidean n-space. Here f_t or $(f + g)_t$ denote the t-th component of f or $f + g$ respectively.

As is mentioned in the next subsection in theorem 2.1.5, addition of elements $f, g \in \mathbb{R}^n$, scalar multiplication with $\alpha \in \mathbb{R}$ and formation of inner products are continuous operations given the norm topology on $\mathbb{R}^n$. The same holds for the projections $\pi_t : \mathbb{R}^n \ni f \mapsto f_t \in \mathbb{R}$ as mappings on vectors. Therefore it is reasonable to expect no pathologies where standard operations on vectors in $\mathbb{R}^n$ are concerned.

Definition 1.1.2 If in a sequence $\{f_k\}_{k=1}^{m} \subset \mathbb{R}^n$ the implication $(\sum_{k=1}^{m} \alpha_k f_k = 0) \Rightarrow (\alpha_k = 0, k = 1, ...m)$ holds, then the elements $f_k \in \mathbb{R}^n$ of that sequence are said to be linearly independent. A sequence $\{e_k\}_{k=1}^{m}$ is called a basis for $\mathbb{R}^n$ iff its elements are linearly independent and $\forall f \in \mathbb{R}^n \ \exists \alpha \in \mathbb{R}^m$ such

[1] The standard notation $f = \{f_k\}_{k=1}^{m}$ indicates f to be an ordered sequence of elements $(f_1, ..., f_m)$ where f_1 is the first and f_m is the m-th element. The individual elements are themselves allowed to be arbitrary objects and can therefore be sequences or Euclidean vectors themselves. See the remark following definition 1.1.2 for more details.

that $\sum_{k=1}^{m} \alpha_k e_k = f$.

A basis of $\mathbb{R}^n$ has always n elements [180, p. 40]. Caution is advised when handling series expansions as in the above definition since the subscript notation is overloaded and is used to denote sequences of elements $f_k \in \mathbb{R}^n, k = 1, ..., m$ as well as sequences of real numbers $f_t \in \mathbb{R}, t = 1, ..., n$ that can be assembled to a vector $f = \sum_{t=1}^{n} f_t e_t \in \mathbb{R}^n$. This notational inconvenience is unavoidable in this context but will become both less cumbersome and less frequent when the motivational subsection closes and Hilbert spaces are treated in an infinite-dimensional setting in subsection 2.1.1 without references to components of a vector in some auxiliary and arbitrarily fixed space.

Definition 1.1.3 The sequence $\{e_k\}_{k=1}^{n} \subset \mathbb{R}^n$ where $e_k = (0, ..., 0, 1, 0, ..., 0) \in \mathbb{R}^n$ is a vector containing a 1 at position k and 0 otherwise is called the canonical Euclidean basis. The $e_k, k \in \{1, ..., n\}$ are the canonical Euclidean basis vectors.

The sequence $\{e_k\}_{k=1}^{n}$ as defined above is indeed a basis of $\mathbb{R}^n$ as is proven e.g. in [180, p. 39]. From $(e_k)_t = \delta_{kt}$, with δ_{kt} the usual Kronecker delta, it follows that $\langle e_k, e_l \rangle = \sum_{j=1}^{n} \delta_{kj}\delta_{jl} = \delta_{kk}\delta_{kl} = \delta_{kl}$ and the canonical Euclidean basis is seen to be even an orthonormal basis (ONB) for $\mathbb{R}^n$. When drawing upon the image of $f \in \mathbb{R}^n$ as a function $f : T \ni t \mapsto f(t) = f_t \in \mathbb{R}$, e_t takes the role of evaluating f at t since $\langle e_t, f \rangle = \langle e_t, \sum_{k=1}^{n} f(k) e_k \rangle = \sum_{k=1}^{n} f(k)\delta_{tk} = f(t)$. There are several different bases for $\mathbb{R}^n$ when $n \geq 2$ and an element $f \in \mathbb{R}^n$ is represented by a different sequence of expansion coefficients depending on the chosen basis.

Theorem 1.1.4 *Given an orthonormal basis $\{e_k\}_{k=1}^{n}$ of $\mathbb{R}^n$, the expansion coefficients α_k for $f \in \mathbb{R}^n$, $f = \sum_{k=1}^{n} \alpha_k f_k$, are given by $\alpha_l = \langle f, e_l \rangle, l = 1, ..., n$.*

Proof: Given the ONB $\{e_k\}_{k=1}^{n}$ by definition $\exists \{\alpha_k\}_{k=1}^{n}$ such that $f = \sum_{k=1}^{n} \alpha_k e_k$. Then it holds that $\langle f, e_l \rangle = \langle \sum_{k=1}^{n} \alpha_k e_k, e_l \rangle = \sum_{k=1}^{n} \alpha_k \delta_{kl} = \alpha_l$. □

This shows that given the ONB $\{e_k\}_{k=1}^{n} \subset \mathbb{R}^n$ the series expansion of any $f \in \mathbb{R}^n$ is simply given by $\sum_{k=1}^{n} \langle f, e_k \rangle e_k$. The expansion coefficients $\alpha_k = \langle f, e_k \rangle$ can consequently be calculated explicitly and efficiently. They are also named Fourier coefficients of f with respect to the ONB $\{e_k\}_{k=1}^{n}$, the reason for which becomes apparent after consulting figure 1.5

Not necessarily orthogonal bases enjoy many of the same convenient properties as ONB's but the calculation of expansion coefficients is computationally and conceptionally more involved. They are treated in subsection 2.1.2.

To talk about differentiability or continuity of a function $f : T \ni t \mapsto f(t) \in \mathbb{R}$ in the variable t demands t to be a topological space and for $n \to \infty$, it might be expected that the domain $T = \{1, ..., n\}$ of $f : T \to \mathbb{R}$ is interpretable as a real interval $[a, b] \subset \mathbb{R}$. It turns out that there is a different way to facultate the transition from finite dimensional n-tuples of real numbers to functions on topological spaces. The key steps include the introduction of the abstract concept of a vector space

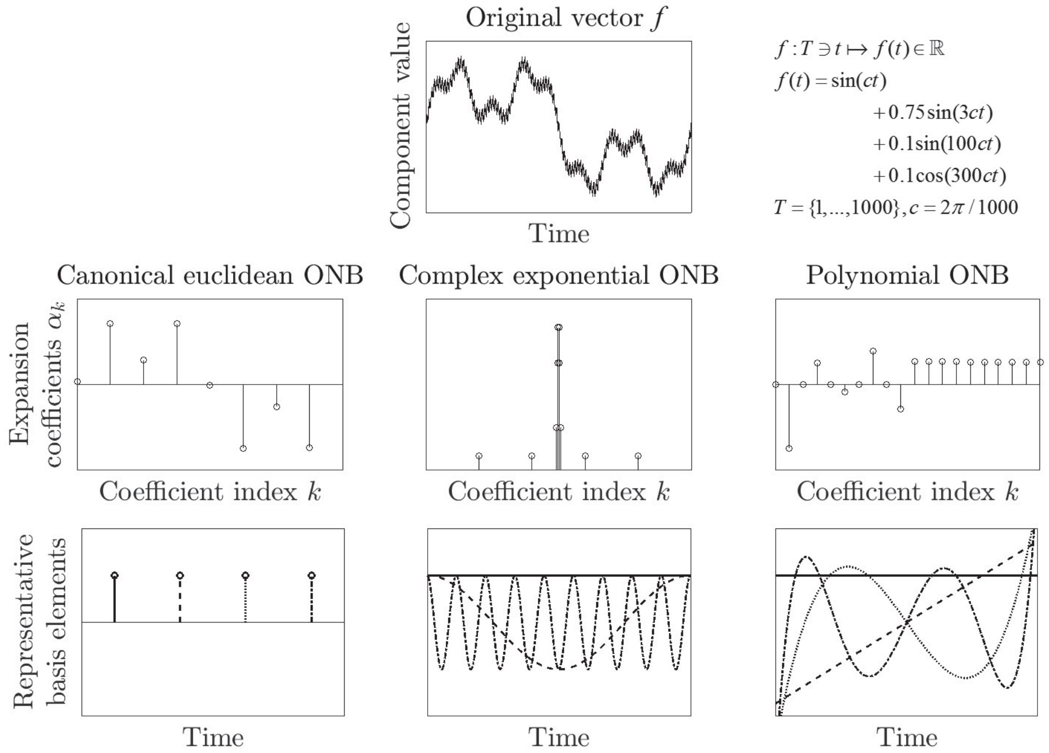


Figure 1.5: Expansions of the same function f — a superposition of four sinusoids with different wavelengths as indicated in the top-right panel — in terms of different ONB's. The underlying indexset is interpreted as time T with $T = \{1, ..., 1000\}$ implying $f \in \mathbb{R}^{1000}$. On the left, the canonical Euclidean basis is used; four exemplary basis elements corresponding to the canonical Euclidean basis elements $e_k, k = 125, 375, 625, 875$ having value 1 at index k and being zero otherwise are plotted in the third row. The expansion coefficients in this basis are simply the point evaluations of f at the indicated times as documented in the second row. In the middle, a Fourier-like basis was used, whose exemplary basis elements are sinusoidals of different frequencies that are again plotted in the bottom row and whose expansion coefficients as recorded in the middle row show sharp peaks at those indices corresponding to the frequencies used in the construction of f. The same investigations have been carried out on the right for orthogonal (Legendre) polynomials. The expansion coefficients are calculated according to theorem 1.1.4; for complex coefficients only their modulus is plotted. Different linestyles have been used to distinguish different basis elements. The construction of the ONB's themselves can be found in subsection 2.1.2.

without referral to sequences of numbers and a careful analysis of different notions of convergence under some distance measure closely related to the norm $\|f\| = \left(\sum_{t=1}^{n} f_t^2\right)^{1/2}$ in $\mathbb{R}^n$.

§ **Vector spaces**

Recall the general definition of a vector space V_{vs} over a field $\mathbb{F}$ as a set V together with an addition operation $+_V : V \times V \to V$ and scalar multiplication $\cdot : \mathbb{F} \times V \to V$. It is demanded that V is a commutative group under addition, and scalar multiplication preserves linear structure [180, p. 31]. This notion of a vector space encompasses the previous examples and does not depend on any arbitrary choice of basis or representability of its elements as sequences of numbers. A vector space can be augmented with additional structures like an inner product $\langle \cdot, \cdot \rangle$ or a norm $\| \cdot \|$, but these are not part of the original definition.

The concern of this monograph lies exclusively with vector spaces V over the fields of real or complex numbers. Examples are $V = \mathbb{R}^n$ or $V = \mathbb{C}^n$ for finite $n \in \mathbb{N}$

[180, p. 34] and spaces of real or complex valued functions on some set T with in general infinite cardinality [181, p. 4]. The latter ones are central for the further development.

As Shilov remarks in his introductory notes on functional analysis [181, p. 4], it is straightforward to construct a tower structure of vector spaces given any field $\mathbb{F}$:

$\mathbb{F}$ is a vector space under addition and multiplication inherited from $\mathbb{F}$.

$\mathbb{F}^n := \mathbb{F} \times ... \times \mathbb{F}$ is a vector space of n-tuples under pointwise addition and multiplication inherited from $\mathbb{F}$.

$\mathbb{F}(T) := \{f : T \to \mathbb{F}\}$, for T some set, is a vector space of functions under pointwise addition and multiplication inherited from $\mathbb{F}$.

$\mathbb{F}^n(T) := \{f : T \to \mathbb{F}^n\}$, for T some set, is a vector space of vector valued functions under pointwise addition and multiplication inherited from $\mathbb{F}^n$.

The vector space $\mathbb{F}^n$ of vector-valued functions on T is already a space quite interesting for applications. Different examples are given in figure 1.6.

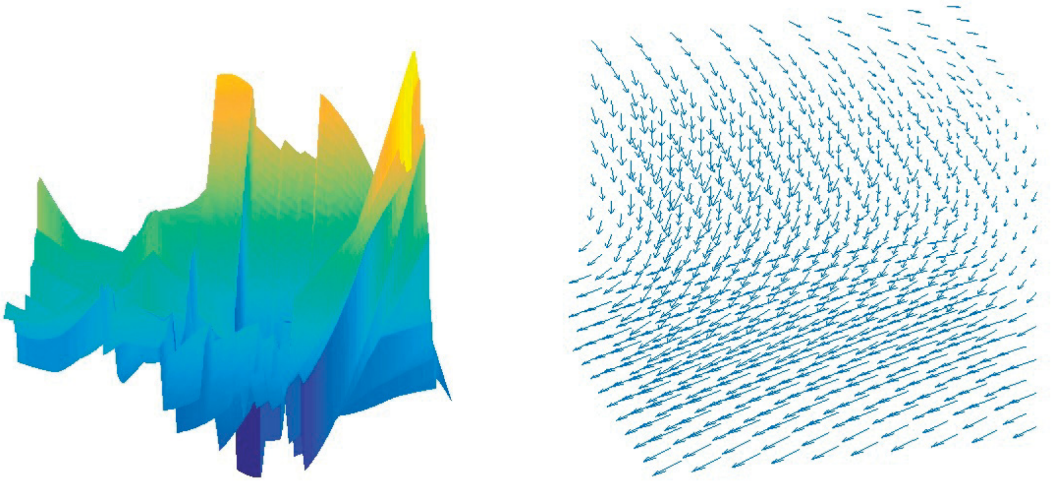


Figure 1.6: On the left an element of $\mathbb{F}(T_1)$ is shown where $T_1 \subset \mathbb{R}^2$. A vector in this space is a function defined on a subset of the Euclidean plane whereas on the right side a vector field is visible, i.e. an element of $\mathbb{F}^3(T_2), T_2 \subset \mathbb{R}^3$. In both cases $\mathbb{F} = \mathbb{R}$.

§ Inner products and norms

Definition 1.1.5 Given a vector space V over $\mathbb{C}$ an inner product $\langle \cdot, \cdot \rangle$ is a map from $V \times V$ to $\mathbb{C}$ that is positive definite (i), conjugate symmetric (ii), homogeneous w.r.t scalar multiplication (iii) and additive (iv). That is, the relations

i) $f \neq 0_V \Rightarrow \langle f, f \rangle > 0, \;\; \langle 0_V, 0_V \rangle = 0$

ii) $\langle f, g \rangle = \overline{\langle g, f \rangle} \;\; \forall f, g \in V$

iii) $\langle \alpha f, g \rangle = \alpha \langle f, g \rangle \;\; \forall f, g \in V, \alpha \in \mathbb{C}$

iv) $\langle f +_V g, h \rangle = \langle f, h \rangle + \langle g, h \rangle \;\; \forall f, g, h \in V$

hold [181, p. 63]. The overline denotes complex conjugation. A vector space together with an inner product is called a pre-Hilbert space or an inner product space.

Examples of inner product spaces are $(\mathbb{R}, \langle\alpha,\beta\rangle_{\mathbb{R}} := \alpha\beta), (\mathbb{C}, \langle\alpha,\beta\rangle_{\mathbb{C}} := \alpha\overline{\beta}), (\mathbb{R}^n, \langle f,g\rangle_{\mathbb{R}^n} := \sum_{k=1}^n f_k g_k)$ and $(\mathbb{C}^n, \langle f,g\rangle_{\mathbb{C}^n} := \sum_{k=1}^n f_k \overline{g_k})$. Note that the inner products are results of a choice and different choices of inner products may lead to different inner product spaces. They all however satisfy the Cauchy-Schwarz inequality [93, p. 10].

Theorem 1.1.6 (Cauchy-Schwarz) *For any inner product $\langle\cdot,\cdot\rangle$, it holds that*

$$|\langle f,g\rangle|^2 \leq \langle f,f\rangle\langle g,g\rangle \ \ \forall f,g \in V$$

The notion of the length or magnitude of a vector is axiomatized with the help of what is called a norm.

Definition 1.1.7 Given a vector space V over $\mathbb{C}$, a norm $\|\cdot\|$ is a map from V to the positive reals $\mathbb{R}_+$ that is non-degenerated, homogeneous w.r.t. absolute values and subadditive. This means the relations

$$\begin{aligned} \text{Non-degeneracy:} &\ \ V \ni f \neq 0 \Rightarrow \|f\| > 0, \|0\| = 0 \\ \text{Positive homogeneity:} &\ \ \|\alpha f\| = |\alpha|\|f\| \ \ \forall f \in V, \alpha \in \mathbb{C} \\ \text{Triangle inequality:} &\ \ \|f+g\| \leq \|f\| + \|g\| \end{aligned}$$

hold [180, p. 53]. A vector space V together with a norm $\|\cdot\|$ is called a normed linear space.

Given any inner product space $(V, \langle\cdot,\cdot\rangle)$, $(V, \|\cdot\|)$ is a normed linear space when the norm is chosen to be the positive squareroot of the inner product of f with itself, $\|f\| = \sqrt{\langle f,f\rangle}$. This implies that every inner product space is also a normed linear space under the induced norm. However, not every normed linear space is also an inner product space [93, p. 14]. Unless otherwise specified, when V is an inner product space and a norm is mentioned then this norm is the induced norm.

By the above remarks, normed linear spaces are a proper generalization of inner product spaces. Although normed linear spaces that are not inner product spaces enter only in chapter 4, most approximation and estimation problems will be stated in terms of the norm even when an inner product is available. The reason for that decision is that a norm $\|\cdot\|$ on a vector space V provides a sensible way to talk about closeness of two vectors $f, g \in V$. Therefore when approximating f by sums of given vectors $g_k, k = 1, ..., n$ expressions like

$$d(f,\hat{f}) = \|\sum_{k=1}^n \alpha_k g_k - f\|$$

naturally arise where the discrepancy measure $d(\cdot,\cdot) : V \times V \rightarrow \mathbb{R}$ is demanded to be small for the approximation $\hat{f} = \sum_{k=1}^n \alpha_k g_k$ to be considered appropriate. It even holds that for any normed linear space V, the function $d(f,g) = \|f - g\|$ is a metric and the topological structure induced by the open balls in that metric is enough to determine answers to questions pertaining to convergence and continuity of sequences and sums of vectors.

The last ingredient missing to upgrade inner product spaces to Hilbert spaces and normed linear spaces to Banach spaces is topological in nature and necessary to guarantee that every sequence of elements exhibiting shrinking pairwise distances converges to a unique limit point lying in the space. This property is called completeness; without it the transition to infinite dimensional spaces (e.g. spaces of functions on arbitrary sets T) would see many of the convenient properties encountered in finite dimensional spaces lost.

Before the notion of a Hilbert space is made entirely formal in the next chapter, we review its main features in an axiomatic list of statements that highlight the concept's usefulness and expected connections to data analytic questions. The theory of Hilbert spaces

- is well established as a central tool in functional analysis and provides a framework in which to efficiently reason about whole sets of functions simultaneously by employing geometric arguments;
- allows to inject objects from some base space into infinite dimensional spaces in which easily analysable linear operations can achieve the same results as highly nonlinear operations in the original base space;
- is general enough to include results valid for, among others, functions, measures, random variables and stochastic processes but still admits specific non-trivial conclusions about the objects of investigation;
- due to its essentially linear nature often generates solutions that are readily adoptable in conceptually simple and computationally efficient ways by standard linear algebra routines;
- is both statistical and physical in nature as is illustrated by Hilbert spaces arising as sets of solutions to differential equations governing the time evolution of a system's state given limited information while at the same time exhibiting a measure of closeness that can be motivated stochastically.

This by no means exhaustive collection of rather empirical observations serves to show that geodetic data analysis has much to gain from a Hilbert space based approach as it regularly acts in the area of conflict lying in the middle grounds between deterministically driven physical systems and uncertainty in observations that is best modelled probabilistically. This claim is further substantiated by several successful applications of the theory of Hilbert spaces to problems from space geodesy that range from satellite orbit determination [179] to the derivation of physically justified estimators for the earth's gravity field based on measurements of e.g. line-of-sight acceleration data [144]. Several investigations have been carried out towards incorporating the framework of Hilbert spaces into the mathematical foundations of geodesy by noting the intrinsic ties between Hilbert spaces, boundary value problems and their respective variational formulations [pp. 29-45][24], [145] and by introducing Hilbert spaces as the natural, infinite-dimensional generalizations of Euclidean spaces into which least-squares methods can easily be transported to grant them wider applicability [1]. Notions like stochastically motivated measures of elastic energy quantifying uncertainty in geodetic networks [24, pp. 159-178] or unwanted deviations from a low-distortion state in the design of new map projections [80] provide further proof of both the unifying generality of Hilbert space-based optimization problems and their practical relevance for very specific problems. Although Hilbert spaces seem to not have permeated fully the instruments and theory of the rather practical discipline of engineering geodesy, it seems safe to assume their usefulness even in this context.

Chapter 2

Theory of Hilbert spaces

This chapter focuses on the theory of Hilbert spaces and establishes basic theorems needed later in the monograph. Geometric intuition is provided regarding important concepts such as superposition, projections and bases, before questions of decomposability of a Hilbert space into subspaces are investigated. Linear operators on Hilbert spaces are introduced along with properties such as boundedness, compactness, selfadjointness and positivity, all of which allow to abstractly connect the subclasses of operators enjoying these properties to operations prevalent in signal processing. After further surveying the relationship between a linear operator and the Hilbert space it acts on, focus is shifted onto the spectral theory of not necessarily bounded selfadjoint operators. This line of inquiry culminates in a statement of Stones theorem on strongly continuous one parameter unitary semigroups describing the time evolution of a dynamical system and lays the groundwork for incorporation of physical knowledge into the mathematical models representing real-world processes. Reproducing kernel Hilbert spaces are defined and their connections to energy minimization and feature representations are made explicit.

2.1 Topology and geometry of Hilbert spaces

In this section, Hilbert spaces are introduced as typically infinite dimensional vector spaces of functions with a topology on them. Their definition is complemented by a number of classical results pertaining to the connection between a Hilbert space and the subspaces contained within it. The operation of creating a Hilbert space from elementary building blocks is generalized to account for the more abstract constructions of direct sums and tensor products. This reveals time series, vector fields, features and spatiotemporal distributions of measurements to be elements of their respective Hilbert spaces as well. As this section covers introductory material largely unknown in the geodetic community, explanations and presentation will be extensive. Theorems are furnished with proofs, if these highlight the use of an important concept.

2.1.1 Hilbert spaces

The norm induced by an inner product is interpreted as a distance function on a normed vector space V and the topology induced by this metric can be deduced unambiguously. Demanding sequences $\{f_k\}_{k=1}^{\infty} \subset V$, whose consecutive elements eventually differ only marginally, to converge to an element f in the normed space V amounts to a property called completeness. Complete inner product spaces are called Hilbert spaces and complete normed vector spaces are called Banach spaces. Some instructive examples of both Hilbert and Banach spaces include objects known from time series analysis and differential equations and the classical ℓ^p and L^p spaces. A set of canonical signal processing problems will be introduced; these will be revisited and upgraded throughout the monograph to illustrate new concepts in a familiar setting.

§ **Topological considerations**

Definition 2.1.1 A family $\mathcal{O}$ of sets in X is called a topology on X iff it satisfies the following conditions:

I $\emptyset, X \in \mathcal{O}$

II $\mathcal{O}_j \in \mathcal{O}\ \forall j \in J \Rightarrow \bigcup_{j\in J} \mathcal{O}_j \in \mathcal{O}$

III $\mathcal{O}_1, \mathcal{O}_2 \in \mathcal{O} \Rightarrow \mathcal{O}_1 \cap \mathcal{O}_2 \in \mathcal{O}$

The set X together with the topology $\mathcal{O}$ is called a topological space. The elements of $\mathcal{O}$ are called open sets. Sets F, whose complements $F^C := \{x \in X : x \notin F\}$ are open, are called closed. Sets may be open, closed, both or neither. For introductory examples see [141, pp. 70-74].

Definition 2.1.2 In a normed linear space V, for any $\epsilon > 0$ and $f_0 \in V$ the set $B_\epsilon(f_0) := \{f \in V : \|f - f_0\| < \epsilon\}$ is called the open ball of radius ϵ around f_0 . Examples are found in figure 2.3.

Theorem 2.1.3 *I For any normed linear space V the family of sets $\mathcal{O}$ with elements $\emptyset \in \mathcal{O}$ and otherwise $O \in \mathcal{O} \Leftrightarrow \forall f \in O \exists \epsilon > 0 : B_\epsilon(f) \subset O$ is a topology on V, named the norm topology [207, p. 15].*

II On $\mathbb{R}^n$ and $\mathbb{C}^n, n \in \mathbb{N}$ all norms are equivalent, where two norms are called equivalent if they induce the same norm topologies [206, p. 26].

As an immediate corollary, the topologies on $\mathbb{R}^n$ and $\mathbb{C}^n$ are independent of the chosen norm. This follows directly from theorem 2.1.3.II and asserts that on finite dimensional vector spaces the norm topologies (and therefore notions of continuity and convergence) all coincide. In infinite dimensional vector spaces such as spaces of functions from arbitrary sets T to $\mathbb{C}$ this convenient situation does not hold. The norm topology admits a characterization of continuity in terms of the norm used in its construction.

Definition 2.1.4 I A function $A : V \to W$ is called continuous iff $A^{-1}(O) \in \mathcal{O}_V\ \ \forall O \in \mathcal{O}_W$. Here $\mathcal{O}_V$ and $\mathcal{O}_W$ are topologies on the topological spaces V and W.

II For a normed linear space V with norm $\|\cdot\|$, a sequence $\{f_k\}_{k=1}^{\infty} \subset V$ is said to converge to an element $f \in V, \lim_{k\to\infty} f_k = f$ if $\forall \epsilon > 0 \;\; \exists n_0 \in \mathbb{N} : n \geq n_0 \Rightarrow f_n \in B_{\epsilon}^{V}(f)$ [108, p. 75].

At this point various technical lemmas from point set topology intervene [141, 108, 71, 207] and establish instructive ways to rewrite the continuity condition making use of the role of the ϵ-open balls in the definition of the norm topology. Furthermore, topology, continuity and convergence are all interrelated as indicated by the sequence of statements below.

Lemma 2.1.5 *I The composition of continuous functions is continuous [141, p. 39].*

II A function $A : V \to W$ between normed linear spaces V, W is continuous, i.e. $A^{-1}(O) \in \mathcal{O}_V \;\; \forall O \in \mathcal{O}_W$ iff $\forall \epsilon > 0 \;\; \exists \delta > 0 : \|f - g\|_V < \delta \Rightarrow \|A(f) - A(g)\|_W < \epsilon$ [141, p. 36].

III For any sequence $\{f_k\}_{k=1}^{\infty}$ such that $\lim_{k\to\infty} f_k = f$ one has $\lim_{k\to\infty} \|f - f_k\| = 0$ [71, p. 31].

IV Limits in normed spaces are unique [108, p. 89].

V A function $A : V \to W$ between normed linear spaces V and W is continuous if and only if $\forall f \in V$ one has that $\lim_{k\to\infty} f_k = f$ implies $\lim_{k\to\infty} A(f_k) = A(f)$, i.e. limits and continuous functions commute [93, p. 73].

VI Vector addition, scalar multiplication and taking norms are continuous operations in the norm topology. If the norm is induced by an inner product, the inner product is continuous too [71, p. 32].

This means that the usual notion of continuity coincides with the topological one — at least for functions $A : V \to W$ and the choice of norm topologies on both domain and codomain. Limits are well behaved and the limiting process commutes with continuous operations. In later chapters other topologies designed in less obvious manners to make certain sets of functions continuous will be employed as well.

Definition 2.1.6 A sequence $\{f_k\}_{k=1}^{\infty}$ in a normed linear space V is said to be a Cauchy sequence — or simply Cauchy — if $\forall \epsilon > 0 \;\; \exists n_0$ such that $m, n \geq n_0 \Rightarrow \|f_n - f_m\| < \epsilon$.

It is trivial to show that the pointwise sum of two Cauchy sequences and the pointwise multiplication of a Cauchy sequence with a scalar is again Cauchy. If a sequence $\{f_k\}_{k=1}^{\infty} \subset V$ converges to $f \in V$ then it is Cauchy. This follows almost directly from the triangle inequality as

$$0 \leq \|f_n - f_m\| \leq \|f_n - f\| + \|f_m - f\|$$

and both terms on the right hand side can be bounded by $\epsilon/2 \;\; \forall \epsilon > 0$ by virtue of $\{f_k\}_{k=1}^{\infty}$ converging to f. More formally $\exists n_0 : n, m \geq n_0 \Rightarrow \|f_n - f\| < \epsilon/2$

and $\|f_m - f\| < \epsilon/2$ implying $\|f_n - f_m\| < \epsilon \ \forall n, m \geq n_0$ and therefore $\{f_k\}_{k=1}^{\infty}$ is Cauchy. The converse does not hold in general, i.e. there exist sequences that are Cauchy but do not converge. Any normed linear space V in which all Cauchy sequences $\{f_k\}_{k=1}^{\infty}$ converge to an element $f \in V$ in the sense of the norm topology is called complete. Completeness is a central ingredient in the following definition.

Definition 2.1.7 (Banach and Hilbert spaces) A complete normed linear space is called a Banach space. An inner product space that is also complete w.r.t. the induced norm is called a Hilbert space.

If a normed linear space / pre-Hilbert space V is given, its completion $\overline{V}$ [182, Section 3.8] is a Banach space / Hilbert space [181, p. 51], [181, p. 65].

§ Examples

The archetypal examples of Hilbert spaces are ℓ^2, the Hilbert space of square summable sequences and $L^2(T)$, the space of (equivalence classes of) square integrable functions on a set T. When $T = \mathbb{N}$ and the counting measure on T is used then unsurprisingly $L^2(T) = \ell^2$. Every Hilbert space $\mathcal{H}$ admitting a countable basis being in one to one correspondence to ℓ^2 is less intuitive but explains the pervasive importance of this particular sequence space.

Example 1 The space $\ell^2 := \{f = \{f_k\}_{k=1}^{\infty} : f_k \in \mathbb{C}, \sum_{k=1}^{\infty} |f_k|^2 < \infty\}$ together with pointwise addition of elements, scalar multiplication and inner product defined as follows is a Hilbert space [93, p. 23].

$$f + g = \{f_k + g_k\}_{k=1}^{\infty} \qquad f, g \in \ell^2 \tag{2.1}$$

$$\alpha f = \{\alpha f_k\}_{k=1}^{\infty} \qquad f \in \ell^2, \alpha \in \mathbb{C} \tag{2.2}$$

$$\langle f, g \rangle_{\ell^2} = \sum_{k=1}^{\infty} f_k \overline{g_k} \qquad f, g \in \ell^2 \tag{2.3}$$

Verification of the vector space axioms and the properties of $\langle \cdot, \cdot \rangle_{\ell^2}$ is straightforward. Helmberg [93, p. 24] gives a constructive argument to show completeness of ℓ^2. ■

Example 2 The space $L^2(T)$ of (equivalence classes of) square integrable functions on a measurable set $T \subset \mathbb{R}^n$ is a Hilbert space. $L^2(T) := \{f : T \to \mathbb{C} \text{ measurable } : \|f\|_{L^2} < \infty\}$ where the inner product is given as $\langle f, g \rangle_{L^2(T)} := \int_T f(t)\overline{g(t)} \, \mathrm{d}\, \lambda(t)$ w.r.t. integration against the standard Lebesgue measure λ on $\mathbb{R}^n$. A rigorous explanation of all involved terms including an account of the role of measurability can be found in [36, p. 131]. ■

Remark In this definition two functions f, g are equivalent if they differ only on a set of Lebesgue measure 0. This implies $\int_T |f(t) - g(t)|^2 \, \mathrm{d}\, \lambda = 0$ but f being equivalent to g in this mean square sense does not equate to f taking the same values as g everywhere. If this stronger notion of equivalence is desired one may invoke the ∞-norm: $\|f - g\|_\infty = \sup_{t\in T} |f(t) - g(t)|$ but the resulting norm does not satisfy the parallelogram law, is therefore not induced by any inner product, and $L^\infty(T)$ is consequently not a Hilbert space.

For the case where $T = [a, b] \subset \mathbb{R}$ is a finite closed interval on the real line [93, p. 33] proves that $L^2(T)$ is the completion of $\mathcal{C}(T) := \{f : T \to \mathbb{C} : f \text{ is continuous}\}$. One may introduce a different measure μ than the Lebesgue measure λ on $\mathbb{R}^n$. The spaces $L^2(T, \mu) := \{f : T \to \mathbb{C}\ \mu\text{-measurable} : \|f\|_{L^2(T,\mu)} < \infty\}$ together with the modified inner product $\langle f, g\rangle_{L^2(T,\mu)} := \int_T f(t)g(t)\mu(dt)$ and $\mu(T_{\text{sub}}) \geq 0\ \ \forall T_{\text{sub}} \subset T$ are Hilbert spaces in their own right [206, p. 205]. It is noted in passing that if $\mu = w\lambda$ with $w \geq 0$ some weight function then putting a norm constraint on functions in $L^2(T, \mu)$ will penalize functions f exhibiting high function values in regions $T_{\text{sub}} \subset T$ where $w(t), t \in T_{\text{sub}}$ is extraordinarily high and amounts to the inclusion of prior knowledge. The practicing geodesist already sees a similarity between the measures introduced here and the precision matrix employed in classical adjustment theory.

The space $L^2(T)$ is of central importance in this monograph. As was alluded to in the previous comment, $L^2(T)$ provides an infinite dimensional setting for least squares problems where the objects to be estimated are functions and not finite dimensional vectors. Hilbert space methods involving $L^2(T)$ are also prevalent in solving linear differential equations as often encountered when the dynamics of a physical system is sought. This is desirable in the context of estimating quantities subjected to physical constraints — from gravity fields in physical geodesy to bending plates and buckling beams in engineering geodesy. When T is not just an arbitrary subset of $\mathbb{R}^n$ on which functions are defined but instead the phase space $Q \times P$ of generalized coordinates and generalized momenta of a physical system, unit vectors $f \in \mathbb{S} \subset L^2(T)$ can be related to probability densities $\rho(\cdot, \cdot)$ on that systems phase space. By setting $\rho(q, p) = |f(q, p)|^2$ one recovers a probabilistic formulation of mechanics reminiscent to the Born rule fundamental to quantum mechanics, [135, p. 34]. The time evolution of a systems state is then determined by the Liouville operator instead of the Hamiltonian; more is found in subsection 2.2.4. For now a simpler approach is used to establish a relationship between Hilbert spaces and stochastic quantities. This is the content of the next example.

Example 3 Given a probability space (Ω, Σ, P) [9, p. 12], denote by $L^2(\Omega, \Sigma, P)$ all random variables $f : \Omega \to \mathbb{R}$ satisfying $E[f] = 0$ and $E[f^2] < \infty$. Here $E[\cdot]$ denotes the expectation operator, Ω is the sample space, Σ is a σ-algebra of events and P is a probability measure. Then $L^2(\Omega, \Sigma, P)$ together with the inner product

$$\langle f, g\rangle = E[fg]\ \ \forall f, g \in L^2(\Omega, \Sigma, P)$$

and obvious addition and multiplication is a Hilbert space [159, p. 18]. By applying the Cauchy-Schwarz inequality one calculates immediately

$$|\langle f, g\rangle|^2 = |E[fg]|^2 \leq E[f^2]E[g^2] = \|f\|^2\|g\|^2$$

This implies the correlation coefficient $\rho = \langle f, g\rangle/\|f\|\|g\|, \rho \in [-1, 1]$ as a suitable normalized measure of linear stochastic dependence of f on g and vice versa. Comparing this to equation 1.6 ρ is seen to be the cosine between f and $g \in L^2(\Omega, \Sigma, P)$. The inner product

$$\langle f, g\rangle = E[fg] = \int_\Omega fgdP = \int_\mathbb{R} fgdP_{(fg)} = \int_{\mathbb{R}^2} fgdP_{(f,g)}$$

identifies random variables $f, g \in L^2(\Omega, \Sigma, P)$ if $\|f - g\| = 0$, i.e. $f - g = 0 \Leftrightarrow \int_\Omega |f - g|^2 dP = \int_{\mathbb{R}^2} |f - g|^2 dP(f, g) = 0$ which implies that f and g are allowed to differ only on a set of measure 0 to be considered the same under the notion of equivalence induced by the norm. In the above, $P(fg)$ denotes $P \circ (fg)^{-1}$ and $P(f, g)$ is the measure $P \circ (f, g)^{-1}$ so that $P(fg)(B_1) = P(\omega \in \Omega : f(\omega)g(\omega) \in B_1)$ and $P(f, g)(B) = P(\omega \in \Omega : (f(\omega), g(\omega)) \in B_2)$ for B_1 and B_2 in the Borel σ-algebras of $\mathbb{R}^1$ and $\mathbb{R}^2$ respectively [36, p. 173]. ■

With the space $L^2(\Omega, \Sigma, P) =: L^2(\Omega)$ a notion of randomness was introduced for the first time in this monograph. Hilbert spaces of finite variance random variables are of prime importance in data analysis and three specific examples of their applications to modelling and inference for time series are given below.

Example 4 (White noise process) Given a finite set T, let for any $t \in T$ $N_t \in L^2(\Omega)$ be a random variable of unit variance that furthermore satisfies

$$E[N_s N_t] = \delta_{ts} \;\; \forall t, s \in T.$$

Then the sequence $\{N_t\}_{t \in T}$ is said to be (weakly) white noise [140, p. 213] where the adjective in parentheses is usually omitted. This convention will also be adopted here - if the white noise follows a Gaussian distribution, this will be explicitly mentioned.

Apart from thermally induced noise, few real-world processes exhibit random behavior well modelled by white noise. It is rather an elementary building block helpful for constructing models of autocorrelated processes. If T is interpreted as discretized time $T \subset \mathbb{Z}$ this can be done in at least two ways: Define a new process X_t either as a weighted sum of white noise variables or as a weighted sum of its preceding process variables X_{t-k} $k \geq 1$ and a random perturbation that is known as the innovation term. In the former case one arrives at what are called moving average processes in time series analysis and in the latter case autoregressive processes result. ■

Remark Typically the symbol W_t is reserved for the Wiener process at time t and white noise — defined as the time derivative of the Wiener process in the distributional sense — is simply denoted by dW_t. To not overcomplicate matters in this first example, the initially introduced symbolism will be kept and later changed to adhere to standard nomenclature.

Some realizations of white noise, moving average and autoregressive processes are plotted in figure 2.1. More rigorous definitions and proofs related to stochastic processes can be found in chapter 3.

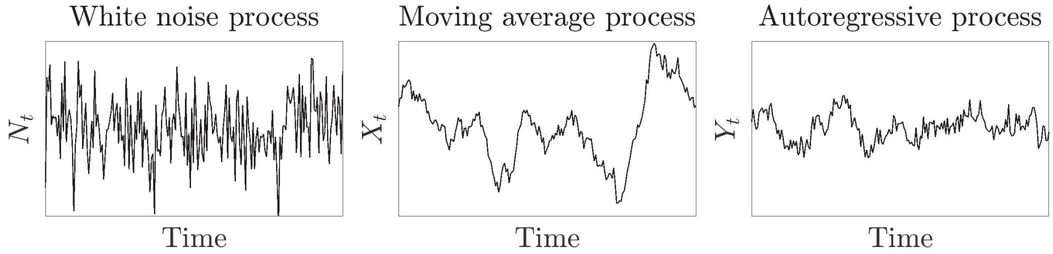


Figure 2.1: Two realizations of the three exemplary processes introduced above. The model coefficients $\alpha := \{\alpha_k\}$ are all $1/10$ for the MA(10) and $\{0.6, 0.4, 0.2, 0.1\}$ for the AR(4).

As the monograph progresses, the following three problems will be encountered regularly in different disguises. As the catalog of available theoretical results grows, they will be upgraded to more generality and then revisited.

Problem I : (Interpolation) Given some values $\{x_t\}_{t\in T_{\text{sub}}}$, $T_{\text{sub}} \subset T$ of some realization of process $\{X_t\}_{t\in T}$ estimate the missing values $\{x_t\}_{t\in T\setminus T_{\text{sub}}}$.

Problem II : (Separation) Given some noisy measurements $\{y_t\}_{t\in T} = \{x_t + n_t\}_{t\in T}$ where n_t are realizations of a white noise process, estimate the values $\{x_t\}_{t\in T}$.

Problem III : (Representation) Given values $\{x_t\}_{t\in T}$ of some realization of a process $\{X_t\}_{t\in T}$ provide a meaningful and efficient representation of its main features in terms of elementary or non-elementary functions.

Figure 2.2 illustrates the simplest instances of Problems I, II and III very crudely. Later iterations of these problems feature the full spatiotemporal setting including physical constraints.

Banach spaces that are not at the same time Hilbert spaces will not be approached in the context of application to signal processing until chapter 4, in which the computational foundations necessary to deal with optimization problems in them are laid out. Nonetheless, for reasons of motivation some examples of Banach spaces are given below. Note that by definition all of the above examples for Hilbert spaces are also examples of Banach spaces.

Example 5 The L^p spaces for $1 \leq p < \infty$ of equivalence classes of functions f from a measurable set $T \subset \mathbb{R}^n$ to the complex numbers satisfying a finiteness

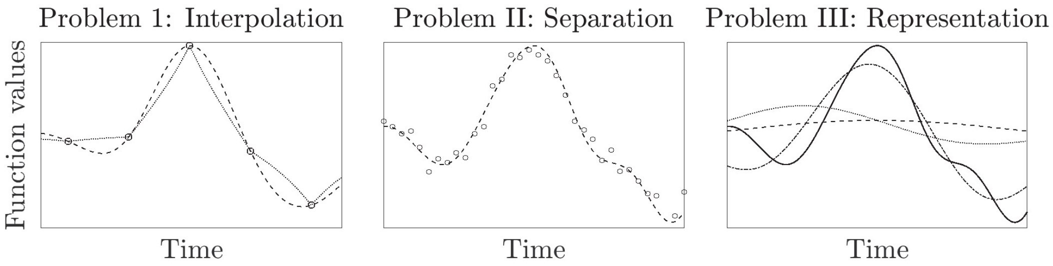


Figure 2.2: One dimensional interpolation, signal separation and representation problems. Unbroken lines or circles are problem data and the broken lines are part of a possible solution. They can be found by employing Hilbert space methods in which the inner product is modified to act on whitened data, see section 3.1

property similar to the L^2 spaces are Banach spaces. They are defined as

$$L^p(T) := \{f : T \to \mathbb{C} \text{ measurable } : \|f\|_p < \infty\}$$

where the norm is $\|f\|_p = \left[\int_T |f|^p d\lambda\right]^{1/p}$ [22, p. 46]. ■

Figure 2.3 illustrates different elements of the unit spheres $\mathbb{S}_p := \{f \in L^p(T) : \|f\|_p = 1\}$ with $T = \mathbb{R}$ for different p.

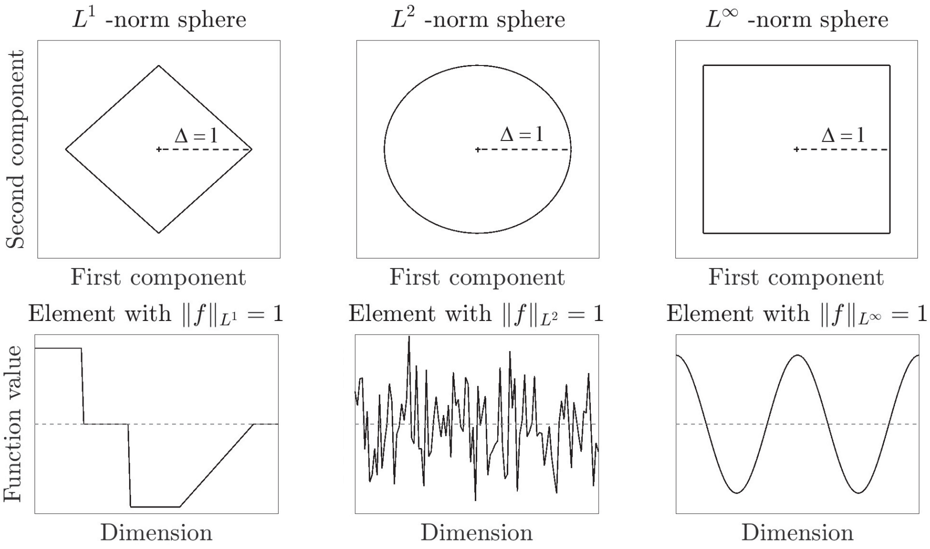


Figure 2.3: The boundaries of three different open balls $B_1(f_0)$ corresponding to different norms on the vector space $\mathbb{R}^2$. Typical unit norm elements of $L^p(\mathbb{R})$ are plotted in the second row, where vaguely a function f is called typical for L^{p_*} if it has a smaller norm in L^{p_*} than in any of the other two L^p spaces for $p \neq p_*$. $\Delta = 1$ is the coordinate difference between the ball's center point and its rightmost border; it is numerically equal to an analogously defined coordinate difference in the direction of the second component.

Since $\|f\|_p$ is a norm $\rho_p(f) = c\exp(-\|f\|_p)$ is a valid probability density function for c a normalizing constant. For $p = 1$ one recovers a multidimensional Laplacian distribution, whereas $p = 2$ corresponds to a multivariate Gaussian. The limiting case of $p \to \infty$ has no specific name to the best of the authors knowledge. The specific form of those probability density functions and their level sets has important implications: Minimization of the ℓ^1-norm promotes sparse solutions, the ℓ^2-norm minimizers are maximum likelihood estimators for Gaussian distributed random variables and the sup-norm measures worst-case performance. More is to be found in chapter 4 in subsection 4.4.1 dealing with applications of ℓ^p-norm based estimation.

2.1.2 Geometry of Hilbert spaces

Norms and inner products are related to each other via the polarization identity and the parallelogram identity. They furthermore satisfy the triangle, reverse triangle and Cauchy-Schwarz inequality and can be used to imbue the set of subspaces of a Hilbert space with additional structure by measuring the angle between them. Subspaces that are orthogonal to each other satisfy certain optimality properties finding their expression in simple construction rules for minimal-norm decompositions of their elements. A complete decomposition of a Hilbert space $\mathcal{H}$ into a sequence of orthogonal subspaces $\{\mathcal{H}_k\}_{k=1}^{\infty}$ can be achieved with the help of an orthonormal basis for $\mathcal{H}$. If the elements of $\mathcal{H}$ are identified with functions in the time domain (=”signals”), the representation in a certain basis can be interpreted as a signal expansion.

§ Subspaces

The classical equations and inequalities are often ingredients to proofs and provide tools for investigating the global algebraic structure of Hilbert spaces. However, to solve the canonical problems I-III stated on page 21, the idea of decomposition is central: either data is to be split into a noise and signal part or a signal needs to be written as a superposition of elementary waves. The decomposition of Hilbert space elements $f \in \mathcal{H}$ presupposes decomposability of the Hilbert space $\mathcal{H}$ itself. Unless specified differently, $\mathcal{H}$ will be a Hilbert space over the complex numbers in what follows during the rest of this section.

Definition 2.1.8 If a (nonempty) subset $\mathcal{M}$ of the Hilbert space $\mathcal{H}$ is closed under linear operations, i.e. if $f + g \in \mathcal{M} \;\; \forall f, g \in \mathcal{M}$ and $\alpha f \in \mathcal{M} \;\; \forall f \in \mathcal{M}, \alpha \in \mathbb{C}$ then it is said to be a linear manifold in $\mathcal{H}$ [93, p. 36].

The definition here applies to Hilbert spaces over the field of complex numbers. For an analogous definition suitable to cover the real case, just relax the requirement regarding closedness under multiplication by scalars to $\alpha f \in \mathcal{M} \;\; \forall f \in \mathcal{M}, \alpha \in \mathbb{R}$. As $-1 \in \mathbb{R} \subset \mathbb{C}$, one always has $f + (-1)f = 0$ as an element of any linear manifold $\mathcal{M}$ in $\mathcal{H}$ and may visualize $\mathcal{M}$ as a hyperplane through the origin of $\mathcal{H}$. This motivates the nonstandard notation $\mathcal{M} \boxslash \mathcal{H}$ for ”$\mathcal{M}$ is a linear manifold in $\mathcal{H}$”. For all Hilbert spaces $\mathcal{H}$, clearly $\{0\} \boxslash \mathcal{H}$ and $\mathcal{H} \boxslash \mathcal{H}$. A linear manifold $\mathcal{M}$ in a Hilbert space is not necessarily a Hilbert space in its own right as it may fail to be complete. A condition precluding erratic behavior of this type is the topological closure of $\mathcal{M}$ in $\mathcal{H}$.

Definition 2.1.9 A topologically closed and therefore complete linear manifold $\mathcal{H}_1$ in a Hilbert space $\mathcal{H}$ is called a subspace and this relation will be denoted by $\mathcal{H}_1 \overline{\boxslash} \mathcal{H}$. If two subspaces $\mathcal{H}_1, \mathcal{H}_2$ of $\mathcal{H}$ satisfy $\langle f, g \rangle = 0 \;\; \forall; f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$ they are said to be orthogonal subspaces and one writes $\mathcal{H}_1 \perp \mathcal{H}_2$ to indicate this relationship.

The topological closure of $\mathcal{H}_1 \boxslash \mathcal{H}$ implies that $\mathcal{H}_1$ contains all its limit points [141, p. 55], i.e. those $f \in \mathcal{H}_1$ with $(U \setminus \{f\}) \cap \mathcal{H}_1 \neq \emptyset$ for all neighborhoods U containing f [108, p. 176]. $\mathcal{H}_1$ is closed in $\mathcal{H}$ iff for any sequence $\{f_k\}_{k=1}^{\infty} \subset \mathcal{H}_1, \lim_{k\to\infty} f_k = f$ implies $f \in \mathcal{H}_1$ [141, p. 55] and $\mathcal{H}_1$ as a subset of a complete normed space $\mathcal{H}$ is closed iff it is complete. A mere linear manifold $\mathcal{H}_1 \boxslash \mathcal{H}$ can be closed by adjoining its limit points.

The above remarks can be augmented with a verification of the closure's closure not only in the topological sense but also under linear operations. It then follows that for any linear manifold $\mathcal{M}$ in a Hilbert space $\mathcal{H}$, its closure $\mathrm{cl}(\mathcal{M}) \overline{\boxslash}\mathcal{H}$, i.e. $\mathrm{cl}(\mathcal{M})$ is a subspace. Because the notation $\mathrm{cl}(\mathcal{M})$ can be cumbersome and in the case of subspaces closure and completion are almost interchangeable, the closure of a subset of $\mathcal{H}$ will also be denoted by an overbar. From now on $\overline{\mathcal{M}} := \mathrm{cl}(\mathcal{M})$ and if a completion of $\mathcal{M}$ is to be constructed this is mentioned explicitly.

§ Orthogonal complements and multiple subspaces

Two operations are available to combine subspaces and linear manifolds to provide new objects. The ordinary vector sum of two subspaces consists of all sums $f + g$ where f and g are from the respective subspaces. Due to topological complications the vector sum of two subspaces rarely forms a subspace and its closure has to be taken instead. For a given subset U of a vector space $\mathcal{H}$, the shorthand notation $\bigvee U$ will stand for $\overline{\mathrm{span}}\, U$ and $U_1 \bigvee U_2$ is used to symbolize $\overline{\mathrm{span}}\, U_1 \cup U_2$ where $\overline{\mathrm{span}}\, U$ denotes the closure of the linear span generated by all finite linear combination of elements in U as described for example in [172, p. 104]. The closure of the span of countable sequences of subsets is denoted by $\bigvee_{k=1}^{\infty} U_k = \overline{\mathrm{span}} \bigcup_{k=1}^{\infty} U_k$.

Definition 2.1.10 For any two linear manifolds $\mathcal{H}_1$ and $\mathcal{H}_2$ in $\mathcal{H}$ the set

$$\mathcal{H}_1 + \mathcal{H}_2 := \{f \in \mathcal{H} : f = f_1 + f_2, f_1 \in \mathcal{H}_1 \text{ and } f_2 \in \mathcal{H}_2\}$$

is called the vector sum of $\mathcal{H}_1$ and $\mathcal{H}_2$. For countable sequences $\{\mathcal{H}_k\}_{k=1}^{\infty}$ of linear manifolds the symbol $\sum_{k=1}^{\infty} \mathcal{H}_k$ denotes the set $\{f \in \mathcal{H} : f = \sum_{k=1}^{\infty} f_k, f_k \in \mathcal{H}_k\}$.

The vector sum of two linear manifolds is again a linear manifold. When $\mathcal{H}_k \overline{\boxslash}\mathcal{H}$ for $k = 1, ...$ one can show that the relationship

$$\mathrm{cl}\left(\sum_{k=1}^{\infty} \mathcal{H}_k\right) = \bigvee_{k=1}^{\infty} \mathcal{H}_k$$

holds [93, p. 39]. Then $\mathrm{cl}\left(\sum_{k=1}^{\infty} \mathcal{H}_k\right) \overline{\boxslash}\mathcal{H}$. When two subspaces $\mathcal{H}_1$ and $\mathcal{H}_2$ are orthogonal to each other and for any subset $F \subset \mathcal{H}$ of a Hilbert space $\mathcal{H}$, $F^{\perp}$ is used to denote the orthogonal complement $F^{\perp} := \{g \in \mathcal{H} : \langle f, g\rangle_{\mathcal{H}} = 0 \;\; \forall f \in F\}$ of F in $\mathcal{H}$, then the following collection of statements holds [93, pp. 40-47].

Theorem 2.1.11 *For $\mathcal{H}_1 \perp \mathcal{H}_2$, $\mathcal{H}_1 \overline{\boxslash} \mathcal{H}$ and $\mathcal{H}_2 \overline{\boxslash} \mathcal{H}$ as well as $F \subset \mathcal{H}$ one finds*

I $(\mathcal{H}_1 + \mathcal{H}_2) \overline{\boxslash} \mathcal{H}$

II Every element $f \in \mathcal{H}_1 + \mathcal{H}_2$ has a unique decomposition $f = f_1 + f_2$ where $f_1 \in \mathcal{H}_1$ and $f_2 \in \mathcal{H}_2$.

III $F^\perp \overline{\boxslash} \mathcal{H}$

IV $(F^\perp)^\perp \overline{\boxslash} \mathcal{H}$ and since $(F^\perp)^\perp := \{g \in \mathcal{H} : \langle g, f \rangle = 0 \, \forall f \in F^\perp\}, (F^\perp)^\perp \supset F$, i.e. it is a subspace containing F.

V $(\mathcal{H}_1^\perp)^\perp = \mathcal{H}_1$

The minimum distance between a point f in a Hilbert space $\mathcal{H}$ and one of its subspaces $\mathcal{H}_1$ is $\delta_f = \inf_{g \in \mathcal{H}_1} \|f - g\|$. As is proven in [3, pp. 8-9] there exists a unique vector $P_{\mathcal{H}_1} f$ in $\mathcal{H}_1$ closest to $f \in \mathcal{H}$ which means that $\|P_{\mathcal{H}_1} f - f\| = \delta_f$. It will be called the projection $P_{\mathcal{H}_1} f$ of f onto $\mathcal{H}_1$. The following theorem is stated without proof. Helmberg [93, pp. 38-47] and the introduction to Akhiezer and Glazmans treatise on linear operators in Hilbert spaces [3] provide further details.

Theorem 2.1.12 *I If $P_{\mathcal{H}_1} f$ denotes the projection of f onto $\mathcal{H}_1 \overline{\boxslash} \mathcal{H}$ then $P_{\mathcal{H}_1} f - f \in \mathcal{H}_1^\perp$.*

II If $f \in \mathcal{H}_1$ then $P_{\mathcal{H}_1^\perp} f = 0$ since $0 = \langle f, P_{\mathcal{H}_1^\perp} f \rangle = \langle P_{\mathcal{H}_1} f + P_{\mathcal{H}_1^\perp} f, P_{\mathcal{H}_1^\perp} f \rangle = \|P_{\mathcal{H}_1^\perp}\|^2$ and $P_{\mathcal{H}_1^\perp} f = 0$ by the non-degeneracy of the norm.

III (Projection theorem) If $\mathcal{H}_1 \overline{\boxslash} \mathcal{H}$ then $\mathcal{H}_1 + \mathcal{H}_1^\perp = \mathcal{H}$.

IV Given $\mathcal{H}_1 \overline{\boxslash} \mathcal{H}$, $\exists!$ decomposition of any $f \in \mathcal{H}$ into $f = f_1 + f_2$ where $f_1 \in \mathcal{H}_1$ and $f_2 \in \mathcal{H}_1^\perp$.

Example 6 Let $\mathcal{H} = L^2[0,1]$ be the Hilbert space of square integrable complex valued functions on $[0,1]$ and $U_k := \{\exp(2\pi i k \cdot), \exp(-2\pi i k \cdot)\}$[1]. Define $\mathcal{H}_1^m$ as

$$\mathcal{H}_1^m := \bigvee_{k=1}^{m} U_k = \overline{\text{span}} \bigcup_{k=1}^{m} \{\exp(\pm 2\pi i k \cdot)\}$$

and $\mathcal{H}_2^m$ as $(\mathcal{H}_1^m)^\perp$. $\mathcal{H}_1^m \overline{\boxslash} \mathcal{H}$ and $\mathcal{H}_2^m \overline{\boxslash} \mathcal{H}$ by construction. $\mathcal{H}_1^m$ corresponds to the subspace of functions with frequencies up until m oscillations per unit interval. All the high frequency dominated members of $L^2[0,1]$ are subsumed in $\mathcal{H}_2^m$. Projecting $f \in \mathcal{H}$ onto $P_{\mathcal{H}_1^m} f \in \mathcal{H}_1^m$ is the same as performing low pass filtering. The decomposition of $f \in \mathcal{H}$ into two different components $f_1 \in \mathcal{H}_1^m$ and $f_2 \in \mathcal{H}_2^m$ is illustrated in figure 2.4.

If one knew that f was noisily observed with the noise being purely high frequency

[1] By replacing a function's argument by a dot and therefore referring to a function f on T as $f(\cdot)$ instead of $f(t)$, one emphasizes that the function f is not simply a collection of values $f(t)$ associated to index elements $t \in T$ but that the whole rule of assigning values to index elements is the object of consideration.

and the signal purely low frequency as measured via oscillations per unit interval, then one would have already derived a solution to problem II posed on page 21. ■

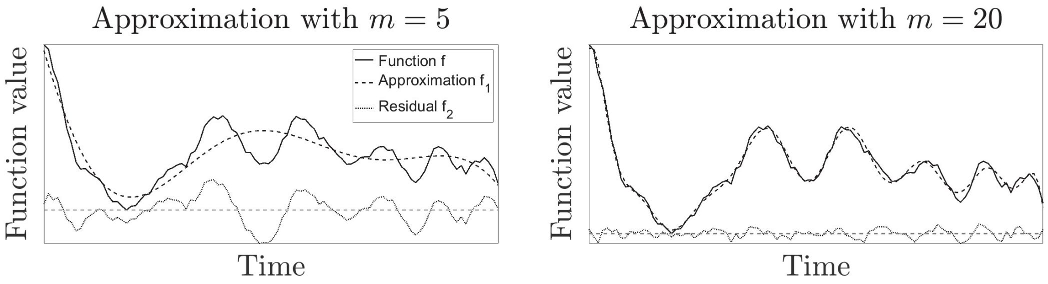


Figure 2.4: On both sides of this figure a decomposition of $f \in L^2[0,1]$ based on the two subspaces $\mathcal{H}_1^m$ and $\mathcal{H}_2^m$ is shown. On the left side $m = 5$ and only low frequent functions are found in $\mathcal{H}_1^m$ whereas on the right side $m = 20$ and $\mathcal{H}_1^m$ contains functions with much less smooth behavior. The indexset $T = [0,1]$ is interpreted as time.

Another way of writing $P_{\mathcal{H}_1} f$ is as the unique minimizer in $\mathcal{H}_1$ of $\|g - f\|$ for some $f \in \mathcal{H}$, that is

$$P_{\mathcal{H}_1} f = \underset{g \in \mathcal{H}_1}{\operatorname{argmin}} \ \|g - f\| \tag{2.4}$$

or in other words $P_{\mathcal{H}_1} f$ is that element $g \in \mathcal{H}_1$ whose deviation from some target vector $f \in \mathcal{H}$ has minimum length.

Theorem 2.1.13 *Let* $\{\mathcal{H}_k\}_{k=1}^{\infty}$ *be such that* $\mathcal{H}_k \,\overline{\boxslash}\mathcal{H} \;\; \forall k \in \mathbb{N}$ *and* $\mathcal{H}_k \perp \mathcal{H}_l \;\; \forall k \neq l$*. Then* $(\sum_{k=1}^{\infty} \mathcal{H}_k) \,\overline{\boxslash}\mathcal{H}$ *and* $\forall f \in \sum_{k=1}^{\infty} \mathcal{H}_k \;\; \exists! f_k \in \mathcal{H}_k : \sum_{k=1}^{\infty} f_k = f$*.*

As a tribute to the special situation encountered when dealing with mutually orthogonal subspaces $\mathcal{H}_k \,\overline{\boxslash}\mathcal{H}$ — for which it always holds that $\sum_{k=1}^{n} \mathcal{H}_k \,\overline{\boxslash}\mathcal{H}$ — often the symbol $\bigoplus$ is used instead of Σ to denote the vector sum. The choice of symbols is no coincidence and it can be shown (see subsection 3.1.3) that the vector sum $\mathcal{H}_1 + \mathcal{H}_2$ of two orthogonal subspaces $\mathcal{H}_1 \perp \mathcal{H}_2$ is the same as the external direct sum $\mathcal{H}_1 \oplus \mathcal{H}_2$ known from abstract algebra. The general situation, in which the $\mathcal{H}_k$ do not need to be contained in an apriori known auxiliary Hilbert space $\mathcal{H}$ will be cleared up in subsection 2.1.3.

§ Bases

Several alternative but in the end interchangeable approaches to define bases for infinite dimensional Hilbert spaces exist. The focus will be on Hilbert spaces, for which a countably infinite orthonormal basis can be found as the general situation without a constraint of this type admits many pathologies and requires a more advanced apparatus than space permits to present here.

Definition 2.1.14 I A Hilbert space $\mathcal{H}$ is said to be separable iff there exists a countable, everywhere dense subset $H \subset \mathcal{H}$, i.e. $\mathrm{cl}(H) = \mathcal{H}$ or in other words $\forall f \in \mathcal{H}$ and $\forall \epsilon > 0 \;\; \exists g \in H : g \in B_\epsilon(f)$[3, p. 5].

II An orthonormal system (ONS) $\{e_k\}_{k\in K} \subset \mathcal{H}$ is called maximal in $\mathcal{H}$ if there does not exist a nonzero vector $f \in \mathcal{H}$ with $\langle f, e_k\rangle = 0\, \forall k \in K$[3, p. 17].

Theorem 2.1.15 *I An orthonormal system $\{e_k\}_{k\in K}$ is linearly independent.*

II If $\mathcal{H}$ is separable then for every orthonormal system $\{e_k\}_{k\in K} \subset \mathcal{H}$, K is finite or countably infinite.

III $\mathcal{H}$ is separable if and only if there exists a maximal orthonormal sequence $\{e_k\}_{k=1}^{\infty} \subset \mathcal{H}$.

IV Any two maximal orthonormal systems in a Hilbert space $\mathcal{H}$ (separable or not) have the same cardinal number; this cardinal number is also called the dimension of $\mathcal{H}$.

Corollary 2.1.15.1 *Since $\{e_k\}_{k=1}^{\infty}$ with $e_k = (0, ...0, 1, 0, ...)$ with a 1 at the k-th position is countably infinite and dense in ℓ^2, ℓ^2 is separable.*

The proof of I is trivial — $\{e_k\}_{k\in K}$ ONS and $\sum_{k\in K} \alpha_k e_k = 0 \Rightarrow \| \sum_{k\in K} \alpha_k e_k\|^2 = 0$ but this is equivalent to $\sum_{k\in K} |\alpha_k|^2 = 0$ which is only the case if the α_k identically vanish — and the proofs of II,III,IV are found in [3, pp. 17-20].

Part II of theorem 2.1.15 asserts that if one finds a maximal orthonormal sequence in $\mathcal{H}$, $\mathcal{H}$ is separable but it also says that if one starts with an orthonormal sequence $\{e_k\}_{k=1}^{\infty}$ and then takes the subspace $\mathcal{H}_1 := \bigvee_{k=1}^{\infty} = \overline{\text{span}\ \bigcup_{k=1}^{\infty}\{e_k\}}$, $\mathcal{H}_1$ is separable by construction. As orthonormal systems in inseparable Hilbert spaces feature uncountably many elements and are both less well investigated and from the perspective of practical computability less tame than the systems for separable Hilbert spaces, inseparable Hilbert spaces are disregarded for the rest of the monograph. Since ℓ^2 and $L^2(T)$ as the most relevant Hilbert spaces encountered in the context of this monograph are separable, this omission is uncritical. It is also for this special case that the definition of a basis is given.

Definition 2.1.16 For any Hilbert space $\mathcal{H}$ an orthonormal sequence $\{e_k\}_{k=1}^{\infty} \subset \mathcal{H}$ is called a basis of $\mathcal{H}$ iff the only vector in $\mathcal{H}$ orthogonal to all $e_k, k = 1, ...$ is the zero vector 0. Often the acronym ONB (orthonormal basis) is used.

This definition is standard [3, p. 19] for infinite dimensional Hilbert spaces but stricter than the classical linear algebraic notions of bases encountered in finite dimensional settings due to the additional requirement of orthonormality. A more liberal definition would not incorporate any extra properties of the sequence $\{e_k\}_{k=1}^{\infty}$. One such relaxed notion of basis is called an exact frame [40, p. 121] but it is rarely used outside of applied harmonic analysis and compressive sensing.

Theorem 2.1.17 (Basis theorem) *The following six statements are equivalent [40, p. 79].*

I $\{e_k\}_{k=1}^{\infty}$ is a basis of $\mathcal{H}$.

II *If* $f \in \mathcal{H}$ *and* $\langle f, e_k \rangle = 0 \ \forall k \in \mathbb{N}$ *then* $f = 0$.

III $\mathcal{H} = \overline{\text{span}} \cup_{k=1}^{\infty} \{e_k\}$.

IV *If* $f \in \mathcal{H}$ *then* $f = \sum_{k=1}^{\infty} \langle f, e_k \rangle e_k$.

V *If* $f \in \mathcal{H}$ *then* $\langle f, g \rangle = \sum_{k=1}^{\infty} \langle f, e_k \rangle \overline{\langle g, e_k \rangle}$.

VI *If* $f \in \mathcal{H}$ *then* $\|f\|^2 = \sum_{k=1}^{\infty} |\langle f, e_k \rangle|^2$.

Remark The construction in statement IV is known as the Fourier expansion in terms of the ONB $\{e_k\}_{k=1}^{\infty}$ and statement VI is also often called Parsevals identity. From VI Bessels inequality $\|f\|^2 \geq \sum_{k=1}^{n} |\langle f, e_k \rangle|^2$ immediately follows.

Corollary 2.1.17.1 (Optimal approximation) *Let* $\{e_k\}_{k=1}^{\infty}$ *be an ONB in a separable Hilbert space* $\mathcal{H}$ *and let* $\mathcal{H}_n^{\Sigma} = \bigoplus_{k=1}^{n} \mathcal{H}_k$ *where the* $\mathcal{H}_k$ *are again the subspaces associated to the* e_k; $\mathcal{H}_k = \overline{\text{span}} \{e_k\}$. *If the optimal approximation to* $f \in \mathcal{H}$ *by elements of* $\mathcal{H}_n^{\Sigma}$ *is denoted by* f^{*n}, *that is*

$$f^{*n} = \underset{g \in \mathcal{H}_n^{\Sigma}}{\text{argmin}} \ \|f - g\|_{\mathcal{H}} = P_{\mathcal{H}_n^{\Sigma}} f$$

then the basis theorem implies the following three instructive consequences.

i) *The projection of* f *onto* $\mathcal{H}_n^{\Sigma}$ *can be written explicitly as* $P_{\mathcal{H}_n^{\Sigma}} f = \sum_{k=1}^{n} \langle f, e_k \rangle e_k$.

ii) *If* $n \geq m$ *then* f^{*n} *approximates* f *better or equally well as* f^{*m}.

iii) *If* $n > m$ *then* $f^{*n} - f^{*m} \in \bigoplus_{k=m+1}^{n} \mathcal{H}_k$.

In the preceeding statements, approximation quality is measured by the norm $\|f^{*n} - f\|_{\mathcal{H}}$ *of the approximation residuals with higher norm indicating lower quality.*

Proof: Once i) is shown ii) and iii) follow trivially.

i) Show that if f^{*n} had any other coefficients α_k than $\langle f, e_k \rangle$, the norm of the residual $f - f^{*n}$ would increase.

$$\begin{aligned}
\|f - f^{*n}\|_{\mathcal{H}}^2 &= \| \sum_{k=1}^{\infty} \langle f, e_k \rangle e_k - \sum_{j=1}^{n} \alpha_j e_j \|_{\mathcal{H}}^2 \\
&= \| \sum_{k=1}^{n} (\langle f, e_k \rangle - \alpha_k) e_k \|_{\mathcal{H}}^2 + \| \sum_{k=n+1}^{\infty} \langle f, e_k \rangle e_k \|_{\mathcal{H}}^2 \\
&= \|\{\langle f, e_k \rangle - \alpha_k\}_{k=1}^{n}\|_{\ell^2}^2 + \|\{\langle f, e_k \rangle\}_{k=n+1}^{\infty}\|_{\ell^2}^2
\end{aligned}$$

But both right hand side terms are nonnegative and only the left one does depend on the expansion coefficients. Since it is zero iff $\alpha_k = \langle f, e_k \rangle, k = 1, ..., n$ this choice of expansion coefficients is optimal. By the non degeneracy of the norm it is the only one to achieve that, hence unique.

ii) $$\|f - f^{*n}\|^2_{\mathcal{H}} = \sum_{k=n+1}^{\infty} |\langle f, e_k\rangle|^2 \le \sum_{k=n+1}^{\infty} |\langle f, e_k\rangle|^2 + \sum_{k=m+1}^{n} |\langle f, e_k\rangle|^2 = \|f - f^{*m}\|^2_{\mathcal{H}}$$

iii) $$\begin{aligned} f^{*n} - f^{*m} &= \sum_{k=1}^{n}\langle f, e_k\rangle e_k - \sum_{k=1}^{m}\langle f, e_k\rangle e_k \\ &= \sum_{k=m+1}^{n}\langle f, e_k\rangle e_k \in \bigoplus_{k=m+1}^{n} \mathcal{H}_k \end{aligned}$$

□

One might originally be unsure if adding more terms by which $f \in \mathcal{H}$ is approximated could potentially decrease the quality of the approximation in some way. Statement ii) asserts that the approximation quality monotonically increases if the terms e_k form an ONB. The implications of statement iii) are particularly clear if $n = m+1$. Then $f^{*m+1} = f^{*m} + \langle f, e_{m+1}\rangle e_{m+1}$ and the constancy of the first m expansion coefficients $\{\alpha_k\}_{k=1}^{m}$ in $f^{m}_{\text{approx}} = \sum_{k=1}^{m} \alpha_k e_k$ is guaranteed when the degree of approximation is increased by one or in fact by way of induction by any natural number. Figure 2.5 illustrates these comments and the fact that approximation quality depends on the ONB.

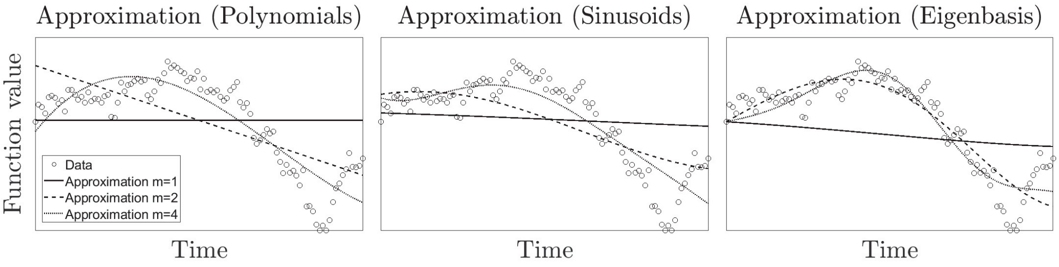


Figure 2.5: The three plots depict approximations of a fixed function f given in terms of data points with different orthonormal bases of $L^2[0,1]$. From the left to the right orthogonal polynomials, complex polynomials and a basis chosen numerically from the set of eigenfunctions of a certain linear operator (the covariance operator, see subsection 2.1.4) were used. The RMSE for f^{*4} are 0.25, 0.24 and 0.19 respectively.

Corollary 2.1.17.2 *(Riesz-Fischer) If $\mathcal{H}$ is separable with orthonormal basis $\{e_k\}_{k=1}^{\infty}$ then $\mathcal{H} = \{\sum_{k=1}^{\infty} \alpha_k e_k : \{\alpha_k\}_{k=1}^{\infty} \in \ell^2\}$.*

Proof: If $f \in \mathcal{H}$ then $\|f\|^2 < \infty$ and $\|f\|^2 = \sum_{k=1}^{\infty} |\langle f, e_k\rangle|^2 = \sum_{k=1}^{\infty} |\alpha_k|^2 < \infty$ implying $\{\alpha_k\}_{k=1}^{\infty} \in \ell^2$. If $f = \sum_{k=1}^{\infty} \alpha_k e_k : \{\alpha_k\}_{k=1}^{\infty} \in \ell^2$ then $\sum_{k=1}^{\infty} |\alpha_k|^2 < \infty$ proves $\sum_{k=1}^{\infty} |\langle f, e_k\rangle|^2 = \|f\|^2 < \infty$ and $f \in \mathcal{H}$. □

The Riesz-Fischer theorem states in a very formal sense that there is only one separable Hilbert space ℓ^2 and all other infinite dimensional separable Hilbert spaces $\mathcal{H}$ can be mapped one to one onto ℓ^2 by identifying the expansion coefficients $\{\alpha_k\}_{k=1}^{\infty}$ in some given orthonormal basis $\{e_k\}_{k=1}^{\infty}$ with the sequence $\{\alpha_k\}_{k=1}^{\infty}$. This seems to be a very specific theorem with limited applicability but as for example the sequence $\{e_k\}_{k=1}^{\infty}$ of elements $e_k = \exp(2\pi k\cdot)$ is a countable basis for $L^2[0,1]$ [93, p. 48] this space is seen to be isomorphic to ℓ^2 as well. This is not obvious intuitively since an $f \in \ell^2$ is a function $\mathbb{N} \to \mathbb{C}$ with countable domain whereas a $g \in L^2[0,1]$ is an equivalence class of functions from the uncountable continuum $[0,1]$ to $\mathbb{C}$. The Riesz-Fischer theorem holds even under more surprising conditions. Ahmed [197, p. 41.] proves a version of the theorem for a series of nonlinear operators defined by integration against the classical Wiener-Fourier kernels arising in Volterra series or Wiener-G-functional expansions of nonlinear operators

Most theorems in this paragraph assume an ONB is already given. When this is not apriori the case and only a sequence $\{v_k\}_{k=1}^{\infty} \subset \mathcal{H}$ of linearly independent vectors is accessible, then one can construct an ONB $\{e_k\}_{k=1}^{\infty}$ of $\mathcal{H}$ out of the sequence $\{v_k\}_{k=1}^{\infty}$ with the help of the Gram-Schmid orthonormalization process [3, pp. 10-13].

Theorem 2.1.18 *Let a sequence $\{v_k\}_{k=1}^{\infty}$ of nonzero linearly independent vectors in a Hilbert space $\mathcal{H}$ be given. Then the sequence $\{e_k\}_{k=1}^{\infty}$ constructed iteratively via*

$$e_k = \frac{u_k}{\|u_k\|} \quad u_k = v_k - \sum_{j=1}^{k-1} \frac{\langle v_k, u_j \rangle}{\langle u_j, u_j \rangle} u_j \tag{2.5}$$

is an orthonormal sequence in $\mathcal{H}$. In this formulation sums of type $\sum_{j=n}^{m} f_j$ with $n > m$ are set to zero.

The properties of the Gram-Schmidt orthonormalization process suggest to construct ONB's for a separable infinite dimensional $\mathcal{H}$ by designing linearly independent sequences $\{v_k\}_{k=1}^{\infty}$ of vectors v_k that have dense span in $\mathcal{H}$ and then apply Gram-Schmidt orthonormalization to them. The next example illustrates the orthonormalization procedure.

Example 7 (Legendre polynomials) Let $\{v_k\}_{k=0}^{2}$ be the monomials x^0, x^1, x^2 as functions from $T = [-1, 1]$ to $\mathbb{R}$. They are linearly independent because any polynomial $p = \sum_{k=1}^{n} \alpha_k x^k$ does not have more than n roots precluding $\sum_{k=1}^{n} \alpha_k x^k = 0$ for coefficients α_k not all zero. The inner product on this space is set to $\langle f, g \rangle = \int_{-1}^{1} f(t)g(t)dt$. The polynomials constructed from monomials by way of 2.1.18 are simply the first 3 Legendre polynomials appropriately rescaled to unit norm [153, p. 443]. ■

The relationship discovered between orthogonal polynomials of up to degree 2 on $[-1, 1]$ and Legendre polynomials holds for arbitrary degrees and it can be shown that the Legendre polynomials form a countable ONB for $L^2[-1, 1]$ which is therefore seen to be separable. Analogue theorems hold for $L^2[\alpha, \beta]$ and scaled Legendre polynomials, $L^2[0, \infty)$ and Chebychev-Laguerre functions, and $L^2(\infty, \infty)$ and Chebychev-Hermite functions as well as for domains that are Cartesian products, $T = T_1 \times T_2, T_k \subset \mathbb{R}$ [93, pp. 55-66].

2.1.3 Abstract constructions with Hilbert spaces

Several abstract constructions are available to combine two Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$ into a new one that inherits properties from both $\mathcal{H}_1$ and $\mathcal{H}_2$. It is possible to form two different types of sum — external and internal— of Hilbert spaces leading to a Hilbert space $\mathcal{H}_1 \oplus \mathcal{H}_2$ being populated by sums of elements of $\mathcal{H}_1$ and $\mathcal{H}_2$. Taking the tensor product of $\mathcal{H}_1$ with $\mathcal{H}_2$ results in a Hilbert space $\mathcal{H}_1 \otimes \mathcal{H}_2$ of sums of products of elements in $\mathcal{H}_1$ and $\mathcal{H}_2$ whereas also a construction rule exists to create a Hilbert space $\mathcal{H}_1/\mathcal{H}_2$ of equivalence classes of elements in $\mathcal{H}_1$ that differ only by an element of $\mathcal{H}_2$ if $\mathcal{H}_2$ is a subspace of $\mathcal{H}_1$.

§ Quotient spaces

Quotient spaces $\mathcal{H}_2/\mathcal{H}_1$ are sets containing equivalence classes of vectors. The concept is made useful in a Hilbert space setting by endowing the quotient space itself with a vector space structure and an inner product. This pre-Hilbert space turns out to be isometrically isomorphic to the orthogonal complement of the subspace that was collapsed to zero and is shown to be complete. Quotient space constructions with certain minimality properties will be useful later. Their elements may be interpreted as classes of signals that differ only by elements of a supposedly irrelevant space $\mathcal{H}_1$ which is to be eliminated from further consideration.

Lemma 2.1.19 *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces with $\mathcal{H}_1 \subset \mathcal{H}_2$. Then the relation $f \sim g \Leftrightarrow f - g \in \mathcal{H}_1$ is an equivalence relation for elements of $\mathcal{H}_2$.*

Proof. Reflexivity follows from $f \sim f \Leftrightarrow f - f = 0 \in \mathcal{H}_1$ since $\mathcal{H}_1$ is a subspace of $\mathcal{H}_2$ and contains 0. Symmetry is a consequence of the closure of $\mathcal{H}_1$ under linear operations by $f \sim g \Leftrightarrow f - g \in \mathcal{H}_1 \Leftrightarrow g - f \in \mathcal{H}_1 \Leftrightarrow g \sim f$ and transitivity is obvious as well since if $f \sim g$ and $g \sim h$ then $f - h = (f - g) + (g - h)$ is the sum of two elements in $\mathcal{H}_1$ and therefore again an element of $\mathcal{H}_1$ implying $f \sim h$. □

Definition 2.1.20 Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces with $\mathcal{H}_1 \overline{\boxslash} \mathcal{H}_2$. Then the set of all equivalence classes $[f] := \{g \in \mathcal{H}_2 : f - g \in \mathcal{H}_1\}, f \in \mathcal{H}_2$ is called the quotient space of $\mathcal{H}_2$ by $\mathcal{H}_1$ and denoted by $\mathcal{H}_2/\mathcal{H}_1$. The map $\pi : \mathcal{H}_2 \ni f \mapsto [f] \in \mathcal{H}_2/\mathcal{H}_1$ sending Hilbert space elements to their corresponding equivalence class is called the canonical projection.
Remark The canonical projection $\pi : \mathcal{H}_1 \to \mathcal{H}_2/\mathcal{H}_1$ is surjective because $\mathcal{H}_2/\mathcal{H}_1 = \{[f], f \in \mathcal{H}_2\} = \pi\mathcal{H}_2$.

Theorem 2.1.21 *The quotient space $\mathcal{H}_2/\mathcal{H}_1$ together with the linear algebraic operations*

$$[f] + [g] = [f + g] \quad [f], [g] \in \mathcal{H}_2/\mathcal{H}_1 \tag{2.6}$$

$$\alpha[f] = [\alpha f] \quad [f] \in \mathcal{H}_2/\mathcal{H}_1, \alpha \in \mathbb{C} \tag{2.7}$$

is a vector space. This vector space can be embedded as a subspace into $\mathcal{H}_2$.

Proof: It will be shown in detail that the operations of addition and multiplication by scalars are well defined. Proving that the vector space axioms hold is trivial.

Suppose that $f_1, f_2 \in [f]$ and $g_1, g_2 \in [g]$. Then $f_1 - f_2 \in \mathcal{H}_1$ and $g_1 - g_2 \in \mathcal{H}_1$ and $(f_1 + g_1) - (f_2 + g_2) = f_1 - f_2 + g_1 - g_2 \in \mathcal{H}_1$ and $[f_1 + g_1] = [f_2 + g_2]$ establishing well-definedness of addition. To prove well-definedness of multiplication by scalars, note that if $f_1, f_2 \in [f]$ then $\forall \alpha \in \mathbb{C}$ one has $f_1 - f_2 \in \mathcal{H}_1 \Rightarrow \alpha(f_1 - f_2) = \alpha f_1 - \alpha f_2 \in \mathcal{H}_1 \Rightarrow \alpha f_1 \sim \alpha f_2$ and $[\alpha f_1] = [\alpha f_2]$.
$\mathcal{H}_2/\mathcal{H}_1$ is a commutative group under addition. It is commutative as shown by $[f] + [g] = [f + g] = [g + f] = [g] + [f]$ and associative as $([f] + [g]) + [h] = [f + g] + [h] = [(f + g) + h] = [f + (g + h)] = [f] + [g + h] = [f] + ([g] + [h])$. Any $p \in \mathcal{H}_1$ satisfies $f + p - f \in \mathcal{H}_1$ implying $[f] + [p] = [f]$ and $[p]$ is the zero element in $\mathcal{H}_2/\mathcal{H}_1$. As for inverses, any element $g \in [f]$ satisfies $[f] + [-g] = [f - g] = [0]$ since $f - g \in \mathcal{H}_1$ and $[f]^{-1} = [-g]$ for any $g \in [f]$. Scalar multiplication preserves linear structure because $\alpha\beta[f] = \alpha[\beta f] = [\alpha\beta f], 1[f] = [1f] = [f]$ and distributivity over $\mathcal{H}_2/\mathcal{H}_1$ and $\mathbb{C}$ are inherited from distributivity over $\mathcal{H}_2$ and $\mathbb{C}$. Consequently $\mathcal{H}_2/\mathcal{H}_1$ is a vector space. Finally note that the map $\psi : \mathcal{H}_2/\mathcal{H}_1 \ni [f] \mapsto \psi[f] = P_{\mathcal{H}_1^\perp} f \in \mathcal{H}_1^\perp$ embeds this quotient space into $\mathcal{H}_2$. □

Definition 2.1.22 Let $\langle \cdot, \cdot \rangle_{P^\dagger}$ be a positive semidefinite, conjugate symmetric form that is homogeneous with respect to scalar multiplication in the first argument and additive. It induces a form $| \cdot |_{P^\dagger}$ by $\sqrt{\langle \cdot, \cdot \rangle_{P^\dagger}} = | \cdot |_{P^\dagger}$. The form $| \cdot |_{P^\dagger}$ is called a seminorm on $\mathcal{H}_2$ if

i) $|f|_{P^\dagger} \geq 0 \quad \forall f \in \mathcal{H}_2$

ii) $|f|_{P^\dagger} = 0 \Leftrightarrow f \in P$

iii) $|\alpha f| = |\alpha| |f|_{P^\dagger} \quad \forall f \in \mathcal{H}_2, \alpha \in \mathbb{C}$

iv) $|f + g|_{P^\dagger} \leq |f|_{P^\dagger} + |g|_{P^\dagger} \quad \forall f, g \in \mathcal{H}_2$.

If $| \cdot |_{P^\dagger}$ is induced by $\langle \cdot, \cdot \rangle_{P^\dagger}$ as described above, the bilinear function $\langle \cdot, \cdot \rangle_{P^\dagger}$ is called a semi-inner product on $\mathcal{H}_2$.

Theorem 2.1.23 *If $\mathcal{H}_1, \mathcal{H}_2$ are separable Hilbert spaces with $\mathcal{H}_1 \overline{\boxslash} \mathcal{H}_2$ then $\mathcal{H}_2/\mathcal{H}_1$ is complete with respect to the norm topology induced by the inner product $\langle [f], [g] \rangle_{\mathcal{H}_2/\mathcal{H}_1} = \langle P_{\mathcal{H}_1^\perp} f, P_{\mathcal{H}_1^\perp} g \rangle_{\mathcal{H}_2}$ and therefore a Hilbert space.*

Proof: It was already proven that $\mathcal{H}_2/\mathcal{H}_1$ is a vector space. What is left to show is that $\langle [f], [g] \rangle_{\mathcal{H}_2/\mathcal{H}_1}$ is i) well defined, ii) an inner product and iii) $\mathcal{H}_2/\mathcal{H}_1$ is complete w.r.t. the induced norm.

i) Suppose $f_1, f_2 \in [f]$ and $g_1, g_2 \in [g]$ then $f_1 - f_2 \in \mathcal{H}_1$ and $P_{\mathcal{H}_1^\perp}(f_1 - f_2) = 0 = P_{\mathcal{H}_1^\perp}(g_1 - g_2)$ since $g_1 - g_2 \in \mathcal{H}_1$. Therefore $P_{\mathcal{H}_1^\perp} f_1 = P_{\mathcal{H}_1^\perp} f_2, P_{\mathcal{H}_1^\perp} g_1 = P_{\mathcal{H}_1^\perp} g_2$ and $\langle P_{\mathcal{H}_1^\perp} f_1, P_{\mathcal{H}_1^\perp} g_1 \rangle_{\mathcal{H}_2} = \langle P_{\mathcal{H}_1^\perp} f_2, P_{\mathcal{H}_1^\perp} g_2 \rangle_{\mathcal{H}_2}$ so that $\langle [f], [g] \rangle_{\mathcal{H}_2/\mathcal{H}_1}$ is well defined independent of the choice of representer for the equivalence classes $[f]$ and $[g]$.

ii) If $[f] \neq [0]$ then $\langle [f], [g] \rangle_{\mathcal{H}_2/\mathcal{H}_1} = \langle P_{\mathcal{H}_1^\perp} f, P_{\mathcal{H}_1^\perp} f \rangle_{\mathcal{H}_2} > 0$ by argument of $[f] \neq [0]$ implying $f - 0 \notin \mathcal{H}_1$ and $f = P_{\mathcal{H}_1} f + P_{\mathcal{H}_1^\perp} f$ with $P_{\mathcal{H}_1^\perp} f = f - P_{\mathcal{H}_1} f \neq 0$. Finally this shows $P_{\mathcal{H}_1^\perp} f \neq 0$ and $\langle [f], [f] \rangle_{\mathcal{H}_2/\mathcal{H}_1} > 0$ byt the non-degeneracy of $\langle\cdot,\cdot\rangle_{\mathcal{H}_2}$. Together with $\langle[0],[0]\rangle_{\mathcal{H}_2/\mathcal{H}_1} = \langle 0,0\rangle_{\mathcal{H}_2}$ this establishes positive definiteness. Conjugate symmetry follows from

$$\langle[f],[g]\rangle_{\mathcal{H}_2/\mathcal{H}_1} = \langle P_{\mathcal{H}_1^\perp}f, P_{\mathcal{H}_1^\perp}g\rangle_{\mathcal{H}_2} = \overline{\langle P_{\mathcal{H}_1^\perp}g, P_{\mathcal{H}_1^\perp}f\rangle}_{\mathcal{H}_2} = \overline{\langle[g],[f]\rangle}_{\mathcal{H}_2/\mathcal{H}_1}$$

and linearity is seen to be correct because

$$\langle\alpha[f],[g]\rangle_{\mathcal{H}_2/\mathcal{H}_1} = \langle[\alpha f],[g]\rangle_{\mathcal{H}_2/\mathcal{H}_1} = \langle\alpha P_{\mathcal{H}_1^\perp}f, P_{\mathcal{H}_1^\perp}g\rangle_{\mathcal{H}_2} = \alpha\langle[f],[g]\rangle_{\mathcal{H}_2/\mathcal{H}_1}$$

where the linearity of the projection was used. Lastly additivity holds as

$$\begin{aligned}\langle[f]+[g],[h]\rangle_{\mathcal{H}_2/\mathcal{H}_1} &= \langle[f+g],[h]\rangle_{\mathcal{H}_2/\mathcal{H}_1}\\ &= \langle P_{\mathcal{H}_1^\perp}(f+g), P_{\mathcal{H}_1^\perp}h\rangle_{\mathcal{H}_2}\\ &= \langle P_{\mathcal{H}_1^\perp}f, P_{\mathcal{H}_1^\perp}h\rangle_{\mathcal{H}_2} + \langle P_{\mathcal{H}_1^\perp}g, P_{\mathcal{H}_1^\perp}h\rangle_{\mathcal{H}_2}\\ &= \langle[f],[h]\rangle_{\mathcal{H}_2/\mathcal{H}_1} + \langle[g],[h]\rangle_{\mathcal{H}_2/\mathcal{H}_1}\end{aligned}$$

iii) The quotient space $\mathcal{H}_2/\mathcal{H}_1$ is complete since every Cauchy-sequence converges. This can be seen by explicit construction of any Cauchy-sequences $\{[f]_k\}_{k=1}^\infty \subset \mathcal{H}_2/\mathcal{H}_1$ limit $[f]$. If $\{[f]_k\}_{k=1}^\infty$ is Cauchy then $\forall\epsilon > 0\ \exists n_0 \in \mathbb{N} : n,m \geq n_0 \Rightarrow \|[f]_n - [f]_m\|_{\mathcal{H}_2/\mathcal{H}_1} < \epsilon$. Now let $\{f_k\}_{k=1}^\infty \subset \mathcal{H}_2$ be any sequence such that $[f_k] = [f]_k$; it follows from $\{[f]_k\}_{k=1}^\infty$ being Cauchy that

$$\begin{aligned}&\forall\epsilon > 0\ \exists n_0 \in \mathbb{N} : n,m \geq n_0 \Rightarrow\\ &\quad \|[f]_n - [f]_m\|_{\mathcal{H}_2/\mathcal{H}_1} = \|[f_n] - [f_m]\|_{\mathcal{H}_2/\mathcal{H}_1} = \|P_{\mathcal{H}_1^\perp}f_n - P_{\mathcal{H}_1^\perp}f_m\|_{\mathcal{H}_2} < \epsilon\end{aligned}$$

and $\{P_{\mathcal{H}_1^\perp}f_k\}_{k=1}^\infty$ is a Cauchy sequence lying in $\mathcal{H}_1^\perp$ by construction. But $\mathcal{H}_1^\perp$ is a Hilbert space and complete which guarantees that $\exists f \in \mathcal{H}_1^\perp \subset \mathcal{H}_2$ with $\lim_{k\to\infty} P_{\mathcal{H}_1^\perp}f_k = f$. Then the element $[f] \in \mathcal{H}_2/\mathcal{H}_1$ satisfies

$$\lim_{k\to\infty}\|[f]-[f]_k\|_{\mathcal{H}_2/\mathcal{H}_1} = \lim_{k\to\infty}\|f - P_{\mathcal{H}_1^\perp}f_k\|_{\mathcal{H}_2} = 0.$$

The equivalence class $[f]$ of f in $\mathcal{H}_1^\perp$ lies in $\mathcal{H}_2/\mathcal{H}_1$ and is the unique limit of the Cauchy sequence $\{[f]_k\}_{k=1}^\infty$. Consequently $\mathcal{H}_2/\mathcal{H}_1$ is a complete inner product space, hence a Hilbert space.

□

Corollary 2.1.23.1 *I The quotient Hilbert space $\mathcal{H}_2/\mathcal{H}_1$ is isometrically isomorphic to $\mathcal{H}_1^\perp$, i.e. there exists a bijective linear map from $\mathcal{H}_1^\perp$ to $\mathcal{H}_2/\mathcal{H}_1$ which preserves inner products.*

II The inner product $\langle[f],[g]\rangle_{\mathcal{H}_2/\mathcal{H}_1}$ and norm $\|[f]\|_{\mathcal{H}_2/\mathcal{H}_1}$ correspond to semi-inner products and seminorms on $\mathcal{H}_2$ that annihilate $\mathcal{H}_1$.

§ Direct sums and tensor products

Direct sums and tensor products are both ways in which to combine two spaces $\mathcal{H}_1, \mathcal{H}_2$ to generate a third, enlarged one. As the name direct sum implies, the resultant object can be interpreted as containing sums of elements $f_1 \in \mathcal{H}_1$ and $f_2 \in \mathcal{H}_2$. However, that interpretation only makes sense when f_1 and f_2 are summable, e.g. because they are functions on the same index set T, and $\mathcal{H}_1, \mathcal{H}_2$ are apriori subspaces of some other Hilbert space. In all other cases, the direct sum has more similarity with the Cartesian product. Both notions of direct sum coexist and are useful at different times.

Definition 2.1.24 Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces with inner products denoted by $\langle \cdot, \cdot \rangle_{\mathcal{H}_1}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_2}$.

i) The set $\mathcal{H}_1 \oplus_e \mathcal{H}_2 := \{f = (f_1, f_2) : f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2\}$ together with the inner product $\langle f, g \rangle_{\mathcal{H}_1 \oplus_e \mathcal{H}_2} = \langle (f_1, f_2), (g_1, g_2) \rangle_{\mathcal{H}_1 \oplus_e \mathcal{H}_2} = \langle f_1, g_1 \rangle_{\mathcal{H}_1} + \langle f_2, g_2 \rangle_{\mathcal{H}_2}$ $\forall f, g \in \mathcal{H}_1 \oplus_e \mathcal{H}_2$ is called the external direct sum or the orthogonal sum of $\mathcal{H}_1$ and $\mathcal{H}_2$ [21, p. 157].

ii) If there exists an $\mathcal{H}_3$ with $\mathcal{H}_1 \overline{\boxslash} \mathcal{H}_3$ and $\mathcal{H}_2 \overline{\boxslash} \mathcal{H}_3$, then the set $\mathcal{H}_1 \oplus_i \mathcal{H}_2 := \{f_1 + f_2 : f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2\}$ together with the squared norm $\|f\|^2_{\mathcal{H}_1 \oplus_i \mathcal{H}_2} = \min_{f = f_1 + f_2, f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2} \|f_1\|^2_{\mathcal{H}_1} + \|f_2\|^2_{\mathcal{H}_2}$ is called the internal direct sum of $\mathcal{H}_1$ and $\mathcal{H}_2$ [20, p. 24].

Whereas the external direct sum of $\mathcal{H}_1$ and $\mathcal{H}_2$ contains vectors with entries in $\mathcal{H}_1$ and $\mathcal{H}_2$, the internal direct sum is to be interpreted rather as a superposition of $\mathcal{H}_1$ and $\mathcal{H}_2$. When both component spaces are not orthogonal to each other, the constructions lead to significantly different results. Since this situation is explored most instructively when the underlying Hilbert spaces have additional structure, a systematic account of the interactions between external and internal direct sums is postponed until reproducing kernel Hilbert spaces have been introduced. If $\mathcal{H}_1 \perp \mathcal{H}_2$, then external and internal direct sums coincide, see theorem 2.1.11. In this case, the subscripts e or i indicating a distinction will be omitted. In comparison to the direct sum that roughly results in added output dimensions, the direct product of $\mathcal{H}_1$ and $\mathcal{H}_2$, frequently also called the tensor product, leads to added input dimensions.

Definition 2.1.25 Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces of functions from T_1, T_2 to $\mathbb{R}$ with inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}_1}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_2}$ respectively. The direct product $\mathcal{H}_1 \otimes \mathcal{H}_2$ of $\mathcal{H}_1$ and $\mathcal{H}_2$ is the completion of the space of functions that are of form

$$f^{\otimes}(s,t) = \sum_{j=1}^{n} f_j^1(s) f_j^2(t) \qquad\qquad s \in T_1, t \in T_2, f_j^1 \in \mathcal{H}_1, f_j^2 \in \mathcal{H}_2$$

together with the inner product

$$\langle f^{\otimes}, g^{\otimes} \rangle_{\mathcal{H}_1 \otimes \mathcal{H}_2} = \sum_{i=1}^{m} \sum_{j=1}^{n} \langle f_i^1, g_j^1 \rangle_{\mathcal{H}_1} \langle f_i^2, g_j^2 \rangle_{\mathcal{H}_2}$$

where the completion is executed w.r.t the induced norm $\|f^{\otimes}\|^2_{\mathcal{H}_1 \otimes \mathcal{H}_2}$ [20, p. 31].

If $\mathcal{H}_1$ and $\mathcal{H}_2$ are spaces of functions on T_1, T_2 then $\mathcal{H}_1 \otimes \mathcal{H}_2$ contains functions $f : T_1 \times T_2 \to \mathbb{R}$, see [127, p. 10] for additional explanations. One may define spaces of functions of several inputs, e.g. spatiotemporal functions, by tensoring together simpler spaces. Often operations in the tensor product space factorize through the base spaces and this makes them both easier to interpret and more efficiently computable.

2.1.4 Linear operators on Hilbert spaces

In this subsection the concept of a linear operator on a Hilbert space is introduced. Bounded linear operators constitute the most common linear operators encountered in practice and it is possible to establish that linear operators are bounded iff they are continuous. Bounded linear operators on a Hilbert space $\mathcal{H}$ form a space $\mathcal{B}(\mathcal{H})$ themselves and properties of those operators can be related to the structure of $\mathcal{H}$. Compact operators form a subclass of $\mathcal{B}(\mathcal{H})$ exhibiting finiteness properties paralleling those of matrices and appear naturally in Hilbert space embeddings of random variables. Orthogonal projection operators map a Hilbert space onto its subspaces and form a subclass of $\mathcal{B}(\mathcal{H})$ containing slightly transformed versions of pseudoinverses. They will be immediately used to provide solutions to simple least squares problems.

§ Bounded linear operators

Bounded linear operators between Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ are topologically well-behaved objects that interact with the natural notions of convergence in a non-complicated manner. The case where $\mathcal{H}_2$ is $\mathbb{R}$ or $\mathbb{C}$ leads one to consider bounded linear functionals forming themselves a Hilbert space called the dual space $\mathcal{H}_1^*$. For them the important Riesz representation theorem is available.

Definition 2.1.26 i) A map $A : \mathcal{H}_1 \to \mathcal{H}_2$ between two vector spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ is called a linear operator if it satisfies $A(\alpha f + \beta g) = \alpha A f + \beta A g \ \forall f, g \in \mathcal{H}_1$ and $\alpha, \beta \in \mathbb{C}$ (or $\mathbb{R}$)[2].

ii) If the codomain of a linear operator is the underlying field $\mathbb{F}$ of the vector space — $\mathbb{C}$ or $\mathbb{R}$ in most cases — then $A : \mathcal{H}_1 \to \mathbb{F}$ is called a linear functional.

Remark If $A : \mathcal{H}_1 \to \mathcal{H}_1$ then A is called a linear operator on $\mathcal{H}_1$ and if $A : \mathcal{M} \to \mathcal{H}_1$ with $\mathcal{M} \boxslash \mathcal{H}_1$ then A is called linear operator in $\mathcal{H}_1$.

Definition 2.1.27 i) A linear operator $A : \mathcal{H}_1 \to \mathcal{H}_2$ between two normed lin-

[2] When applying a linear operator A to an element f, one may either write $A(f)$ or simply Af to reduce the number of parentheses. There is rarely potential for any confusion.

ear spaces is called bounded if $\exists c \in \mathbb{R}$ such that

$$\|Af\|_{\mathcal{H}_2} \leq c\|f\|_{\mathcal{H}_1} \text{ for all } f \in \mathcal{H}_1. \tag{2.8}$$

ii) The smallest such $c \in \mathbb{R}$ in equation 2.8 is called the operator norm of A.

$$\|A\|_{op} = \sup_{f \in \mathcal{H}_1 \backslash \{0\}} \frac{\|Af\|_{\mathcal{H}_2}}{\|f\|_{\mathcal{H}_1}} \tag{2.9}$$

One arrives at this specific formulation by reordering $\|Af\|_{\mathcal{H}_2} = c\|f\|_{\mathcal{H}_1}$ to $c \geq \|Af\|_{\mathcal{H}_2}/\|f\|_{\mathcal{H}_1}$ and realizing that the smallest c still satisfying this equation can be found by taking the supremum of this particular ratio.

Example 8 A matrix $A = \{a_{ij}\}_{j=1,i=1}^{m\ \ n}$ is a linear operator from $\mathbb{R}^m$ to $\mathbb{R}^n$ with the usual matrix-vector multiplication. Its operator norm $\|A\|_{op}$ depends on the norms with which one endows $\mathcal{H}_1$ and $\mathcal{H}_2$; if for example the Euclidean standard norms are used $\|A\|_{op} = \sqrt{\lambda_{\max}}$ is the square root of the maximum eigenvalue of A^*A [30, p. 876]. Here again the asterisk denotes transposition and complex conjugation, i.e. $(A^*)_{ij} = \overline{a_{ji}}$. ■

Theorem 2.1.28 *The operator norm is a norm on bounded linear operators and furthermore it holds that [93, p. 80]*

$$\|A\|_{op} = \sup_{f \in \mathcal{H}_1 \backslash \{0\}} \frac{\|Af\|_{\mathcal{H}_2}}{\|f\|_{\mathcal{H}_1}} = \sup_{f \in \mathcal{H}_1, \|f\|_{\mathcal{H}_1} \leq 1} \|Af\|_{\mathcal{H}_2} = \sup_{f \in \mathcal{H}_1, \|f\|_{\mathcal{H}_1} = 1} \|Af\|_{\mathcal{H}_2}.$$

Corollary 2.1.28.1 *Linear bounded operators on a Hilbert space $\mathcal{H}_1$ form a Banach space under the operator norm $\|\cdot\|_{op}$ [93, p. 80].*
Remark The Banach space of bounded linear functionals $\varphi : \mathcal{H}_1 \ni f \mapsto \varphi(f) \in \mathbb{C}$ is called the dual space of $\mathcal{H}_1$ and denoted by $\mathcal{H}_1^*$.

Bounded linear functionals $\varphi : \mathcal{H}_1 \to \mathbb{C}$ acting on a Hilbert space can be written explicitly as inner products with certain elements $f_\varphi \in \mathcal{H}_1$. This is the content of the Riesz representation theorem.

Theorem 2.1.29 (Riesz representation) *Let $\varphi : \mathcal{H}_1 \to \mathbb{C}$ be a bounded linear functional. Then $\exists! f_\varphi \in \mathcal{H}_1$ with $\varphi g = \langle f_\varphi, g\rangle_{\mathcal{H}_1}$ for all $g \in \mathcal{H}_1$ and furthermore $\|\varphi\|_{op} = \|f_\varphi\|_{\mathcal{H}_1}$ [93, p. 76].*

This representation will be important in section 2.3 where reproducing kernel Hilbert spaces are introduced as those spaces of functions for which evaluation $e_t : \mathcal{H}_1 \ni f \mapsto e_t(f) = f(t) \in \mathbb{C}$ at a point $t \in T$ is continuous therefore guaranteeing the existence of a function $K_t(\cdot) \in \mathcal{H}_1$ with $e_t f = \langle K_t, f\rangle_{\mathcal{H}_1}\ \ \forall f \in \mathcal{H}_1$.

§ Special properties of operators

Matrices are finite-dimensional linear operators and compact operators as limiting cases of finite rank operators can be interpreted as their infinite-dimensional generalizations. Similarly, selfadjoint positive definite operators are likened to covariance matrices for which they provide infinite dimensional analogues in form of the covariance operator. Apart from this, often operators arise whose adjoints are their inverses — these unitary operators have a basic role as representers of basis change.

Any linear map $L : \mathbb{R}^n \to \mathbb{R}^n$ applied to $f \in \mathbb{R}^n$ has the result Lf with $(Lf)_i = \sum_{j=1}^n \alpha_{ij} f_j$ as any output dimension $i = 1, ..., n$ necessarily depends linearly on all the input dimensions $j = 1, ..., n$. The application of L to the vector f can be represented as multiplying a coefficient matrix A with the coefficient vector containing the components f_j of f. Another way to write this is as

$$Lf = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle f, e_j \rangle_{\mathbb{R}^n} \langle Ae_i, e_j \rangle_{\mathbb{R}^n} e_i = \sum_{j=1}^{n} g_j \langle f, e_j \rangle_{\mathbb{R}^n} \tag{2.10}$$

where $(A)_{ij} = \alpha_{ij}, g_j = \sum_{i=1}^n \langle Ae_i, e_j \rangle_{\mathbb{R}^n} e_i$ and $\{e_i\}_{i=1}^n$ is an ONB of $\mathbb{R}^n$. There is no reason why one may not define for a fixed $n < \infty$ a linear operator on an arbitrary Hilbert space $\mathcal{H}$ in a fashion similar to equation 2.10, i.e.

$$L : \mathcal{H} \ni f \mapsto \sum_{j=1}^{n} u_j \langle f, h_j \rangle_{\mathcal{H}} \in \mathcal{H}, \tag{2.11}$$

where $\{u_j\}_{j=1}^n$ and $\{h_j\}_{j=1}^n$ are arbitrary sequences of elements in $\mathcal{H}$. Operators of this type are called finite rank operators. Their limits for $n \to \infty$ are compact operators [164, p. 204] which themselves contain subclasses of operators satisfying different finiteness criteria and whose further properties require first defining the concept of adjoint operators.

Definition 2.1.30 For a bounded linear operator $A : \mathcal{H}_1 \to \mathcal{H}_2$, the operator $A^* : \mathcal{H}_2 \to \mathcal{H}_1$ satisfying $\langle Af, g \rangle_{\mathcal{H}_2} = \langle f, A^* g \rangle_{\mathcal{H}_1}$ is called the adjoint operator.

Definition 2.1.31 Let $L : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator on the Hilbert space $\mathcal{H}$ and let $\{e_i\}_{i=1}^\infty$ be an ONB of $\mathcal{H}$.

i) L is called compact iff it maps every weakly convergent sequence[3] into a strongly converging one [93, p. 187].

ii) L is called Hilbert-Schmidt iff $\sum_{j=1}^\infty \|Le_j\|^2 < \infty$. This is equivalent to $\sum_{j=1}^\infty \lambda_j < \infty$ for $\{\lambda_j\}_{j=1}^\infty$ the sequence of eigenvalues of $L^* L$ where L^* is the adjoint operator [63, p. 278].

iii) L is called trace class iff $\sum_{j=1}^\infty \sqrt{\lambda_j} < \infty$ where $\{\lambda_j\}_{j=1}^\infty$ is the sequence of eigenvalues of $L^* L$ [63, p. 276].

[3] A sequence$\{x_k\}_{k=1}^\infty$ is said to converge weakly, iff there exists an element x with $\lim_{k\to\infty} \langle x_k, y \rangle_{\mathcal{H}} = \langle x, y \rangle_{\mathcal{H}}$ for all $y \in \mathcal{H}$.

Both Hilbert-Schmidt operators and trace class operators are compact operators [63, pp. 276-278]. The finiteness conditions on the sum of the spectrum can be related to finiteness conditions on the energy of a sequence of random variables.

Theorem 2.1.32 *The adjoint A^* of A always exists, is unique [50, p. 514] and satisfies the following equations [93, p. 97].*

i) $(A^*)^* = A$ *ii)* $(\lambda A)^* = \overline{\lambda} A^* \quad \forall\ \lambda \in \mathbb{C}$

iii) $(A + B)^* = A^* + B^*$ *and* $(BA)^* = A^* B^*$

iv) $\|A^*\|_{op} = \|A\|_{op}$ *v) If* $\exists A^{-1}$ *: then* $(A^{-1})^* = (A^*)^{-1}$

The adjoint operator A^* is a generalization of the transpose A^T of a matrix A or in the complex case the Hermitian transpose $A^H = \overline{(A^T)}$. The adjoint x^* of a finite-dimensional column vector x is x^H as implied by $\langle \alpha x, y\rangle_{\mathbb{R}^n} = \alpha\langle x, y\rangle_{\mathbb{R}^n} = \langle \alpha, x^H y\rangle_{\mathbb{R}}$. The following two classes of operators are defined by demanding simple relationships between an operator and its adjoint.

Definition 2.1.33 Let $L : \mathcal{H} \to \mathcal{H}$ be a linear operator on the Hilbert space $\mathcal{H}$. Then [93, p. 98]

i) L is called unitary iff $L^* = L^{-1}$.

ii) L is called selfadjoint iff $L^* = L$.

Since unitary operators satisfy $\langle f, g\rangle_{\mathcal{H}} = \langle Lf, Lg\rangle_{\mathcal{H}}\ \ \forall f, g \in \mathcal{H}$, they do not change a vectors length or two vectors relative orientation to each other — they are isometries. As such they essentially represent a relabeling of the space $\mathcal{H}$ and are related to basis changes similar to their finite-dimensional counterparts, the orthogonal matrices. The Fourier transform is an example of a unitary operator. It allows representing a function f in a basis of complex exponentials, i.e. as a superposition of elementary waves and has been used to generate signal decompositions as for example already encountered in figure 2.5. If the selfadjoint operator L satisfies $\langle Lf, f\rangle_{\mathcal{H}} \geq 0\ \forall f \in \mathcal{H}$, then it is called positive and it is suitable to define a new semi-inner product on $\mathcal{H}$ via $\langle Lf, f\rangle_{\mathcal{H}} = \langle f, f\rangle_{\tilde{\mathcal{H}}}$. Positive definite matrices often arise as the covariance matrices of random vectors and their infinite dimensional analogues are called covariance operators.

Definition 2.1.34 Let $X_{\cdot} : \Omega \ni \omega \mapsto X_{\cdot}^{\omega} \in \mathcal{H}_K$ be a Hilbert space valued random variable having finite energy and therefore satisfying $E[\|X_{\cdot}\|^2_{\mathcal{H}_K}] = \int_{\Omega} \|X_{\cdot}^{\omega}\|^2_{\mathcal{H}_K} dP(\omega) < \infty$. Let $\mathcal{H}_k$ have the special property that it is a space of functions on T on which evaluation at $t \in T$ is continuous and represented by $K(t, \cdot) \in \mathcal{H}_K$. Then for each $\omega \in \Omega$, $X_{\cdot}^{\omega} : T \ni t \mapsto X_t^{\omega} \in \mathbb{R}$ is a real valued function and the unique linear operator C_X defined by

$$\langle C_X f, g\rangle_{\mathcal{H}_K} = E\left[\langle X_{\cdot}, f\rangle_{\mathcal{H}_K} \langle X_{\cdot}, g\rangle_{\mathcal{H}_K}\right] \tag{2.12}$$

is called the covariance operator of X [20, p. 28].

Note the defining equation's similarity to the relationship $\langle \Sigma f, g\rangle_{\mathbb{R}^n} = E[\langle xx^H f, g\rangle_{\mathbb{R}^n}] = E[\langle x, f\rangle_{\mathbb{R}^n}\langle x, g\rangle_{\mathbb{R}^n}]$ that is upheld between a random vector x, its covariance matrix Σ and arbitrary, but fixed vectors f and g. In this sense, C_X is the infinite-dimensional, functional analogue of the covariance matrix Σ of a random vector in $\mathbb{R}^n$. The covariance operator C_X is a selfadjoint kernel operator from $\mathcal{H}_K$ to $\mathcal{H}_K$ in the sense that $\forall f \in \mathcal{H}_K$

$$\begin{aligned}
(C_X f)_{(t)} = \langle K(t,\cdot), C_X f(\cdot)\rangle_{\mathcal{H}_K} &= \langle C_X K(t,\cdot), f(\cdot)\rangle_{\mathcal{H}_K} = \langle u(t,\cdot), f(\cdot)\rangle_{\mathcal{H}_K} \\
u(s,t) = \langle K(s,\cdot), u(t,\cdot)\rangle_{\mathcal{H}_K} &= E\left[\langle K(s,\cdot), X^{\cdot}_{\cdot}\rangle_{\mathcal{H}_K}\langle K(t,\cdot), X^{\cdot}_{\cdot}\rangle_{\mathcal{H}_K}\right] \\
&= E\left[X^{\cdot}_{s} X^{\cdot}_{t}\right] \qquad (2.13)
\end{aligned}$$

with kernel $u(s,t)$ equal to the covariance function of the process $\{X^{\cdot}_{t} : t \in T\}$ [20, p. 29] which is assumed to have zero mean for reasons of notational convenience. The operator C_X can therefore be interpreted as a weighted integration against a covariance function with the exact nature of the weights being determined by choice of inner product on $\mathcal{H}_K$.

Theorem 2.1.35 *The covariance operator C_X of a stochastic process $X^{\cdot}_{\cdot} : \Omega \times T \to \mathbb{R}^2$ with finite energy as measured by some norm $\|\cdot\|_{\mathcal{H}_K}$ as in definition 2.1.34 is a compact linear operator.*

Proof: First note that C_X is bounded. For this, recall that $E[XY] = \langle X, Y\rangle_{L^2(\Omega)}$ is a valid inner product for square integrable random variables X and $Y \in L^2(\Omega)$ on some probability space Ω. It then holds that

$$\begin{aligned}
\|C_X f\|_{\mathcal{H}_K} &= \left\|E\left[\langle X^{\cdot}_{\cdot}, f\rangle_{\mathcal{H}_K} X^{\cdot}_{\cdot}\right]\right\|_{\mathcal{H}_K} \\
&\leq \|f\|_{\mathcal{H}_K}\left\|E\left[\|X^{\cdot}_{\cdot}\|_{\mathcal{H}_K} X^{\cdot}_{\cdot}\right]\right\|_{\mathcal{H}_K} \\
&\leq \|f\|_{\mathcal{H}_K} E[\|X^{\cdot}_{\cdot}\|^2_{\mathcal{H}_K}] \qquad = \alpha\|f\|_{\mathcal{H}_K} \qquad (2.14)
\end{aligned}$$

with $\alpha < \infty$ where Jensen's inequality was used to derive $\alpha = E\|X^{\cdot}_{\cdot}\|^2_{\mathcal{H}_K}$ as the finite upper bound for the operator norm $\|C_X\|_{\text{op}}$. C_X is furthermore selfadjoint and positive as implied by $\langle C_X f, f\rangle_{\mathcal{H}_K} = E\left[\langle X^{\cdot}_{\cdot}, f\rangle^2_{\mathcal{H}_K}\right] \geq 0$ from which it follows that $|C_X| = \sqrt{C_X^* C_X}$ satisfies $\sum_{k=1}^{\infty}\langle |C_X| e_k, e_k\rangle_{\mathcal{H}_K} = \sum_{k=1}^{\infty}\langle C_X e_k, e_k\rangle_{\mathcal{H}_K}$ [63, p. 276]. Then clearly C_X is trace class since for any ONB $\{e_k\}_{k=1}^{\infty}$ for $\mathcal{H}_K$

$$\begin{aligned}
\sum_{k=1}^{\infty}\langle C_X e_k, e_k\rangle_{\mathcal{H}_K} &= \sum_{k=1}^{\infty} E\left[\langle X^{\cdot}_{\cdot}, e_k\rangle^2_{\mathcal{H}_K}\right] \\
&= E\left[\langle \sum_{k=1}^{\infty}\langle X^{\cdot}_{\cdot}, e_k\rangle_{\mathcal{H}_K} e_k, \sum_{l=1}^{\infty}\langle X^{\cdot}_{\cdot}, e_l\rangle_{\mathcal{H}_K} e_l\rangle_{\mathcal{H}_K}\right] \\
&= E\left[\langle X^{\cdot}_{\cdot}, X^{\cdot}_{\cdot}\rangle_{\mathcal{H}_K}\right] \\
&= E[\|X^{\cdot}_{\cdot}\|^2_{\mathcal{H}_K}] \qquad < \infty
\end{aligned}$$

where the Fourier expansion in terms of the ONB $\{e_k\}_{k=1}^{\infty}$ was used, recall theorem 2.1.17. As C_X is trace class, C_X is compact [63, p. 272]. □

§ Projections

Orthogonal projections are those idempotent bounded linear operators which are also selfadjoint. The interaction between different projections $P_1, P_2 : \mathcal{H} \to \mathcal{H}$ is determined by the relation their images $\mathrm{im}(P_1), \mathrm{im}(P_2)$ have as subspaces of $\mathcal{H}$. The orthogonal projection onto a subspace can be shown to have the same minimum norm property that is demanded during least squares adjustment. Therefore the solution to these problems is given by orthogonal projections that according to the general formula for the construction of orthogonal projections from their image and nullspace turns out to be related to the classical Moore-Penrose pseudoinverse.

Definition 2.1.36 A bounded linear operator $P : \mathcal{H} \to \mathcal{H}$ on a Hilbert space $\mathcal{H}$ is called an (orthogonal) projection iff it is idempotent $(PP = P)$ and selfadjoint $(P = P^*)$.

Remark The image $\mathcal{H}_1 := P\mathcal{H} = \{Pf : f \in \mathcal{H}\}$ is a subspace , $\mathcal{H}_1 \overline{\boxslash} \mathcal{H}$ [93, pp. 102-104] and $P : \mathcal{H} \to \mathcal{H}$ is called the projection onto $\mathcal{H}_1$. This is often emphasized by writing $P_{\mathcal{H}_1}$ instead of P. Oblique projections do exist but will be of no further concern herein; all projections are implicitly understood to be orthogonal if nothing to the contrary is mentioned.

i) There is a straightforward relationship between projections and least squares estimators. For a design matrix $A \in \mathbb{R}^n \otimes \mathbb{R}^m$ of full column rank the choice of parameters $x \in \mathbb{R}^m$ minimizing the residuals $Ax - b$ for some vector of observations $b \in \mathbb{R}^n =: \mathcal{H}$ is given by

$$||Ax - b||^2_{\mathbb{R}^n} \to \min \qquad \hat{x} = (A^T A)^{-1} A^T b = A^+ b \tag{2.15}$$

where A^+ is the Moore-Penrose pseudoinverse of A. The best reconstruction of $b \in \mathbb{R}^n$ by means of the model Ax is then $A\hat{x}$;

$$A\hat{x} = AA^+ b = A(A^T A)^{-1} A^T b = P_{\mathrm{im}(A)} b.$$

Here $A(A^T A)^{-1} A^T = P_{\mathrm{im}(A)}$ is idempotent and selfadjoint and its range is that of A implying $P_{\mathrm{im}(A)}$ to be the projection onto the column space of A as indicated by choice of symbols. A direct interpretation is possible by recalling that $P_{\mathcal{H}_1} f = \mathrm{argmin}_{g \in \mathcal{H}_1} \|f - g\|^2_{\mathcal{H}}$. Then in the above least squares problem $\mathcal{H}_1 = \mathrm{im}(A) = \{Ax : x \in \mathbb{R}^m\}$ is the set of all observations possible under the model $y = Ax$ for observations y and $P_{\mathrm{im}(A)} b = \hat{b}$ is the projection of any observation b onto one that is actually explainable by the model. If b were actually generated from some $q \in \mathbb{R}^m$ via $b = Aq$, then

$$\hat{b} = A\hat{x} = A[(A^T A)^{-1} A^T A q] = AIq = b$$

where $(A^T A)^{-1} A^T A = A^+ A = I$ is the identity projection, $\hat{b}$ reconstructs b and $\hat{x}$ recovers the underlying true parameters $q \in \mathbb{R}^m$.

ii) If A has full row rank then one has enough parameters $\{x_i\}_{i=1}^m$ to for any $b \in \mathbb{R}^n$ find a suitable $x \in \mathbb{R}^m$ with $Ax = b$. The pseudoinverse is then $A^+ = A^T(AA^T)^{-1}$ and $\hat{x} = A^+b$; consequently

$$A\hat{x} = AA^T(AA^T)^{-1}b = Ib$$

and the space of observations generated by the model $Ax = y$ contains the actual space of observations. The vector $\hat{x}$ is the solution to $Ax = b$ with the least norm [190, p. 404], i.e.

$$\hat{x} = \underset{x \in A^{-1}b}{\operatorname{argmin}} \ \|x\|^2_{\mathbb{R}^m}.$$

In comparison to case i), it is typically not possible anymore to recover the true value of a parametervector x from the measurements. If $b = Aq$ for some $q \in \mathbb{R}^m$, then

$$\hat{x} = A^+Aq = A^T(AA^T)^{-1}Aq = P_{\mathrm{im}(A^T)}q$$

and the true underlying q is projected onto a guess $\hat{x}$ that is of the same length or shorter since $\|P\|_{op} \leq 1$ generically. Whereas in case i) $\hat{b}$ was the projection of the observations b onto the ones consistent with the model, in case ii) $\hat{x}$ is the projection of the parameters x onto the nontrivial ones. They are orthogonal to those parameters that predict trivial observations as $\ker(A) \perp \mathrm{im}(A^T)$ according to the fundamental theorem of linear algebra [190, p. 198]. One may think of least squares, minimum norm solutions and the involved model equations entirely in terms of projections. The next theorem has further implications.

Theorem 2.1.37 *Let P_1, P_2 be projections on $\mathcal{H}$ and denote the corresponding subspaces by $\mathcal{H}_1$ and $\mathcal{H}_2$. Then [93, pp. 105-107]*

i) $I - P_{\mathcal{H}_1} = P_{\mathcal{H}_1^\perp}$ is the projection onto $\mathcal{H}_1^\perp \cong \mathcal{H}/\mathcal{H}_1$.

ii) $\mathcal{H}_1 \perp \mathcal{H}_2$ iff $P_1P_2 = 0$ and this is equivalent to $P_1 + P_2$ being the projection onto the subspace $(\mathcal{H}_1 \oplus \mathcal{H}_2) \,\overline{\boxslash}\mathcal{H}$.

iii) $\mathcal{H}_1 \,\overline{\boxslash}\mathcal{H}_2$ iff $P_1P_2 = P_2P_1 = P_1$ and this is equivalent to $P_2 - P_1$ being the projection onto the subspace $\mathcal{H}_2/\mathcal{H}_1 \cong \mathcal{H}_2 \cap \mathcal{H}_1^\perp$.

When Hilbert spaces are considered as containing signals, then the theorem admits an easy interpretation. Suppose that $\mathcal{H}$ contains all observable signals and choose two subspaces $\mathcal{H}_1 \,\overline{\boxslash}\mathcal{H}$ and $\mathcal{H}_2 \,\overline{\boxslash}\mathcal{H}$ that correspond to two models for the underlying process generating elements in $\mathcal{H}$. For example $\mathcal{H}_1$ and $\mathcal{H}_2$ may be spaces of polynomials, functions of explanatory variables or bandlimited functions. The solutions

$$\hat{b}_1 = \underset{b_1 \in \mathcal{H}_1}{\operatorname{argmin}} \ \|b_1 - b\|_{\mathcal{H}} \qquad\qquad \hat{b}_2 = \underset{b_2 \in \mathcal{H}_2}{\operatorname{argmin}} \ \|b_2 - b\|_{\mathcal{H}}$$

are $P_{\mathcal{H}_1}b$ and $P_{\mathcal{H}_2}b$, the best reconstructions of b via elements in $\mathcal{H}_1$ and $\mathcal{H}_2$. If one supposes finite dimensionality and that each element in $\mathcal{H}_1$ and $\mathcal{H}_2$ can be written in terms of the two models

$$b_1 \in \mathcal{H}_1 \Rightarrow b_1 = A_1 x_1 \qquad\qquad b_2 \in \mathcal{H}_2 \Rightarrow b_2 = A_2 x_2,$$

the best guesses for the parameter vectors $x_1 \in \mathcal{H}_1^X$ and $x_2 \in \mathcal{H}_2^X$ are given by $\hat{x}_1 = A_1^+ b$ and $\hat{x}_2 = A_2^+ b$ respectively. Theorem 2.1.37 implies statements i) to iii).

i) $(I - P_{\mathcal{H}_1})b$ is the projection of b onto $\mathcal{H}_1^\perp$. Since $\mathcal{H}_1$ is a model for the (interesting part of the) signal, $P_{\mathcal{H}_1^\perp} b$ is the best guess for the noise inherent in the observations b. Analogously for $I - P_{\mathcal{H}_2} = P_{\mathcal{H}_2^\perp}$.

ii) If $\mathcal{H}_1 \perp \mathcal{H}_2$ then the two models $\mathcal{H}_1$ and $\mathcal{H}_2$ do not contain redundant components as no nontrivial prediction of any observation by model $\mathcal{H}_1$ can be recreated by model $\mathcal{H}_2$ and vice versa. As $\mathcal{H}_1 \,\overline{\boxslash}\, \mathcal{H}_2^\perp$ and $\mathcal{H}_2 \,\overline{\boxslash}\, \mathcal{H}_1^\perp$, what one model considers signal, the other one considers noise; if even $\mathcal{H}_1^\perp = \mathcal{H}_2$ this induces a splitting of $\mathcal{H}$ into $\mathcal{H}_1 \oplus \mathcal{H}_2$ and $b = \hat{b}_1 + \hat{b}_2$. Note furthermore that $(\mathcal{H}_1 \oplus \mathcal{H}_2) \,\overline{\boxslash}\, \mathcal{H}$ implying that both models can be added yielding a new model in which the best guess for b is $\hat{b} = P_{\mathcal{H}_1 \oplus \mathcal{H}_2} b = (P_{\mathcal{H}_1} + P_{\mathcal{H}_2})b$.

iii) If $\mathcal{H}_1 \,\overline{\boxslash}\, \mathcal{H}_2$ then the second model is more expressive than the first one as any observation predicted with model $\mathcal{H}_1$ can also be recreated under model $\mathcal{H}_2$. $P_{\mathcal{H}_2} - P_{\mathcal{H}_1} = P_{\mathcal{H}_2/\mathcal{H}_1}$ is a projection onto that part of $\mathcal{H}_2$ which is not explainable by $\mathcal{H}_1$.

Figure 2.6 illustrates these statements. The explanations were embedded into the context of linear adjustment theory only for the sake of easier interpretability. The Hilbert space approach to signal reconstruction and estimation does not require finite dimensionality and solutions can be stated in terms of projections without reference to linear algebraic equations of type $Ax = b$ that need to be solved for the parameters x. Since the previous interpretations make use primarily of the Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ they stay valid even in this more general situation.

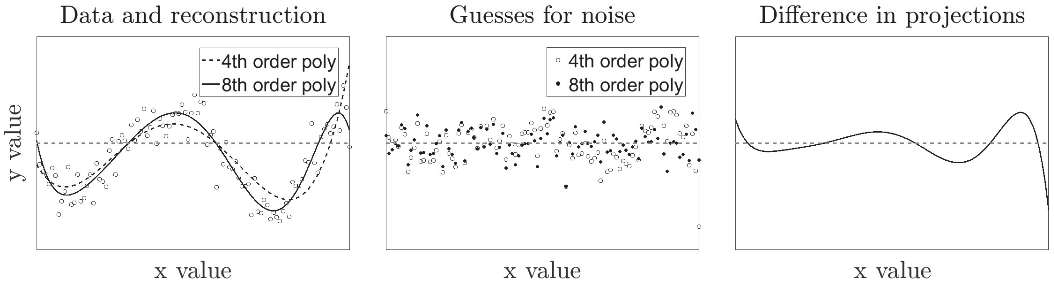


Figure 2.6: Data is fit by a fourth order polynomial and an eighth order polynomial (first panel). The guesses for the noise using these two models are plotted in the center. Note that $\|I - P_{\mathcal{H}_1} b\| > \|I - P_{\mathcal{H}_2} b\|$ even though $\mathcal{H}_2$ is not necessarily a 'better' model. The right hand side plots the projection $(P_{\mathcal{H}_2} - P_{\mathcal{H}_1})b$ to show the reconstructive benefit of the fifth to eighth order terms. The scale is equal for all plots.

It should be clear from the basis theorem 2.1.17 that given an ONB $\{e_i\}_{i=1}^\infty$ of $\mathcal{H}$, the sequence of $\mathcal{H}_i := \{\alpha e_i : \alpha \in \mathbb{R}\}$ together with the inner product $\langle \alpha e_i, \beta e_i \rangle_{\mathcal{H}_i} = \alpha\beta$ is a sequence of orthogonal spaces and the direct sums $V_n = \bigoplus_{i=1}^n \mathcal{H}_i$ form a sequence of increasing Hilbert spaces. Then $V_n \,\overline{\boxslash}\, V_{n+1}$ and the approximation error monotonically decreases with n as

$$\begin{aligned}\|P_{V_n}f - f\|_{\mathcal{H}}^2 = \| \sum_{j=n+1}^{\infty} \langle f, e_i\rangle_{\mathcal{H}} e_i\|_{\mathcal{H}}^2 &= \sum_{j=n+1}^{\infty} |\langle f, e_i\rangle_{\mathcal{H}}|^2 \\ &\geq \sum_{j=n+2}^{\infty} |\langle f, e_i\rangle_{\mathcal{H}}|^2 = \|P_{V_{n+1}}f - f\|_{\mathcal{H}}^2.\end{aligned}$$

Clearly, in the limit $\lim_{n\to\infty} \|P_{V_n}f - f\|_{\mathcal{H}}^2$ is 0 as $\mathcal{H} = \bigoplus_{i=1}^{\infty} \mathcal{H}_i$. The gain in approximation accuracy by adding $\mathcal{H}_{n+1}$ to V_n is

$$\begin{aligned}\|P_{V_n}f - f\|_{\mathcal{H}}^2 - \|P_{V_{n+1}}f - f\|_{\mathcal{H}}^2 = |\langle f, e_{n+1}\rangle_{\mathcal{H}}|^2 &= \|(P_{V_{n+1}} - P_{V_n})f\|_{\mathcal{H}}^2 \\ &= \|P_{\mathcal{H}_{n+1}}f\|_{\mathcal{H}}^2. \qquad (2.16)\end{aligned}$$

Different heuristics for choosing decompositions of $\mathcal{H}$ into an infinite sum of orthogonal basis Hilbert spaces arise from the eigendecomposition of positive definite operators. In section 2.2 these spectral theoretic aspects are treated in more detail and ways of choosing a mean-square optimal basis by factorizing covariance operators are provided in subsection 3.2.1.

2.2 Spectral theory of linear operators

In this section the analysis of linear operators commences by means of a generalization of the familiar eigendecomposition of a symmetric matrix to infinite dimensional settings. The spectrum replaces the set of eigenvalues and eigenfunctions take the role the eigenvectors had before. The spectra of orthogonal projections and positive definite linear operators satisfy some simple equations that allow the definition of an order structure on them. By manipulation of the spectra of selfadjoint operators a functional calculus can be established and nontrivial conclusions regarding solutions to operator-valued equations are deduced. Furthermore, the thoughts leading up to this functional calculus can be extended to include a decomposition of a selfadjoint operator into a weighted integral of maximal projections onto one dimensional subspaces. This is known as the spectral theorem and the set of methods associated with it will be applied immediately to calculate the time evolution of physical systems by means of unitary state transition operators.

2.2.1 The spectrum of an operator

The spectrum of a linear operator A is defined as the set of complex values $\lambda \in \mathbb{C}$ such that A acts similar to multiplication by λ in at least one of the subspaces spanning its domain. Even though the spectrum is therefore related principally to questions of non-invertibility of A, it gives insights into extremal properties upheld by A and indeed a variational characterization of a subset of the spectrum exists. Its use implies the topological properties of the spectrum via the Neumann series and the spectral radius formula gives a first hint at how a measure of uncertainty may be extracted from a covariance operator. Some immediate consequences can be drawn for the spectra of covariance operators and a natural partial order can be established via a comparison of spectra.

Definition 2.2.1 Let a not necessarily bounded operator A on a Hilbert space $\mathcal{H}$ be given. The values $\lambda \in \mathbb{C}$ such that $(A - \lambda I)$, I the identity operator, is either not injective or $(A - \lambda I)^{-1}$ is unbounded are said to constitute the spectrum $\sigma(A) \subset \mathbb{C}$ of A. If $\lambda \in \mathbb{C} \setminus \sigma(A)$ then λ is called a regular value.

It can be shown that $\sigma(A)$ is closed and the set of regular values is open [93, p. 163]. The spectrum need not consist only of discrete points as is the case in the finite dimensional setting of linear operators on $\mathbb{R}^n$ represented my matrices but generally may also include continuous parts like entire intervals. Under certain circumstances, the resolvent operator $R(\lambda) := (A - \lambda I)^{-1}$ is expressible as a power sum in A.

Theorem 2.2.2 *If for $A \in \mathcal{B}(\mathcal{H})$, $\|A\|_{op} < |\lambda|$ then one has $\lambda \notin \sigma(A)$ and*

$$(A - \lambda I)^{-1} = -\frac{1}{\lambda}\sum_{k=0}^{\infty}\left(\frac{A}{\lambda}\right)^k \tag{2.17}$$

Proof: If $\|A\|_{op} < |\lambda|$ then $\|A\|_{op}/|\lambda| < 1$ and $\|\sum_{k=0}^{\infty}\left(\frac{A}{\lambda}\right)^k\|_{op} \leq \sum_{k=0}^{\infty}\left(\frac{\|A\|_{op}}{|\lambda|}\right)^k$. The latter term is a geometric series and therefore converges. It then holds that

$$\sum_{k=0}^{\infty}\left(\frac{A}{\lambda}\right)^k (A - \lambda I) = \sum_{k=0}^{\infty}\left(\frac{A^{k+1}}{\lambda^k} - \frac{A^k}{\lambda^{k-1}}\right) = \sum_{k=1}^{\infty}\frac{A^k}{\lambda^{k-1}} - \sum_{k=0}^{\infty}\frac{A^k}{\lambda^{k-1}} = -\lambda I$$

This implies equation 2.17. Technical details augmenting the outline of the proof may be found in [93, p. 161]. Folland [63, p. 3] lists some corollaries. □

Corollary 2.2.2.1 *I For $A, B \in \mathcal{B}(\mathcal{H})$, if $A^{-1} \in \mathcal{B}(\mathcal{H})$ and $\|B\|_{op} \leq 1/\|A^{-1}\|_{op}$ then $(A - B)^{-1} \in \mathcal{B}(\mathcal{H})$ with $(A - B)^{-1} = A^{-1}\sum_{k=0}^{\infty}(BA^{-1})^k$.*

II For $A, B \in \mathcal{B}(\mathcal{H})$ if $A^{-1} \in \mathcal{B}(\mathcal{H})$ and $\|B\|_{op} \leq 1/(2\|A^{-1}\|_{op})$ then $\|(A - B)^{-1} - A^{-1}\|_{op} \leq 2\|A^{-1}\|_{op}\|B\|_{op} \leq \|A^{-1}\|_{op}$.

III Inversion as a map from the open set of invertible elements of $\mathcal{B}(\mathcal{H})$ to $\mathcal{B}(\mathcal{H})$ is continuous.

IV For $A \in \mathcal{B}(\mathcal{H})$ the spectral radius $\rho(A) = \sup\{|\lambda| : \lambda \in \sigma(A)\}$ can be written alternatively as $\rho(A) = \lim_{k\to\infty}\sqrt[k]{\|A^k\|_{op}}$.

V For $\lambda, \mu \notin \sigma(A), \lambda \neq \mu$ one finds $[\mu - \lambda]^{-1}[R(\mu) - R(\lambda)] = -R(\mu)R(\lambda)$ which implies in the limit case $\mu \to \lambda$ that $\frac{\partial}{\partial\lambda}[A - \lambda I]^{-1} = -[A - \lambda I]^{-2}$.

Remark It is possible to show that theorem 2.2.2 and corollaries 2.2.2.1 hold even if not $\|\cdot\|_{op}$ is used but any norm on $\mathcal{B}(\mathcal{H})$ under which it is a Banach-algebra [63, p. 3].

Corollaries I to V are useful later on for error estimates when uncertainty and inference are limited no longer to vectors $x \in \mathcal{H}$ but encompass objects like for instance the covariance operator. Subsequently investigations into robustness of the inversion process under perturbations of the involved operators become important. The covariance operator C_X was a bounded selfadjoint positive definite compact kernel operator of trace class. It is to be expected that this assortment of characteristic properties has some bearing on the spectrum of C_X. This is indeed the case as can be deduced from the set of statements below.

Theorem 2.2.3 *Let A and B be not necessarily bounded linear operators on the Hilbert space $\mathcal{H}$ and denote by $\sigma(A), \sigma(B)$ their spectra.*

I If A is bounded then $\sigma(A) \subset \overline{B}_{\|A\|_{op}}(0)$, the closed ball in $\mathbb{C}$ with radius equal to A's operator norm.

II If A is selfadjoint then $\sigma(A) \subset \mathbb{R}$.

III If A is bounded and positive then $\sigma(A) \subset [0, \infty)$.

IV If A is compact then $\sigma(A) = \{0\} \cup \lambda(A)$ where $\lambda(A)$ is the set of discrete eigenvalues satisfying $Av = \lambda v$ for some eigenfunction $v \in \mathcal{H}$. If there are infinitely many eigenvalues then $\lim_{k\to\infty} \lambda_k = 0$.

V If A is compact, selfadjoint and nonzero, then $\sigma(A) \neq \emptyset$ and $\exists \lambda \in \sigma(A)$: $|\lambda| = \|A\|_{op}$.

The assertions are taken from [3, pp .117-132], [93, pp. 162-196] and [63, p. 274], where also their proofs can be found. By combining I to V it is possible to deduce that the spectrum of the covariance operator C_X of a finite energy stochastic process is indeed expressible as a norm-decreasing countable sequence $\{\lambda_k\}_{k=1}^{\infty} \subset \mathbb{R}_+$ that converges to 0 and is entirely contained in the interval $[0, \|C_X\|_{op}]$. The inversion of an operator A essentially requires disassembling and then reassembling it with inverted spectra $\sigma(A^{-1}) = \{\lambda^{-1} : \lambda \in \sigma(A)\}$ — a procedure that will be made precise in the next subsection. It is therefore to be expected, that formation of C_X^{-1} is not entirely unproblematic as the eigenvalues of C_X^{-1} will diverge. As a matter of fact this behavior is already visible in most finite dimensional covariance matrices; for an example see figure 2.7.

The findings are documented in the following theorem. It already foreshadows that some work will need to be invested in later chapters to deal with inevitably arising numerical questions.

Theorem 2.2.4 *The inverse (if it exists) of the covariance operator C_X acting in an infinite dimensional Hilbert space $\mathcal{H}$ is unbounded, i.e. $\exists f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \in \mathbb{R}$ but $\|C_X^{-1} f\|_{\mathcal{H}} \nless \alpha \|f\|_{\mathcal{H}}$ for any positive finite real number α.*

Proof: As is proven in [93, p. 189] the compact operators on $\mathcal{H}$ form a topologically closed symmetric two sided ideal[4], in $\mathcal{B}(\mathcal{H})$. This ideal will be denoted by $\mathcal{K}(\mathcal{H})$ and as a simple consequence $\forall B \in \mathcal{B}(\mathcal{H})$ and $A \in \mathcal{K}(\mathcal{H})$, $AB, BA \in \mathcal{K}(\mathcal{H})$ by

[4] Roughly, a subset $\mathcal{K}$ of an algebra $\mathcal{B}$ is called a two sided ideal iff it is a subgroup under addition and $\mathcal{K}$ is absorbing w.r.t multiplication with elements from $\mathcal{B}(\mathcal{H})$. For a rigorous definition, consult [94, p. 469].

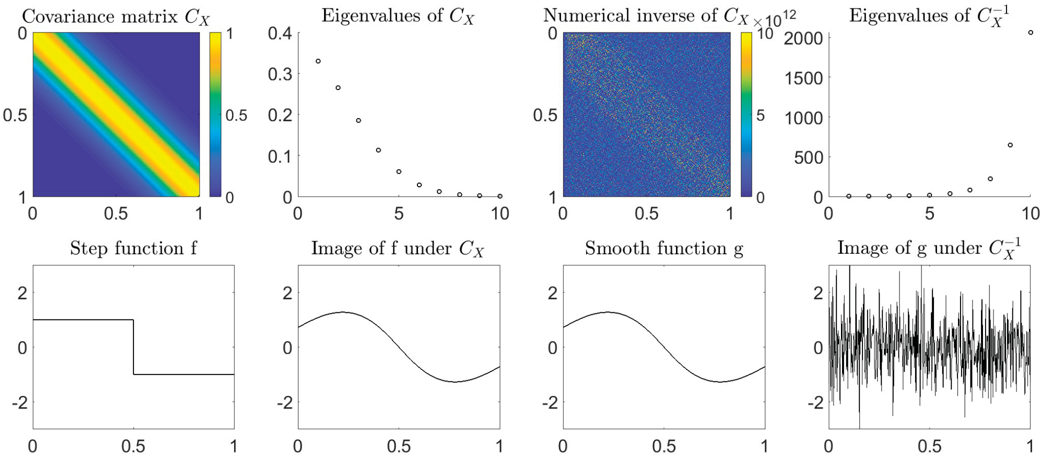


Figure 2.7: A covariance operator C_X of a random vector X of dimension 1000 whose entries are associated to the unit interval $T = [0, 1]$; consequently C_X is a matrix whose elements quantify the covariance between X at different locations on the unit interval. The second order moment function generating C_X is $E[X_s X_t] = K(s,t) = \exp(-\|s-t\|^2/d^2)$ and eigenvalues of C_X (only the first ten are shown) converge rapidly to 0 leading to numerical problems during inversion as illustrated by the plots showing the numerical inverse and its diverging spectrum. This behavior is symptomatic for covariance operators. The bottom row shows that the action of C_X on a step function results in an extremely smooth step-like wave. The inverse C_X^{-1} consequently would need to transform a smooth function into a step function — an operation that behaves very irregularly as testified to by the results on the bottom right. The functions in the bottom row have been rescaled to unit norm to increase comparability.

the absorption property of ideals [94, p. 469]. Let C_X be the covariance operator. Then C_X is compact and $C_X C_X^{-1} = I$. But I is not compact in infinite dimensions because for any ONB $\{e_k\}_{k=1}^{\infty}$ of $\mathcal{H}_K$, $\{e_k\}_{k=1}^{\infty}$ is weakly convergent to zero by $\|g\|^2 = \sum_{k=1}^{\infty} |\langle g, e_k\rangle|^2 \ \forall g \in \mathcal{H}$ implying $\lim_{k\to\infty}\langle g, e_k\rangle = \langle g, 0\rangle = 0$. However, $I\{e_k\}_{k=1}^{\infty}$ is not even Cauchy since $\|e_k - e_{k-1}\| = \sqrt{2} \ \forall k \in \mathbb{N}$ [93, p. 185]. This implies that $C_X^{-1} \notin \mathcal{B}(\mathcal{H})$. It is therefore unbounded. □

It is convenient to establish an order relation — called the Loewner partial order in the finite dimensional matrix case — on the set $\mathcal{P}_+(\mathcal{H})$ of positive operators acting on a Hilbert space $\mathcal{H}$. If an operator A is positive, this is indicated by $A \succeq 0$ and if $A, B \in \mathcal{P}_+(\mathcal{H})$ then $A \succeq B$ is defined to hold iff $A - B \succeq 0$. Note that by Rayleigh's principle (theorem 2.2.5.I) and the Courant minimax principle (theorem 2.2.5.II) the eigenvalues of a compact positive selfadjoint operator admit a variational characterization.

Theorem 2.2.5 *Let a compact positive selfadjoint operator A on a Hilbert space $\mathcal{H}$ be given and denote by $\mathcal{H}_n$ any one n-dimensional subspace of $\mathcal{H}$. Then [85, pp. 278-287]*

I For $\lambda_{\max} = \sup\limits_{\lambda\in\sigma(A)} \lambda$ it holds that $\lambda_{\max} = \sup\limits_{f\in\mathcal{H}} \frac{\langle Af,f\rangle_{\mathcal{H}}}{\langle f,f\rangle_{\mathcal{H}}} = \sup\limits_{f\in\mathcal{H},\|f\|_{\mathcal{H}}=1} \langle Af, f, \rangle_{\mathcal{H}}$.

II For the n-th eigenvalue λ_n in descending order, a weaker version of I holds.

$$\lambda_n = \sup_{\mathcal{H}_n \boxslash \mathcal{H}} \ \min_{f\in\mathcal{H}_n,\|f\|_{\mathcal{H}}=1} \langle Af, f\rangle_{\mathcal{H}} = \min_{\mathcal{H}_{n-1} \boxslash \mathcal{H}} \ \sup_{f\in\mathcal{H}_{n-1}^{\perp},\|f\|_{\mathcal{H}}=1} \langle Af, f\rangle_{\mathcal{H}}$$

Alternatively $\lambda_n = \sup_{f \in (\bigvee_{k=1}^{n-1}\{e_k\})^{\perp}, \|f\|_{\mathcal{H}}=1} \langle Af, f\rangle_{\mathcal{H}}$ where $\{e_k\}_{k=1}^{n-1}$ are the $n-1$ first eigenfunctions corresponding to the biggest eigenvalues and e_n will be chosen to maximize $\langle Ae_k, e_k\rangle$ subject to being orthogonal to the space already spanned by $\{e_k\}_{k=1}^{n-1}$. Note that for the covariance operator C_X, $\lambda_{\max}(C_X)$ gives the highest variance achievable by any linear combination $\langle f, X_{\cdot}^{\cdot}\rangle$, $\|f\| = 1$ of random variables belonging to the stochastic process $X_{\cdot}^{\cdot} : \Omega \times T \to \mathbb{R}$ and can therefore be interpreted as the maximum uncertainty while the corresponding eigenfunction $e_{\max}$ associated with $\lambda_{\max}$ provides the 'direction' of maximal stochastic volatility. On the one hand this allows one to identify particularly relevant subspaces of $\mathcal{H}$ that account for much of the processes' total variance and can be employed to derive sparse representations and tackle the canonical problem III posed on page 21.

On the other hand theorem 2.2.3.V states that $\lambda_{\max} = \|C_X\|_{op}$ which implies that $\lambda_{\max}$ as a representative for the size $\|\cdot\|_{op}$ of the covariance operator is a worthwhile minimization target when C_X depends on parameters that are to be chosen as to limit the upper bound on a stochastic processes' uncertainty [198].

2.2.2 Functional calculus

For each selfadjoint linear operator A there exists a continuous functional calculus and the map from continuous functions to operators is an algebra homomorphism. Consequently formation of an inverse A^{-1} can be written as an operation on the spectrum of A implying large condition numbers for compact operators and a problematic situation where inversion of covariance operators is needed for statistical inference. The functional calculus can be used to not only define polar decompositions and square roots of positive operators enabling whitening of random vectors but also allows for operator-valued functional equations to be stated and solved. As such it can help to illuminate the connection between a random variable and some of its transforms.

Theorem 2.2.6 *Let a bounded selfadjoint operator $A \in \mathcal{B}(\mathcal{H})$ be given and denote its spectrum by $\sigma(A) \subset \mathbb{R}$. If f is any continuous complex valued function from $\sigma(A)$ to $\mathbb{C}$, write $f \in C(\sigma(A))$. Then $\exists! f(A) \in \mathcal{B}(\mathcal{H})$ such that $f(A)$ is a limit of polynomials in A and the mapping $\psi_A : C(\sigma(A)) \ni f(\cdot) \mapsto f(A) \in \mathcal{B}(\mathcal{H})$ is a homomorphism from the algebra of continuous complex valued functions with pointwise multiplication to the Banach algebra of bounded operators on a Hilbert space $\mathcal{H}$ with multiplication as composition of operators. Practically, this means that [93, pp.232-235]*

I For any sequence of polynomials $\{f_k\}_{k=1}^{\infty} \subset C(\sigma(A))$ with $\lim_{k\to\infty} \sup\{|f_k(\lambda) - f(\lambda)| : \lambda \in \sigma(A)\} = 0 \ \exists! \lim_{k\to\infty} f_k(A) = f(A) \in \mathcal{B}(\mathcal{H})$.

II The mapping $f \mapsto f(A)$ preserves linear structure and is multiplicative, i.e. $\psi_A(f+g) = \psi_A(f)+\psi_A(g)$, $\psi_A(\alpha f) = \alpha\psi_A(f)$ and $\psi_A(fg) = \psi_A(f)\psi_A(g)$ for any $f, g \in C(\sigma(A))$ and $\alpha \in \mathbb{C}$.

III The mapping $f \mapsto f(A)$ is actually a $*$-homomorphism, i.e $\psi_A(\overline{f}) = \psi_A(f)^*$ where the overline denotes complex conjugation and the asterisk denotes taking the adjoint of a bounded linear operator.

IV *The norm of $f(A)$ is upper bounded by the supremum of f applied to the spectrum of A by $\|f(A)\|_{op} \leq 2\sup\{|f(\lambda)| : \lambda \in \sigma(A)\}$ and one even finds $\sigma(f(A)) = f(\sigma(A))$.*

Since the polynomials on $[\lambda_{\min}, \lambda_{\max}] \subset \mathbb{R}$ are dense in $C[\lambda_{\min}, \lambda_{\max}]$ by the Weierstrass approximation theorem [206, p. 29], statement I asserts that for any continuous function on $\mathbb{R}$, $f(A)$ may be approximated to arbitrary precision by polynomials in A because the existence of a polynomial sequence $\{f_k\}_{k=1}^{\infty}$ with $\lim_{k\to\infty} f_k = f$ is guaranteed for f a continuous complex function. Results II and III can be generalized to larger classes of functions that include real valued step function and characteristic functions as well [93, pp. 244-250] when the topologies are relaxed appropriately.

Even though no constructive procedure to arrive at $f(A)$ given $f \in C(\sigma(A))$ and $A \in \mathcal{B}(\mathcal{H})$ with $A^* = A$ was given, the homomorphism properties II and III allow some interesting conclusions — especially in the case of covariance operators. Note that any algebraic property of functions $f, g \in C(\mathbb{R})$, definable solely in terms of multiplication and addition, carries over unperturbed to the operators $f(A)$ and $g(A)$ since if $\sum_{k=0}^{n_f} \sum_{l=0}^{n_g} \alpha_{kl} f^k g^l = c$, then

$$\begin{aligned} cI = \psi_A(c) = \psi_A\left(\sum_{k=0}^{n_f}\sum_{l=0}^{n_g} \alpha_{kl} f^k g^l\right) &= \sum_{k=0}^{n_f}\sum_{l=0}^{n_g} \alpha_{kl}\psi_A(f)^k \psi_A(g)^l \\ &= \sum_{k=0}^{n_f}\sum_{l=0}^{n_g} \alpha_{kl} f(A)^k g(A)^l. \end{aligned} \tag{2.18}$$

It is then entirely possible to give existence guarantees for and analyze the spectra of inverses A^{-1} $(f = x^{-1}, g = x, fg = 1)$, resolvents $(A-\lambda I)^{-1}$ $(f = (x-\lambda)^{-1}, g = (x-\lambda), fg = 1)$ and square roots $\sqrt{A}$ $(f = \sqrt{x}, g = x, f^2 - g = 0)$.

For any positive semidefinite selfadjoint operator A, $f(A)$ is obviously positive semidefinite and selfadjoint again by theorem 2.2.6.III and IV if $f \in C(\sigma(A))$ maps to the positive reals in the sense that $f(\sigma(A)) \subset \mathbb{R}_+ \cup \{0\}$. Let now $A \in \mathcal{B}(\mathcal{H})$ be positive semidefinite and selfadjoint. Furthermore suppose it to be invertible for the time. Then $\psi_A : C(\sigma(A)) \mapsto \mathcal{B}(\mathcal{H})$ exists, is usually called the functional calculus of A and indicates via $fx = 1 \Leftrightarrow f = x^{-1}, x \neq 0$ that

$$f(A)A = \psi_A(f)\psi_A(x) = \psi_A(fx) = \psi_A(1) = I. \tag{2.19}$$

Since $f(A)$ and A commute by $f(A)A = \psi_A(fx) = \psi_A(xf) = Af(A)$, $f(A) = \psi_A(x^{-1})$ is the unique two sided inverse of A. By theorem 2.2.6.IV

$$\sigma(f(A)) = \{\lambda^{-1} : \lambda \in \sigma(A)\}. \tag{2.20}$$

This shows that $\lambda_{\max}(A^{-1}) = 1/\lambda_{\min}(A)$ and $\lambda_{\min}(A^{-1}) = 1/\lambda_{\max}(A)$ and recalling theorem 2.2.3.V implies $\|A^{-1}\|_{op} = 1/\lambda_{\min}(A)$. This inspires the provably correct [93, p. 239] conjecture that A is invertible iff $0 \notin \sigma(A)$. In the context of covariance operators C_x this explains, why no bounded inverse exists — $\lim_{k\to\infty} \lambda_k = 0$ and $\sigma(C_x)$ is closed implying $0 \in \sigma(C_x)$. This results in noninvertibility of C_x inside the algebra of bounded operators and, even in the finite but high-dimensional case, unfavorable condition numbers for both C_x and C_x^{-1} because for n indicating the number of dimensions:

$$\begin{aligned} \lim_{n\to\infty} \mu(C_x) &= \lim_{n\to\infty} \lambda_{\max}(C_x)/\lambda_{\min}(C_x) \\ &= \lim_{n\to\infty} \lambda_{\max}(C_x^{-1})/\lambda_{\min}(C_x^{-1}) = \lim_{n\to\infty} \mu(C_x^{-1}). \end{aligned} \tag{2.21}$$

Here $\mu(A)$ is the condition number of A [106, p. 37] and the central terms in the chain equality both diverge for nonzero C_x. One may slightly generalize this sequence of comments and investigate the functions $f_\alpha \in C(\sigma(A))$ satisfying $f_\alpha(x - \alpha) = 1 \;\; \forall \alpha \notin \sigma(A)$ to conclude for the spectrum of the resolvent $R(\lambda)$

$$\begin{aligned} \sigma(R(\lambda)) &= f_\alpha(\sigma(A)) \\ &= \{(\lambda - \alpha)^{-1} : \lambda \in \sigma(A)\}. \end{aligned} \tag{2.22}$$

In light of the preceeding comments regarding the spectra of positive selfadjoint operators (theorem 2.2.3.III) like the covariance operator C_x, one has

$$\sigma(\alpha I - C_x) = \{\alpha - \lambda : \lambda \in \sigma(C_x)\} \tag{2.23}$$

for $\alpha \in \mathbb{R}$ and this allows to deduce that $\lambda_{\max} = \inf \alpha : \alpha I - C_x \succeq 0$. This is a well known result in semidefinite programming enabling minimization of maximum eigenvalues [112, p. 4].

Similarly useful is the construction of square roots of positive semidefinite operators $A \in \mathcal{B}(\mathcal{H})$. Let again $w := W_\cdot$ be white noise and $x := X_\cdot$ be a mean-zero finite energy process, both finite dimensional to avoid compactness problems. The covariance operators C_x, C_w can be written as [26, p. 96]

$$C_x = E[x \otimes x^*] \quad C_w = E[w \otimes w*] = I. \tag{2.24}$$

For C_x apply the functional calculus ψ_{C_x} to the functions $f = \sqrt{t}$ and $g = \sqrt{t}^{-1}$ for which obviously

$$f^2 = t \quad \text{and} \quad g^2 = 1/t$$

for $t \neq 0$. Then $\psi_{C_x}(f) = f(C_x)$ and $\psi_{C_x}(g) = g(C_x)$ satisfy

$$\begin{aligned} y = \sqrt{C_x} w \Rightarrow C_y &= E[\sqrt{C_x} w \otimes w^* \sqrt{C_x}^*] = \sqrt{C_x} I \sqrt{C_x}^* \qquad (2.25) \\ &= \psi_{C_x}(f)\psi_{C_x}(f) = \psi_{C_x}(f^2) = \psi_{C_x}(t) = C_x \\ v = \sqrt{C_x}^{-1} x \Rightarrow C_v &= E[\sqrt{C_x}^{-1} x \otimes x^* \sqrt{C_x}^{-*}] = \sqrt{C_x}^{-1} C_x \sqrt{C_x}^{-1} \qquad (2.26) \\ &= \psi_{C_x}(g)\psi_{C_x}(t)\psi_{C_x}(g) = \psi_{C_x}(gtg) = \psi_{C_x}(1) = I, \end{aligned}$$

i.e. $\psi_{C_x}(g)$ is the whitening and $\psi_{C_x}(f)$ is the de-whitening operator. There exists more than one de-whitening operator that maps white noise into a process with prespecified covariance operator C_x as a short investigation reveals. Let A now be any bounded normal operator on $\mathcal{H}$. Since A^*A satisfies $\langle A^*Af, f\rangle_{\mathcal{H}} = \langle Af, Af\rangle_{\mathcal{H}} = \|Af\|^2_{\mathcal{H}} \geq 0 \;\; \forall f \in \mathcal{H}$ one finds $\sigma(A^*A) \subset \mathbb{R}_+ \cup \{0\}$. Then $\sqrt{A^*A} =: |A|$ can be constructed via the functional calculus $\psi_{A^*A}(\cdot)$ applied to $f = \sqrt{t}$. A unitary operator U exists [93, p. 242] such that

$$A = U|A| = |A|U. \tag{2.27}$$

This is called the polar decomposition of A and it is clear that, irrespective of the unitary operator U,

$$E[|C_x|Uw \otimes w^*U^*|C_x|^*] = |C_x|UIU^*|C_x|^* = C_x$$

implying that the de-whitening operator is unique only up to a unitary transformation. It is interesting to apply complex exponential functions to selfadjoint operators A to generate operators of type e^{itA}. By the homomorphism property of the functional calculus, if $f_t(\cdot) = \exp(it\cdot)$ and $f_s(\cdot) = \exp(is\cdot)$ then

$$f_t(A)f_s(A) = \psi_A(f_t)\psi_A(f_s) = \psi_A(f_tf_s)\psi_A(f_{t+s}) = f_{t+s}(A).$$

They behave just like the exponential function and arise frequently as solutions to time evolution equations; this will be postponed however to the next sections and as a last example for the application of $\psi_A : C(\sigma(A)) \to \mathcal{B}(\mathcal{H})$ we mention decompositions of covariance operators. This is the property of the next theorem, whose proof depends on the Karhunen-Loewe expansion that is not rigorously formulated until subsection 2.3.1 theorem 2.3.5 since it requires spectral theory and additional nomenclature.

Theorem 2.2.7 (Functional calculus for covariance operators) *Let C_x denote the covariance operator of the finite energy stochastic process $x := X^{\cdot}_{\cdot} : \Omega \times T \to \mathbb{R}$. For any placeholder expression y, interpret C_y in the same way.*

I *For any $g \in C(\sigma(C_x))$ set $f(\lambda) = \lambda|g(\lambda)|^2$. Then $C_{[g(C_x)]x} = f(C_x)$.*

II *Let $\varphi_1, \varphi_2 \in C(\sigma(C_x))$ be a partition of unity on $\sigma(C_x)$; i.e. $\varphi_1(\lambda)+\varphi_2(\lambda) = 1 \;\; \forall\lambda \in \sigma(C_x)$. Then*

$$C_x = [\lambda\varphi_1(\lambda)](C_x) + [\lambda\varphi_2(\lambda)](C_x) = C_{[\sqrt{\varphi_1(\cdot)}](C_x)x} + C_{[\sqrt{\varphi_2(\cdot)}](C_x)x}$$

III *Let $g_1, g_2 \in C(\sigma(C_x))$ be real valued and set $f_k(\lambda) = \lambda|g_k(\lambda)|^2$ for $k = 1, 2$. Then*

$$C_{[g_1(\cdot)+g_2(\cdot)](C_x)x} = f_1(C_x) + f_2(C_x)$$

if and only iff $\operatorname{supp} g_1 \cap \operatorname{supp} g_2 \subset \{0\}$. *In that case also $[g_1(C_x)]x$ and $[g_2(C_x)]x$ are uncorrelated.*

IV *If* $\operatorname{supp} g_1 \cap \operatorname{supp} g_2 = \emptyset$ *with* $g_1(\lambda), g_2(\lambda) \in \{0, 1\} \quad \forall \lambda \in \sigma(C_x)$, *then* $\psi_{C_x}(g_1)$ *and* $\psi_{C_x}(g_2)$ *are mutually orthogonal projections. When furthermore* g_1 *and* g_2 *form a partition of unity and* $\operatorname{supp} g_1 \cup \operatorname{supp} g_2 = \sigma(C_x)$ *then they are exhaustive and* $\psi_{C_x}(g_1) + \psi_{C_x}(g_2) = I$.

Proof: Statement I follows from the Karhunen Loewe decomposition of a stochastic process and statements II-IV are simple corollaries of I, the homomorphism property of functional calculus and some straightforward computation. As always, x is assumed to be zero-mean and C_x is then $C_x = E[x \otimes x^*] = \sum_{k=1}^{\infty} \lambda_k e_k \otimes e_k^*$ by Mercers theorem. If $g \in C(\sigma(C_x))$ then it follows that $g(C_x) = \sum_{k=1}^{\infty} g(\lambda_k) e_k \otimes e_k^*$ and consequently

$$\begin{aligned} E\left[g(C_x)x \otimes x^* g(C_x)^*\right] &= g(C_x)C_x g(C_x)^* \\ &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} g(\lambda_i) e_i \otimes e_i^* \lambda_j e_j \otimes e_j^* \overline{g(\lambda_k)} e_k \otimes e_k^* \\ &= \sum_{k=1}^{\infty} \lambda_k |g(\lambda_k)|^2 e_k \otimes e_k^* \\ &= f(C_x) \end{aligned}$$

with $f(\lambda) = \lambda |g(\lambda)|^2$. This proves I. Using the property that φ_1, φ_2 are partitions of unity, II follows immediately by applying I to $C_{[\sqrt{\varphi_1(C_x)}]x}$ and $C_{[\sqrt{\varphi_2(C_x)}]x}$. For III note that I implies

$$\begin{aligned} C_{[g_1(\cdot)+g_2(\cdot)](C_x)x} &= [g_1(\cdot) + g_2(\cdot)](C_x)C_x[g_1(\cdot) + g_2(\cdot)]^*(C_x) \\ &= g_1(C_x)C_x g_1(C_x)^* + g_2(C_x)C_x g(C_x)^* \\ &\qquad g_1(C_x)C_x g_2(C_X)^* + g_2(C_x)C_x g_1(C_x)^* \\ &= f_1(C_x) + f_2(C_x) + 2C_x g_1(C_x) g_2(C_x) \end{aligned}$$

where in the last step selfadjointness of $g_k(C_x)$ for g_k real-valued and the commutativity of functions of C_x was used. Now $C_x g_1(C_x) g_2(C_x) = 0$ if and only if $\operatorname{supp} g_1 \cap \operatorname{supp} g_2 \subset \{0\}$ as follows from the calculations below.

$$\begin{aligned} &i) \operatorname{supp} g_1 \cap \operatorname{supp} g_2 \subset \{0\} && \Rightarrow C_x g_1(C_x) g_2(C_x) \\ & && = \psi_{C_x}(\lambda g_1(\lambda) g_2(\lambda)) = \psi_{C_x}(0) = 0 \\ &ii) \neg \operatorname{supp} g_1 \cap \operatorname{supp} g_2 \subset \{0\} && \Rightarrow C_x g_1(C_x) g_2(C_x) \\ & && = \psi_{C_x}(\underbrace{\lambda g_1(\lambda) g_2(\lambda)}_{q(\lambda)}) = \psi_{C_x}(q(\cdot)) \neq 0 \end{aligned}$$

That for nonzero $q(\cdot)$ also $\psi_{C_x}(q) \neq 0$ follows from $\|\psi_{C_x}(q)\|_{op} = \sup\{|q(\lambda)| : \lambda \in \sigma(C_x)\} > 0$ for $q(\cdot)$ not vanishing identically on $\sigma(C_x)$. By the non-degeneracy of the operator norm, $\psi_{C_x}(q) \neq 0$. Since $E[g_1(C_x)x \otimes x^* g_2(C_x)] = C_x g_1(C_x) g_2(C_x) = 0$ for $\operatorname{supp} g_1 \cap \operatorname{supp} g_2 \subset \{0\}$ the uncorrelatedness assertion is

proven as well. Therefore II holds in its entirety.
If even $\operatorname{supp} g_1 \cap \operatorname{supp} g_2 = \emptyset$ holds, then $\psi_{C_x}(g_1)\psi_{C_x}(g_2) = \psi_{C_x}(g_1 g_2) = 0$ and if in addition $g_k(\lambda) \in \{0,1\}$ for $\lambda \in \sigma(C_x)$ then

i) $\psi_{C_x}(g_k)\psi_{C_x}(g_k) = \psi_{C_x}(g_k^2) = \psi_{C_x}(g_k)$

ii) $\psi_{C_x}(g_k)^* = \psi_{C_x}(\overline{g_k}) = \psi_{C_x}(g_k)$

Consequently $\psi_{C_x}(g_k), k = 1,2$ are orthogonal projections and their ranges are orthogonal by theorem 2.1.37. When $g_1(\lambda)+g_2(\lambda) = 1 \ \forall\lambda \in \sigma(C_x)$ then $\psi_{C_x}(g_1)+\psi_{C_x}(g_2) = \psi_{C_x}(g_1+g_2) = \psi_{C_x}(1) = I$. □

Remark The theorem is quite easily understood in the context of low-pass or high-pass filtering. Suppose w is a Wiener process. Then C_w has as eigenfunctions the Fourier-basis and the eigenvalues decay monotonically with the spectrum [38]. Suppose, C_w is split into a low-frequency part $C_L = \psi_{C_w}(\lambda 1_{[0,B)})$ and a high-frequency part $C_H = \psi_{C_w}(\lambda 1_{[B,\infty)})$ where λ indicates the independent variable and 1_E is the characteristic function of the set E evaluated at λ. One then has $C_w = C_L + C_H$ and it is implied by theorem 2.2.7.III that knowing the high-frequent part of a signal does not allow forecasting the low-frequent part. See figure 2.8 for an illustration.

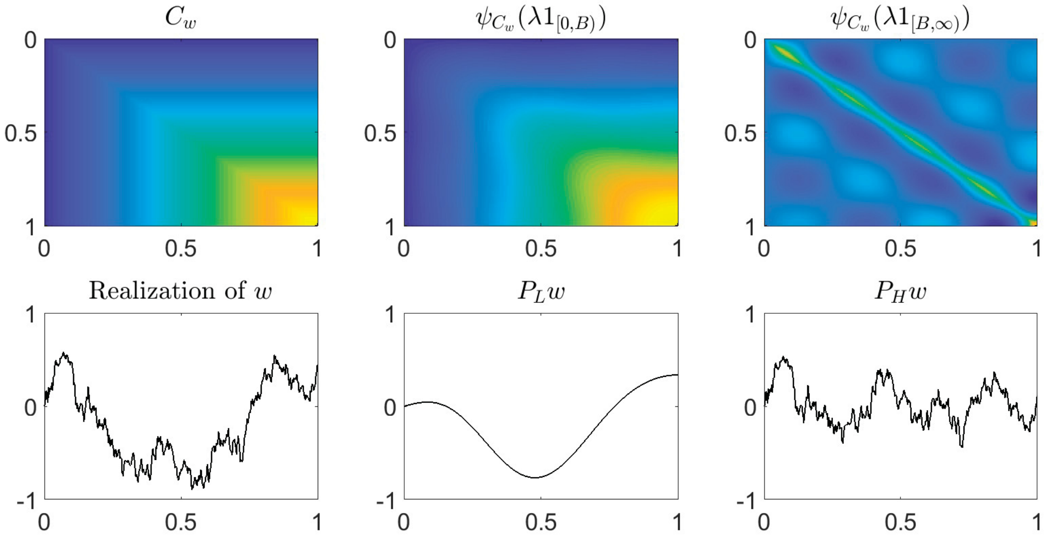


Figure 2.8: The covariance matrix C_w of a discrete Wiener process w on $T = [0,1]$ has covariance function $E[W_t W_s] = \min(t,s)$; $C_w = C_L + C_H$ and $E[[P_L w](s)[(P_H w](t)] = 0$ for all $s,t \in T$ as claimed in the preceeding remarks based on theorem 2.2.7. Here P_L, P_H are the orthogonal projections $1_{[0,B)}(C_w)$ and $1_{[B,\infty)}(C_w)$ as constructed according to theorem 2.2.7.IV.

For a fixed selfadjoint operator $A \in \mathcal{B}(\mathcal{H})$ the size of the set $\{f(A) : f \in C(\sigma(A))\}$ and its algebra structure can be surveyed quite easily, when A is the covariance operator C_x of a finite energy stochastic process x. To this end, note that the algebra $\mathcal{A}$ generated by C_x and I has a spectrum $\sigma(\mathcal{A})$ consisting of multiplicative linear functionals on $\mathcal{A}$ that is topologically isomorphic to $\sigma(C_x)$ by [63, p. 8]. Since $\mathcal{A}$ consists of bounded selfadjoint operators, it is a unital C^*-algebra and as for any

$f \in C(\sigma(C_x)),\ \ f(A)$ commutes with A by

$$Af(A) = \psi_A(x)\psi_A(f) = \psi_A(xf) = \psi_A(fx) = f(A)A \tag{2.28}$$

for all $A \in \mathcal{A}$, it is also commutative. As such the commutative version of the Gelfand-Naimark theorem [63, p. 11] intervenes and guarantees that $C(\sigma(C_x))$ and $\mathcal{A}$ are isometrically $*$-isomorphic as Banach algebras and there exists a one to one correspondence between continuous functions on the spectrum of C_x and elements of $\mathcal{A}$. If C_x has simple spectrum — the multiplicity of each eigenvalue is 1 — then a bounded operator $S \in \mathcal{B}(\mathcal{H})$ commutes with C_x if and only if $S = f(C_x)$ [4, p. 77] and consequently the set $\{f(C_x) : f \in C(\sigma(C_x))\}$ is isomorphic to the set of all $S \in \mathcal{B}(\mathcal{H})$ that commute with C_x.

Again under the assumption of a simple spectrum, one last comment is made. It is well known [178, p. 46] that if two selfadjoint operators with simple spectrum commute, then they have the same eigenfunctions. Since by equation 2.28 C_x and $f(C_x)$ commute, C_x and $f(C_x) = C_y$ share the same eigenfunctions if $f : \sigma(C_x) \to \mathbb{R}_+ \cup \{0\}$ is injective. In light of $f(C_x)$ as a selfadjoint operator being completely described by spectrum and eigenfunctions, this suggests that application of f can change any operator only as far as its spectrum is concerned and leaves its eigenvectors unperturbed.

2.2.3 The spectral theorem

A functional calculus for function classes more general than continuous functions is reviewed briefly for bounded selfadjoint operators at first and it is shown how projection valued measures can be assembled to yield this type of functions of an operator. Inversion of this construction leads to the spectral theorem whose implications are illustrated. The more general spectral theorem for not necessarily bounded selfadjoint operators — needed in the context of unbounded observables occurring in the description of physical systems — is mentioned but not stated. The theorem will be applied to the differentiation and Fourier-Plancherel operators and to some of the usual constituents of the Schrödinger equation, of which an interpretation is offered in subsection 2.2.4. Intuition is provided regarding the relationship between approximately satisfied differential equations and Hilbert spaces in which the induced norm provides a discrepancy measure quantifying a function's departure from being in the DE's solution set.

As indicated in the last subsection, it is possible to extend the functional calculus $\psi_A : C(\sigma(A)) \to \mathcal{B}(\mathcal{H})$ associated to a selfadjoint operator $A\mathcal{H} \to \mathcal{H}$ to functions that are no longer continuous on $\sigma(A)$ but merely bounded and Borel-measurable; this set is denoted by $\mathcal{B}(\sigma(A))$. These functions include certain step functions and characteristic functions of Borel-sets on $\sigma(A)$ [36, p. 59], [63, p. 20]. In theorem 2.2.7 it was shown that if $f(\sigma(A)) \subset \{0, 1\}$ then $\psi_A(f)$ is an orthogonal projection. A Hilbert space valued measure on $\sigma(A) \subset \mathbb{R}$ might be constructed by

$$P(E) = \psi_A(1_E)$$

where $E \subset \sigma(A)$ and 1_E is the characteristic function of E. Since $1_E(\lambda)$ is real for all values of λ and $1_E^2 = 1_E$, $\psi_A(1_E)$ is selfadjoint and idempotent; i.e. an orthogonal projection onto its range. Choosing any real-valued $f \in \mathcal{B}(\sigma(A))$, the equation

$$B = \int_{-\infty}^{\infty} f(\lambda) dP(\lambda) \tag{2.29}$$

is interpreted in the Riemann-Stieltjes sense [164, p. 172] with $P(\lambda) = \psi_A(1_{(-\infty,\lambda]})$. It results in a selfadjoint bounded operator with operator norm $\|B\|_{op} = \|f_{\sigma(A)}\|_\infty$ [63, p. 19]. The statement guaranteeing the possibility of assembling a self adjoint operator in this way is sometimes referred to as the inverse spectral theorem.

Clearly, the more interesting case is when an operator is given and its decomposition in terms of projections is to be determined. That this is always possible for a not necessarily bounded selfadjoint operator is the content of the spectral theorem. We present here a version due to Folland [63, pp. 20-27], which asserts under certain conditions diagonalizability of a commutative algebra of operators in the sense of equation 2.29 employing only a single family of projections.

Theorem 2.2.8 (Spectral theorem) *Let $\mathcal{A}$ be a commutative C^*-algebra identified with some subalgebra of bounded operators on a Hilbert space $\mathcal{H}$. Let the spectrum be $\sigma(\mathcal{A})$ on which the standard Borel σ-algebra Σ is constructed. Denote for $A \in \mathcal{A}$ by $\hat{A}$ any A's Gelfand transform $\Gamma_{\mathcal{A}} : \mathcal{A} \ni A \mapsto \Gamma_{\mathcal{A}}(A) = \hat{A} \in C(\sigma(\mathcal{A}))$ with $\hat{A} : \sigma(\mathcal{A}) \ni \lambda \mapsto \lambda(A) \in \mathbb{C}$. Then there $\exists!\ P_{\mathcal{A}}(\cdot) : \Sigma \ni E \mapsto P_{\mathcal{A}}(E) \in \mathcal{B}(\mathcal{H})$, called projection valued measure, such that*

I For all $A \in \mathcal{A}$ and $f \in \mathcal{B}(\sigma(\mathcal{A}))$, $A = \int_{\sigma(\mathcal{A})} \hat{A} dP_{\mathcal{A}}$, $f(A) = \int_{\sigma(\mathcal{A})} f \circ \hat{A} dP_{\mathcal{A}}$ where the map $\psi_A : f \mapsto f(A)$ coincides with the inverse Gelfand transform $\Gamma_{\mathcal{A}}^{-1}$ for $f \in C(\sigma(\mathcal{A}))$. For an algebra generated by a single selfadjoint operator A and the identity I, this trivially implies $A = \int_{\sigma(A)} \lambda dP(\lambda)$ and $f(A) = \int_{\sigma(A)} f(\lambda) dP(\lambda)$ where the homeomorphism $\hat{A} : \sigma(\mathcal{A}) \to \sigma(A)$ was used to simplify the range of integration.

II For all $E \in \Sigma$, $P_{\mathcal{A}}(E) \in \mathcal{B}(\mathcal{H})$ is an orthogonal projection onto some subspace $\mathcal{H}_E \overline{\boxslash} \mathcal{H}$.

III The projection valued measure is increasing in the sense that for $E, F \in \Sigma, E \subset F$ implies $P_{\mathcal{A}}(E) \prec P_{\mathcal{A}}(F)$ with respect to the partial order on positive selfadjoint operators. Furthermore for arbitrary $E, F \in \Sigma$ one has $P_{\mathcal{A}}(\emptyset) = 0, P_{\mathcal{A}}(\sigma(\mathcal{A})) = I, P_{\mathcal{A}}(E \cap F) = P_{\mathcal{A}}(E) P_{\mathcal{A}}(F)$ and if $E_i \cap E_j = \emptyset\ \forall i \neq j$ then $P_{\mathcal{A}}\left(\cup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P_{\mathcal{A}}(E_i)$.

IV For $T \in \mathcal{B}(\mathcal{H})$, the conditions i)$[T, A] = 0\ \ \forall A \in \mathcal{A}$, ii) $[T, P_{\mathcal{A}}(E)] = 0\ \ \forall E \in \Sigma$ and iii) $[T, f(A)] = 0\ \ \forall f \in \mathcal{B}(\sigma(\mathcal{A}))$ are all equivalent. Here $[\cdot, \cdot]$ denotes the commutator $[T, A] = TA - AT \in \mathcal{B}(\mathcal{H})$.

For the case of one selfadjoint operator A acting on $\mathbb{R}^n$ — i.e. a symmetric matrix — the theorem simplifies to asserting that A can be diagonalized and that there exists a family of mutually orthogonal projections $\{P_\lambda\}_{\lambda\in\sigma(A)}$ onto the eigenspace corresponding to λ such that [63, p. 17]

$$I = \sum_{\lambda\in\sigma(A)} P_\lambda \quad A = \sum_{\lambda\in\sigma(A)} \lambda P_\lambda \quad f(A) = \sum_{\lambda\in\sigma(A)} f(\lambda)P_\lambda. \tag{2.30}$$

In the finite dimensional matrix case it is entirely standard to numerically generate the spectral family $\{P_\lambda\}_{\lambda\in\sigma(A)}$ given some symmetric matrix A via eigendecomposition and the functional calculus introduced in the previous subsection becomes applicable for sufficiently small dimensions.

If A is a compact positive definite operator — say the covariance operator $C_x : \mathcal{H} \to \mathcal{H}$ of a finite energy stochastic process x — then there is an ONB consisting of eigenvectors of C_x [63, p. 27]. Furthermore positivity and countability of C_x's spectrum lead to the expression $C_x = \sum_{\lambda\in\sigma(C_x)} \lambda P_\lambda$. Selfadjoint compact operators may always be written in the form $C_x f = \sum_{k=1}^{\infty} \alpha_k \langle f, \overline{g_k}\rangle_{\mathcal{H}} g_k$ [164, p. 233], the one dimensional projections P_λ are of finite rank, therefore compact and might be written as $P_\lambda f = \langle e_\lambda, f\rangle_{\mathcal{H}} e_\lambda$ instead with $e_\lambda \in \mathcal{H}$ a unit vector spanning $P_\lambda(\mathcal{H})$. This implies with the usual identifications of $e_{\lambda_k} = e_k$ and $e_k^* \in \mathcal{H}^*, e_K^* f = \langle e_k, f\rangle_{\mathcal{H}}$ as an element of the dual space of $\mathcal{H}$ the decomposition

$$C_x = \sum_{k=1}^{\infty} \lambda_k e_k \otimes e_k^* \quad \{\lambda_k\}_{k=1}^{\infty} \subset \mathbb{R}_+ \cup \{0\}. \tag{2.31}$$

If x is a stochastic process on index set T then one may evaluate $C_x f$ directly as

$$(C_x f)(s) = \sum_{k=1}^{\infty} \lambda_k e_k(s) \langle e_k(\cdot), f\rangle_{\mathcal{H}} = \langle u(s,\cdot), f\rangle_{\mathcal{H}} \tag{2.32}$$

$$u(s,t) = \sum_{k=1}^{\infty} \lambda_k e_k(s) e_k(t) \;\; \forall s,t \in T \tag{2.33}$$

with $K(\cdot,\cdot)$ the kernel of the operator C_x. Since kernels of operators are unique given certain conditions on $\mathcal{H}$ [20, p. 27] and with the remarks from subsection 2.1.4 one sees that $E[X_s X_t] = \sum_{k=1}^{\infty} \lambda_k e_k(s) e_k(t)$ is the second order moment function of the stochastic process x. This is also known as the Mercer decomposition [164, p. 245].

Remark Note that $u(s,t)$ is not the reproducing kernel of $\mathcal{H}$ but the kernel of the operator C_x. Both uses of the word "kernel" are unfortunately standard and not to be confused.

The expected energy $\|x\|_{\mathcal{H}}^2$ of the process is given by the trace of the operator C_x.

$$E[\|x\|_{\mathcal{H}}^2] = E\left[\sum_{k=1}^{\infty} |\langle x, e_k\rangle_{\mathcal{H}}|^2\right] = \sum_{k=1}^{\infty} \langle C_x e_k, e_k\rangle_{\mathcal{H}} = \operatorname{tr} C_x = \sum_{k=1}^{\infty} \lambda_k \tag{2.34}$$

We might furthermore derive a low rank approximation to C_x that is optimal for finite dimensional $\mathcal{H}$ in the sense of both the Frobenius and the spectral norm. For a given n, $C_x^{\text{approx}} = \sum_{k=1}^{n} \lambda_k e_k \otimes e_k^*$ is the minimizer of the problem

$$\min_{\{\alpha_k\}_{k=1}^n \subset \mathbb{R}, \{\varphi_k\}_{k=1}^n \subset \mathcal{H}} \qquad \| \sum_{k=1}^{n} \alpha_k \varphi_k \otimes \varphi_k^* - C_x \|_{F,2} \tag{2.35}$$

where $\|A\|_F = \sqrt{\operatorname{tr} A^* A}$ and as usual $\|A\|_2 = \sup\{|\lambda| : \lambda \in \sigma(A)\}$ [49]. It has other convenient properties related to the Karhunen Loewe expansion of a stochastic process that create opportunities for sparse representations and are surveyed later.

When investigating the dynamical behavior of physical systems, one often encounters selfadjoint differential operators A which are usually not bounded. By first subjecting A to the Cayley transform $U = \psi_A\left([x - i][x + i]^{-1}\right)$ and thereby mapping it into a unitary operator from whose spectral decomposition $U = \int_0^{2\pi} e^{i\mu} dP(\mu)$ the selfadjoint operator A may be extracted, an analogue the spectral theorem can be derived for unbounded operators [93, pp. 292-300]. Apart from some restrictions regarding the mode of convergence, the appropriate representation is

$$A = \int_{-\infty}^{\infty} \lambda dP(\lambda).$$

Differentiation $D = i\partial/\partial t$ and the Laplacian $\Delta = D^* D$ are selfadjoint on $\mathcal{H} = L^2(-\infty, \infty)$ as can be concluded by partial integration and observing that necessarily $\lim_{t\to\pm\infty} f(t) = 0$ for $f \in \mathcal{H}$. As is shown in [3, pp. 112-113], the differentiation operator is unitarily equivalent to the multiplication operator with the Fourier transform mediating between time/space and frequency domain.

Theorem 2.2.9 *Let $\mathcal{F} : \mathcal{H}_{abs} \ni f \mapsto \mathcal{F}f \in L^2(-\infty, \infty)$ with $\mathcal{F}f(\omega) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} e^{i\omega s} f(s) ds$ be the Fourier-Plancherel operator acting on the space $\mathcal{H}_{abs}$ of absolutely integrable functions. Furthermore let $D = i = i\partial/\partial t$ be the differentiation operator and $M : f(t) \mapsto tf(t)$ be multiplication by the independent variable with the domain of both operators sensibly restricted as for example in [4, p. 84]. Then*

$$D = \mathcal{F} M \mathcal{F}^* \tag{2.36}$$

i.e differentiation corresponds to multiplication in the frequency domain and the differentiation operator is diagonalizable by the Fourier transform.

As the projection valued measure for the multiplication operator $M : \operatorname{dom}_M \to L^2(-\infty, \infty)$,$\operatorname{dom}_M = \{f \in L^2(-\infty, \infty) : Mf \in L^2(-\infty, \infty)\}$ is seen to be [93, p. 303]

$$P_M([\alpha, \beta])f = \begin{cases} f(s) & s \in [\alpha, \beta] \\ 0 & s \notin [\alpha, \beta] \end{cases} \quad [\alpha, \beta] \subset \mathbb{R}^2$$

and the projection valued measure P_D for differentiation satisfies $P_D([\alpha,\beta]) = \mathcal{F}P_M([\alpha,\beta])\mathcal{F}^*$. A quick calculation [3, p. 84] reveals $P_D([\alpha,\beta]) = P_D(\beta) - P_D(\alpha)$ to be

$$\begin{aligned}
P_D([\alpha,\beta])f = \mathcal{F}P_M([\alpha,\beta])\underbrace{\mathcal{F}^*f}_{g} &= \frac{1}{2\pi}\int_\alpha^\beta g(s)e^{i\omega s}ds \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_\alpha^\beta f(\gamma)e^{is\gamma}e^{i\omega s}d\gamma ds \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{i\beta(\gamma-\omega)} - e^{i\alpha(\gamma-\omega)}}{i(\gamma-\omega)}f(\gamma)d\gamma.
\end{aligned}$$

This shows that the spectral family of the differentiation operator is explicitly known. By an extension of the Riesz-Dunford functional calculus to sectorial operators it is possible to define powers of selfadjoint unbounded operators [16] analogously to the bounded case. A detailed study of the many technical intricacies is found for example in [89]. From

$$\begin{aligned}
-\Delta &= DD = \mathcal{F}M^2\mathcal{F}^* \\
&= \int_{-\infty}^{\infty}\lambda^2 dP_D(\lambda) = \int_0^\infty \lambda d\left(P_D(\sqrt{\lambda}) - P_D(\sqrt{\lambda})\right)
\end{aligned}$$

the negative Laplacian is seen to be a positive operator with projection valued measure [4, p.87]

$$\begin{aligned}
P_\Delta([\alpha,\beta]) &= P_\Delta(\beta) - P_\Delta(\alpha) \\
P_\Delta(\lambda)f(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\sin\sqrt{\lambda}(\gamma-\omega)}{\gamma-\omega}f(\gamma)d\gamma. \qquad (2.37)
\end{aligned}$$

Clearly, the differential operators D and Δ are important since they feature prominently in the description of many physical systems. More generally, if a systems state f is supposed to satisfy the differential equation

$$\begin{aligned}
Af &= g \qquad (2.38) \\
A &= \psi_D(p(\cdot))
\end{aligned}$$

where A is a generally unbounded polynomial $p(\cdot)$ in the differential operator $D = i\partial/\partial t$ and g some suitably chosen function, then one may be tempted to postulate a relaxed version of 2.38 as

$$Af - g = w \quad w \text{ white noise} \qquad (2.39)$$

and enter into the following calculation. Here as before, the author wants to avoid the complications of white noise in infinite dimensional spaces and reverts to the intuitive case of $\mathcal{H} = \mathbb{R}^n$, n finite with the operator A some suitable discretization of the original differential operator discussed above.

$$\begin{aligned} I = E[w \otimes w^*] &= A \underbrace{E[f \otimes f^*]}_{C_f} A^* + \underbrace{E[g \otimes g^*]}_{G} \\ AC_f A^* &= I - G \end{aligned} \tag{2.40}$$

If the problem were homogeneous ($g = 0$ deterministically) and A invertible, equation 2.40 is invertible and

$$\begin{aligned} C_f &= (A^*A)^{-1} \\ \|w\|^2_{\ell^2} = \|Af\|^2_{\ell^2} &= \langle A^*Af, f\rangle_{\ell^2} = \langle f, f\rangle_{\mathcal{H}} \end{aligned}$$

with $\langle h, g\rangle_{\mathcal{H}} = \langle A^*Ah, g\rangle_{\ell^2} = \langle C_f^{-1}h, g\rangle_{\ell^2}$. Then $\|w\|^2_{\ell^2} = \langle f, f\rangle_{\mathcal{H}}$ measures how much f deviates from a solution to the differential equation 2.39 with $g = 0$ and establishes a link between inner products $\langle \cdot, \cdot\rangle_{\mathcal{H}}$ and Hilbert spaces that satisfy some linear equation only approximately. More on the relation between covariance operators, kernels and regularization is collected in the survey paper by Steinke and Schölkopf [188] from which also the following insight is taken. From $D = \mathcal{F}M\mathcal{F}^*$ and $A = \psi_D(p(\cdot))$ follows $A = \mathcal{F}M_p\mathcal{F}^*$ and $A^*A = \mathcal{F}M_{|p(\cdot)|^2}\mathcal{F}^*$ where $M_{p(\cdot)}$ denotes multiplication by $p(t)$ where t is the independent variable. Then

I Apparently to calculate the inner product $\langle f, f, \rangle_{\mathcal{H}}$ one may instead evaluate the expression by looking at the inverse Fourier transforms and multiplying these with the polynomial $p(\cdot)$ to derive $p(\gamma)\check{f}(\gamma)$. This is readily seen by $\|w\|^2_{\ell^2} = \langle A^*Af, f\rangle_{\ell^2} = \langle \mathcal{F}M_{|p(\cdot)|^2}\mathcal{F}^*f, f\rangle_{\ell^2} = \langle M_{|p(\cdot)|}\check{f}, M_{|p(\cdot)|}\check{f}\rangle_{\ell^2}$ where the inverse Fourier transform of a function f is denoted by $\check{f}$.
As polynomials typically grow unboundedly on $\mathbb{R}$ the value $p(\gamma)$ increases for large γ leading to a penalization of higher frequencies as the discrepancy measure $\|w\|^2_{\ell^2} = \langle f, f, \rangle_{\mathcal{H}} = \langle p\check{f}, p\check{f}\rangle_{\ell^2} = \|p\check{f}\|^2_{\ell^2}$ contains an amplification factor of $|p(\gamma)|^2 >> 1$ for large frequencies γ.

II As I shows, the modification of the inner product by inclusion of $A^*A = C_f^{-1}$ has the effect of a regularization. As $C_f = (A^*A)^{-1} = \mathcal{F}\psi_D(|p(\cdot)|^{-2})\mathcal{F}^*$, the corresponding kernel function of the operator C_f is given in the discrete case as

$$\begin{aligned} (C_f g)(t) &= \sum_{j=1}^{n} u(s_j, t)g(s_j) \\ u(s_k, s_l) &= \sum_{j=1}^{n} \frac{1}{|p(\lambda_j)|^2} \exp\left(\frac{2\pi i}{n} j(k-l)\right) \end{aligned}$$

For more details and a proof as well as interpretations of the above facts when $(C_f g)(t) = \int_{-\infty}^{\infty} u(s, t)g(s)ds$ with $u(\cdot, \cdot)$ typical kernel functions like the squared exponential, see [188] and [75].

Therefore approximate satisfaction of a differential equation $\psi_D(p)f =$

$\sum_{j=1}^{n} \alpha_j D^j f = 0$ prescribes a certain covariance structure for the process f whose second order moments are related to the (discrete) Fourier transform of a polynomial acting on D's spectrum. We summarize our findings in the following conjecture.

Conjecture 2.2.1 (Approximate determinism) *The positive definite linear operator C_f is the covariance operator of a stochastic process f satisfying $Af = w$ with w weakly white noise and A some differential operator iff C_f is diagonalizable by the Fourier transform. In this case f is second order stationary and C_f is an integral operator with translation invariant kernel. $|A|$ is given by $\mathcal{F} M_{1/\sqrt{q}} \mathcal{F}^*$ where $M_q = \mathcal{F}^* C_f \mathcal{F}$.*

2.2.4 One parameter unitary semigroups

The variational formulation of classical Newtonian mechanics is reviewed and the basic concepts of Hamiltonian mechanics are interpreted in the sense of Gauss' principle of least contraint. An adaptation of Schrödingers formulism opens up the possibility to cast the search for the time evolution of a mechanical system as a second order partial differential equation in which a self adjoint linear operator appears. Employing functional calculus and Stones theorem, the time evolution of a system through state space can be written with the help of one parameter unitary semigroups of operators. Diffusion processes are investigated for illustrative purposes and the question is raised, how partial knowledge of a systems Hamiltonian could be incorporated into estimation problems.

Newtons second law of motion states that $F = \dot{p}$ where F is the force, the dot signifies differentiation w.r.t time and p is the momentum consisting of the product of mass and velocity. As is well known from classical mechanics, there exist several reformulations of the laws of motion that shift the attention away from a differential interpretation focusing on incremental particle interactions and take a more holistic perspective via a variational approach asserting global optimality properties of the trajectories actually occurring in nature. Hamilton's principle for example states that for a motion between times t_1 and t_2 the action integral

$$\int_{t_1}^{t_2} T(t) - V(t) dt \qquad T: \text{ kinetic Energy } \; V: \text{potential Energy}$$

takes on a stationary value [205, p. 74]. From this, d'Alembert's principle can be derived of which Gauss gave a statistical reinterpretation that postulates a least-squares approach to mechanics [123, pp. 106-107]. For a system of n particles with masses m_k, accelerations $\ddot{q}_k$ and acted upon by forces F_k he proposed the investigation of the objective function

$$z = \sum_{k=1}^{n} \frac{1}{2m_k} \left(F_k - m_k \ddot{q}_k\right)^2$$

and showed that the actual motion occurring in nature minimizes z out of all possible continuous trajectories potentially traceable by the particles. If no constraints are present, $F = m\ddot{q}$ as one would expect. In all other cases this idea closely resembles the perspective put forward in the preceding section, where we allowed a differential equation to be satisfied only approximately as measured by the discrepancy term $Af = w$ whose severity we subsequently quantified via the quadratic scalar $\|w\|^2_{\mathcal{H}}$.

In the context of variational principles it is especially elegant to introduce the Hamiltonian function $H(q,p), q, p \in \mathbb{R}^n$ whose inputs are generalized coordinates q and generalized momenta p that signify the systems position and velocity in phase space Φ — the space of all possible configurations and configuration-changes of the system which is represented as a point in this $2m$ dimensional not necessarily geometrical space. For conservative systems, the Hamiltonian $H(q,p) = T + V$ signifies total energy and is the only quantity necessary to determine the systems time evolution via the canonical equations of motion [205, p. 78]

$$\dot{q}_k = \frac{\partial H}{\partial p_k} \qquad\qquad \dot{p}_k = -\frac{\partial H}{\partial q_k} \tag{2.41}$$

or in more compact notation

$$\dot{x} = \begin{bmatrix} \dot{q} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \operatorname{grad} H = J\nabla H \tag{2.42}$$

which looks a lot like a gradient system but with the symplectic gradient $J\nabla$ instead of a regular one. This guarantees a flow orthogonal to the direction of steepest descent and along the isolines of total energy.

Since $\frac{dH}{dt} = \sum_{k=1}^n \frac{\partial H}{\partial q_k}\dot{q}_k + \frac{\partial H}{\partial p_k}\dot{p}_k = \sum_{k=1}^n \frac{\partial H}{\partial q_k}\frac{\partial H}{\partial p_k} - \frac{\partial H}{\partial p_k}\frac{\partial H}{\partial q_k} = 0$, the Hamiltonian is furthermore a constant of the motion. It can be shown that for a Hamiltonian system, the Euler-Lagrange equation 2.43 as well as Liouville's theorem 2.44 towards the incompressibility of the flow in phase space Φ hold [123, pp. 164-167, pp. 178-180]. The following two statements hold.

I Lagrange equations of motion

$$\left[\frac{\partial}{\partial q_k} - \frac{d}{dt}\frac{\partial}{\partial \dot{q}_k}\right]\mathcal{L} = 0 \qquad \text{where } \mathcal{L}(t,q,\dot{q}) = T - V, p_k = \frac{\partial \mathcal{L}}{\partial \dot{q}_l}, q_k = \frac{\partial H}{\partial p_k} \tag{2.43}$$

II Liouvilles theorem

$$\begin{aligned} \operatorname{div}(\dot{q}_1, ..., \dot{q}_n, \dot{p}_1, ..., \dot{p}_n) &= \sum_{k=1}^n \frac{\partial}{\partial q_k}\dot{q}_k + \frac{\partial}{\partial p_k}\dot{p}_k \\ &= \sum_{k=1}^n \left[\frac{\partial}{\partial q_k}\frac{\partial}{\partial p_k} - \frac{\partial}{\partial p_k}\frac{\partial}{\partial q_k}\right] H = 0 \end{aligned} \tag{2.44}$$

Both theorems are of importance to us. The Euler-Lagrange equations allow to swap

between formulations in terms of often quadratic energy functionals of the form $\int_{t_1}^{t_2} T(t) - V(t)dt = \frac{1}{2}\|D_t q\|_{\mathcal{H}_M}^2 - \langle 1, V(q)\rangle_{L^2([t_1,t_2])}$ where $\mathcal{H}_M$ is a non-orthogonal sum of n copies of $L^2([t_1,t_2])$ with inner product $\langle f,g\rangle_{\mathcal{H}_M} = \int_{t_1}^{t_2} f^T(t)Mg(t)dt$, $M \in \mathbb{R}^n \otimes \mathbb{R}^n \succeq 0$, and equivalent differential equations for which efficient numerical computation is possible.
Liouville's theorem implies that time evolution in phase space Φ is actually measure preserving in the sense that the map $\Phi \ni x_0 \mapsto x_t \in \Phi$ where $x_t = (q(t), p(t))$ are position and momentum at time t leaves invariant the integral [164, p. 388]

$$\int_\Phi f(x_0)dx = \int_\Phi f(x_t)dx \qquad t \in \mathbb{R}, f \in L^1(\Phi).$$

In the Koopman-von Neumann interpretation of classical mechanics [118] the motion of a single point through phase space Φ is replaced by the time evolution of a probability density $\rho : \Phi \to \mathbb{R}$ whose normalization condition may be written as $\rho = \psi\overline{\psi}$ with $\psi \in L^2(\Phi)$ and $\|\psi\|_{L^2(\Phi)} = 1$. The ψ's are wavefunctions that populate the unit sphere of $L^2(\Phi)$ and satisfy a first-order-version of Schrödinger's equation with the Hamiltonian substituted by the Liouvillian. It can be shown [176, p. 13] that Liouville's theorem implies the total time derivative of the time dependent probability density $\rho(q,p,t)$ to vanish.

$$\frac{d\rho}{dt} = \frac{\partial\rho}{\partial t} + \sum_{k=1}^n \left[\frac{\partial\rho}{\partial q_k}\frac{\partial q_k}{\partial t} + \frac{\partial\rho}{\partial p_k}\frac{\partial p_k}{\partial t}\right] = 0$$

But then

$$\frac{\partial\rho}{\partial t} = -\sum_{k=1}^n \left[\frac{\partial\rho}{\partial q_k}\frac{\partial H}{\partial p_k} - \frac{\partial\rho}{\partial p_k}\frac{\partial H}{\partial q_k}\right] = \sum_{k=1}^n \left[\frac{\partial H}{\partial q_k}\frac{\partial}{\partial p_k} - \frac{\partial H}{\partial p_k}\frac{\partial}{\partial q_k}\right]\rho.$$

Finally by renaming the operatorial part of the right side to L/i, the evolution equation

$$i\dot{\rho} = L\rho \tag{2.45}$$

is recovered. Since the same equation governs the wavefunctions ψ via $i\dot{\psi} = L\psi$ as well [135, p. 10], one suspects that a family of operators $U_t, t \in T$ of the form $U_t = \exp(-itL)$ exists with $U_t\psi(\cdot,\cdot,0) = \psi(\cdot,\cdot,t)$. Due to the properties of time evolution $U_tU_s = U_{t+s}$ should hold and U_t would necessarily have to be unitary to conserve the property of $\rho(\cdot,\cdot,t) = \rho_t$ being a probability density integrating to one because

$$\int_\Phi \rho_t dx = \langle U_t\psi_0, U_t\psi_0\rangle_{L^2(\Phi)} = \langle U_t^*U_t\psi_0, \psi_0\rangle_{L^2(\Phi)}$$

for all $t \in T, \psi_0 \in L^2(\Phi)$ which is fulfilled if $U_t^*U_t = I$. The existence of a unitary operator with these properties is indeed guaranteed by Stones theorem on one parameter unitary semigroups which is postulated without any reference to preceding physical considerations, see [164, pp. 380-388] for more details.

Theorem 2.2.10 (Stones theorem) *Suppose that a family $\{U_t\}_{t\in\mathbb{R}}$ of linear operators on $\mathcal{H}$ is given and depends continuously on t in the sense that $\lim_{s\to t}\|U_s - U_t\|_{op} = 0$. Suppose furthermore that $U_0 = I$ and $U_tU_s = U_{t+s}$. This one parameter unitary semigroup has a simple representation in terms of complex exponentials and is generated by a selfadjoint operator, that is*

$$U_t = \int_{-\infty}^{\infty} e^{it\lambda} dP(\lambda) \qquad\qquad U_t = \exp(itA) \qquad A^* = A$$

where $P(\lambda)$ defines a spectral family, the elements of which commute with $U_t \ \forall t \in \mathbb{R}$. The generator A is a selfadjoint not necessarily bounded linear operator satisfying $iA = \lim_{t\to 0} t^{-1}(U_t - I)$.

Note that for some $f_0 \in \mathcal{H}$ the expression for iAf_0 is actually the rate of change of f_0 if it evolves over time according to the law $f_t(\cdot) = U_t f_0(\cdot)$ with f_t then being a solution to the initial value problem $\partial/\partial t f_t(\cdot) = iAf_t(\cdot), f_0(\cdot)$ given. With some more work and by the usual procedure of variation of parameters one arrives at the solution to the inhomogeneous abstract Cauchy problem for positive times $t \geq 0$

$$\begin{aligned} &\dot{f}(t) = Af(t) + g(t) && g(t) \in \mathcal{H} \ \forall t \in \mathbb{R}_+ \cup \{0\} \\ &f(0) = f_0 && f_0 \in \mathcal{H} \end{aligned} \tag{2.46}$$

where $\mathcal{H}$ is given, A is a linear operator from $\mathcal{H}$ to $\mathcal{H}$ that generates a strongly continuous semigroup U_t in the sense of Stones theorem [52, p.166] and the solution $f(t)$ is to be found in $\mathcal{H}$ for all $t \geq 0$. It is given by the formula [52, p. 437]

$$f(t) = U_t f_0 + \int_0^t U_{t-s} g(s) ds. \tag{2.47}$$

This formula includes as special cases the solution to the heat, Schrödinger and wave equations since they may all be written as

$$\dot{f}(t) = Af(t) \qquad\qquad f(0) = f_0$$

where bar some constants A is either the Laplacian $-\Delta$, the Hamiltonian $H = \Delta + M_V$ for some potential V or the positive semidefinite square root $\sqrt{-\Delta}$ [57, p. 412] [99, p.334].

With an explicit functional calculus ψ_{C_x} available, it is possible to find families of covariance operators $f_t(C_x)$ that satisfy differential equations inspired by physical constraints. In analogy to the trivial scalar differential equation $\partial_t x = \lambda x$ solved by $x = c\exp(\lambda t)$, problems of type $\partial_t x = Ax$ with $x(t)$ an element of a Hilbert space $\mathcal{H}$ and $A : \mathcal{H} \to \mathcal{H}$ a linear operator, have solutions

$$x = x_0 \exp(At), x(0) = x_0.$$

This identity is a simple consequence of formula 2.47 with $g(\cdot) = 0$ and already this simple formula allows injection of physical knowledge into a stochastic process as illustrated by the following example.

Example 9 (Heat equation) Let $u(\cdot,\cdot) : T \times X \ni (t,x) \mapsto u(t,x) \in \mathbb{R}$ be the function that assigns to each location $x \in X$ its heat at time $t \in T$. It is determined completely by the initial value problem [57, p. 412]

$$\frac{\partial}{\partial t}u - \alpha \sum_{j=1}^{n} \frac{\partial^2}{\partial x_j^2} u = 0 \qquad n = \dim X$$
$$u(0,x) = u_0(x) \tag{2.48}$$

where α in this context is thermal diffusivity which we set to 1 for simplicity's sake and $u_0(\cdot)$ is known. If we introduce the shorthand notation $\Delta = \sum_{k=1}^{n} \partial^2/\partial x_k^2$ for the Laplace operator it is clear by the preceding comments that

$$u(t,\cdot) = \exp(t\Delta)u_0(\cdot) = \sum_{k=0}^{\infty} \frac{(t\Delta)^k}{k!} u_0(\cdot).$$

Suppose now that the initial condition is actually random and the stochastic process $u_0(\cdot)$ in space has covariance operator $C_{u_0} = C_u^{t=0}$. The covariance C_u^t of the full spatiotemporal process $u(t,x)$ given the random initialization $u_0(x)$ is then simply

$$C_u^t = E[u(t,\cdot) \otimes u(t,\cdot)^*] = e^{t\Delta} C_{u_0} e^{t\Delta}.$$

This allows accurate forecasting and reconstruction of a systems time evolution and even closed form solutions given only very sparsely distributed measurements. See figure 2.9 for an example which deals with heat dissipation in a rod. Applications to the more mechanically inclined dynamical systems encountered in geodesy are straightforward.

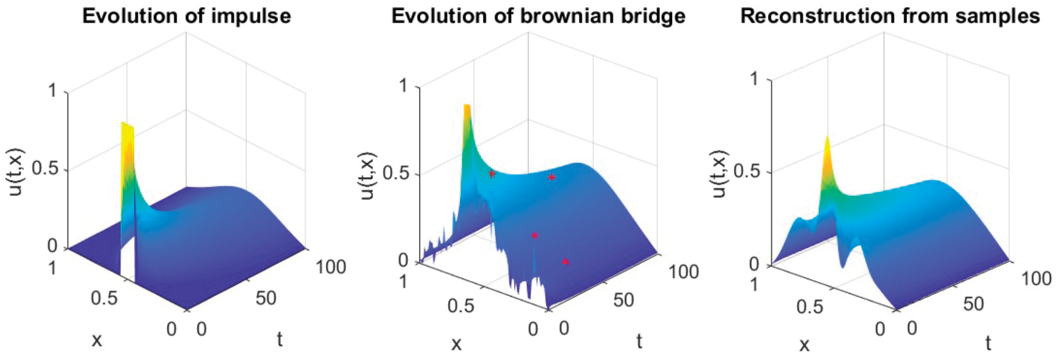


Figure 2.9: On the left the time evolution of a finite width impulse under the dynamics implied by the heat equation is plotted, whereas in the middle column the initial value $u_0(\cdot)$ was chosen randomly and the subjected to the time evolution operator $e^{t\Delta}$. The red asterisks mark measurements that were used to estimate the behavior exhibited in the middle column via $u(t,\cdot) = \exp(t\Delta)u_0^*, u_0^* = \operatorname{argmin}_{v \in \mathcal{H}_K} \|Av - u_m\|_{\ell^2}^2 + \|v\|_{\mathcal{H}_K}^2$ where u_m is the vector of measurements and A a linear operator consisting of point evaluations and exponentials of the Laplacian. The result of that estimation is plotted on the right, only five measurements were used to derive it.

Just from the properties of Δ, the limiting behavior of the system can be inferred. The Laplacian on a periodic domain $[0,1]$ with vanishing boundary terms is negative definite with eigenvalues $\lambda_k = -(k\pi)^2$ and (unnormalized) eigenfunctions $\varphi_k = \sin(k\pi x)$ for $k = 1, ...$ [57, p. 66]. The largest eigenvalue is still negative and associated to the function $\sin(\pi x)$, consequently $\exp(t\Delta) = \sum_{k=1}^{\infty} e^{t\lambda_k} P_\Delta(\lambda_k)$ is dominated by the first term of the infinite sum. For large t it is equal to $c\varphi_1(\cdot) \otimes \varphi_1(\cdot)^*$ for some small c. This also finds its analogy in the time evolution of the Covariance operator C_u^t, which is virtually indistinguishable from a rank-one operator after a sufficient amount of time, see figure 2.10. Obviously, in the limit $t \to \infty$, $\exp(t\Delta) = 0$ due to the negativity of the eigenvalues; all heat dissipates in the end.

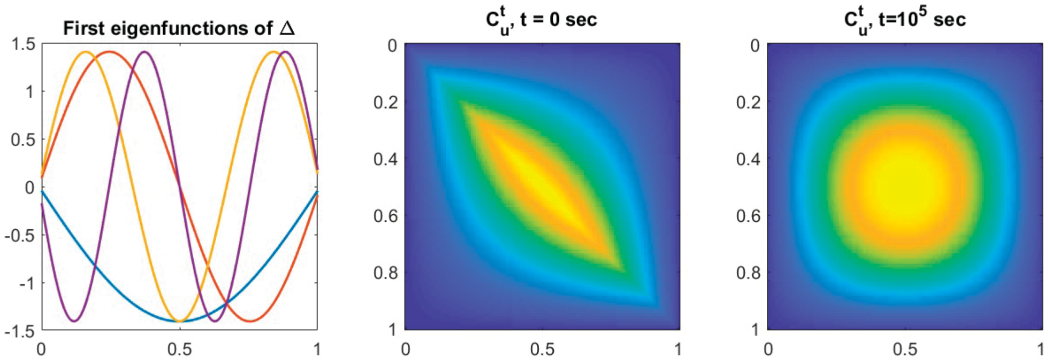


Figure 2.10: The time evolution of $u_t(\cdot)$ is dominated by the first eigenfunction $\varphi_1(\cdot)$, the Covariance operator $C_u^t = E[u_t \otimes u_t^*]$ as a function of t behaves like having kernel $K(\cdot,\cdot) = \varphi_1(\cdot)\varphi_1(\cdot)$ for large t.

■

After this illustrative example linking statistical mechanics, deterministic dynamical systems and covariance operators in Hilbert spaces, physical considerations will be largely disregarded for the rest of the monograph and a stochastic perspective is assumed. Although a deterministic prior on a systems time evolution is interesting especially in the context of optimal control to integrate monitoring and manipulation of objects, we leave this for future work. Lastly we remark that Stones theorem on one parameter unitary semigroups and time evolution is closely related to von Neumanns statistical ergodic theorem [93, p. 279] which states that for any unitary operator $U \in \mathcal{B}(\mathcal{H})$ with spectral family $P(\lambda)$

$$\lim_{n\to\infty} \|\frac{1}{n}\sum_{k=0}^{n-1} U^k f - P_I f\|_{\mathcal{H}} = 0 \tag{2.49}$$

$\forall f \in \mathcal{H}$. P_I is the projection onto the invariant subspace $\mathcal{H}_I \overline{\boxslash} \mathcal{H}$ with $f \in \mathcal{H}_I \Leftrightarrow Uf = f$ and $\mathcal{H}_I$ possibly trivial. When U represents time evolution by some fixed amount t then $U^k = U_{kt}$ and $f(kt) = U^k f_0 \in \mathcal{H}$. If the initial condition f_0 is written as $f_0 = P_I f_0 + P_I^\perp f_0 = f_I + f_D$ the theorem asserts that the evolution of the dynamic part f_D is zero on average and over long times only the invariant part

f_I matters as fluctuations $U^k f_D$ satisfy $\lim_{n\to\infty} \|\frac{1}{n}\sum_{k=0}^{n-1} U^k f_D\|_{\mathcal{H}} = 0$.

Note furthermore that the invariant subspace $P_I\mathcal{H}$ is a property of the dynamics U_t and independent of the initial condition f_0. When the subspace $\mathcal{H}_I$ is trivial, the average converges to zero and this is obviously exactly the case if $1 \notin \sigma(U)$ as then $(U - I)f = 0$ has no nonzero solutions $f_I \in \mathcal{H}_I$ and $Uf_I = f_I$ never holds apart from $f_I = 0$ and $\mathcal{H}_I = \{0\}$.

2.3 Reproducing kernel Hilbert spaces

In this section reproducing kernel Hilbert spaces (the acronym RKHS will be used for both the singular and plural form) are introduced as a subclass of function spaces containing elements with structural properties inherited from a central object that is called the kernel. Apart from the more technical definition involving the continuity of evaluation functionals in the norm topology of the function space, insight is given as to how a kernel determines a corresponding function space and vice versa. Since RKHS are central to the whole monograph, special effort is undertaken to demonstrate their usefulness by emphasizing the properties most convenient in the context of estimation and providing examples of objects embeddable into an RKHS. Keeping in line with the idea of an RKHS as a function space, splines enter the fray naturally as solutions to optimization problems in RKHS, in which the objective function to be minimized parallels the notion of mechanical energy in a quite real sense. To complement this rather concrete perspective, it will also be shown how RKHS can be interpreted as feature spaces in which nonlinear features can be represented and manipulated in a linear fashion.

2.3.1 Definition and properties of RKHS

RKHS are first and foremost Hilbert spaces, i.e. usually infinite dimensional vector spaces augmented with an inner product and complete with respect to the topology induced by the norm inherited from the inner product. Specifically, they are those Hilbert spaces $\mathcal{H}$, for which all the evaluation functionals $e_t \in \mathcal{H}^*, t \in T$ mapping functions $f(\cdot) : T \to \mathbb{C}$ to function values $f(t)$ are continuous and therefore admit a unique representer $K_t \in \mathcal{H}, \langle K_t, f\rangle_{\mathcal{H}} = e_t f$ by the Riesz representation theorem. Surprisingly far reaching consequences follow from this property —normwise convergence implies pointwise convergence, all bounded linear operators can be written as kernel operators and the representer theorem holds. It remains to be clarified, how $K_t \in \mathcal{H}$ can actually be found given $e_t \in \mathcal{H}^*$. Here the Moore-Aronszajn theorem fills in the missing links by stating that every positive definite kernel $K(\cdot,\cdot)$ uniquely determines an RKHS, whose representer of evaluation at $t \in T$ is $K(t,\cdot)$ and it furthermore asserts that this correspondence is indeed a bijection. In recognition of the central role played by positive definite kernels, their properties are briefly surveyed.

Definition 2.3.1 A reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is a Hilbert space of real- or complex-valued functions on an index set T such that for all $t \in T$ the evaluation functionals $e_t : \mathcal{H} \ni f \mapsto e_t f = f(t) \in \mathbb{F}, \mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ are continuous [20, p. 9].

Since in an RKHS $\mathcal{H}$, e_t is continuous and linear, it is also bounded. Then the Riesz representation theorem [93, p. 76] guarantees the existence of an element $K_t \in \mathcal{H}$ such that

$$e_t f = \langle K_t, f \rangle_{\mathcal{H}} \;\; \forall f \in \mathcal{H} \tag{2.50}$$

i.e. $\langle K_t, f \rangle_{\mathcal{H}} = f(t)$ and K_t is the representer of evaluation at $t \in T$. The latter equality for K_t is called the reproducing property. K_t as an element of $\mathcal{H}$ is itself a function from T to $\mathbb{C}$ (the restriction to $\mathbb{R}$ is trivial as [7] notes) and for all $s \in T$, $K_t(s) = \langle K_t(\cdot), K_s(\cdot) \rangle_{\mathcal{H}}$ which is also often expressed as $K(s,t) = \langle K(s,\cdot), K(t,\cdot) \rangle_{\mathcal{H}}$. The notation $K_s(t) = K(s,t)$ was used to emphasize the dependence of the number $K_s(t)$ both on the location $s \in T$ on which $\langle K_s, \cdot \rangle_{\mathcal{H}}$ is evaluated and the location $t \in T$ on which K_s is evaluated itself.

It can be shown that the properties of K_t are indicative for the properties of all functions in $\mathcal{H}$ in the sense that one may always write $\mathcal{H}$ as the closure of translates of K_t, i.e.

$$\mathcal{H} = \text{cl}\left\{ f : T \to \mathbb{C} : f = \sum_{i=1}^{\infty} \alpha_i K_{t_i}(\cdot) \text{ and } \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \alpha_i \overline{\alpha_j} K(t_i, t_j) < \infty \right\} \tag{2.51}$$

where the closure is to be executed as described in [7]. Properties of K_t invariant under linear combination like for example continuity pass down to all $f \in \mathcal{H}$ [20, p. 35]. As a two-variable function, $K(\cdot,\cdot) : T \times T \to \mathbb{C}$ provides a meaningful, equivalent description of $\mathcal{H}$ that is more easily understood than an abstract characterization and can be analyzed and manipulated using functional analytic methods. This function is termed the reproducing kernel (RK) of the RKHS $\mathcal{H}$; when this relationship is to be made evident, the notation $\mathcal{H}_K$ is used.

Definition 2.3.2 A two-variable function $K(\cdot,\cdot) : T \times T \to \mathbb{C}$ satisfying

i) $K(t,\cdot) \in \mathcal{H} \;\; \forall t \in T$

ii) $\langle f(\cdot), K(t,\cdot) \rangle_{\mathcal{H}} = f(t) \;\; \forall t \in T$ and $\forall f \in \mathcal{H}$

is called the reproducing kernel of the Hilbert space $\mathcal{H}$ of functions from T to $\mathbb{C}$.

From the reproducing property one may immediately derive that any reproducing kernel $K(\cdot,\cdot)$ is conjugate symmetric and positive definite because

$$K(s,t) = K_s(t) = \langle K_s, K_t \rangle_{\mathcal{H}} = \overline{\langle K_t, K_s \rangle_{\mathcal{H}}} = \overline{K_t(s)} = \overline{K(t,s)} \tag{2.52}$$

and

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \overline{\alpha_j} K(t_i, t_j) = \langle \sum_{i=1}^{n} \alpha_i K(t_i,\cdot), \sum_{j=1}^{n} \alpha_j K(t_j,\cdot) \rangle_{\mathcal{H}} = \| \sum_{i=1}^{n} \alpha_i K(t_i,\cdot) \|_{\mathcal{H}}^2 \geq 0 \tag{2.53}$$

where $\{\alpha_i\}_{i=1}^n \subset \mathbb{C}$ and $\{t_i\}_{i=1}^n \subset T$ are arbitrary sequences of complex coefficients and index elements. The positive definiteness demonstrated in equation 2.53 is significant and indeed the characterizing feature of a reproducing kernel as the next theorem (taken from [7]) asserts.

Theorem 2.3.3 (Moore-Aronszajn) *To any positive definite function $K(\cdot,\cdot) : T \times T \to \mathbb{C}$, one may associate a unique RKHS $\mathcal{H}_K$ with reproducing kernel K. The span $\mathcal{H}^0$ of the set $\{K(t,\cdot)\}_{t\in T}$ satisfies $\mathcal{H}^0 \boxslash \mathcal{H}_K$ and $\mathrm{cl}\,\mathcal{H}^0 \overline{\boxslash}\mathcal{H}_K$.*

Apart from an RKHS $\mathcal{H}$ of functions on T being uniquely associated to a positive definite kernel K with domain $T \times T$, the elements $f \in \mathcal{H}$ formed as superpositions of kernel functions themselves exhibit some convenient properties that simplify calculation in RKHS.

Theorem 2.3.4 *Let $\mathcal{H}_K$ be an RKHS of functions from T to $\mathbb{C}$ and denote by $K(\cdot,\cdot) : T \times T \to \mathbb{C}$ its reproducing kernel. Then the following statements hold.*

1. *Cauchy-Schwarz implies that $\forall f, g \in \mathcal{H}_K, |f(t) - g(t)| = |\langle f - g, K_t\rangle_{\mathcal{H}_K}| \leq \|f - g\|_{\mathcal{H}_K}\sqrt{K(t,t)}$. As a consequence, normwise convergence implies pointwise convergence [20, p. 10] and if one finds $\lim_{n\to\infty}\|\hat{f}_n - f\|_{\mathcal{H}_K} = 0$ for any sequence of approximations $\hat{f}_n \in \mathcal{H}_K$ for f then also $\lim_{n\to\infty}\hat{f}_n(t) = f(t)\ \forall t \in T$.*

2. *For any bounded linear operator $L : \mathcal{H}_K \to \mathcal{H}_K$, the function $u_L^t := L^*K_t \in \mathcal{H}_K$ satisfies $\langle u_L^t, f\rangle_{\mathcal{H}_K} = \langle L^*K_t, f\rangle_{\mathcal{H}_K} = \langle K_t, Lf\rangle_{\mathcal{H}_K} = (Lf)(t)$. The unique two-variable function $u_L^{\cdot} : T \times T \to \mathbb{C}$ is called the kernel of the operator L [20, p. 27].*

3. *(Representer theorem) For an RKHS $\mathcal{H}_K$ of real valued functions on T, a strictly monotonically increasing function $\varphi : \mathbb{R} \to \mathbb{R}, b \in \mathbb{R}^n$ and a linear evaluation operator $A : \mathcal{H}_K \ni f \mapsto \{f(t_j)\}_{j=1}^n \in \mathbb{R}^n$, the solution f_{opt} to the regularized reconstruction problem*

$$f_{opt} = \underset{f\in\mathcal{H}_K}{\operatorname{argmin}}\ \|Af - b\|_{\mathcal{H}_K}^2 + \varphi\left(\|f\|_{\mathcal{H}_K}\right)$$

can be written as $f_{opt} = \sum_{j=1}^n \alpha_j K(t_j,\cdot)$, a linear superposition of kernels centered around the points $t_j, j = 1, ..., n$ at which f was observed [173].

Properties 1-3 are useful especially from a perspective focused on estimation and computability. They guarantee that if an estimator $\hat{f}_n$ converges to the ground truth f in the norm topology, then $\hat{f}_n$'s values will eventually coincide with the ones of f. Furthermore, optimal reconstruction problems have solutions that can be explicitly written down, consist of a number of terms scaling linearly with the measurements and have a coefficient vector $\{\alpha_j\}_{j=1}^n \in \mathbb{R}^n$ as the only unknown quantity. Linear transformations by operators $L : \mathcal{H}_K \to \mathcal{H}_K$ can be handled without leaving the RKHS framework by reducing them to easily computable inner products. It is also possible to start with a kernel function $u(\cdot,\cdot)$ and construct a linear operator L_u

from it. When one starts from the second moment function $K(s,t) = E[X_s X_t]$ for a set of random variables $X_t, t \in T$, the resultant operator L_K and RKHS $\mathcal{H}_K$ have a particularly telling decomposition that is summarized in the next theorem (taken from [20, pp. 68-70]).

Theorem 2.3.5 *Let $K : T \times T \to \mathbb{R}, T = [a,b] \subset \mathbb{R}$ be the continuous second order moment function of the set of random variables $X_t, t \in T$ and denote by L_K the linear operator $L_K f = \int_T K(\cdot,t) f(t) dt$ with domain and codomain $L^2(T)$.*

1. *The Mercer decomposition $K(s,t) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(s) \varphi_j(t)$ holds where $\{\lambda_j\}_{j=1}^{\infty} \subset \mathbb{R}^+$ and $\{\varphi_j\}_{j=1}^{\infty} \subset L^2(T)$ are the sequence of eigenvalues and eigenfunctions of L_K. The eigenfunctions are orthonormal, i.e. $\langle \varphi_i, \varphi_j \rangle_{L^2(T)} = \delta_{ij}$.*
2. *The random function $X_{\cdot}^{\cdot} : \omega \mapsto X_{\omega}^{\cdot}$ can be written in form of the so-called Karhunen-Loewe expansion (KLE) as $X_{\cdot}(t) = \sum_{j=1}^{\infty} \xi_j^{\cdot} \sqrt{\lambda_j} \varphi_j(t)$ where $\xi_j^{\cdot} : \Omega \to \mathbb{R}$ is a weakly white noise random variable with $E[\xi_i^{\cdot} \xi_j^{\cdot}] = \delta_{ij}$.*

Even though the two theorems were presented only for the simple case of $T \subset \mathbb{R}$ being a closed interval, they hold more generally. The Mercer decomposition of an operator $L_k : L^2(T,\mu) \to L^2(T,\mu)$ is possible when K is a continuous positive definite kernel and T is only supposed to be a topological Hausdorff space with finite Borel measure μ [117, p. 145], [60]. When T is a subset of $\mathbb{R}^n$, a Karhunen-Loewe expansion is possible as well [131, p. 19]; the simulations in this monograph are generated via KLE.

The Mercer decomposition allows to express the inner product and norm of $\mathcal{H}_K$ as a modified inner product in $L^2(T)$. For the inner product in $\mathcal{H}_K$, it is demanded that $\langle K(s,\cdot), K(t,\cdot) \rangle_{\mathcal{H}_K} = K(s,t)$ and this implies

$$\langle K(s,\cdot), K(t,\cdot) \rangle_{\mathcal{H}_K} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \lambda_i \lambda_j \varphi_i(s) \varphi_j(t) \langle \varphi_i, \varphi_j \rangle_{\mathcal{H}_K} \overset{!}{=} \sum_{i=1}^{\infty} \lambda_i \varphi_i(s) \varphi_i(t).$$

The easiest way to achieve this is by having $\langle \varphi_i, \varphi_j \rangle_{\mathcal{H}_K} = \lambda_i^{-1} \langle \varphi_i, \varphi_j \rangle_{L^2(T)}$ which then leads to a Hilbert space $\mathcal{H}_K$ with inner product

$$\langle f, g \rangle_{\mathcal{H}_K} = \langle L_K^{-1} f, g \rangle_{L^2(T)} = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} \langle f, \varphi_j \rangle_{L^2(T)} \langle g, \varphi_j \rangle_{L^2(T)} \ \forall f, g \in \mathcal{H}_K. \quad (2.54)$$

By the uniqueness property mentioned in the Moore-Aronszajn theorem, there is only one Hilbert space $\mathcal{H}_K$ for which K is the reproducing kernel and therefore this construction of inner product is not only valid but the only one possible. More on this topic can be found for example in [148] where it is also shown that $f \in \mathcal{H}_K$ if and only if

$$\|f\|^2_{\mathcal{H}_K} = \sum_{j=1}^{\infty} \lambda_j^{-1} \langle f, \varphi_j \rangle^2_{L^2(T)} = \sum_{j=1}^{\infty} \frac{f_j^2}{\lambda_j} < \infty \tag{2.55}$$

i.e. $\mathcal{H}_K$ is that subset of functions f in $L^2(T)$ whose coefficients in the ONB $\{\varphi_i\}_{i=1}^{\infty}$ decay significantly faster than the sequence $\{\lambda_i\}_{i=1}^{\infty}$. This mirrors closely theorems relating differentiability and the rate of decay of Fourier transforms. The finiteness demand in equation 2.55 can be interpreted as the statement that eigenfunctions φ_j with small λ_j are atypical for elements of $\mathcal{H}_K$ — the norm $\|f\|_{\mathcal{H}_K}$ in this sense gives a continuous quantification of the degree to which f belongs to $\mathcal{H}_K$ and can be considered to be a generalization of the sum of squares $\|f\|^2_{\mathbb{R}^n} = \sum_{i=1}^n f_i^2$ of an n-dimensional vector.

Example 10 Suppose $T = \{1, ..., n\}$, then $K : T \times T \to \mathbb{R}$ satisfies $\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j K(t_i, t_j) \geq 0 \; \forall$ vectors $\{\alpha_i\}_{i=1}^m \in \mathbb{R}^m$ and positions $\{t_i\}_{i=1}^m \subset T$. From it one constructs the positive definite operator

$$L_K : L^2(T) \ni f \mapsto \int_T K(\cdot, t) f(t) dt \in L^2(T)$$

where the right hand side is $(L_k f)(s) = \int_T K(s,t) f(t) dt = \sum_{i=1}^n K(s, t_i) f(t_i)$. But then L_k is just the matrix with entries $(L_K)_{ij} = K(t_i, t_j)$. According to Mercers theorem, L_K and its inverse have the decomposition

$$(L_K)_{st} = \sum_{j=1}^n \lambda_j \varphi_j(s) \varphi_j(t) \qquad (L_K^{-1})_{st} = \sum_{j=1}^n \frac{1}{\lambda_j} \varphi_j(s) \varphi_j(t).$$

The inner product is then

$$\langle f, g \rangle_{\mathcal{H}_K} = \sum_{j=1}^n \frac{1}{\lambda_j} \langle f, \varphi_j \rangle_{\mathbb{R}^n} \langle g, \varphi_j \rangle_{\mathbb{R}^n} = \langle L_K^{-1} f, g \rangle_{\mathbb{R}^n}$$

and the norm is $\|f\|^2_{\mathcal{H}_K} = \langle L_K^{-1} f, f \rangle_{\mathbb{R}^n} = \|L_K^{-1/2} f\|^2_{\mathbb{R}^n}$ where it was used that positive definite selfadjoint operators have well-defined square roots and inverses as explained in section 2.2.2. If K were a covariance function, then L_K would be the covariance matrix and $-\|L_K^{-1/2} f\|^2_{\mathbb{R}^n}$ would be the term appearing in the exponent of a Gaussian probability density. This gives further intuition as to why one may consider the norm as a measure of atypicality. ■

By the Moore-Aronszajn theorem, positive definite (p.d.) kernels and RKHS can be identified — operations on p.d. kernels leading to new p.d. kernels have analogues in terms of new RKHS being assembled from base RKHS. The properties of the convex cone [11, p. 39] of reproducing kernels listed in the next theorem are proven in [7] and [20, pp. 24-30].

Theorem 2.3.6 *Let K, K_1, K_2 be real valued positive definite kernels. Denote by $\mathcal{H}_K, \mathcal{H}_{K_1}, \mathcal{H}_{K_2}$ the corresponding reproducing kernel Hilbert spaces.*

1. *For $T_1 \subset T$, the restriction $K_r := K\,|_{T_1 \times T_1}$ of K to the subset $T_1 \times T_1$ is the reproducing kernel of the space $\mathcal{H}_{K_r}$ consisting of restrictions $f_1 := f\,|_{T_1}$ of functions $f \in \mathcal{H}_K$ together with the norm*

$$\|f_1\|^2_{\mathcal{H}_{K_r}} = \min_{f_1 = f|_{T_1}, f \in \mathcal{H}_K} \|f\|^2_{\mathcal{H}_K}.$$

2. *The sum $K = K_1 + K_2$ is a reproducing kernel with corresponding RKHS $\mathcal{H}_K = \mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2}$. It has elements $f \in \mathcal{H}_K$ such that $f = f_1 + f_2$ with $f_1 \in \mathcal{H}_{K_1}, f_2 \in \mathcal{H}_{K_2}$ and the norm is*

$$\|f\|^2_{\mathcal{H}_K} = \min_{f = f_1 + f_2, f_j \in \mathcal{H}_{K_j}} \|f_1\|^2_{\mathcal{H}_{K_1}} + \|f_2\|^2_{\mathcal{H}_{K_2}}.$$

3. *If $K = K_2 - K_1$ is a positive definite kernel, it is the reproducing kernel of the RKHS $\mathcal{H}_K := \mathcal{H}_{K_2} \ominus \mathcal{H}_{K_1}$. One has $\mathcal{H}_{K_1} \subset \mathcal{H}_{K_2}, \|f\|_{\mathcal{H}_{K_1}} \geq \|f\|_{\mathcal{H}_{K_2}}$ and $\mathcal{H}_K$ is the Hilbert space formed by completing the everywhere dense set $\mathcal{H} := \left(I - P_{\mathcal{H}_{K_1}}\right)\mathcal{H}_{K_2}$ where $(P_{\mathcal{H}_{K_1}} f_2)(t) = \langle f_2(\cdot), K_1(t, \cdot)\rangle_{\mathcal{H}_{K_2}} \ \ \forall f_2 \in \mathcal{H}_{K_2}$ is the projection onto $\mathcal{H}_{K_1}$. $\mathcal{H}$ is equipped with the norm*

$$\begin{aligned}\|f\|^2_{\mathcal{H}} = \|\sqrt{I - P_{\mathcal{H}_{K_1}}} f_2\|^2_{\mathcal{H}_{K_2}} &= \langle f_2, f_2 - P_{\mathcal{H}_{K_1}} f_2\rangle_{\mathcal{H}_{K_2}} \\ &= \|f_2\|^2_{\mathcal{H}_{K_2}} - \|P_{\mathcal{H}_{K_1}} f_2\|^2_{\mathcal{H}_{K_1}} \qquad (2.56)\end{aligned}$$

   for all f in the set $\mathcal{H}$. The element f_2 is chosen such that $f_2 - P_{\mathcal{H}_{K_1}} f_2 = f$.

4. *For $T_j \times T_j$ being the domain of K_j, $j = 1, 2$, the direct product kernel $K = K_1 \otimes K_2$ acting on $(T_1 \times T_2) \times (T_1 \times T_2)$ via $K(s_1, s_2, t_1, t_2) = K_1(s_1, t_1)K_t(s_2, t_2)$ corresponds to the RKHS $\mathcal{H}_K = \mathcal{H}_{K_1} \otimes \mathcal{H}_{K_2}$ which is the completion of the span of functions $f = f_1 \otimes f_2 : T_1 \otimes T_2 \ni (t_1, t_2) \mapsto f_1(t_1)f_2(t_2) \in \mathbb{R}$, $f_j \in \mathcal{H}_{K_j}$.*

The above theorem and the general construction fo RKHS from basis functions are illustrated in the next two examples.

Example 11 (RKHS of polynomials) Starting with the space spanned by monomials on $[-1, 1]$, one can obtain the Legendre polynomials via diagonalization and assemble them into a positive definite kernel yielding the first example of a functional RKHS readily available for concrete computations. Define $T = [-1, 1]$ and $\mathcal{H}_{\mathbb{P}_n}$ as the RKHS of polynomials up to order n with the inner product of $L^2(T)$. Starting from the monomials

$$p_1(t) = 1 \qquad p_2(t) = t \qquad p_3(t) = t^2 \quad \forall t \in T$$

Gram-Schmidt orthonormalization leads to the orthonormal basis vectors

$$e_1(t) = \sqrt{\frac{1}{2}} \qquad e_2(t) = \sqrt{\frac{3}{2}}t \qquad e_3(t) = \sqrt{\frac{5}{2}}\frac{1}{2}\left(3t^2 - 1\right) \quad \forall t \in T$$

which are just rescaled versions of the first three Legendre polynomials [153, p. 443]. Creating the kernel $K(s,t) = \sum_{j=1}^{3} e_j(s)e_j(t) = (1/2) + (3/2)st + (5/8)(3s^2-1)(3t^2-1)$, it is easy to check that $K(\cdot,\cdot)$ is reproducing for $\mathcal{H}_{\mathbb{P}_2}$ as for $f \in \mathcal{H}_{\mathbb{P}_2}$ it holds that $f = \sum_{i=1}^{3} \widetilde{f}_i p_i(\cdot) = \sum_{i=1}^{3} f_i e_i(\cdot)$ and since $\langle e_i(\cdot), e_j(\cdot)\rangle_{\mathcal{H}_{\mathbb{P}_2}} = \delta_{ij}$, one has

$$\begin{aligned}\langle K(s,\cdot), f(\cdot)\rangle_{\mathcal{H}_{\mathbb{P}_2}} &= \sum_{i=1}^{3} e_i(s)\langle e_i(\cdot), f(\cdot)\rangle_{\mathcal{H}_{\mathbb{P}_2}} \\ &= \sum_{i=1}^{3} e_i(s)\langle e_i(\cdot), \sum_{j=1}^{3} f_j e_j(\cdot)\rangle_{\mathcal{H}_{\mathbb{P}_2}} = \sum_{i=1}^{3} f_i e_i(s)\end{aligned}$$

which is nothing else than $f(s)$ as required. ■

Example 12 (Brownian bridge and Wiener process) Suppose one knows from observation that a good measure of atypicality of a function f is the L^2 norm of its derivative and that it always vanishes at the boundary points 0 and 1 of $T = [0,1]$. The function space $\mathcal{H}_1$ containing these f is

$$\mathcal{H}_1 = \{f \in L^2(T) : f(0) = f(1) = 0, \|\partial_t f\|_{L^2(T)} < \infty\}$$

with inner product $\langle f, g\rangle_{\mathcal{H}_1} = \int_T (\partial_t f)(s)(\partial_t g)(s)ds$ where the derivatives ∂_t are defined in a suitably weak sense. Atteia [11, pp. 7-11] then shows that $\mathcal{H}_1$ is an RKHS with the kernel K_1

$$K_1(s,t) = \min(s,t) - st \tag{2.57}$$

being a solution to a certain differential equation in order to guarantee the reproducing property. Functions in $\mathcal{H}_1 = \mathcal{H}_{K_1}$ are called Brownian bridges since they start and end at the same height and their increments have much of the characteristics of white noise. In a similar way, the original Wiener process as integrated white noise can be associated to the Hilbert space

$$\mathcal{H}_2 = \{f \in L^2(T) : f(0) = 0, \|\partial_t f\|^2_{L^2(T)} < \infty\}$$

with inner product $\langle f, g\rangle_{\mathcal{H}_2} = \int_T (\partial_t f)(s)(\partial_t g)(s)ds$ and kernel

$$K_2(s,t) = \min(s,t) \tag{2.58}$$

as [11, pp. 7-11] proves. The increments of functions $f \in \mathcal{H}_2 = \mathcal{H}_{K_2}$ look like white noise, the norm $\|f\|_{\mathcal{H}_2} = \|\partial_t f\|_{L^2(T)}$ therefore measures atypicality of f by reducing it to the atypicality of its derivative. Figure 2.11 shows the kernels

$K_1, K_2, K_3 = K_1 + K_2, K_4 = K_2 - K_1, K_5 = A \otimes AK_1 = \int_0^s \int_0^t K_1(u,v)dudv$ and $K_6 = K_1 \otimes K_2$ as well as some randomly drawn elements from the RKHS $\mathcal{H}_{K_1}, ..., \mathcal{H}_{K_6}$. It is notable that the kernel $K_5 = A \otimes AK_1$ is related to the space $\mathcal{H}_5 = \{Af_1 : f \in \mathcal{H}_{K_1}\}$ although the relationship is not entirely straightforward in general, see [170, pp. 89-91]. ■

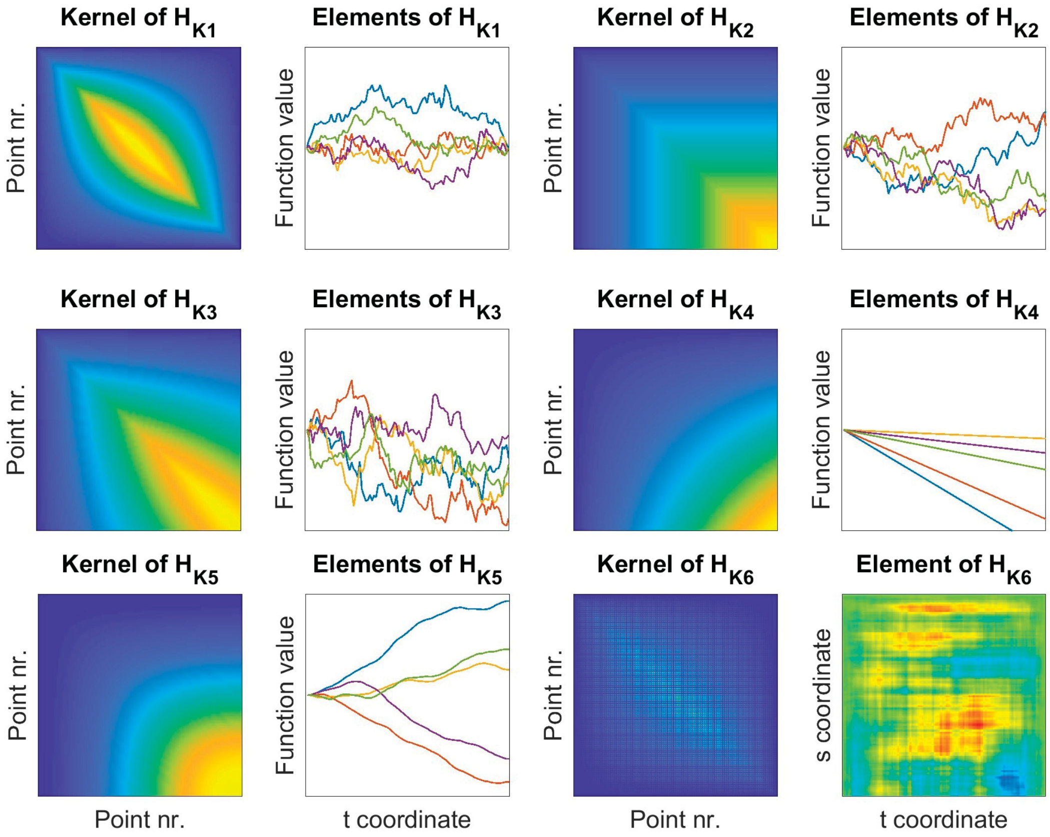


Figure 2.11: The kernels and some representative elements of the Hilbert spaces $\mathcal{H}_{K_1}, ..., \mathcal{H}_{K_6}$ whose definitions are given in the text. Whereas the scaling on the plots of the elements is equal, the colormap used to display the kernels varies to guarantee good visibility of their main features. Notice that elements of $\mathcal{H}_{K_3}$ are sums of elements from $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$ and an analysis of $\mathcal{H}_{K_4}$ reveals that a Wiener process is essentially just a Brownian bridge plus a randomly drawn line. Since $\mathcal{H}_{K_6}$ as a tensor product contains 2-dimensional functions that are Brownian bridges in x-direction and Wiener processes in y-direction, only one of its elements is plotted. Simulations were generated via Karhunen Loewe expansion as described in theorem 2.3.5.

As a last comment, note that for $\mathcal{H}_K = \mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2}, K = K_1 + K_2$, the inner product of a function with one of the component kernels, i.e. the expression

$$\langle f(\cdot), K_1(\cdot,\cdot)\rangle_{\mathcal{H}_K} = \langle (L_{K_1+K_2}^{-1} f)(\cdot), K_1(\cdot,\cdot)\rangle_{L^2(T)} = \int_T K(\cdot,s)(L_{K_1+K_2}^{-1} f)(s)ds$$

is the projection of $f \in \mathcal{H}_K$ onto $\mathcal{H}_{K_1}$ [7]. This is easily seen to be $\Sigma_1(\Sigma_1+\Sigma_2)^{-1}f$ in the discrete case where $f \in \mathbb{R}^n$ and $(\Sigma_l)_{ij} = K_l(i,j), l = 1,2$ are positive definite matrices. As is shown later (91, theorem 3.1.5) this latter equation coincides with a statistically motivated estimator for f_1 based on observation of $f = f_1 + f_2$. Even though this reasoning is made watertight only in section 3.1, we proceed to give examples of RKHS containing infinitely differentiable and continuous but nowhere differentiable functions and tie their elements behaviour to the form of their kernel via correlation based arguments. For the RKHS of smooth functions $\mathcal{H}_{K_1}, K_1 = \exp(-|s-t|^2)$ [39, p. 89] and everywhere continuous but nowhere differentiable functions forming $\mathcal{H}_{K_2}, K_2(s, t = \exp(-|s-t|)$ [162, p. 86] an exemplary result of a projection is plotted in figure 2.12.

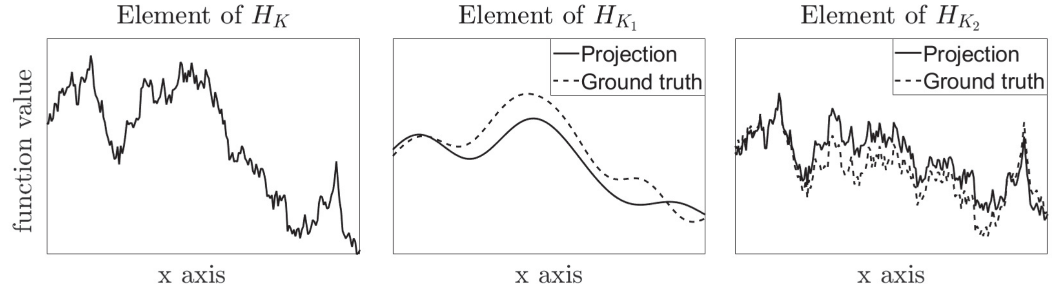


Figure 2.12: The element f of $\mathcal{H}_K$ plotted in the left panel is a sum of the elements $f_1 \in \mathcal{H}_{K_1}$ and $f_2 \in \mathcal{H}_{K_2}$ plotted in dashed lines in panel two and three. The projections $\langle f, K_1 \rangle_{\mathcal{H}_K}$ and $\langle f, K_2 \rangle_{\mathcal{H}_K}$ are close to f_1 and f_2 respectively.

2.3.2 Interpretation of RKHS

RKHS can be defined as the closure of the span of kernel translates. This suggests that K might play the role of an impulse response function linking superpositions of impulse-type inputs to linear superpositions of kernel translates. To showcase this approach, a familiar problem of theoretical mechanics is now attacked from a Newtonian perspective and the solution is tied via its Hamiltonian formulation to the minimization of an energy functional featuring spatial derivatives. This energy term is related to a norm in an RKHS and the solution can be written as a superposition of basis functions related to K. On the other hand, RKHS can also be perceived as purely abstract feature spaces. This perspective is illustrated with the help of the archetypal X-OR example showcasing a 2-D classification problem that can only be solved by a linear classifier after being nonlinearly injected into $\mathbb{R}^3$. The role of kernels as tools to describe and manipulate high-dimensional features is emphasized.

§ RKHS as spaces of functions

Suppose a weightless elastic string is clamped at constant height 0 at both endpoints of the interval $T = [0, 1]$ in such a way that at its equilibrium state it is just a horizontal line. It is then subjected to a spatially distributed load that is described by the density function $w(s), s \in T$. The displacement function $f(\cdot) \in L^2(T)$ quantifying for each position $s \in T$ the deviation from the trivial $(w = 0)$ horizontal position of rest is to be determined. This situation is illustrated in figure 2.13 where also the expected solutions to very simple load configurations are plotted.

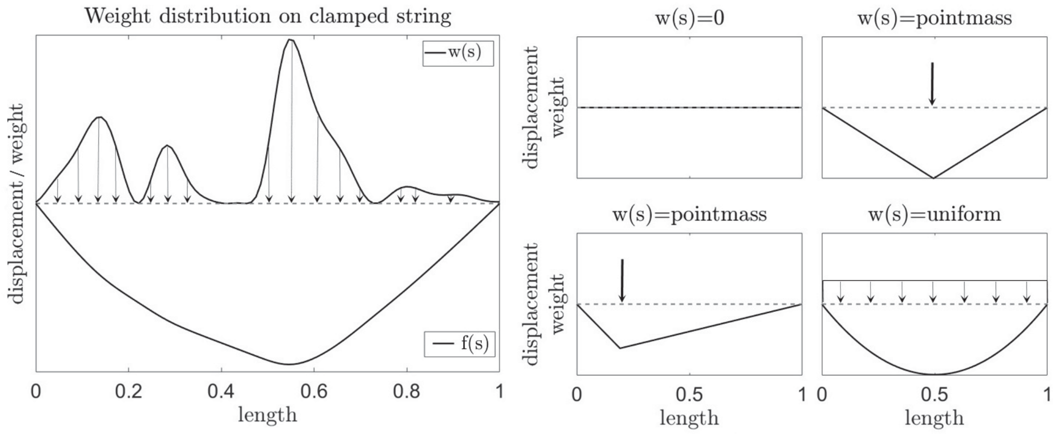


Figure 2.13: An elastic string is clamped in at $s = 0$ and $s = 1$. Under the load $w(\cdot)$ it deforms; the displacement is given by the function $f(\cdot) \in L^2(T)$. On the basis of intuitive physical reasoning, one would expect $f(\cdot)$ to behave like plotted in the panels on the right. The arrows represent the force being exerted by load w.

The function $w(s)$ is a density and should be interpreted in such a way that $\int_{[s-\Delta,s+\Delta]} w(\tilde{s})d\tilde{s}$ quantifies the total weight of the load in the interval $[s-\Delta, s+\Delta]$. The differential equation for $f(s)$ in terms of the density function $w(s)$ is then given as the second order ODE [81, p. 22]

$$\frac{\partial^2}{\partial s^2} f(s) = w(s) \qquad f(0) = f(1) = 0 \tag{2.59}$$

where all the material constants have been chosen as to multiply to one for the sake of simplicity. This equation can be derived using a purely Newtonian interpretation [81, p.26] or as the Euler Lagrange equation associated to minimizing the energy H

$$H = \langle w, f\rangle_{L^2(T)} + \frac{1}{2}\|f\|^2_{\mathcal{H}} = \int_T w(s)f(s)ds + \frac{1}{2}\int_T |\partial_t f|^2 ds \tag{2.60}$$

[33, p. 49] where $\|\cdot\|_{\mathcal{H}}$ is the norm in the Sobolev space from the examples on page 71. This follows trivially from the fact — stated in equation 2.43 — that to minimize $\int_T g(s, f, f_s)ds$ one may equivalently solve

$$\frac{\partial g}{\partial f} - \frac{d}{ds}\left(\frac{\partial g}{\partial f_s}\right) \tag{2.61}$$

where in the above the subscript denotes differentiation, i.e. $f_s = \partial_s f$. Setting $g = (1/2)f_s^2 - fw$, clearly $H = \int_T g ds$ and

$$\frac{\partial g}{\partial f} = w \qquad \frac{d}{ds}\left(\frac{\partial g}{\partial f_s}\right) = f_{ss}$$

implying the Euler Lagrange equation 2.61 to be fulfilled iff $\partial_s^2 f - w = 0$. How the energy functional H arises through consideration of elastic forces and potential energy is explained for example in [205, pp. 95-97]. A standard method for determining solutions to linear differential equations $Qf = w$ like equation 2.59 is to derive Greens function for the problem, i.e. a two variable function $G(\cdot, \cdot) : T \times T \to \mathbb{R}$ such that $(QG)(s, t)$ acts like the delta distribution $\delta(s - t)$ [81, p. 50]. Then

$$Q \int_T G(s, t)w(t)dt = \int_T (QG)(s, t)w(t)dt = w(s)$$

and $\int_T G(s, t)w(t)dt = f(s)$ is a fundamental solution satisfying $Qf = w$. The integral operator $L_G : w \mapsto \int_T G(s, t)w(t)dt$ maps weights distributions as quantified by the density $w(\cdot)$ to the displacement function; $G(\cdot, \cdot)$ is therefore some type of impulse-response function. In the case of the loaded string and $Q = \partial_s^2$, G can be calculated explicitly to be

$$G(s, t) = min(s, t) - st$$

[81, p. 24], which is just the kernel for the Brownian bridge. In this way the operator L_G of integration against a certain type of covariance is providing

- a solution to a differential equation $Qf = w$ with boundary conditions.
- a solution to a Hamiltonian energy minimization problem $\langle f, w\rangle_{L^2(T)} + (1/2)\|f\|^2_{\mathcal{H}} \to \min$ in Hilbert space.
- the means of mapping input functions $w(\cdot)$ to outputs $f(\cdot)$.

Note however, that if one sets w to finite dimensional white noise and calculates the covariance matrix $\Sigma_f = E[f \otimes f^*]$ for $f = L_G w$ one gets $\Sigma_f = L_G L_G^*$ rather than L_G so the solutions f are not realizations of the Brownian bridge process themselves. Instead of being elements of $\mathcal{H}_G$, they are elements of $\mathcal{H}_{\tilde{G}}$ where

$$\tilde{G}(s, t) = \int_T G(s, u)G(u, t)du$$

and conversely if one defines the positive semidefinite and selfadjoint operator $\sqrt{L_G}$ via functional calculus, then $\sqrt{L_G}w$ has covariance matrix L_G. It is therefore reasonable to think of an RKHS $\mathcal{H}_K$ as the function space containing the solutions of $Qf = w$ for w white noise and Q a selfadjoint differential operator satisfying $Q\sqrt{L_K} = I$, e.g $Q = [L_K]^{-1/2}$. Not all Greens functions are positive semidefinite kernels themselves and no identification can be made [8]. The relationship between kernels K, Greens functions G, differential operators Q, and the function spaces $\mathcal{H}_K$ is nonetheless interesting especially from the perspective of differential equations governing physical systems and more can be found for example in [161, pp. 349-357] and [58]. Figure 2.14 concludes this example.

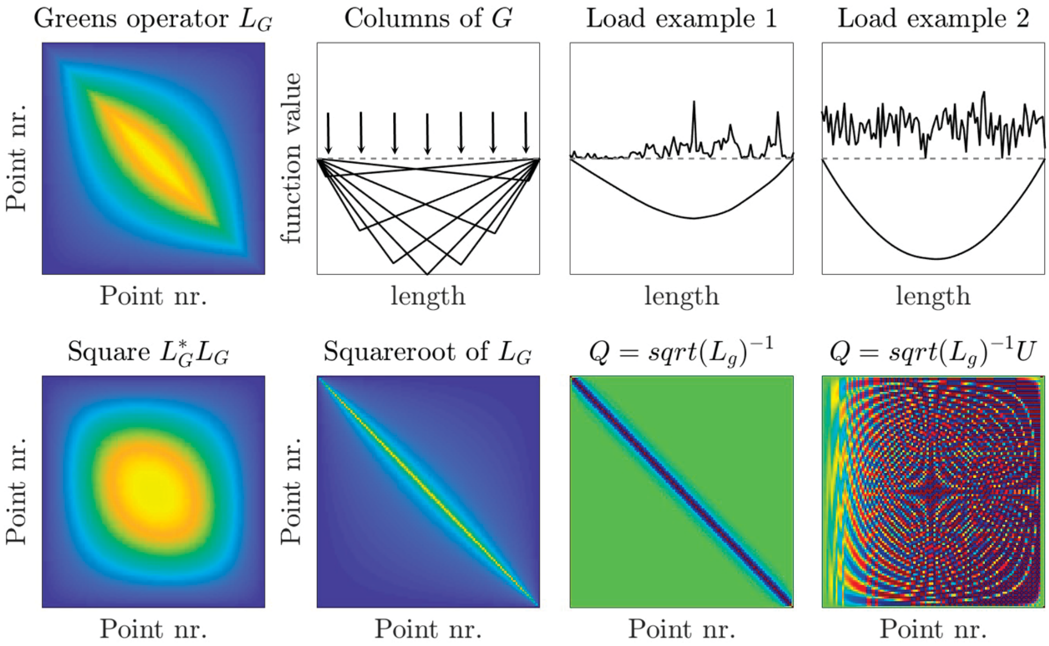


Figure 2.14: Illustration of Greens function as a matrix. In this discretized version, $(L_G w)_i = \sum_{j=1}^n G(t_i, t_j) w_j$ for $w \in \mathbb{R}^n$. The last three panels of the top row contain example solutions for different load configurations that are indicated in red; in the second panel they are simply point masses as represented by the canonical euclidean basis $e_k, k = 1, ..., n$. The bottom row shows $\tilde{G} = L_G G$ such that $L_{\tilde{G}}$ is the covariance matrix describing the solutions of the differential equations. The other three panels show the matrix $F = \sqrt{L_G}$ such that $Fw \sim \mathcal{N}(0, L_G)$ for w white noise and different choices of differential operators Q such that $\mathcal{H}_G$ contains solutions to $Qf = w$. Whereas the first Q is similar to a discretized second derivative, the second Q is arbitrarily transformed using some unitary matrix U.

Solutions to optimization problems involving norms and inner products in RKHS are called splines. This class of objects does not only contain abstract constructions consisting of superpositions of kernel functions but also the very simple piecewise cubic splines for which Holladay proved that they minimize the curvature among all sufficiently often differentiable interpolants going through a predefined set of points $\{(s_i, f_i)\}_{i=1}^n$ [98]. As a matter of fact, this constrained minimization of the global curvature term $\|\partial_s^2 f\|_{L^2(T)}^2$ is equivalent to minimizing the mechanical energy of a thin elastic rod forced through the various positions whose x and y values are given by some set $\{(s_i, f_i)\}_{i=1}^n$. These were the original physical real world objects used extensively in shipbuilding to which the name 'spline' was originally associated. We could have equivalently carried out the previous investigations in the setting of Euler-Bernoulli beam theory but we decided to avoid dealing with 4-th order derivatives to keep the presentation simple.

§ RKHS as spaces of features

The reason for the popularity of RKHS in the context of pattern recognition and classification is that they can be interpreted not only as spaces of functions but from a slightly different perspective do form spaces of potentially infinite-dimensional features. This often allows to solve high-dimensional and nonlinear discrimination tasks with what are essentially linear methods in Hilbert spaces.

Figure 2.15 shows the archetypal X-OR example that is often considered as an introductory toy problem in machine learning related literature (see e.g. [132]). If the goal is to separate linearly the labeled data points

$$p^i = \{(x^i, y^i)\}_{i=1}^m \qquad x^i \in \mathcal{X} = \mathbb{R}^2, y^i \in \{-1, 1\}$$

using a $(2-1)$-dimensional hyperplane $P \subset \mathbb{R}^2$, i.e. to find a translation t_P and a normal vector n_P such that

$$\operatorname{sign}\left[\langle n_P, x^i - t_P\rangle_{\mathbb{R}^2}\right] = y^i$$

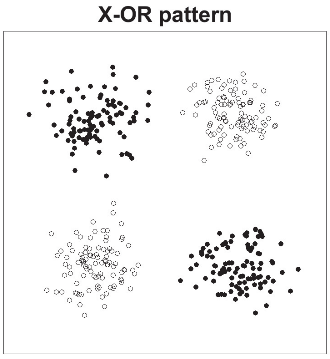


Figure 2.15: The two classes represented by filled and empty circles are not separable via a line in $\mathbb{R}^2$.

then figure 2.15 convinces the reader that this is impossible. However, if one were to embed the problem into three-dimensional space by introducing feature maps ϕ

$$\phi : \mathcal{X} \ni x = [x_1, x_2]^T \mapsto [x_1, x_2, x_1 x_2] = \phi(x) \in \mathbb{R}^3,$$

then the data points $p_\phi^i = \{(\phi(x^i), y^i)\}_{i=1}^m \subset \mathbb{R}^3 \times \{-1, 1\}$ are easily linearly separable by the trivial $(3-1)$ dimensional hyperplane with $n_P = [0, 0, 1]^T$ and $t_P = [0, 0, 0]^T$. In this case a handcrafted nonlinear feature map $\phi : \mathcal{X} \to \mathbb{R}^n, n > \dim \mathcal{X}$ has revealed the dataset to have an only nonlinear structure in $\mathcal{X}$ but an exploitable linear structure in $\mathbb{R}^n$. The usual way to automatize this procedure and circumvent the need to manually craft features is by introducing infinite-dimensional nonlinear feature maps $\phi : \mathcal{X} \to \mathcal{H}_K$ taking values in some RKHS $\mathcal{H}_K$. If one sets

$$\phi(x) = K(x, \cdot) \in \mathcal{H}_K \quad \forall x \in \mathcal{X},$$

the reproducing property $\langle f(\cdot), K(x, \cdot)\rangle_{\mathcal{H}_K} = f(x)$ guarantees that all inner products between features are efficiently evaluable as

$$\langle \phi(x_i), \phi(x_j)\rangle_{\mathcal{H}_K} = \langle K(x_i, \cdot), K(x_j, \cdot)\rangle_{\mathcal{H}_K} = K(x_i, x_j) \quad \forall x_i, x_j \in \mathcal{X}.$$

Finding the optimal set of parameters describing the hyperplane separating the two sets $X^- = \{x^i : y^i = -1\}$ and $X^+ = \{x^i : y^i = +1\}$ can be cast as a quadratic program whose dimension is of the order of available observations thanks to the representer theorem. This method of deriving a data driven classification rule is termed a support vector machine (SVM) [174]. Given enough data consisting of measurements $x \in \mathcal{X}$ and labels $y \in \{-1, 1\}$, it is therefore possible to potentially extract the apriori unknown linear or nonlinear functionals

$$\psi : \mathcal{X} \ni x \mapsto \psi(x) \in \mathbb{R}$$

whose sign determines the label corresponding to $x \in \mathcal{X}$. Unlike in the two-dimensional illustration in figure 2.15, the underlying decision rule can not simply be guessed from datasets in $\mathbb{R}^n, n \geq 4$, demonstrating that this approach delivers a substantial increase in discriminatory power compared to visual inspection by a practitioner.

Being able to manipulate an expressive set of nonlinear features is helpful also when one wants to quantitatively characterize probability distributions. Second and first order moments are by far the most popular descriptors in this type of situation. However, they are often not sufficient to distinguish even severely different distributions and a correlation coefficient of zero between two random variables X and Y is at times mistakenly considered indicative of independence between X and Y ignoring nonlinear dependencies and potential non-Gaussianity. A brief overview of typical pitfalls and problems is given in figure 2.16, which has been inspired by examples provided in a talk held by Arthur Gretton at the Tübingen machine learning summer school in 2015.

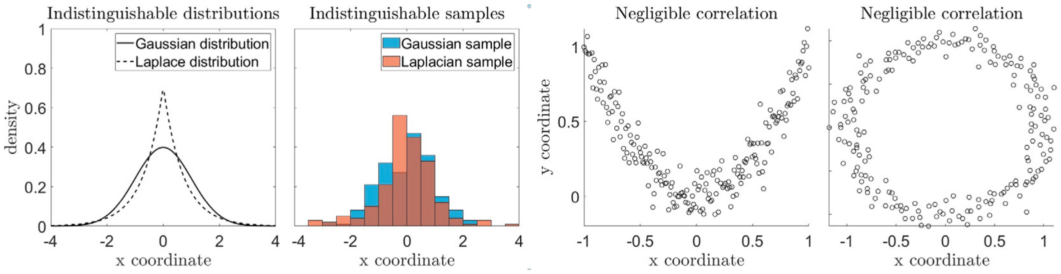


Figure 2.16: A Gaussian with parameters $\mu = 0$ and $\sigma = 1$ has the same first two moments as a Laplacian probability distribution with location parameter $\mu = 0$ and scale $b = 1/\sqrt{2}$. They are indistinguishable when no higher order moments are used for description and the problem naturally carries over to the comparison of finite samples. The panels on the right hand side showcase two situations in which there is an obvious functional dependence of Y on X but the correlation coefficient (almost) vanishes, $|\operatorname{corr}(X,Y)| < 10^{-1}$ in both cases. The observed nonzero correlation is induced by noise only as point distributions on the true underlying models (parabola, circle) have vanishing central second order moments.

Smola et al. [184] and Le Song et al. [186] have established a systematic way to embed probability distributions into RKHS of features and found a link between linear operators on these spaces and typical operations occurring during Bayesian inference. They show that the embedding is injective under mild conditions on the kernel K of the embedding space $\mathcal{H}_K$ and derive several independence criteria that generalize the correlation coefficient towards quantities that detect nonlinear functional dependencies. The ideas are as follows [184] , [186].

Definition 2.3.7 Suppose P and Q are probability distributions describing random variables X and Y with realizations x_j, y_j respectively. $\mathcal{H}_{K_P}$ and $\mathcal{H}_{K_Q}$ are RKHS with kernels K_P and K_Q.

1. The functions

$$\mu_P := E_X[K_P(X,\cdot)] \quad \text{and} \quad \widehat{\mu}_P - \frac{1}{m}\sum_{j=1}^{m} K_P(x_j,\cdot)$$

   are called the kernel mean embedding and the empirical kernel mean embedding.

2. The operator

$$C_{XY} := E_{[PQ]}[K_P(X,\cdot) \otimes K_Q(Y,\cdot)^*] - E_P[K_P(X,\cdot)] \otimes E_Q[K_Q(Y,\cdot)]^*$$

   is called the cross covariance operator; the covariance operators C_{XX} and C_{YY} are constructed analogously. $[PQ]$ is the joint probability distribution of X and Y.

3. The names of Hilbert Schmidt independence criterion (HSIC) and constrained covariance (COCO) are given to the terms [83, 84]

$$HSIC([PQ], \mathcal{H}_{K_P}, \mathcal{H}_{K_Q}) : = \|C_{XY}\|_{HS}^2 \tag{2.62}$$

$$COCO([PQ], \mathcal{H}_{K_P}, \mathcal{H}_{K_Q}) : = \sup_{f \in \mathcal{H}_{K_P}, g \in \mathcal{H}_{K_Q}} \frac{\operatorname{cov}(f(X)g(Y))}{\|f\|_{\mathcal{H}_{K_P}} \|g\|_{\mathcal{H}_{K_Q}}} \tag{2.63}$$

   They are measures of X and Y's possibly nonlinear statistical dependence.

The subjects of the above definitions have some convenient properties.

Theorem 2.3.8 *With the same notation as in definition 2.3.7, the following holds.*

1. *If $E_X[K_P(X,X)] < \infty$, then $\mu_P \in \mathcal{H}_{K_P}$ and μ_P is reproducing in the sense that [186]*

$$\langle \mu_P, f \rangle_{\mathcal{H}_P} = E_X[f(X)].$$

2. *If $E_X[K_P(X,X)] < \infty$ and $E_Y[K_Q(Y,Y)] < \infty$, then $C_{XX} \in \mathcal{H}_{K_P} \otimes \mathcal{H}_{K_P}^*$, $C_{YY} \in \mathcal{H}_{K_Q} \otimes \mathcal{H}_{K_Q}^*$ and $C_{XY} \in \mathcal{H}_{K_P} \otimes \mathcal{H}_{K_Q}^*$. Furthermore C_{XY} is the centered kernel mean embedding of the joint probability distribution $[PQ]$ of X and Y, i.e.*

$$C_{XY} = \mu_{[PQ]} - \mu_P \otimes \mu_Q^*$$

   and it is the unique operator satisfying $\langle f, C_{XY} g \rangle_{\mathcal{H}_{K_P}} = \operatorname{cov}(f(X)g(Y))$ [146, p. 35]. Obviously, similar statements hold for C_{XX} and C_{YY}.

3. *IF K_P and K_Q are characteristic kernels [69] on compact domains, then*

$$HSIC = 0 \Leftrightarrow X \coprod Y \Leftrightarrow COCO = 0.$$

   HSIC is efficiently estimable from a finite amount of data by calculating

$$\widehat{HSIC} = [m-1]^{-2} \operatorname{tr}\left(L_{K_P} H L_{K_Q} H\right)$$

   where $H = I - m^{-1}\mathbf{1} \otimes \mathbf{1}^*$, I is the unit matrix, $\mathbf{1}$ is a vector of ones and $(L_{K_P})_{ij} = K_P(x_i, x_j)$, $(L_{K_Q})_{ij} = K_Q(y_i, y_j)$. COCO is equal to the spectral norm of C_{XY}, i.e. $COCO = \|C_{XY}\|_{\mathcal{B}(L^2,L^2)} = \sigma_{\max}(C_{XY})$.

Part 3 of theorem 2.3.8 is interesting in the sense that it shows a solution to the problems of the standard linear correlation exhibited in figure 2.16. Whereas many other measures for detecting and quantifying nonlinear dependence between random variables exist, few have a simple computable representation in terms of linear operators. Figure 2.17 aims to illustrate the procedure with which COCO identifies functional dependence ignored by the linear correlation coefficient. Practical implementation of an estimator resorts to reducing the computation of $\|C_{XY}\|_{\mathcal{B}(L^2,L^2)}$ with infinite-dimensional C_{XY} to solving for the largest singular value of an $n \times n$ matrix

$$COCO_{emp} = \frac{1}{n}\sqrt{\sigma_{\max}\left(HL_{K_P}HHL_{K_Q}H\right)}$$

where L_{K_P} and L_{K_Q} are $n \times n$ kernel matrices and H is the centering matrix as before [84].

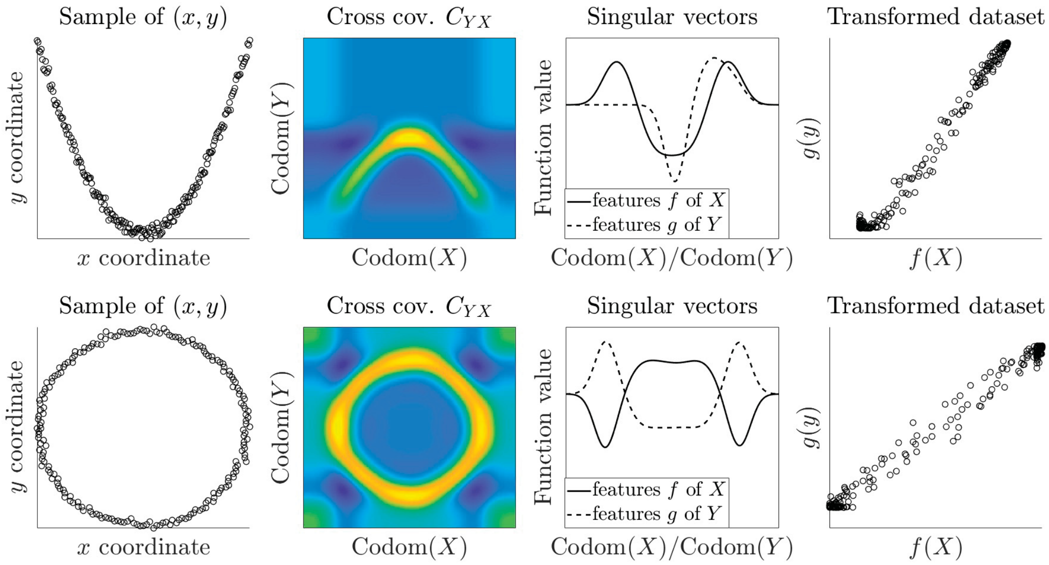


Figure 2.17: The samples from the previous figures can be transformed nonlinearly to reveal a linear dependence in feature space. For this, the left and right singular eigenvectors of the cross covariance operator C_{XY} are used. The value of COCO is 0.0045 for the example in row 1, 0.0038 for row 2, 0.0018 for white noise and 0.045 for a straight line. The squared exponential kernel was used.

As previously indicated, marginal probability distributions P_X, P_Y, joint probability distributions P_{XY} and conditional probability distributions $P_{X|Y}$ can all be treated in an RKHS framework. Le Song and Fukumizu show that marginalization and conditioning correspond to applications of linear operators acting on RKHS; more specifically it is even possible to find kernel analogues of the sum rule, the chain rule and the Bayes rule [186, 68].

In section 3.3 the methodology of Hilbert space embeddings of probability distributions will be applied to compare samples and check independence in the context of some simple geodetic problems featuring non-Gaussianity and nonlinearity. Other than in this subsection, however, the view of RKHS as function spaces with physical meaning will stay the predominant perspective.

Chapter 3

Statistical inference and splines

The goal of this chapter is to provide an account of practically useful information that endows the functional analytic parts of chapter 2 with a stochastic background and employs both for optimal estimation. Minimum norm problems as naturally arising during investigations of physical systems like in subsection 2.3.2 will be connected to maximum likelihood estimation. The end result of this line of thought will be a seemingly simple formula for abstract splines that is linear in the data and relies on the covariance matrices, and therefore the kernel, being known. Different ways to handcraft kernels by transferring the Hilbert space operations of subsection 2.1.3 to a functional level are discussed and reveal the abstract spline formula developed before as the basis case to which most problems can be reduced. Focus will then shift towards imbuing direct sum and tensor product constructions on Hilbert spaces with concrete and interpretable meaning. The chapter is closed with fully worked examples from geodesy that cover process modelling as well as high dimensional inference and statistical testing.

3.1 Hilbert spaces and estimation

This section is meant as an exposition on how the exchange of an estimation problem for a norm minimization problem in RKHS can be justified. As a stochastic perspective is assumed and signals are considered as collections of random variables that are mutually correlated, the validity of the arguments put forward will depend explicitly on the involved probability distributions. It will prove therefore necessary to first define what is understood by a stochastic process and a random field before the task of best linear unbiased estimation is tackled. Several different interpretations of the standard Kriging equation are provided and enable effortless swapping to the perspective most convenient and intuitive under the circumstances given. Independently of that, however, the mathematical formalism of choice will always be the one associated to RKHS and abstract splines — solutions to optimization problems in RKHS posed in terms of measurement operators and energy operators that significantly generalize the idea of measurements as point evaluations and energy as curvature.

3.1.1 Stochastic processes and random fields

The arithmetic mean $\hat{\mu}$ of i.i.d. Gaussian distributed random variables is the maximum likelihood estimator for the common expected value μ of those random variables. During the derivation, a quadratic form of the type $\sum_{i=1}^n (X_i - \hat{\mu})^2 \sigma^{-2}$ is minimized explaining the name least squares estimator. This establishes equality between LS- and ML-estimators for this exemplary case and hints at connections to be discovered later in more general settings. Examining the problem of interpolation instead of parameter estimation, one is lead to consider as interesting objects assignments of random variables to an index set T rather than single random variables. Such objects are called stochastic processes or random fields and while each random variable $X_t, t \in T$ can be described with a one dimensional probability distribution, it is the global behavior of all X_t together as given by the joint probability distribution that is needed for inference. For a stochastic process and its covariance function $K(\cdot,\cdot)$, Loewe's theorem asserts constructively the existence of an isometric isomorphism into an RKHS for which minimum norm problems of the form $\|x\|_{\mathcal{H}_K}$ arise naturally and have an interpretation as maximum likelihood estimators for jointly Gaussian distributed random variables.

Suppose a function $f(\cdot) = \sum_{i=1}^m \alpha_i g_i(\cdot) : T \to \mathbb{R}$ was measured at locations $\{t_i\}_{i=1}^n$ and denote the vector of observations as $y \in \mathbb{R}^n$. If the measurements are subject to uncertainty, then the $y_i, i = 1, ..., n$ can be interpreted as realizations of random variables

$$Y_i = f(t_i) + N_i \tag{3.1}$$

where $N_i : \Omega \to \mathbb{R}, i = 1, ..., n$ for some probability space Ω is a mean zero noise random variable subsuming the irreproducible random deviations of Y_i from $f(t_i)$ as induced by e.g. thermal noise or quantization errors in the measurement device. If it is assumed that the set $N_1, ..., N_n$ is jointly Gaussian distributed with covariance matrix Σ_{NN}, then one may assemble them into the random vector $N : \Omega \to \mathbb{R}^n$ with joint probability density function

$$p_N(\epsilon) = (2\pi)^{-n/2} |\Sigma_{NN}|^{-1/2} \exp\left(-\frac{1}{2}\|\Sigma_{NN}^{-1/2}(\mu_N - \epsilon)\|_{\mathbb{R}^n}\right) \quad \forall \epsilon \in \mathbb{R}^n \tag{3.2}$$

where $\mu_N = 0$ and $|\Sigma_{NN}|$ denotes the determinant of Σ_{NN} [160, p. 68]. This is also written as $N \sim \mathcal{N}(0, \Sigma_{NN})$. From equation 3.1 and the linearity of expectation it follows that $Y : \Omega \to \mathbb{R}^n, (Y)_i = Y_i$, is also a Gaussian random vector and

$$\mu_Y := E[Y] = E[F + N] = F \in \mathbb{R}^n \quad (F)_i = f(t_i)$$
$$\Sigma_{YY} := E[Y \otimes Y^*] - E[Y] \otimes E[Y]^* = E[N \otimes N^*] = \Sigma_{NN}$$

establishing $Y \sim \mathcal{N}(\mu_Y, \Sigma_{NN})$. Noticing the explicit dependence of $p_Y(y)$ on μ_Y and setting $L(\mu_Y|y) := p_Y(y)$, L measures how probable any concrete observation vector $y \in \mathbb{R}^n$ is given the expected value μ_Y. L is termed the likelihood and maximizing it w.r.t. μ_Y is akin to finding that μ_Y which has the highest probability of generating the observations. The resulting maximum likelihood estimator $\hat{\mu}_Y$ is

$$\hat{\mu}_Y = \underset{\mu_Y \in \mathbb{R}^n}{\operatorname{argmax}} L(\mu_Y|y)$$

$$
\begin{aligned}
&= \underset{\mu_Y \in \mathbb{R}^n}{\operatorname{argmin}} \; -\log L(\mu_Y|y) \\
&= \underset{\mu_Y \in \mathbb{R}^n}{\operatorname{argmin}} \; \frac{1}{2}\|\Sigma_{YY}^{-1/2}(y-\mu_Y)\|_{\mathbb{R}^n}^2 \qquad (3.3)
\end{aligned}
$$

where in the last step it was used that the expression $(n/2)\log(2\pi)+(1/2)\log|\Sigma_{YY}|$ does not depend on the parameter μ_Y. If $\mu_Y = \{f(t_i)\}_{i=1}^n$ and $f(\cdot)$ has the decomposition $f(\cdot) = \sum_{i=1}^m \alpha_i g_i(\cdot)$ then μ_Y may be written as $\mu_Y = G\alpha$ with $G \in \mathbb{R}^n \otimes \mathbb{R}^n, (G)_{ij} = g_j(t_i)$ and $\alpha \in \mathbb{R}^m$ is the parameter vector to be determined. The least squares problem 3.3 can then be reformulated as $\hat{\mu}_Y = G\hat{\alpha}$ with

$$
\hat{\alpha} = \underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}} \; \|\widetilde{G}\alpha - \widetilde{y}\|_{\mathbb{R}^n}^2 \qquad \widetilde{y} = \Sigma_{YY}^{-1/2} y, \;\; \widetilde{G} = \Sigma_{YY}^{-1/2} G \qquad (3.4)
$$

which, assuming $m < n$ and G of full column rank, has as its solution $\hat{\alpha} = \widetilde{G}^+\widetilde{y}$ [18, p. 108] with the $^+$-sign denoting the pseudoinverse:

$$
\hat{\alpha} = \left(\widetilde{G}^*\widetilde{G}\right)^{-1}\widetilde{G}^*\widetilde{y} = \left(G\Sigma_{YY}^{-1}G\right)^{-1}G^*\Sigma_{YY}^{-1}y \qquad (3.5)
$$

$$
\hat{\mu}_Y = G\left(G^*\Sigma_{YY}^{-1}G\right)^{-1}G^*\Sigma_{YY}^{-1}y \qquad (3.6)
$$

Equations 3.5 and 3.6 are the well known least squares estimators widely used in adjustment theory [151, p. 133]. Notice that the matrix $P = G\left(G^*\Sigma_{YY}^{-1}G\right)^{-1}G^*\Sigma_{YY}^{-1}$ satisfies $P^2 = P^* = P : \mathbb{R}^n \to \mathbb{R}^n$ and is therefore an orthogonal projection w.r.t. the inner product $\langle f, g\rangle_{\mathcal{H}} = \langle \Sigma_{YY}^{-1} f, g\rangle_{\mathbb{R}^n}$ hinting already at a Hilbert space approach to the problem.

Model 3.1 presupposes a parametric form $f = \sum_{i=1}^m \alpha_i g_i$ for the function f to be extracted from the data y and the subsequent calculations operate under the assumption that $E[f] = f$, i.e. f is deterministic. This restricts the class $\mathcal{F}$ of f's, for which inference is possible under this model and places on the practitioner the burden of choosing a proper model. As additionally there is no way to prefer certain parameter combinations and all $f \in \mathcal{F}$ are considered a priori to be equally likely, the simple least squares procedure just outlined is both too rigid in its assumptions and too weak in its ability to integrate prior knowledge. Typically, real world phenomena are not described by a class $\mathcal{F}$ of deterministic functions that are all equally likely and for reasons of robustness and flexibility one should allow the class $\mathcal{F}$ to be rather broad and furnished with a probability measure. The concept of a stochastic process [140, p. 190] achieves exactly that.

Definition 3.1.1 A map $X_\cdot(\cdot) : \Omega \times T \ni (\omega, t) \mapsto X_\omega(t) \in \mathbb{C}$ is called a stochastic process on T. Here Ω is some probability space furnished with both a sigma algebra and a probability measure [140, p. 20] . If $E[\|X_\cdot(t)\|_{\mathbb{C}}^2] < \infty \; \forall t \in T$, then the process $X_\cdot(\cdot)$ is called second order. If for all finite $\{t_1, ..., t_n\}$ the joint probability distribution of $[X_\cdot(t_1), ..., X_\cdot(t_n)]^T$ is Gaussian, one speaks of a Gaussian process.

Remark The property of being second order should not be confused with second order stationarity which states that the first two moments of a process are translation invariant: $E[X_t] = \mu \, \forall t \in T$ and $E[X_{t_1+s}X_{t_2+s}] = E[X_{t_1}X_{t_2}] \ \forall t_1, t_2, s \in T$.
Remark When $T \in \mathbb{R}^m, m > 1$, then the term 'random field' is often preferred.

The definition given above is not the most general one but sufficient for the purposes of this monograph. A stochastic process has two different interpretations. It is possible to split the map $X_\cdot(\cdot) : \Omega \times T \to \mathbb{C}$ into

$$\begin{aligned} a) \ \ X_\cdot(\cdot) :& T \ni t \mapsto [X_\cdot(t) : \Omega \to \mathbb{C}] \\ b) \ \ X_\cdot(\cdot) :& \Omega \ni \omega \mapsto [X_\omega(\cdot) : T \to \mathbb{C}] \end{aligned}$$

Expression a) suggests that $X_\cdot(\cdot)$ may be seen as a set $\{X_\cdot(t)\}_{t\in T}$ of mutually dependent random variables whereas b) corresponds to equating $X_\cdot(\cdot)$ to randomly drawing functions $X_\omega(\cdot)$ from T to $\mathbb{C}$. Both interpretations are valid and useful at different times.

Example 13 Set $T = \{1, ..., n\}$ and define $N_i, i = 1, ..., n$ to be independent white noise variables $N_i \sim \mathcal{N}(0, \sigma_0^2)$. The set $\{X_t\}_{t\in T}$ with $X_t = \sum_{j=1}^t N_j$ is a stochastic process with

$$E[X_t] = 0 \tag{3.7}$$
$$E[X_sX_t] = \sigma_0^2 \min(s,t). \tag{3.8}$$

As a discretization of the Wiener process it is not second order stationary. It arises in many practical applications in which random deviations accumulate during measurements and is a reasonable stochastic model for the error in leveling as illustrated in figure 3.1. ■

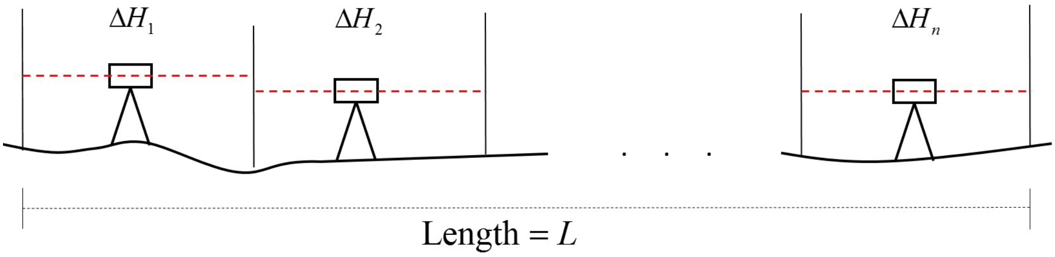


Figure 3.1: A sketch of the leveling procedure. The vertical bars symbolize the position of leveling rods and the rectangles indicate the positions of the leveling instrument. Assuming each measured height difference ΔH_j to be influenced by white noise with variance σ_0^2, one recovers equations 3.7 and 3.8 for the leveling error $X_t = \sum_{j=1}^t \Delta H_j - E[\sum_{i=1}^t \Delta H_j]$. Under the assumption of constant distances between rods, the usually claimed relationship $\sigma^2 \propto L$ linking leveling variance σ^2 and length L [209, p. 272] follows.

Example 13 showcases that nontrivial processes can be generated from white noise. The latter takes on a special role in stochastic calculus. To define it formally on continuous domains, one would need to introduce generalized stochastic processes — random linear functionals rather than random functions. The construction is given in e.g. [97] or [139]. When dealing with white noise we will either revert to the finite dimensional case of a white noise random vector or argument informally.

Formally, the relationship between an RKHS $\mathcal{H}_K$ and a Hilbert space of random variables is one of being isometrically isomorphic rather than strict equality. Given a mean-zero second order stochastic process $X := \{X_t\}_{t\in T}$, the Hilbert space

$$\overline{\mathcal{L}}(X) := \mathrm{cl}\left\{\sum_{j=1}^{n} \alpha_j X_{t_j} : \ \ t_j \in T, \alpha_j \in \mathbb{R} \ \ \text{for } j \in \{1, ..., n\}\right\}$$
$$\langle X_1, X_2\rangle_{\overline{\mathcal{L}}(X)} := \int_\Omega X_1(\omega)X_2(\omega)\mu(d\omega) = E[X_1X_2]$$

with Ω a probability space with probability measure μ is called the Hilbert space $\overline{\mathcal{L}}(X)$ generated by the process X [20, p. 62]. In what follows, stochastic processes will always assumed to be mean-zero if nothing to the contrary is mentioned.

Theorem 3.1.2 (Loewe representation theorem) *The Hilbert space $\overline{\mathcal{L}}(X)$ generated by the second order stochastic process $\{X_t\}_{t\in T}$ on T is isometrically isomorphic to the RKHS $\mathcal{H}_K$, $K(s,t) = E[X_sX_t]$ $\forall s,t \in T$ via (the extension of) the mappings*

$$\psi : \overline{\mathcal{L}}(X) \ni \sum_{i=1}^{n} \alpha_i X_{t_i} \mapsto \sum_{i=1}^{n} \alpha_i K(t_i, \cdot) \in \mathcal{H}_K \tag{3.9}$$
$$\psi^{-1} : \mathcal{H}_K \ni \sum_{i=1}^{n} \alpha_i K(t_i, \cdot) \mapsto \sum_{i=1}^{n} \alpha_i X_{t_i} \in \overline{\mathcal{L}}(X). \tag{3.10}$$

A proof can be found in [20, p.65]. This indirect isometry argument is necessary because — as [120] notes — it is impossible to construct Gaussian probability measures on infinite dimensional Hilbert spaces $\mathcal{H}$ directly. Classically, one follows the construction proposed in [86, 87] and completes $\mathcal{H}$ w.r.t to the uniform norm to create a slightly enlarged Banach space $\mathcal{B}$ on which Gaussian measures are definable and directly related to the reproducing kernel of $\mathcal{H}$ (if it exists). Whereas only then one might rigorously speak of having a probability distribution over the function space $\mathcal{B} \supset \mathcal{H}$, it is conceptually convenient to speak of the likelihood $p(f)$ of $f \in \mathcal{H}$

$$f \in \mathcal{H} \quad p(f) \propto \exp\left(-\|f\|^2_{\mathcal{H}}\right)$$

as then norm-minimization in $\mathcal{H}$ corresponds to minimum variance and maximum likelihood estimation [124]. For clarifying comments and to provide intuition, we will regularly argument as though $\mathcal{H}$ is furnished with a probability measure μ although μ is in reality only defined on the closure of $\mathcal{H}$ w.r.t. some norm. The author hopes that this perspective is helpful even though it is not entirely accurate.

In finite dimensional cases, the isometric isomorphisms ψ and ψ^{-1} from theorem 3.1.2 have explicit representations in terms of covariance matrices and their inverses. Denote by K the $n \times n$ matrix with entries $(K)_{ij} = K(t_i, t_j)$ and by K^{-1} its inverse. Then

$$\psi^{-1} : \mathcal{H}_K \ni f \mapsto \sum_{i=1}^{n} \sum_{j=1}^{n} X_{t_i} (K^{-1})_{ij} \langle K(t_j, \cdot), f \rangle_{\mathcal{H}_K} \tag{3.11}$$

is the inverse to ψ.

Proof: If ψ^{-1} is defined as above, the one finds that ψ, ψ^{-1} are surjective and injective as

$$\begin{tikzcd}
& \mathcal{H}_K \arrow[dr, "\psi^{-1}"] & \\
\overline{\mathcal{L}}(X) \arrow[ur, "\psi"] \arrow[rr, "id_{\overline{\mathcal{L}}(X)}"] & & \overline{\mathcal{L}}(X)
\end{tikzcd}$$

$$\begin{aligned}
&\psi^{-1} \circ \psi(X_{t_l}) \\
&\quad = \psi^{-1}[K(t_l, \cdot)] \\
&\quad = \sum_{i=1}^{n} \sum_{j=1}^{n} X_{t_i} (K^{-1})_{ij} \langle K(t_i, \cdot), K(t_l, \cdot) \rangle_{\mathcal{H}_K} \\
&\quad = X_{t_l} \qquad \forall t_l \in T
\end{aligned}$$

$$\begin{tikzcd}
& \overline{\mathcal{L}}(X) \arrow[dr, "\psi"] & \\
\mathcal{H}_K \arrow[ur, "\psi^{-1}"] \arrow[rr, "id_{\mathcal{H}_K}"] & & \mathcal{H}_K
\end{tikzcd}$$

$$\begin{aligned}
&\psi \circ \psi^{-1}[K(t_l, \cdot)] \\
&\quad = \psi \left(\sum_{i=1}^{n} \sum_{j=1}^{n} X_{t_i} (K^{-1})_{ij} \langle K(t_j, \cdot), K(t_l, \cdot) \rangle_{\mathcal{H}_K} \right) \\
&\quad = \psi(X_{t_l}) \\
&\quad = K(t_l, \cdot) \quad \forall t_l \in T
\end{aligned}$$

Since for all $s, t \in T$ it holds that

$$\langle X_s, X_t \rangle_{\overline{\mathcal{L}}(X)} = K(s,t) = \langle K(s, \cdot), K(t, \cdot) \rangle_{\mathcal{H}_K} = \langle \psi(X_s), \psi(X_t) \rangle_{\mathcal{H}_K}$$
$$\langle K(s, \cdot), K(t, \cdot) \rangle_{\mathcal{H}_K} = K(s,t) = \langle X_s, X_t \rangle_{\overline{\mathcal{L}}(X)} = \langle \psi^{-1}[K(s, \cdot)], \psi^{-1}[K(t, \cdot)] \rangle_{\overline{\mathcal{L}}(X)}$$

and ψ, ψ^{-1} are both seen to be isometric as well. □

For $T = \{t_1, ..., t_n\}$ and denoting $f \in \mathbb{R}^n, (f)_j = f(t_j)$ and $\alpha \in \mathbb{R}^n$ with $X = \sum_{j=1}^{n} \alpha_j X_{t_j}$, the mappings are given by

$$\psi : X \mapsto K\alpha \qquad \psi^{-1} : f \mapsto \sum_{j=1}^{n} (K^{-1} f)_i X_{t_i}$$

This has immediate geometric relevance for the optimal estimation of random variables. Suppose a square integrable random variable $X \in L^2(\Omega)$ is to be projected onto the subspace $\overline{\mathcal{L}}(X)$ generated by observations $X_1, ..., X_n$. Then the minimum variance estimator $\hat{X}$ satisfying

$$E[(\hat{X}-X)^2] = \|\hat{X}-X\|^2_{L^2(\Omega)} \to \min$$

is according to theorem 2.1.17.1 given by

$$\hat{X} = P_{\overline{\mathcal{L}}(X)}X$$

where $P_{\overline{\mathcal{L}}(X)}$ is the orthogonal projection from $L^2(\Omega)$ onto its subspace $\overline{\mathcal{L}}(X)$. As follows from the rules governing projections (see 2.1.4) and figure 3.2,

$$X-\hat{X} = (I-P_{\overline{\mathcal{L}}(X)})X \perp \overline{\mathcal{L}}(X).$$

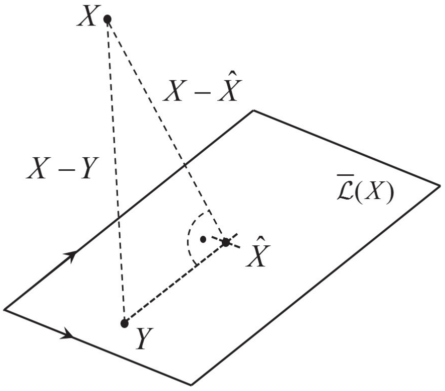


Figure 3.2: Orthogonality relations for minimum variance estimation.

Conversely, if $\|X-\hat{X}\|_{L^2(\Omega)}, \hat{X} \in \overline{\mathcal{L}}(X)$ is supposed to be minimal, then $X-\hat{X} \in \overline{\mathcal{L}}(X)^\perp$ since otherwise $\|X-\hat{X}\|^2_{L^2(\Omega)} = \|X-\hat{X}\|^2_{\overline{\mathcal{L}}(X)^\perp} + \|X-\hat{X}\|^2_{\overline{\mathcal{L}}(X)}$ and the last term is ≥ 0. Therefore the minimum variance estimator $\hat{X} \in \overline{\mathcal{L}}(X)$ for $X \in L^2(\Omega)$ must satisfy $\langle X-\hat{X}, X_{t_j}\rangle_{L^2(\Omega)} = 0 \ \forall X_{t_j}$ spanning $\overline{\mathcal{L}}(X)$. By the properties of the isometric isomorphisms

$$\langle X-\hat{X}, X_{t_i}\rangle_{L^2(\Omega)} = 0 \Leftrightarrow \langle \psi(\hat{X}), K(t_i,\cdot)\rangle_{\mathcal{H}_K} = \rho_X(t_i)$$

with $\rho_X(t_i) = E[XX_{t_i}]$. If $\rho_X \in \mathcal{H}_K$, this is the case iff $\hat{X} = \psi^{-1}\rho_X$ and the estimation variance is $\|\hat{X}\|^2_{\overline{\mathcal{L}}(X)} = \|\rho_X\|^2_{\mathcal{H}_K}$ [20, p. 72]. If $X = X_t$ and $\overline{\mathcal{L}}(X) = \overline{\text{span}}\,\{X_{t_1},...,X_{t_n}\}$, then

$$\begin{aligned}\hat{X}_t = \psi^{-1}\rho_X &= \sum_{i=1}^n\sum_{j=1}^n X_{t_i}(K^{-1})_{ij}\langle K(t_j,\cdot),\rho_x(\cdot)\rangle_{\mathcal{H}_K}\\ &= \sum_{i=1}^n\sum_{j=1}^n (K^{-1})_{ij}K(t_j,t)X_{t_i}\\ &= \sum_{i=1}^n \lambda_i X_{t_i} \qquad (3.12)\end{aligned}$$

with $\lambda_i = \sum_{j=1}^n (K^{-1})_{ij}K(t_j,t)$ is just the simple Kriging estimator as recorded e.g. in [39, p. 151]. This shows that there is a relation between optimization in $\mathcal{H}_K$ and in $\overline{\mathcal{L}}(X)$. It is summarized in the next theorem.

Theorem 3.1.3 *Suppose $\overline{\mathcal{L}} := \overline{\mathcal{L}}(X)$ is the Hilbert space generated by the mean-zero second order process $\{X_t\}_{t\in T}$ with inner product $\langle X_s, X_t\rangle_{\overline{\mathcal{L}}(X)} = E[X_sX_t] = K(s,t)\ \forall s,t \in T$ and $\mathcal{H}_K$ is the corresponding RKHS. Let $\overline{S}(X)$ be the Hilbert space generated by the random variables $X_{t_1},...,X_{t_n} \in \overline{\mathcal{L}}(X)$ and denote by $\psi : \overline{\mathcal{L}} \to \mathcal{H}_K$ the isometric isomorphism from Loewe's theorem. Then the following ways to devise an optimal spatial estimator $\hat{X}_{t_0} \in \overline{S}(X)$ for $X_{t_0} \in \overline{\mathcal{L}}$ are equivalent.*

i) $\hat{X}_{t_0} = \underset{\hat{X}_{t_0} \in \overline{S}(X)}{\text{argmin}} \ \|\hat{X}_{t_0} - X_{t_0}\|^2_{\overline{\mathcal{L}}}$.

ii) $\hat{X}_{t_0} \in \overline{S}(X)$ *solves* $\langle \hat{X}_{t_0}, X_{t_i}\rangle_{\overline{S}(X)} = K(t_0, t_i)$ *for all* X_{t_i} *spanning* $\overline{S}(X)$.

iii) $\hat{f} = \underset{f \in \mathcal{H}_K, f = \sum_{j=1}^n \lambda_j K(t_j, \cdot)}{\text{argmin}} \ \|f(\cdot) - K(t_0, \cdot)\|^2_{\mathcal{H}_K}, \quad \hat{X}_{t_0} = \psi^{-1}\hat{f}$

iv) $\hat{f} = \underset{f \in \mathcal{H}_K, f(t_i) = K(t_0, t_i)}{\text{argmin}} \ \|f\|^2_{\mathcal{H}_K}, \quad \hat{X}_{t_0} = \psi^{-1}(f)$

They coincide with the simple Kriging estimator as known from the geostatistical literature.

Proof: Straightforward computation reveals that all estimators have the same form. A derivation of the optimizers i) and ii) and comments hinting at iv) can be found in [39, p. 151] and [20, p. 72] and [20, p. 77] respectively; they are reproduced below for the readers convenience. Denote by K the matrix with elements $(K)_{ij} = K(t_i, t_j)$ and by K^{-1} its inverse. Let $K_{t_0} \in \mathbb{R}^n$ be the vector with entries $(K_{t_0})_j = K(t_0, t_j)$ and $\lambda = [\lambda_1, ..., \lambda_n]^T$ the vector of coefficients.

i) Finding $\hat{X}_{t_0} = \sum_{j=1}^n \lambda_j X_{t_j}$ to minimize $\|\hat{X}_{t_0} - X_{t_0}\|^2_{\overline{\mathcal{L}}(X)}$ is equivalent to minimizing

$$\begin{aligned}\|\hat{X}_{t_0} - X_{t_0}\|^2_{L^2(\Omega)} &= \|\sum_{j=1}^n \lambda_j X_{t_j} - X_{t_0}\|^2_{\overline{\mathcal{L}}(X)} \\ &= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \langle X_{t_i}, X_{t_j}\rangle_{\overline{\mathcal{L}}(X)} \\ &\qquad - 2\sum_{j=1}^n \lambda_j \langle X_{t_j} X_{t_0}\rangle_{\overline{\mathcal{L}}(X)} + \langle X_{t_0}, X_{t_0}\rangle_{\overline{\mathcal{L}}(X)} \\ &= \lambda^T K \lambda - 2\lambda^T K_{t_0} + \sigma_0^2\end{aligned}$$

Taking partial derivatives and setting them to zero produces $\lambda = K^{-1}K_{t_0}$.

ii) The solution to $\langle \hat{X}_{t_0}, X_{t_i}\rangle_{\overline{S}(X)} = K(t_0, t_i), \hat{X}_{t_0} \in \overline{S}(x)$ is $\psi^{-1}[K(t_0, \cdot)]$. The explicit formula for ψ^{-1} then leads to

$$\begin{aligned}\psi^{-1}[K(t_0, \cdot)] &= \sum_{i=1}^n \sum_{j=1}^n X_{t_i} (K^{-1})_{ij} \langle K(t_j, \cdot), K(t_0, \cdot)\rangle_{\mathcal{H}_K} \\ &= \sum_{j=1}^n \lambda_j X_{t_j} \qquad \lambda = K^{-1}K_{t_0}\end{aligned}$$

which is the same as the term derived for i).

iii) Since $\psi : \overline{\mathcal{L}}(X) \to \mathcal{H}_K$ is an isometry, $f = \psi(\hat{X}_{t_0})$ minimizes $\|f(\cdot) - K(t_0, \cdot)\|^2_{\mathcal{H}_K}$ iff $\|\hat{X}_{t_0} - X_{t_0}\|^2_{\overline{\mathcal{L}}(X)}$ is minimized by $\hat{X}_{t_0}$. Explicitly,

$f = \sum_{j=1}^{n} \lambda_j K(t_j, \cdot)$ inserted into $\|f(\cdot) - K(t_0, \cdot)\|_{\mathcal{H}_K}^2$ produces $\lambda^T K \lambda - 2\lambda^T K_{t_0} + \sigma_0^2$ which has a minimum at $\lambda = K^{-1} K_{t_0}$.

iv) The constrained minimization $\|f\|_{\mathcal{H}_K}^2 \to \min, f \in \mathcal{H}_K, f(t_i) = K(t_0, t_i)$ can be formulated with the help of Lagrangian multipliers $\mu \in \mathbb{R}^n$ as

$$\begin{aligned} &\langle f, f\rangle_{\mathcal{H}_K} + 2\sum_{i=1}^{n} \mu_i \left(f(t_i) - K(t_0, t_i)\right) \to \min \\ \Leftrightarrow &\lambda^T K \lambda + 2\mu^T \left[K\lambda - K_{t_0}\right] \to \min \end{aligned}$$

where the representer theorem was used to express the minimizer f in terms of the kernel. The corresponding SLAE has solution $\lambda = K^{-1} K_{t_0}$.

□

The theorem establishes that minimum variance estimation of a random variable X_{t_0} can be seen from different perspectives: As minimizing an error variance (i), as an orthogonal projection onto a set of observed random variables (ii), as finding an element with a closely aligned correlation structure in RKHS (iii), or as solving a constrained energy minimization problem in RKHS (iv). When the underlying process is Gaussian, minimum variance estimation, maximum likelihood estimation and the formation of conditional expectation coincide [150].

Plugging in numbers x_{t_i} into the estimator derived from the optimization problem

$$\hat{f}(t) = \underset{f(t_i) = K(t, t_i), f(t) = \sum_{j=1}^{n} \lambda_j K(t, t_i)}{\operatorname{argmin}} \|f\|_{\mathcal{H}_K}^2$$

one finds $\hat{f}(t) = \sum_{j=1}^{n} \lambda_j K(t_j, t), \lambda = K^{-1} K_t$ and consequently $\hat{X}_t = \psi^{-1} \hat{f} = \sum_{j=1}^{n} \lambda_j X_{t_j}, \lambda = K^{-1} K_t$, i.e. $\hat{x}_t = \langle K^{-1} K_t, b\rangle_{\mathbb{R}^n}$ where b is the vector of observations $(b)_j = x_{t_j}$. This is the same solution one would recover if one interpreted $\mathcal{H}_K$ itself as a Hilbert space of functions with a Gaussian measure on it in which the negative log likelihood of a function $f \in \mathcal{H}_K$ is given by $c_1 \|f\|_{\mathcal{H}_K}^2 + c_2$. One would then solve

$$\sigma_x = \underset{Ax = b, x \in \mathcal{H}_K}{\operatorname{argmin}} \|x\|_{\mathcal{H}_K}^2$$

where $A : \mathcal{H}_K \to \mathbb{R}^n$ is the linear operator of evaluation at points $t_1, ..., t_n$, b is the vector of observed data $b = [x_{t_1}, ..., x_{t_n}]^T$ and $\sigma_x(t)$ now directly is an estimator for $x_t, t \in T$. This follows trivially from any solution σ_x having form $\sigma_x(t) = \sum_{j=1}^{n} \lambda_j K(t, t_j)$ according to the representer theorem and the fact that the associated Lagrangian is

$$\begin{aligned} &\langle x, x\rangle_{\mathcal{H}_K} + 2\sum_{i=1}^{n} \mu_i \left[\sum_{j=1}^{n} \lambda_j K(t_i, t_j) - b_i\right] \to \min \\ \Leftrightarrow &\lambda^T K \lambda + 2\mu^T \left[K\lambda - b\right] \to \min \end{aligned}$$

The resulting SLAE has solution $\lambda = K^{-1}b$ which implies $\sigma_x(t) = \sum_{j=1}^{n}(K^{-1}b)_j K(t_j,t) = \langle K^{-1}b, K_t\rangle_{\mathbb{R}^n}$. Since K is a symmetric positive semidefinite matrix, $K^{-1} = (K^{-1})^*$ and $\langle K^{-1}b, K_t\rangle_{\mathbb{R}^n} = \langle b, K^{-1}K_t\rangle_{\mathbb{R}^n}$ which is the expression derived from optimization formulation iv). The objects σ_x are called interpolating splines and are covered in more generality in the next section.
Up until now only Gaussian probability distributions have been considered and in the future it will often be assumed that some naturally occurring function can be modelled as a Gaussian process. One may cite the following justifications for this.

1) The Gaussian probability distribution is the maximum entropy distribution given only the first two statistical moments. If these come from observed data, it is the probability distribution presupposing the least amount of additional structure apart from what is observed.

2) Inference with random variables that are distributed according to joint Gaussian laws can be reduced to linear algebraic manipulations involving covariance matrices and observation vectors. This is convenient both computationally and storage-wise and in contrast to inference procedures involving higher order moments.

If the Gaussian process assumption is violated, the consequence is that the minimum norm estimators are not maximum likelihood estimators anymore. Nonetheless, they retain their property of having smallest expected square error.

3.1.2 Abstract splines for estimation

Instead of having as measurements available for estimation of $f(\cdot) \in \mathcal{H}$ just a set of function evaluations $\{f(t_s)\}_{t_s \in T_{sample}}$ it is also possible to have the measurements depending on $f(\cdot)$ via a linear operator $A : \mathcal{H} \rightarrow \mathcal{A}$. Similarly, it might be that the proper measure of energy is not known for f but only for Bf where $B : \mathcal{H} \rightarrow \mathcal{B}$ is called Energy operator. A certain $f^* \in \mathcal{H}$ satisfying $f^* = \operatorname{argmin}_{f\in\mathcal{H}} \|Af - a\|_{\mathcal{A}}^2 + \|Bf\|_{\mathcal{B}}^2$ is then called an abstract smoothing spline and it is clear that this framework contains previously covered methods like Kriging as subcases when A is evaluation and B is the identity on $\mathcal{H}$. Uniqueness and existence of solutions are not guaranteed anymore but under reasonable assumptions a closed form solution can again be found and useful relationships regarding smoothing splines, interpolating splines and commutativity of linear operators with the problem formulation can be established. Example calculations for tomography type problems will showcase the simplicity and usefulness of abstract splines.

Definition 3.1.4 Let $\mathcal{H}, \mathcal{H}_A, \mathcal{H}_B$ be Hilbert spaces and denote by $A \in \mathcal{B}(\mathcal{H}, \mathcal{H}_A), B \in \mathcal{B}(\mathcal{H}, \mathcal{H}_B)$ two bounded linear operators. For $a \in \mathcal{H}_A, r \geq 0$, the object

$$\sigma_f = \underset{f\in\mathcal{H}}{\operatorname{argmin}} \ \|Af - a\|_{\mathcal{H}_A}^2 + r\|Bf\|_{\mathcal{H}_B}^2 \tag{3.13}$$

is called an abstract smoothing spline with measurement operator A, energy operator B and regularization parameter r [20, p. 116].

Example 14 Adjustment can be seen as a subcase of smoothing splines. Suppose in the smoothing spline

$$\sigma_f = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \ \|Af - b\|^2_{\mathcal{H}_A} + \|Bf\|^2_{\mathcal{H}_B}$$

that $\mathcal{H}$ is the RKHS of functions of type $f(\cdot) = \sum_{j=1}^n c_j g_j(\cdot)$ on S for example by setting $K = \sum_{j=1}^n g_j \otimes \overline{g_j}$. If $\mathcal{H}_A = \mathbb{R}^{n_{\text{obs}}}$ with covariance matrix Σ_A, $B : \mathcal{H}_X \to \mathcal{H}_B$ is the trivial operator that sends everything to zero, A is evaluation at $\{s_j\}_{j=1}^{n_{\text{obs}}} \subset S$ and $G \in \mathbb{R}^{n_{\text{obs}}} \otimes \mathbb{R}^n$ is the matrix with entries $(G)_{ij} = g_j(s_i)$, then $Ax = Gc$ and

$$\begin{aligned} A\sigma_f = A\left(\underset{f \in \mathcal{H}}{\operatorname{argmin}} \ \|Af - b\|^2_{\mathcal{H}_A} \right) &= G\left(\underset{c \in \mathbb{R}^n}{\operatorname{argmin}} \ \|Gc - b\|^2_{\mathcal{H}_A} \right) \\ &= G\left(G^T \Sigma_A^+ G\right)^+ G^T \Sigma_A^+ b. \end{aligned} \tag{3.14}$$

Then $A\sigma_f$ are predictions equivalently derivable from the solution to an adjustment problem in the coefficient vector $c \in \mathbb{R}^n$ with solution $\hat{c}$ and $\sigma_f = \sum_{j=1}^n \hat{c}_j g_j(\cdot)$. ■

As just seen, the term $\|Af - a\|^2_{\mathcal{H}_A}$ is similar to the expression $\|Af - a\|^2_{\mathbb{R}^n}$ encountered in classical least squares. Both can be interpreted as penalizing deviations between observations predicted by a model (Af) and the actual observations (a). At the same time as minimizing this discrepancy, the term $\|Bf\|^2_{\mathcal{H}_B}$ in the definition of abstract splines provides some intrinsic measure of energy for f that is absent from least squares formulations. When f is a process driven by a differential equation, B often has concrete physical meaning and is related to the systems Hamiltonian, e.g. $B = \Delta$. From a stochastic perspective this could be considered as jointly maximizing the likelihood of the residuals $Af - a \in \mathcal{H}_A$ and of the function $f \in \mathcal{H}$ as quantified by $Bf \in \mathcal{H}_B$. The relationship is illustrated in figure 3.3.

Under slightly less general conditions than the ones underlying definition 3.1.4, the abstract smoothing spline equation can be solved uniquely.

Theorem 3.1.5 *Let $\mathcal{H} = \mathcal{H}_K$ be an RKHS with RK K and $\mathcal{H}_A = \mathbb{R}^n$ with inner product $\langle u, v\rangle_{\mathcal{H}_A} = \langle \Sigma_A^{-1/2} u, \Sigma_A^{-1/2} v\rangle_{\mathbb{R}^n}$. Let $A : \mathcal{H}_K \to \mathbb{R}^n$ consist of n linearly independent linear functions $l_1, ..., l_n : \mathcal{H}_K \to \mathbb{R}$ and B as in definition 3.1.4.*

i) If $\ker A \cap \ker B = \{0\}$ *and the range of the operator $L : \mathcal{H}_K \ni f \mapsto (Af, Bf) \in \mathcal{H}_A \oplus \mathcal{H}_B$ is closed in $\mathcal{H}_A \oplus \mathcal{H}_B$, then the smoothing spline*

$$\sigma_f = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \ \|Af - a\|^2_{\mathcal{H}_A} + \|Bf\|^2_{\mathcal{H}_B}$$

with measurement operator A, energy operator B and data $a \in \mathbb{R}^n$ exists and is uniquely determined [21, p. 4].

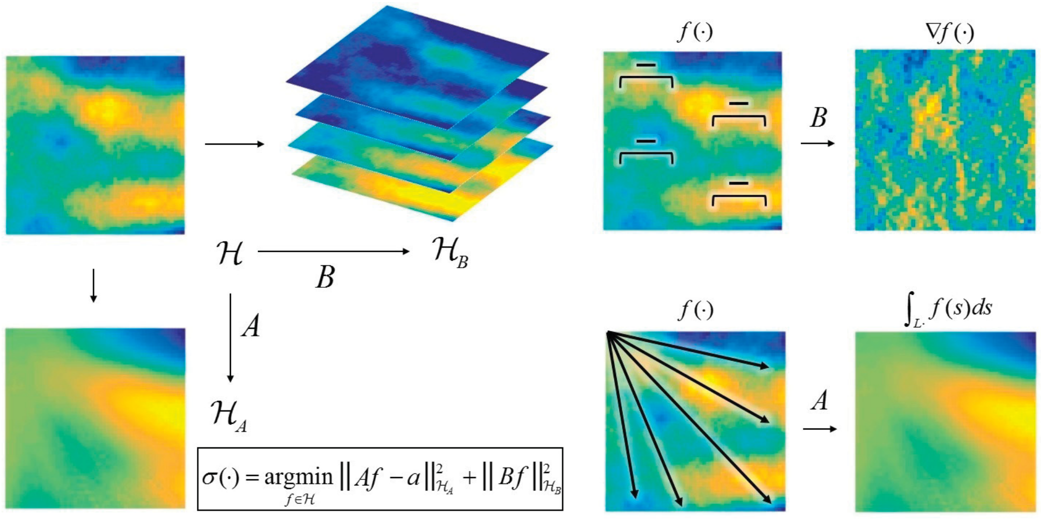

$\mathcal{H}$
B
$\mathcal{H}_B$
A
$\mathcal{H}_A$
$\sigma(\cdot) = \underset{f \in \mathcal{H}}{\text{argmin}} \, \|Af - a\|^2_{\mathcal{H}_A} + \|Bf\|^2_{\mathcal{H}_B}$
$f(\cdot)$
$\nabla f(\cdot)$
B
$f(\cdot)$
$\int_L f(s)ds$
A

Figure 3.3: An interpretation of the smoothing spline formulation. Measurements $a \in \mathcal{H}_A$ are known and should be close to Af. The a priori likelihood of $f \in \mathcal{H}$ itself is not known but only the one for $Bf \in \mathcal{H}_B$. According to the explanations given in the main text, the best $f \in \mathcal{H}$ is the one that minimizes the discrepancy $Af - a$ and the energy Bf as measured by the norms in the respective spaces $\mathcal{H}_A$ and $\mathcal{H}_B$. In the above illustration, A is line integration and the energy of f is quantified by measuring the size of its derivative. B is then representable in two equivalent ways as either differentiation directly as on the right or as a canonical projection onto a Hilbert space of equivalence classes of functions that deviate only by a constant with the differentiation operation built into the norm on $\mathcal{H}_B$. This is symbolized on the left half of the image by equating one element of $\mathcal{H}_B$ to an equivalence class of vectors in the base space $\mathcal{H}$.

ii) The abstract smoothing spline is given by the expression [21, p. 51,p. 167]

$$\sigma_f = \sum_{i=1}^{n} \lambda_i l_i K_B + \sum_{j=1}^{m} \mu_j q_j \tag{3.15}$$

where the q_j form a basis of $\ker B$ and K_B is the reproducing kernel of the semi-Hilbert space $\tilde{B}$ of functions in $\mathcal{H}_K$ with inner product $\langle f, g \rangle_{\tilde{B}} = \langle Bf, Bg \rangle_{\mathcal{H}_B}$. The vectors $\lambda = [\lambda_1, ..., \lambda_n]^T, \mu = [\mu_1, ..., \mu_m]^T$ satisfy

$$\begin{bmatrix} C_K + \Sigma_A & Q \\ Q^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} a \\ 0 \end{bmatrix}. \tag{3.16}$$

The $n \times n$ matrix C_K has entries $(C_K)_{ij} = (A \otimes AK_B)_{ij} = l_i \otimes l_j K_B$ and Q is the $n \times m$ matrix with elements $(Q)_{ij} = l_i q_j$.

Theorem 3.1.5 is helpful in practice. If $\Sigma_A = 0$ and $B = I$, it specializes to the Kriging estimator defined by equation 3.12. The resulting spline has the property that it also provides solutions to linearly perturbed problems in the sense that the best estimator for Lf is $L\sigma_f$ if σ_f is the best estimator for f. This means that for a bounded linear operator $L : \mathcal{H}_K \to \mathcal{H}_g$ it is irrelevant if first f is estimated optimally and L applied subsequently or Lf is estimated in a single step. We have

$$\sigma_{Lf} = L\sigma_f \tag{3.17}$$

under reasonable construction of all participant entries. The proof is via straightforward computation comparing the estimation problems in table 3.1.

According to theorem 3.1.5 the solution to the smoothing spline problem

$$\sigma_f = \underset{f \in \mathcal{H}_f}{\operatorname{argmin}} \; \|ALf - a\|^2_{\mathcal{H}_A} + \|f\|^2_{\mathcal{H}_f}$$
$$\text{is } \sigma_f = \sum_{i=1}^{n} \lambda_i l_i LK_f(\cdot,\cdot) = a^T \left[C_g + \Sigma_A\right]^{-1} l_i LK_f(\cdot,\cdot)$$

with the matrix C_g having entries $(C_g)_{ij} = l_i L \otimes l_j LK_f = l_i \otimes l_j K_g$. By the same theorem, the solution to the smoothing spline problem

$$\sigma_g = \underset{g \in \mathcal{H}_g}{\operatorname{argmin}} \; \|Ag - a\|^2_{\mathcal{H}_A} + \|g\|^2_{\mathcal{H}_g}$$
$$\text{is } \sigma_g = \sum_{i=1}^{n} \mu_i l_i K_g(\cdot,\cdot) = a^T \left[C_g + \Sigma_A\right]^{-1} l_i K_g(\cdot,\cdot)$$

with C_g and Σ_A as before. Now, naming $\sigma_g = \sigma_{Lf}$, one has

$$L\sigma_f = a^T[C_g + \Sigma_A]^{-1} l_i L \otimes LK_f = a^T[C_g + \Sigma_A]^{-1} l_i K_g = \sigma_{Lf}.$$

Estimating for example differentials or integrals of any noisily observed function f is then possible simply by deriving the spline estimate σ_f for f and applying differentiation or integration to it. One therefore does not have to leave the framework of smoothing splines even if linearly transformed quantities are to be estimated.

Besides smoothing splines, one often encounters interpolating splines; solutions to the problem

$$\sigma_f = \underset{f \in A^{-1}a}{\operatorname{argmin}} \; \|f\|^2_{\mathcal{H}_K}. \tag{3.18}$$

With the same notation as in theorem 3.1.5 they have the solution $\sigma_f = \sum_{j=1}^{n} \lambda_j l_j K(\cdot,\cdot)$ with $\lambda = C_K^{-1} a$. If the RK K is the sum of two kernels, $K = K_1 + K_2$, and consequently $\mathcal{H}_K = \mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2}$ then σ_f can be written as the sum of the two smoothing splines $\sigma_{f_1}, \sigma_{f_2}$ as follows immediately from the

Table 3.1: Two related estimation problems

Problem	**To estimate**	**Measured**	**Penalized**	**Data**
a)	f	Lf via A	f via $\mathcal{H}_f$	$a = (AL)f$
b)	$g = Lf$	g via A	g via $\mathcal{H}_g$	$a = A(Lf)$
Comment	$\mathcal{H}_f$ has reproducing kernel K_f and $\mathcal{H}_g = \mathcal{H}_{Lf}$ has RK $K_g = L \otimes LK_f$. A is made up of the linearly dependent linear functionals $l_1, ..., l_n : \mathcal{H}_g \to \mathbb{R}$.			

closed form solutions of

$$\sigma_{f_1} = \underset{f_1 \in \mathcal{H}_{K_1}}{\operatorname{argmin}} \ \|Af_1 - a\|^2_{C_{K_2}} + \|f_1\|^2_{\mathcal{H}_{K_1}} \qquad = \sum_{j=1}^{n} \lambda_j l_j K_1(\cdot, \cdot)$$

$$\sigma_{f_2} = \underset{f_2 \in \mathcal{H}_{K_2}}{\operatorname{argmin}} \ \|Af_2 - a\|^2_{C_{K_1}} + \|f_2\|^2_{\mathcal{H}_{K_2}} \qquad = \sum_{j=1}^{n} \lambda_j l_j K_2(\cdot, \cdot).$$

In both formulations what is considered noise $(Af - a)$ and what is considered signal (f) are switched. In the first expression f_1 is signal and f_2 is noise and in the second expression f_1 is noise and f_2 is signal implying that it is more appropriate to call both f_1 and f_2 signals which simply are to be separated. As $\sigma_{f_1} + \sigma_{f_2} = \sigma_f$, the interpolating spline can be said to be a superposition of a best guess for the signal f_1 and a best guess for the signal f_2 — a smoothing spline is then nothing else than the projection of $\sigma_f \in \mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2}$ onto one of the subspaces. If we slightly reinterpret σ_f as consisting of $[\sigma_{f_1}, \sigma_{f_2}]$ in the external direct sum $\mathcal{H}_{K_1} \oplus_e \mathcal{H}_{K_2}$, then one may write $P_{\mathcal{H}_{K_1}} \sigma_f = \sigma_{f_1}$ and $P_{\mathcal{H}_{K_2}} \sigma_f = \sigma_{f_2}$ where $P_{\mathcal{H}_{K_1}}, P_{\mathcal{H}_{K_1}}$ are orthogonal projections onto $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$ respectively.

Abstract splines have applications in statistics and the physical sciences, see e.g. [203]. To illustrate the possibilities, three problems with varying degrees of generality are outlined together with an analysis of their solution in subsection 3.3.1.

- The simple problem of denoising total station measurements by solving

  $$\sigma_d = \underset{d \in \mathcal{H}_D}{\operatorname{argmin}} \ \|Ad - a\|^2_{\mathbb{R}^n} + \|d\|^2_{\mathcal{H}_D}$$

  where A is evaluation, d is deformation, a is data and $\mathcal{H}_D$ is an RKHS of smooth temporal functions.

- The tomography type problem of inferring a spatial distribution of changes in refractive index n_r from total station measurements towards stable prisms by solving

  $$\sigma_{n_r} = \underset{n_r \in \mathcal{H}_{N_r}}{\operatorname{argmin}} \ \|An_r - a\|^2_{\mathbb{R}^n} + \|n_r\|^2_{\mathcal{H}_{N_r}}$$

  where A is line integration, a is data and $\mathcal{H}_{N_r}$ is an RKHS of smooth spatial functions.

- The problem of inferring the deformation of a clamped elastic string under variable load from noisy measurements by solving

  $$\sigma_d = \underset{d \in \mathcal{H}_D}{\operatorname{argmin}} \ \|Ad - a\|^2_{\mathbb{R}^n} + \|Bd\|^2_{\mathcal{H}_B}$$

  where A is evaluation, d is deformation, and a is data. The operator B is the Laplacian and $\mathcal{H}_D, \mathcal{H}_B$ are RKHS of functions related to the underlying differential equations.

3.1.3 Hilbert space constructions

Hilbert spaces seem to provide a model of only a very restricted class of processes because of the initial interpretation of an RKHS as a space of scalar functions on an index set T. This limitation will be shown to be one of intuition only as the constructions carried out abstractly in subsection 2.1.3 are translated into an RKHS setting providing a correspondence between linear algebraic operations on vector spaces and bilinear operations on positive definite kernels. By investigating the direct sum and tensor product of Hilbert spaces, both the combination of RKHS to a new one and the decomposition of a given RKHS into simpler ones become feasible operations. This last statement implies that a decomposable RKHS' elements are interpretable as vector valued; consequently also the scalar abstract splines $\sigma \in \mathcal{H}_K$ may be seen in that way and $\sigma \cong (u, v) \in \mathcal{H}_1 \oplus \mathcal{H}_2$ is then called vector spline. Tensor splines are a computationally convenient way to handle spatiotemporal problems which normally can not be solved in a reasonable amount of time due to the size of the involved kernel matrices and the cost of inverting them. Their implementation is briefly discussed and an example illustrates the effect of the tensor product factorization on prediction accuracy and runtime.

§ **Direct sums**

Let $\mathcal{H}_A, \mathcal{H}_B$ be two reproducing kernel Hilbert spaces of functions on T with inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}_A}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_B}$ and kernels K_A and K_B. Recall that there are two types of algebraic direct sums, the external and internal direct sums as denoted by $\oplus_e$ and $\oplus_i$ respectively [10, p. 92].

$$\mathcal{H}_e = \mathcal{H}_A \oplus_e \mathcal{H}_B := \{(a, b) : a \in \mathcal{H}_A, b \in \mathcal{H}_B\} \tag{3.19}$$

$$\mathcal{H}_i = \mathcal{H}_A \oplus_i \mathcal{H}_B := \{a + b : a \in \mathcal{H}_A, b \in \mathcal{H}_B\} \tag{3.20}$$

with the inner product $\langle f, g \rangle_{\mathcal{H}_e} = \langle a_f, a_g \rangle_{\mathcal{H}_A} + \langle b_f, b_g \rangle_{\mathcal{H}_B}$ for the elements $f = (a_f, b_f), g = (a_g, b_g) \in \mathcal{H}_e$. $\mathcal{H}_e$ is a Hilbert space [21, p. 157] and $\mathcal{H}_A \perp \mathcal{H}_B$ are orthogonal subspaces $\mathcal{H}_A \overline{\boxslash} \mathcal{H}_e, \mathcal{H}_B \overline{\boxslash} \mathcal{H}_e$. If one introduces the map

$$\psi : \mathcal{H}_e \ni (a, b) \mapsto a + b \in \mathcal{H}_i$$

and denotes by N the nullspace $\ker \psi \boxslash \mathcal{H}_e$, then N has the elements

$$N = \{(a, b) \in \mathcal{H}_e : a + b = 0\} \quad = \quad \{(a, -a) : a \in \mathcal{H}_A \cap \mathcal{H}_B\}.$$

By theorem 2.1.11 one finds $N^\perp \cong \mathcal{H}_e / \ker \psi \overline{\boxslash} \mathcal{H}_e$ and the restriction $\overline{\psi} : N^\perp \to \mathcal{H}_i$ is one-to-one allowing the definition of an inner product on $\mathcal{H}_i$ making it a Hilbert space as well [21, p. 24] via

$$\begin{aligned} \langle f, g \rangle_{\mathcal{H}_i} = \langle \overline{\psi}^{-1}(f), \overline{\psi}^{-1}(g) \rangle_{N^\perp} &= \langle (a_f, b_f), (a_g, b_g) \rangle_{N^\perp} \\ &= \langle a_f, a_g \rangle_{\mathcal{H}_A} + \langle b_f, b_g \rangle_{\mathcal{H}_B}. \end{aligned}$$

This establishes $\mathcal{H}_i \cong N^\perp \cong \mathcal{H}_e / \ker \psi$ and we may always interpret the internal direct sum $\mathcal{H}_i : \mathcal{H}_A \oplus_i \mathcal{H}_B$ containing elements $a + b, a \in \mathcal{H}_A, b \in \mathcal{H}_B$ as a subspace of the vector valued external direct sum $\mathcal{H}_e := \mathcal{H}_A \oplus_e \mathcal{H}_B$ containing elements of the form $(a, b), a \in \mathcal{H}_A, b \in \mathcal{H}_B$. As a direct consequence, if $\mathcal{H}_A \perp \mathcal{H}_B$, then $\ker \psi = \{0\}$ and $\mathcal{H}_e \cong \mathcal{H}_i$.

Lemma 3.1.6 *If $X, Y, \mathcal{H}_1, \mathcal{H}_2$ are Hilbert spaces and $\varphi_1 : X \to \mathcal{H}_1, \varphi_2 : Y \to \mathcal{H}_2$ are isometric isomorphisms then $\varphi := \varphi_1 \oplus \varphi_2 : X \oplus_e Y \ni (x, y) \mapsto (\varphi_1(x), \varphi_2(y)) \in \mathcal{H}_1 \oplus_e \mathcal{H}_2$ is also an isometric isomorphism.*

Proof: Denote by X_e and $\mathcal{H}_e$ the external direct sums $X \oplus_e Y$ and $\mathcal{H}_1 \oplus_e \mathcal{H}_2$ respectively. They are Hilbert spaces. As $\forall q \in \mathcal{H}_e, q = (f_1, f_2)$ and φ_1, φ_2 are isometries, $\exists x \in X$ and $y \in Y$ s.t. $\varphi_1(x) = f_1$ and $\varphi_2(y) = f_2$. Then $(\varphi_1(x), \varphi_2(y)) = \varphi\left((x, y)\right) = (f_1, f_2) = q$ and φ is surjective. The map φ is injective because if $u_1, u_2 \in X_e$ and $u_1 \neq u_2$ then $u_1 - u_2 = (\Delta x, \Delta y)$ and Δx or Δy (or both) are $\neq 0$. Consequently $\varphi(u_1 - u_2) = (\varphi_1(\Delta x), \varphi_2(\Delta y)) \neq 0$ since the components φ_1, φ_2 are injective and φ is therefore injective. For any $u_1, u_2 \in X_e$

$$\begin{aligned}
\langle u_1, u_2\rangle_{X_e} &= \langle \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \rangle_{X_e} \\
&= \langle x_1, x_2\rangle_X + \langle y_1, y_2\rangle_Y \\
\langle \varphi(u_1), \varphi(u_2)\rangle_{\mathcal{H}_e} &= \langle \begin{bmatrix} \varphi_1(x_1) \\ \varphi_2(y_1) \end{bmatrix}, \begin{bmatrix} \varphi_1(x_1) \\ \varphi_2(y_2) \end{bmatrix} \rangle_{\mathcal{H}_e} \\
&= \langle \varphi_1(x_1), \varphi_1(x_2)\rangle_{\mathcal{H}_1} + \langle \varphi_2(y_1), \varphi_2(y_2)\rangle_{\mathcal{H}_2} \\
&= \langle x_1, x_2\rangle_X + \langle y_1, y_2\rangle_Y
\end{aligned}$$

and both of the inner products are equal because the individual components φ_1, φ_2 of the map φ are isometries. This establishes $X_e \cong \mathcal{H}_e$ via the isometric isomorphism $\varphi = \varphi_1 \oplus \varphi_2$. □

An analogous theorem does not hold for internal direct sums. If $\varphi_1 : X \to \mathcal{H}_1, \varphi_2 : Y \to \mathcal{H}_2$ are isometries, $X \oplus_i Y$ and $\mathcal{H}_1 \oplus_i \mathcal{H}_2$ are not necessarily isomorphic via $\overline{\varphi} = \psi_{\mathcal{H}}\varphi_1 \oplus \varphi_2\overline{\psi}_X^{-1}$ if $X \perp Y$ but $\mathcal{H}_1 \cap \mathcal{H}_2 \neq \{0\}$ and consequently $\mathcal{H}_1 \not\perp \mathcal{H}_2$.

This can be seen as follows: Let $X_i := X \oplus_i Y, X_e := X \oplus_e Y, \mathcal{H}_i := \mathcal{H}_1 \oplus_i \mathcal{H}_2$ and let $\mathcal{H}_e$ be the external direct sum $\mathcal{H}_1 \oplus_e \mathcal{H}_2$. Take a nontrivial element $0 \neq f \in \mathcal{H}_1 \cap \mathcal{H}_2$. Construct the two elements $u_1 = \varphi_1^{-1}(f) \in X$ and $u_2 = \varphi_2^{-1}(f) \in Y$. Then $\langle u_1, u_2\rangle_{X_i} = \langle \overline{\psi}_X^{-1}(u_1), \overline{\psi}_X^{-1}(u_2)\rangle_{X_e} = \langle (u_1, 0), (0, u_2)\rangle_{X_e} = 0$ since $X \perp Y$ and $\overline{\psi}_X^{-1}$ is an isometry between $\ker \psi_X^{\perp} = X_e$ and X_i. On the other hand we have

$$\begin{tikzcd}
X_i \arrow[d, "\overline{\psi}_X^{-1}"'] & \mathcal{H}_i \\
X_e \arrow[r, "\varphi_1 \oplus \varphi_2"] & \mathcal{H}_e \arrow[u, "\psi_{\mathcal{H}}"']
\end{tikzcd}$$

$$\langle \overline{\varphi}(u_1), \overline{\varphi}(u_2)\rangle_{\mathcal{H}_i} = \langle (\psi_{\mathcal{H}}[(\varphi_1(u_1), 0)], \psi_H[(0, \varphi_2(u_2)]\rangle_{\mathcal{H}_i} = \langle f, f\rangle_{\mathcal{H}_i} \neq 0$$

since $f \neq 0$ and $\langle \cdot, \cdot\rangle_{\mathcal{H}_i}$ is an inner product. This establishes the existence of elements u_1, u_2 s.t. $\langle u_1, u_2\rangle_{X_i} \neq \langle \overline{\varphi}(u_1), \overline{\varphi}(u_2)\rangle_{\mathcal{H}_i}$. □

Using lemma 3.1.6 and the previous comments, it is straightforward to find the relationships between internal and external direct sums of spaces generated by stochastic processes and their corresponding reproducing kernel Hilbert spaces.

Theorem 3.1.7 *Let $X = \{X_t\}_{t \in T}$ and $Y = \{Y_t\}_{t \in T}$ be two independent mean-zero second order stochastic processes on T and denote by Z the composite stochastic process with $Z_t = X_t + Y_t \;\; \forall t \in T$. Then*

i) $\overline{\mathcal{L}}(Z) \cong \mathcal{H}_{K_Z} \cong \mathcal{H}_{K_X} \oplus_i \mathcal{H}_{K_Y}$

ii) $\overline{\mathcal{L}}(X) \oplus_i \overline{\mathcal{L}}(Y) \cong \overline{\mathcal{L}}(X) \oplus_e \overline{\mathcal{L}}(Y) \cong \mathcal{H}_{K_X} \oplus_e \mathcal{H}_{K_Y}$

where $\overline{\mathcal{L}}(X), \overline{\mathcal{L}}(Y)$ and $\overline{\mathcal{L}}(Z)$ are the Hilbert spaces generated by the respective stochastic processes and $\mathcal{H}_{K_Z}$ is the RKHS with kernel $K_Z(s,t) = K_X(s,t) + K_Y(s,t) \;\; \forall s,t \in T$.

Proof: i) For two random variables Z_s, Z_t using the assumption that $X_s \coprod Y_t \;\; \forall s,t \in T$, it holds that $E[Z_s Z_t] = E[(X_s + Y_s)(X_t + Y_t)] = E[X_s X_t] + E[Y_s Y_t] = K_X(s,t) + K_Y(s,t) = K_Z(s,t)$. By Loewes theorem $\overline{\mathcal{L}}(Z) = \mathcal{H}_{K_z}$. Since $K_Z = K_X + K_Y$, theorem 2.3.6 pertaining to sums of kernels forming new kernels establishes that

$$\mathcal{H}_{K_Z} = \{f_1 + f_2 : f_1 \in \mathcal{H}_{K_X}, f_2 \in \mathcal{H}_{K_Y}\}$$
$$\|f\|_{\mathcal{H}_{K_Z}} = \min_{f = f_1 + f_2, f_1 \in \mathcal{H}_{K_X}, f_2 \in \mathcal{H}_{K_Y}} \sqrt{\|f_1\|^2_{\mathcal{H}_{K_X}} + \|f_2\|^2_{\mathcal{H}_{KY}}}$$

is just the internal direct sum $\mathcal{H}_{K_X} \oplus_i \mathcal{H}_{K_Y}$. This is due to the fact that $\|f\|_{\mathcal{H}_{K_Z}}$ is attained for $(f_1, f_2) \in \ker \psi_{\mathcal{H}}^{\perp}$, i.e. $\|f\|_{\mathcal{H}_{K_Z}} = \|\overline{\psi}_{\mathcal{H}}^{-1} f\|_{\mathcal{H}_{K_X} \oplus_e \mathcal{H}_{K_Y}}$ where $\psi_{\mathcal{H}} : \mathcal{H}_{K_X} \oplus_e \mathcal{H}_{K_Y} \ni (f,g) \mapsto f + g \in \mathcal{H}_{K_X} \oplus_i \mathcal{H}_{K_Y}$ and $\overline{\psi}_{\mathcal{H}}^{-1}$ is the restriction onto $\ker \psi_{\mathcal{H}}^{\perp}$, see [20, p. 25].

ii) Suppose that $\forall s,t \in T \; X_s \coprod Y_t$. Then $E[X_s Y_t] = 0$ and $\overline{\mathcal{L}}(X) \cap \overline{\mathcal{L}}(Y) = \{0\}$ as for subsets of $L^2(\Omega)$ one has $\forall X \in \overline{\mathcal{L}}(X), Y \in \overline{\mathcal{L}}(Y)$

$$\|X - Y\|_{L^2(\Omega)} = \langle X - Y, X - Y \rangle_{L^2(\Omega)} = \|X\|^2_{L^2(\Omega)} + \|Y\|^2_{L^2(\Omega)}.$$

Since $X = Y \Leftrightarrow \|X - Y\|_{L^2(\Omega)} = 0$, this implies $X = Y = 0$. Then clearly $\overline{\mathcal{L}}(X) \oplus_e \overline{\mathcal{L}}(Y) \cong \overline{\mathcal{L}}(X) \oplus_i \overline{\mathcal{L}}(Y)$ since $N = \ker \psi_{\overline{\mathcal{L}}} = \{0\}$ for

$$\psi_{\overline{\mathcal{L}}} : \overline{\mathcal{L}}(X) \oplus_e \overline{\mathcal{L}}(Y) \ni (X,Y) \mapsto X + Y \in \overline{\mathcal{L}}(X) \oplus_i \overline{\mathcal{L}}(Y)$$

and $N^{\perp}$ was isometrically isomorphic to $\overline{\mathcal{L}}(X) \oplus_i \overline{\mathcal{L}}(Y)$ via $\psi_{\overline{\mathcal{L}}}$. By lemma 3.1.6 $\overline{\mathcal{L}}(X)$ being isometrically isomorphic to $\mathcal{H}_{K_X}$ and $\overline{\mathcal{L}}(Y)$ being isometrically isomorphic to $\mathcal{H}_{K_Y}$ implies $\overline{\mathcal{L}}(X) \oplus_e \overline{\mathcal{L}}(Y) \cong \mathcal{H}_{K_X} \oplus_e \mathcal{H}_{K_Y}$. □

The theorem implies that a superposition $\{Z_t\}_{t \in T} = \{X_t + Y_t\}_{t \in T}$ of independent stochastic processes $\{X_t\}_{t \in T}$ and $\{Y_t\}_{t \in T}$ has the RKHS representation $\mathcal{H}_K$ with $K = K_X + K_Y$ the reproducing kernel of $\mathcal{H}_K \cong \mathcal{H}_{K_X} \oplus_i \mathcal{H}_{K_Y}$. One can therefore perform optimal estimation of Z_t by manipulation of $\mathcal{H}_K$ and project elements f

of $\mathcal{H}_K$ onto the component spaces $\mathcal{H}_{K_X}$ and $\mathcal{H}_{K_Y}$ to perform estimation of X_t, Y_t given measurements of Z_t only. The two expressions

$$\begin{aligned} \overline{\mathcal{L}}(Z) &= \overline{\text{span}}\,\{X_t + Y_t : t \in T\} \\ \overline{\mathcal{L}}(X) \oplus_e \overline{\mathcal{L}}(Y) &= \{(X,Y) : X \in \overline{\mathcal{L}}(X), Y \in \overline{\mathcal{L}}(Y)\} \end{aligned}$$

are different in the sense that the bottom one includes the top one and may be interpreted as vector valued opening up the possibility to extend the estimation framework from scalar functions toward vector valued ones.

Example 15 Let $X \sim \mathcal{N}(0, \sigma_X^2), Y \sim \mathcal{N}(0, \sigma_Y^2)$ be independent second order random variables and define $Z = X + Y$. They can be identified with stochastic processes on a one-element set T. The corresponding RKHS are

$$\begin{aligned} \mathcal{H}_{K_X} &= \mathbb{R} \quad K_X = \sigma_X^2 \quad \langle f, g\rangle_{\mathcal{H}_{K_X}} = \frac{fg}{\sigma_X^2} \\ \mathcal{H}_{K_Y} &= \mathbb{R} \quad K_Y = \sigma_Y^2 \quad \langle f, g\rangle_{\mathcal{H}_{K_Y}} = \frac{fg}{\sigma_Y^2}. \end{aligned}$$

The external direct sum $\mathcal{H}_e = \mathcal{H}_{K_X} \oplus \mathcal{H}_{K_Y}$ is given as $\mathbb{R}^2$ and for $f = (f_1, f_2) \in \mathcal{H}_e, g = (g_1, g_2) \in \mathcal{H}_e$ the inner product is

$$\langle f, g\rangle_{\mathcal{H}_e} = \langle f_1, g_1\rangle_{\mathcal{H}_{K_X}} + \langle f_2, g_2\rangle_{\mathcal{H}_{K_Y}} = \frac{fg}{\sigma_X^2} + \frac{fg}{\sigma_Y^2}.$$

$\mathcal{H}_e$ is the space of two dimensional vectors that are outcomes of sampling from $(X,Y) \in \overline{\mathcal{L}}(X) \oplus_e \overline{\mathcal{L}}(Y)$ and the negative log likelihood of $f \in \mathcal{H}_e$ is (bar some constants) given as $\|f\|^2_{\mathcal{H}_e}$.
The map $\psi_{\mathcal{H}}$ maps $f = (f_1, f_2) \in \mathcal{H}_e$ to $f_1 + f_2$. Its kernel is given as $N = \ker \psi_{\mathcal{H}} = \{(a, -a) : a \in \mathbb{R}\} \overline{\boxslash}\mathcal{H}_e$. The orthogonal complement is

$$\begin{aligned} N^\perp &= \{(f,g) \in \mathcal{H}_e : \langle f, a\rangle_{\mathcal{H}_{K_X}} - \langle g, a\rangle_{\mathcal{H}_{K_Y}} = 0 \;\; \forall a \in \mathbb{R}\} \\ &= \{(f,g) \in \mathbb{R}^2 : \frac{fa}{\sigma_X^2} - \frac{ga}{\sigma_Y^2} = 0 \;\; \forall a \in \mathbb{R}\} \\ &= \{(q, q\sigma_Y^2\sigma_X^{-2}) : q \in \mathbb{R}\}. \end{aligned}$$

Since $\mathcal{H}_i = \mathcal{H}_{K_X} \oplus_i \mathcal{H}_{K_Y}$ is defined as $\{f + g : f \in \mathcal{H}_{K_X}, g \in \mathcal{H}_{K_Y}\}$, clearly $\mathcal{H}_i \cong \mathbb{R} \cong N^\perp$ with the inner product

$$\begin{aligned} \langle f, g\rangle_{\mathcal{H}_i} &= \langle \overline{\psi}_{\mathcal{H}}^{-1} f, \overline{\psi}_{\mathcal{H}}^{-1} g\rangle_{\mathcal{H}_e} \\ &= \langle \begin{bmatrix} f\frac{1}{1+c} \\ f\frac{c}{1+c} \end{bmatrix}, \begin{bmatrix} g\frac{1}{1+c} \\ g\frac{c}{1+c} \end{bmatrix} \rangle_{\mathcal{H}_e} \qquad c = \frac{\sigma_Y^2}{\sigma_X^2} \\ &= \left(\frac{1}{1+c}\right)^2 \langle f, g\rangle_{\mathcal{H}_{K_X}} + \left(\frac{c}{1+c}\right) \langle f, g\rangle_{\mathcal{H}_{K_X}} \end{aligned}$$

where $\overline{\psi}_{\mathcal{H}}^{-1}$ is the inverse of $\overline{\psi}_{\mathcal{H}} : N^{\perp} \ni (f_1, f_2) \mapsto f_1 + f_2 \in \mathcal{H}_i$ and given by

$$\overline{\psi}_{\mathcal{H}}^{-1}(f) = \left[f\left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_Y^2}\right), f\left(\frac{\sigma_Y^2}{\sigma_X^2 + \sigma_Y^2}\right)\right] \quad \forall f \in \mathcal{H}_i \cong \mathbb{R}.$$

Simplification yields

$$\langle f, g\rangle_{\mathcal{H}_i} = fg\left[\frac{\sigma_X^2}{(\sigma_X^2 + \sigma_Y^2)^2} + \frac{\sigma_Y^2}{(\sigma_X^2 + \sigma_Y^2)^2}\right] = \frac{fg}{\sigma_X^2 + \sigma_Y^2} = \langle f, g\rangle_{\mathcal{H}_{K_Z}}$$

where $K_Z = E[ZZ] = E[(X+Y)(X+Y)] = \sigma_X^2 + \sigma_Y^2 = K_X + K_Y$. Therefore $\mathcal{H}_i \cong \mathcal{H}_{K_Z}$ and by Loewe's theorem also $\mathcal{H}_i \cong \overline{\mathcal{L}}(Z)$. If for example $\sigma_X^2 = 1$ and $\sigma_Y^2 = 2$ then

i) $\|z\|^2_{\mathcal{H}_i} = z(\sigma_X^2 + \sigma_Y^2)^{-1}$ is proportional to the negative log probability that $Z = z$.

ii) $(x, y) = (z\sigma_X^2(\sigma_X^2 + \sigma_Y^2)^{-1}, z\sigma_Y^2(\sigma_X^2 + \sigma_Y^2)^{-1}) = \overline{\psi}_{\mathcal{H}}^{-1}(z)$ are the best guesses for x and y given a specific value for $Z = z$, i,e, the ones satisfying $\|x\|^2_{\mathcal{H}_{K_X}} + \|y\|^2_{\mathcal{H}_{K_Y}} \to \min$ among all x, y s.t. $x + y = z$.

These relations are illustrated in figure 3.4 ■

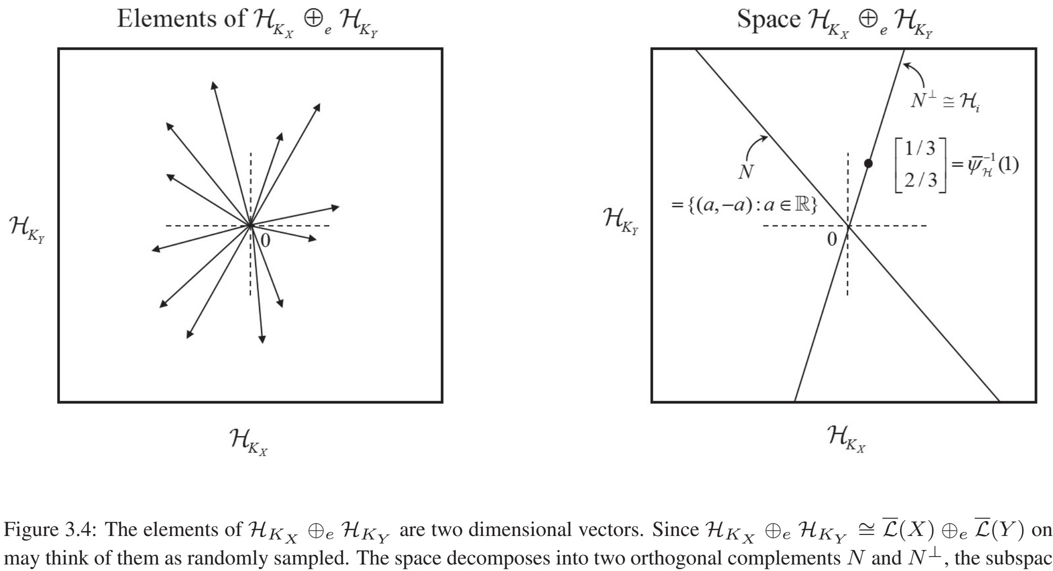

Elements of $\mathcal{H}_{K_X} \oplus_e \mathcal{H}_{K_Y}$
$\mathcal{H}_{K_Y}$
0
$\mathcal{H}_{K_X}$
Space $\mathcal{H}_{K_X} \oplus_e \mathcal{H}_{K_Y}$
$N^{\perp} \cong \mathcal{H}_i$
$\begin{bmatrix}1/3\\2/3\end{bmatrix} = \overline{\psi}_{\mathcal{H}}^{-1}(1)$
N
$= \{(a, -a) : a \in \mathbb{R}\}$
$\mathcal{H}_{K_Y}$
0
$\mathcal{H}_{K_X}$

Figure 3.4: The elements of $\mathcal{H}_{K_X} \oplus_e \mathcal{H}_{K_Y}$ are two dimensional vectors. Since $\mathcal{H}_{K_X} \oplus_e \mathcal{H}_{K_Y} \cong \overline{\mathcal{L}}(X) \oplus_e \overline{\mathcal{L}}(Y)$ one may think of them as randomly sampled. The space decomposes into two orthogonal complements N and $N^{\perp}$, the subspace $N^{\perp} \boxslash \mathcal{H}_e$ is isometrically isomorphic to $\mathcal{H}_i$ and its members are best guesses for X and Y given Z.

The formal definition of internal and external direct sums allows to clarify the relations between interpolating and smoothing splines in a lemma.

Lemma 3.1.8 *Suppose σ_f and $\sigma_{f_1}, \sigma_{f_2}$ are solutions to the interpolating and smoothing spline problems*

$$\sigma_f = \underset{f \in A^{-1}a}{\operatorname{argmin}} \ \|f\|^2_{\mathcal{H}_1 \oplus_i \mathcal{H}_2} \qquad \sigma_{f_j} = \underset{f_j \in \mathcal{H}_j}{\operatorname{argmin}} \ \|Af_j - a\|^2_{A\mathcal{H}_{K_{\bar{j}}}} + \|f_j\|^2_{\mathcal{H}_j}$$

for $j = 1, 2, \bar{j} = 3 - j$ where $a \in \mathbb{R}^n$ is a data vector, the measurement operator $A : \mathcal{H}_1 \oplus_i \mathcal{H}_2 \to \mathbb{R}^n$ is composed of n linearly independent linear functionals $l_1, ..., l_n$ and $\mathcal{H}_1, \mathcal{H}_2$ are RKHS with RK K_1, K_2. Then σ_{f_j} is the orthogonal projection $P_{\mathcal{H}_{K_j}}\sigma_f$ of σ_f if one considers σ_f as injected into the external direct sum $\mathcal{H}_1 \oplus_e \mathcal{H}_2$.

Proof: Denote, as before, by $\psi : \mathcal{H}_1 \oplus_e \mathcal{H}_2 \to \mathcal{H}_1 \oplus_i \mathcal{H}_2$ the addition mapping (f_1, f_2) to $f_1 + f_2$ and use the abbreviations $\mathcal{H}_e, \mathcal{H}_i$ for the external and internal direct sums. Define $q = (\sigma_{f_1}, \sigma_{f_2}) \in \mathcal{H}_e$. If $\overline{\psi}$ is the restriction of ψ onto $\ker\psi^{\perp}$, then $\overline{\psi}^{-1}$ exists. The solutions for σ_f, σ_{f_1} and σ_{f_2} are given by the terms

$$\sigma_f = \sum_{j=1}^{n} \lambda_j l_j (K_1 + K_2) \quad \sigma_{f_1} = \sum_{j=1}^{n} \lambda_j l_j K_1 \quad \sigma_{f_2} = \sum_{j=1}^{n} \lambda_j l_j K_2$$

where the vector λ of coefficients is equal for all expressions, see equation 3.15. Clearly $\psi q = \sigma_f$. Since $\overline{\psi}^{-1}(\sigma_f)$ is the unique element of $\ker\psi^{\perp}\ \overline{\boxslash}\mathcal{H}_e$ such that $\psi \circ \overline{\psi}^{-1}\sigma_f = \sigma_f$ it remains to show that $q \in \ker\psi^{\perp}$ to prove that $q = (\sigma_{f_1}, \sigma_{f_2}) = \overline{\psi}^{-1}\sigma_f$. For this, notice that $\ker\psi = \{n_e = (n, -n) : n \in \mathcal{H}_1 \cap \mathcal{H}_2\}$ is closed and

$$\langle q, n_e \rangle_{\mathcal{H}_e} = \langle \sigma_{f_1}, n \rangle_{\mathcal{H}_1} - \langle \sigma_{f_2}, n \rangle_{\mathcal{H}_2} = 0.$$

This follows from the fact that $n \in \mathcal{H}_1 \cap \mathcal{H}_2$ and K_1, K_2 are both reproducing for n under the inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}_1}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_2}$ as then

$$\begin{aligned}\langle (l_j K_1(\cdot, \cdot), n \rangle_{\mathcal{H}_1} = \langle \sum_{i=1}^{\infty} \lambda_i^M (l_j \varphi_i(\cdot)) \varphi_i(\cdot), n(\cdot) \rangle_{\mathcal{H}_1} &= l_j \sum_{i=1}^{\infty} \lambda_i^M \varphi_i \langle \varphi_i(\cdot), n(\cdot) \rangle_{\mathcal{H}_1} \\ &= l_j \langle K_1, n \rangle_{\mathcal{H}_1} \\ &= l_j n\end{aligned}$$

with $K_1 = \sum_{i=1}^{\infty} \lambda_i^M \varphi_i(\cdot)\varphi_i(\cdot)$ the Mercer decomposition of the kernel K_1. The same steps lead to $\langle l_j K_2(\cdot, \cdot), n(\cdot) \rangle_{\mathcal{H}_2} = l_j n$. Then

$$\begin{aligned}\langle \sigma_{f_1}, n \rangle_{\mathcal{H}_1} - \langle \sigma_{f_2}, n \rangle_{\mathcal{H}_2} &= \sum_{j=1}^{n} \lambda_j \langle l_j K_1, n \rangle_{\mathcal{H}_1} - \sum_{i=1}^{n} \lambda_i \langle l_i K_2, n \rangle_{\mathcal{H}_2} \\ &= \sum_{j=1}^{n} \lambda_j (l_j n - l_j n) \\ &= 0\end{aligned}$$

and $q = (\sigma_{f_1}, \sigma_{f_2}) \in \mathcal{H}_e$ is the unique element $\overline{\psi}^{-1}\sigma_f$. It is in this sense that a smoothing spline is built from orthogonal projections of an interpolating spline as

$$\sigma_{f_1} = \overline{\psi} P_{\mathcal{H}_1} \overline{\psi}^{-1} \sigma_f \qquad (3.21)$$

$$\sigma_{f_2} = \overline{\psi} P_{\mathcal{H}_2} \overline{\psi}^{-1} \sigma_f \tag{3.22}$$

where $P_{\mathcal{H}_1}, P_{\mathcal{H}_2}$ are the orthogonal projections in $\mathcal{H}_1 \oplus_e \mathcal{H}_2$. □

Remark Often we will not specifically mention the map $\overline{\psi}$ translating between $\mathcal{H}_1 \oplus_i \mathcal{H}_2$ and $\ker \psi^{\perp}\ \overline{\boxslash}\mathcal{H}_1 \oplus_r \mathcal{H}_2$ and just call $\sigma_{f_1}, \sigma_{f_2}$ orthogonal projections of σ_f onto $\mathcal{H}_1$ and $\mathcal{H}_2$. This is to be understood formally in the sense of the above equations.

The notion of a quotient space was introduced already in subsection 2.1.3. Quotient spaces of RKHS can be made into RKHS again [21, p. 30]. If one defines a semi-inner product $\langle \cdot, \cdot \rangle$ for elements of $\mathcal{H}_K$, e.g. via

$$\langle f, g \rangle = \langle Bf, Bg \rangle_{\mathcal{H}_K} \qquad B \in \mathcal{B}(\mathcal{H}_K)$$

the resultant seminorm $|\cdot|$ is zero for elements of $\ker B =: P$ and the semi-inner product $\langle \cdot, \cdot \rangle$ is denoted by $\langle \cdot, \cdot \rangle_P$ to reflect this. If for any functional $l \in \mathcal{H}_K^*$, $lK_p \in \mathcal{H}_K$ and for any $l \in \mathcal{H}_K^*$ s.t. $l(f) = 0 \ \ \forall f \in P$ the restricting reproducing property $l(f) = \langle lK_P, f \rangle_P$ holds, then K_P is called a semi-reproducing kernel [21, p. 30]. If $B = I$, then K_P is simply the reproducing kernel K of $\mathcal{H}_K$; otherwise K_P might be different. Since the case $B = I$ will be most common, semi-reproducing kernels will only be used sporadically and a detailed discussion is avoided here apart from an instructive example in which the energy operator B enables unpenalized estimation of a constant mean in an adjustment-like way. More can be found in [21, pp. 26-31] and [20, pp. 40-42].

Example 16 Suppose $\mathcal{H}_K$ is an RKHS of functions on $T = [0, 1]$ with RK K such that $\mathcal{H}_0$, the Hilbert space of finite constants, is a subspace of $\mathcal{H}_K$ and split $\mathcal{H}_K = \mathcal{H}_0 \oplus \mathcal{H}_1$ where the sum is orthogonal and $\mathcal{H}_1 = \mathcal{H}_K / \mathcal{H}_0$. Denote by $P_{\mathcal{H}_1}$ the orthogonal projection onto $\mathcal{H}_1$. From this, build the seminorm $|\cdot|_{\mathcal{H}_0}$ with nullspace $\mathcal{H}_0$ and find

$$\|P_{\mathcal{H}_1} f\|^2_{\mathcal{H}_K} =: |f|^2_{\mathcal{H}_0} = \|[f]\|^2_{\mathcal{H}_k/\mathcal{H}_0}.$$

The semi-inner product is $\langle f, g \rangle_{\mathcal{H}_0} = \langle P_{\mathcal{H}_1} f, P_{\mathcal{H}_1} g \rangle_{\mathcal{H}_K}$. The reproducing kernel for the restricted set $f \in \mathcal{H}_1$ with semi-inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is then still K as $\forall f \in \mathcal{H}_1$

$$\begin{aligned} f(s) = \langle f(\cdot), K(s, \cdot) \rangle_{\mathcal{H}_K} = \langle P^*_{\mathcal{H}_1} P_{\mathcal{H}_1} f(\cdot), K(s, \cdot) \rangle_{\mathcal{H}_K} &= \langle P_{\mathcal{H}_1} f(\cdot), P_{\mathcal{H}_1} K(s, \cdot) \rangle_{\mathcal{H}_K} \\ &= \langle f(\cdot), K(s, \cdot) \rangle_{\mathcal{H}_0}. \end{aligned}$$

The solution to the interpolation problem with energy operator $P_{\mathcal{H}_1}$ and interpolating conditions $f(t_i) = a_i, i = 1, ..., n$ is then simply

$$\sigma_f = \sum_{i=1}^{n} \lambda_i K(t_i, \cdot)$$

$$\begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} C_K & F \\ F^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} a \\ 0 \end{bmatrix}$$

with the matrix $(C_K)_{ij} = K(t_i, t_j), a \in \mathbb{R}^n$ the data and $F \in \mathbb{R}^n$ a vector of ones. If for a certain $t_0 \in T$, the vector $[K(t_1, t_0), ..., K(t_n, t_0)]^T \in \mathbb{R}^n$ is denoted by c, the explicit solution is

$$\begin{aligned} \lambda &= C_K^{-1}\left(a - F(F^T C_K^{-1} F)^{-1} F^T C_K^{-1} a\right) \\ \sigma_f(t_0) &= c^T C_K^{-1}\left(a - F(F^T C_K^{-1} F)^{-1} F^T C_K^{-1} a\right). \end{aligned} \tag{3.23}$$

But this is just the ordinary Kriging solution [43] consisting of estimating a constant mean with a best linear unbiased estimator (BLUE) and performing simple Kriging on the residuals. The statement generalizes and introducing an energy operator $P_{\mathcal{H}_1}$ with nullspace $\mathcal{H}_0$ of functions $g_1, ..., g_m$ allows $g_1, ..., g_m$ to be chosen without any penalization leading to the BLUE for the coefficients of $g_1, ..., g_m$ [20, p. 89]. ■

§ Vector splines

To extend the theory valid for the scalar case (spline $\sigma : T \to \mathbb{R}$) to apply also to vector valued functions, only a few modifications are necessary. They require primarily internal and external direct sums as well as quotient space constructions and in analogy to the real-valued case allow the derivation of a formula suitable for optimal estimation of a spline function $\sigma : T \to \mathbb{R}^d$ based on measurements of $A\sigma$ where optimality is quantified by the energy operator via $\|B\sigma\|_{\mathcal{H}_B}$. The class of functions $\sigma : T \to \mathbb{R}^d$ includes trajectories and vector fields. The theory is due to Bezhaev and Vasilenko, whose chapter on vector splines [21, pp. 157-174] is condensed in what follows to only cover the basic principles of construction and solution. Recall the definition of abstract smoothing and interpolating splines and denote by $\oplus$ the external direct sum.

If $\{\mathcal{H}_{K_j}\}_{j=1}^{n_s}$ is a sequence of reproducing kernel Hilbert spaces with RK K_j, then $\mathcal{H} = \bigoplus_{j=1}^{n_s} \mathcal{H}_{K_j}$ is a Hilbert space and each $f \in \mathcal{H}$ may be written as $f = (f_1, ..., f_n)$ for $f_j \in \mathcal{H}_{K_j}, j = 1, ..., n_s$ The associated inner product is

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{j=1}^{n_s} \langle f_j, g_j \rangle_{\mathcal{H}_{K_j}}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}_{K_j}}$ denotes the inner product in $\mathcal{H}_{K_j}$. For $\mathcal{H}_{K_j}, j = 1, ..., n_s$ introduce now the semi inner products $\langle, \cdot, \cdot \rangle_{P_j}$ with $P_j \overline{\boxslash} \mathcal{H}_{K_j}$. For the semi-inner product, positive definiteness does not hold and $\langle f, f \rangle_{P_j} = 0$ does not imply $f = 0$ but merely $f \in P_j$. Therefore if $f, g \in \mathcal{H}_{K_j}$ then $f = f_0 + f_0^{\perp}$ where $f_0 \in P_j^{\perp}$ and $f_0^{\perp} \in P_j$ with the same decomposition holding for g and

$$\langle f, g \rangle_{P_j} = \langle f_0 + f_0^{\perp}, g_0 + g_0^{\perp} \rangle_{P_j} = \langle f_0, g_0 \rangle_{P_j} = \langle f_0, g_0 \rangle_{\mathcal{H}_{K_j}}.$$

An equivalent way of writing this is via quotient spaces. Since $P_j \overline{\boxslash} \mathcal{H}_{K_j}, \mathcal{H}_{K_j}/P_j$

is a Hilbert space of equivalence classes. For all $f, g \in \mathcal{H}_{K_j}$,

$$\langle [f], [g] \rangle_{\mathcal{H}_{K_j}/P_j} = \langle f_0, g_0 \rangle_{\mathcal{H}_{K_j}} \qquad \|[f]\|^2_{\mathcal{H}_{K_j}/P_j} = \inf_{g \in P_j} \|f + g\|^2_{\mathcal{H}_{K_j}}$$

are the corresponding inner products and norms on the quotient spaces $\mathcal{H}_{K_j}/P_j$. The maps $[\cdot] : \mathcal{H}_{K_j} \to \mathcal{H}_{K_j}/P_j$ are the canonical projections. If one denotes by P the sum $P_1 \oplus ... \oplus P_{n_s} = \bigoplus_{j=1}^{n_s} P_j$ and defines for $\mathcal{H} = \bigoplus_{j=1}^{n_s} \mathcal{H}_{K_j}$ the semi-inner product

$$\langle f, g \rangle_P = \sum_{j=1}^{n_s} \langle f_j, g_j \rangle_{P_j} \tag{3.24}$$

then $\mathcal{H}$ is a semi-Hilbert space w.r.t. $\langle \cdot, \cdot \rangle_P$ and $\mathcal{H}/P \cong \bigoplus_{j=1}^{n_s} \mathcal{H}_{K_j}/P_j$ is the corresponding quotient space with inner product $\langle [f], [g] \rangle_{\mathcal{H}/P} = \sum_{j=1}^{n_s} \langle [f_j], [g_j] \rangle_{\mathcal{H}_{K_j}/P_j}$. Within this framework, one can solve the estimation of vector valued quantities.

Theorem 3.1.9 *The solution to the interpolating vector spline equation*

$$\sigma_f = \operatorname*{argmin}_{f \in \mathcal{H}, f \in A^{-1}a} \|[f]\|^2_{\mathcal{H}/P} \tag{3.25}$$

for $A : \mathcal{H} \ni f \mapsto [l_1(f), ..., l_n(f)]^T \in \mathbb{R}^n$ and P finite dimensional exists and is unique. It has the form $\sigma_f = (\sigma_{f_1}, ..., \sigma_{f_n})$ with

$$\sigma_{f_i}(\cdot) = \sum_{j=1}^{n} \lambda_j l_{ji} K_i(\cdot, \cdot) + \sum_{j=1}^{\dim P_i} \mu_{ji} q_{ji}(\cdot) \tag{3.26}$$

where q_{ji} is the j-th element of a basis for P_i, l_{ji} is the linear functional defined by $l_j(f) = \sum_{i=1}^{n} l_{ji}(f_i)$ and $\lambda, \mu_1, ..., \mu_{n_s}$ satisfy

$$\begin{bmatrix} \Sigma_1 + ... + \Sigma_{n_s} & Q_1 & \cdots & Q_{n_s} \\ Q_1^T & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ Q_{n_s}^T & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu_1 \\ \vdots \\ \mu_{n_s} \end{bmatrix} = \begin{bmatrix} a \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{3.27}$$

The matrices Q_m contain the elements $(Q_m)_{ij} = l_{im} q_{jm}$ and Σ_m is the matrix with entries $(\Sigma_m)_{ij} = l_{im} l_{jm} K_m^{P_m}$ where the $K_m^{P_m}$ are the semi-reproducing kernels of the spaces $\mathcal{H}_{K_m}/P_m$.

The original statement and a proof can be found in [21, p. 160]. If each $\mathcal{H}_{K_j}$ is itself composite, i.e $\mathcal{H}_{K_j} = \mathcal{H}_{X_j} \oplus \mathcal{H}_{Y_j}$, smoothing splines can be found by projecting the interpolating spline orthogonally onto the spaces $\mathcal{H}_{X_j}$ or $\mathcal{H}_{Y_j}$ in accordance to lemma 3.1.8. Some exemplary results are plotted in figure 3.5.

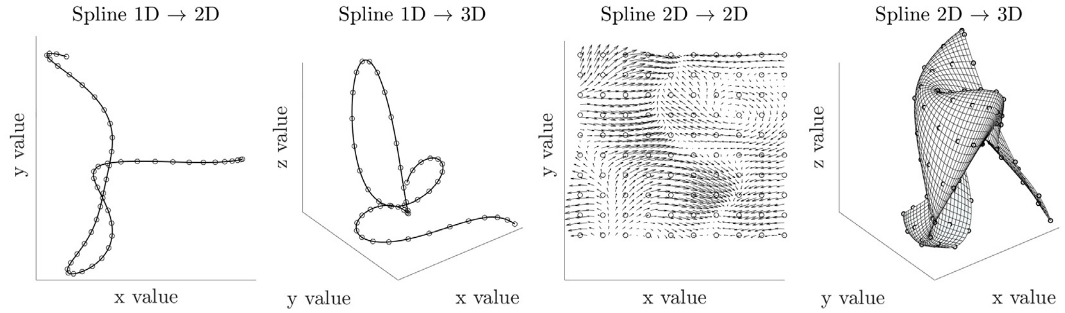


Figure 3.5: Exemplary interpolating splines in two and three dimensions. The circles indicate measurements which were interpolated using the formulas in theorem 3.1.9 and a squared exponential kernel. The interpolating spline was evaluated on a regular grid and has the form $\sigma : T \to F$, where from left to right $(T \subset \mathbb{R}, F \subset \mathbb{R}^2)$, $(T \subset \mathbb{R}, F \subset \mathbb{R}^3)$, $(T \subset \mathbb{R}^2, F \subset \mathbb{R}^2)$ and $(T \subset \mathbb{R}^2, F \subset \mathbb{R}^3)$.

§ Implementation of tensor splines

Tensor product decomposition of the underlying RKHS $\mathcal{H}_Z^\otimes$ enables factorization of the covariance matrices necessary to derive the estimator. We introduce some notation that will help in dealing with the case where $\mathcal{H}_Z^\otimes = \mathcal{H}_Z^s \otimes \mathcal{H}_Z^t$ and $\mathcal{H}_Z^s, \mathcal{H}_Z^t$ are themselves RKHS that are further decomposable. For the moment the reader may interpret the setup as one concerned with signal separation of spatiotemporal functions on the domain $S \times T$ with S, T and superscripts s, t denoting space and time respectively. Set

$$\mathcal{H}_X^\otimes = \mathcal{H}_X^s \otimes \mathcal{H}_X^s \qquad \mathcal{H}_Y^\otimes = \mathcal{H}_Y^s \otimes \mathcal{H}_Y^t \qquad \mathcal{H}_Z^\otimes = \mathcal{H}_Z^s \otimes \mathcal{H}_Z^t$$
$$\mathcal{H}_Z^s = \mathcal{H}_X^s \oplus \mathcal{H}_Y^s \qquad \mathcal{H}_Z^t = \mathcal{H}_X^t \oplus \ \mathcal{H}_Y^t \tag{3.28}$$

with corresponding reproducing kernels $K_X^s, K_X^t, K_Y^s, \ldots$ where we will also write $\mathcal{H}_{K_X^s} = \mathcal{H}_X^s$ when the explicit dependence of the RKHS on the kernels is to be emphasized. Furthermore denote by $\Xi_\wedge^u$ and $\Xi_\sim^u$, $u = s, t$ the solution operators for the interpolating and smoothing spline equations, i.e.

$$\sigma_z^u = \underset{z \in (A^u)^{-1} a^u}{\operatorname{argmin}} \ \|z\|^2_{\mathcal{H}_X^u \oplus \mathcal{H}_Y^u} \qquad =: \Xi_\wedge^u a^u \tag{3.29}$$

$$\sigma_x^u = \underset{x \in \mathcal{H}_X^u}{\operatorname{argmin}} \ \|A^u x - a^u\|^2_{A^u \mathcal{H}_Y^u} + \|x\|^2_{\mathcal{H}_X^u} \qquad =: \Xi_\sim^u a^u \tag{3.30}$$

where $A^u : \mathcal{H}_Z^u \to \mathbb{R}^{n_u}$ are the measurement operators and $A^u \mathcal{H}_Y^u := \mathcal{H}_{\mathcal{A}^u \otimes A^u K_Y^u}$. The next theorem provides a way of inferring a splines' expansion coefficients λ that is both computationally efficient and not too taxing on the memory. It furthermore clears up the relationship between interpolating and smoothing splines that are either only spatial or temporal and jointly spatiotemporal splines.

Theorem 3.1.10 *Let the notation be as introduced above. Then*

I The spatiotemporal tensor product smoothing spline $\sigma_X^\otimes$

$$\sigma_X^\otimes = \underset{x \in \mathcal{H}_X^\otimes}{\operatorname{argmin}} \ \| \underbrace{A^s \otimes A^t}_{A} x - a\|^2_{A[\mathcal{H}_Z^\otimes / \mathcal{H}_X^\otimes]} + \|x\|^2_{\mathcal{H}_X^\otimes} \tag{3.31}$$

is just an orthogonal projection of the spatiotemporal tensor interpolating spline and we have

$$\Xi^{\otimes}_{\sim} = \Xi^{s}_{\sim} \otimes \Xi^{t}_{\sim} = P^{s}_{X} \otimes P^{t}_{X} \Xi^{s}_{\wedge} \otimes \Xi^{t}_{\wedge} = P^{\otimes}_{X} \Xi^{\otimes}_{\wedge}$$

where $\Xi^{\otimes}_{\sim} : \mathbb{R}^{n_s} \otimes \mathbb{R}^{n_t} \to \mathcal{H}^{\otimes}_{X}$ *is the solution operator for the tensor spline 3.31,* P^{s}_{X} *and* P^{t}_{X} *are orthogonal projections onto* $\mathcal{H}^{s}_{X}$ *and* $\mathcal{H}^{t}_{X}$ *respectively and* $\Xi^{\otimes}_{\wedge}$ *is the tensor product of the solution operators of the separate spatial and temporal interpolation problems as defined in 3.29*

II *The smoothing spline* $\sigma^{\otimes}_{X}(u,v)$ *for any* $u \in S, v \in T$*can be given explicitly as a superposition of linear operators* A^s, A^t *applied to the appropriate kernel functions with their second arguments fixed to* s *and* t.

$$\sigma^{\otimes}_{X}(u,v) = \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} \lambda_{ij} (A^s K^s_X(\cdot,u))_i (A^t K^t_X(\cdot,v))_j$$
$$\lambda^{\otimes} = (\Sigma^s_X + \Sigma^s_Y)^{-1} \otimes \left(\Sigma^t_X + \Sigma^t_Y\right)^{-1} a \qquad (\lambda^{\otimes})_{ij} = \lambda_{ij}$$

The data $a \in \mathbb{R}^{n_s} \otimes \mathbb{R}^{n_t}$ *are assumed to lie on a regular grid in* $S \times T$ *as the measurement operator* A *factorizes into two parts* $A^s \otimes A^t$. *The matrices* $\Sigma^u_q := A^u \otimes A^u K_q$ *for* $q = X, Y$ *and* $u = s, t$ *are the covariance matrices induced by the linear operator* A^u *acting on elements of* H^u_q*, i.e.* $\Sigma^u_q = E[A^u q \otimes (A^u q)^*]$.

III *The coefficient tensor* $\lambda^{\otimes} \in \mathbb{R}^{n_s} \otimes \mathbb{R}^{n_t}$ *may be calculated via the matrix product*

$$\lambda^{\otimes} = (\Sigma^s_X + \Sigma^s_Y)^{+} a \left(\Sigma^t_X + \Sigma^t_Y\right)^{+}$$

when the data a *is assembled as a matrix* $a \in \mathbb{R}^{n_s} \otimes \mathbb{R}^{n_t}$. *If the corresponding spline* $\sigma^{\otimes}_{X}$ *is to be evaluated on a regular grid, say* $\sigma^{eval}_{X} = \left\{\sigma^{\otimes}_{X}(s^{eval}_i, t^{eval}_j)\right\}_{i=1,j=1}^{n^{eval}_s n^{eval}_t}$*, then*

$$\sigma^{eval}_{X} = (Q^s)^* \lambda^{\otimes} Q^t \tag{3.32}$$
$$(Q^s)_{ij} = (A^s K^s_X(\cdot, s^{eval}_j))_i \qquad \left(Q^t\right)_{ij} = (A^t K^t_X(t_i, t^{eval}_j))_i.$$

IV *The worst case computational complexity to calculate the full tensor* σ^{eval}_{X} *is approximately* $n_s^3 + n_t^3 + n_s^2 n_t + n_t^2 n_s + n^{eval}_s n_s n_t + n^{eval}_s n_t n^{eval}_t$ *when using the formulas expounded in III. The naive implementation without tensor product factorization roughly has worst case computational complexity of* $(n_s n_t)^3 + n^{eval}_s n^{eval}_t n_s n_t$.

Remark It is worth noting that tensorsplines are not more general than ordinary splines. They are more suitable for large scale computations though, as problem structure is used to simplify, or at least speed up, the solution of the SLAE's arising during their evaluation.

Proof: I The space $\mathcal{H}_X^{\otimes}$ has RK $K_X^s K_X^t =: K_X^{\otimes}$ and if A is written as $l_1, ..., l_n$ then $A[\mathcal{H}_Z^{\otimes}/\mathcal{H}_X^{\otimes}]$ has as kernel the matrix $C_{Z/X}$ with entries

$$\begin{aligned}(C_{Z/X})_{ij} &= \left(A \otimes A\left[(K_X^s + K_Y^s)(K_X^t + K_Y^t) - K_X^{\otimes}\right]\right)_{ij}\\ &= \left(A \otimes A\left[K_X^s K_Y^t + K_Y^{\otimes} + K_Y^s K_X^t\right]\right)_{ij}\\ &= l_i \otimes l_j\left[K_X^s K_Y^t + K_Y^{\otimes} + K_Y^s K_X^t\right]\end{aligned}$$

If the matrix $A \otimes A K_X^{\otimes}$ is denoted by C_X then an alternative formulation of equation 3.31 is $\sigma_X^{\otimes} = \underset{x \in \mathcal{H}_X^{\otimes}}{\operatorname{argmin}} \; \|Ax - a\|^2_{C_{Z/X}} + \|x\|^2_{\mathcal{H}_X^{\otimes}}$ to which the solution is given as

$$\begin{aligned}\sigma_X^{\otimes} &= \sum_{j=1}^{n} \lambda_j l_j K_X^{\otimes} \qquad\qquad \lambda = (C_{Z/X} + C_X)^{-1} a\\ &= \underbrace{\left[l_1 K_X^{\otimes}, ..., l_n K_X^{\otimes}\right](C_{Z/X} + C_X)^{-1}}_{\Xi_{\sim}^{\otimes}}\end{aligned}$$

where the operator $\Xi_{\sim}^{\otimes} : \mathbb{R}^n \to \mathcal{H}_X^{\otimes}$ is the solution operator for the tensor smoothing spline. Now if one just splits the space $\mathcal{H}_Z^{\otimes} = \mathcal{H}_Z^s \otimes \mathcal{H}_Z^t$ into two parts

$$\mathcal{H}_Z^{\otimes} = \underbrace{\mathcal{H}_X^s \otimes \mathcal{H}_X^t}_{\mathcal{H}_X^{\otimes}} \oplus_i \underbrace{\mathcal{H}_X^s \otimes \mathcal{H}_Y^t \oplus_i \mathcal{H}_Y^s \otimes \mathcal{H}_X^t \oplus_i \mathcal{H}_Y^s \otimes \mathcal{H}_Y^t}_{\mathcal{H}_{Z/X}^{\otimes}}$$

then lemma 3.1.8 asserts that $\sigma_X^{\otimes} = P_{\mathcal{H}_X^{\otimes}} \sigma_Z^{\otimes}$ where $P_{\mathcal{H}_X^{\otimes}}$ is the orthogonal projection onto $\mathcal{H}_X^{\otimes}$ as interpreted in lemma 3.1.8 and $\sigma_Z^{\otimes}$ is the interpolating spline

$$\sigma_Z^{\otimes} = \underset{z \in A^{-1}a, z \in \mathcal{H}_Z^{\otimes}}{\operatorname{argmin}} \; \|z\|^2_{\mathcal{H}_Z^{\otimes}}.$$

Then $\sigma_Z^{\otimes}$ has the explicit representation

$$\begin{aligned}\sigma_Z^{\otimes} &= \sum_{j=1}^{n} \lambda_j l_j K_Z^{\otimes} \qquad \lambda = (C_{Z/X} + C_X)^{-1} a\\ &= \underbrace{\left[l_1 K_Z^{\otimes}, ..., l_n K_Z^{\otimes}\right](C_{Z/X} + C_X)^{-1}}_{\Xi_{\wedge}^{\otimes}} a\end{aligned}$$

where $\Xi_{\wedge}^{\otimes} : \mathbb{R}^n \to \mathcal{H}_Z^{\otimes}$ is the solution operator for the interpolating spline. According to [21, p. 181], $\Xi_{\wedge}^{\otimes}$ factors into $\Xi_{\wedge}^{\otimes} = \Xi_{\wedge}^s \otimes \Xi_{\wedge}^t$. It is still left to show that $P_X^{\otimes} = P_{\mathcal{H}_X^{\otimes}}$ factors as $P_{\mathcal{H}_X^s} \otimes P_{\mathcal{H}_X^t}$. This follows from

$$(P_{\mathcal{H}_X^s} \otimes P_{\mathcal{H}_X^t})^* = (P_{\mathcal{H}_X^s} \otimes P_{\mathcal{H}_X^t})(P_{\mathcal{H}_X^s} \otimes P_{\mathcal{H}_X^t}) = P_{\mathcal{H}_X^s} \otimes P_{\mathcal{H}_X^t}$$

establishing it as an orthogonal projection and since its range is $P_{\mathcal{H}_X^s}(\mathcal{H}_Z^s) \otimes P_{\mathcal{H}_X^t}(\mathcal{H}_X^t) = \mathcal{H}_X^s \otimes \mathcal{H}_Z^t$, it is the orthogonal projection onto $\mathcal{H}_X^{\otimes}$. Then $P_{\mathcal{H}_X}^{\otimes} \Xi_\wedge^{\otimes} = P_{\mathcal{H}_X^s} \otimes P_{\mathcal{H}_X^t} \Xi_\wedge^s \otimes \Xi_\wedge^t$ as claimed.

II This follows almost immediately from I. Note that $A = A^s \otimes A^t$ and it is possible to rewrite the result of applying A as a matrix of dimensions $n_s \times n_t$, i.e.

$$(AK_X^{\otimes})_{ij} = (A^s \otimes A^t K_X^s K_X^t)_{ij} = (A^s K_X^s)_i (A^t K_X^t)_j.$$

Similarly for $\lambda^{\otimes}$ one has

$$\begin{aligned}\lambda^{\otimes} = [A \otimes AK_X^{\otimes}]^{-1} a &= \left[(A^s \otimes A^s) \otimes (A^t \otimes A^t) K_Z^s K_Z^t\right]^{-1} a \\ &= \left[A^s \otimes A^s K_Z^s \otimes A^t \otimes A^t K_Z^t\right]^{-1} a \\ &= (\Sigma_X^s + \Sigma_Y^s)^{-1} \otimes \left(\Sigma_X^t + \Sigma_Y^t\right)^{-1} a\end{aligned}$$

III This is just a reformulation of II. For any selfadjoint matrices C^s, C^t and $a \in \mathbb{R}^{n_s} \otimes \mathbb{R}^{n_t}$ it holds that the latter one can be written as $a = \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} a_{ij} e_i^s \otimes (e_j^t)^*$ which enables the simplification

$$\begin{aligned}C^s \otimes C^t a &= \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} a_{ij} (C^s e_i^s) \otimes (C^t e_j^t)^* \\ &= C^s \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} a_{ij} e_i^s \otimes (e_j^t)^* (C^t)^* \qquad = C^s a C^t\end{aligned}$$

and implies $\lambda^{\otimes} = (\Sigma_X^s + \Sigma_X^s)^{-1} a (\Sigma_X^t + \Sigma_Y^t)^{-1} \in \mathbb{R}^{n_s} \otimes \mathbb{R}^{n_t}$. The $n_s^{\text{eval}} \times n_t^{\text{eval}}$ matrix σ_X^{eval} has then the entries

$$\begin{aligned}(\sigma_X^{\text{eval}})_{ij} = &\left[(A^s K_X^s(\cdot, s_i^{\text{eval}}))_1, \ldots, (A^s K_X^s(\cdot, s_i^{\text{eval}}))_{n_s}\right] \lambda^{\otimes} \\ &\left[(A^t K_X^t(\cdot, t_j^{\text{eval}}))_1, \ldots, (A^t K_X^t(\cdot, t_j^{\text{eval}}))_{n_t}\right]^T\end{aligned}$$

and the whole matrix can be written as $(Q^s)^* \lambda^{\otimes} Q^t$ as claimed in the theorem.

IV For the tensor spline evaluation it is necessary to calculate

$$\sigma_X^{\text{eval}} = (Q^s)^* \lambda^{\otimes} Q^t = \underbrace{(Q^s)^*}_{(n_s^{\text{eval}}, n_s)} \underbrace{(\Sigma_X^s + \Sigma_Y^s)^+}_{(n_s, n_s)} \underbrace{a}_{n_s, n_t)} \underbrace{\left(\Sigma_X^t + \Sigma_Y^t\right)^+}_{(n_t, n_t)} \underbrace{Q^t}_{n_t, n_t^{\text{eval}}}$$

where the inversion takes an amount of operations proportional to n_s^3 and n_t^3 respectively, the inner matrix multiplications take $n_s^2 n_t$ and $n_s n_t^2$ and the outer matrix multiplications consume $n_s^{\text{eval}} n_s n_t$ and $n_s^{\text{eval}} n_t n_t^{\text{eval}}$ steps. In total, the worst case amount of operations may be written as $n_s(n_s^2 + n_s n_t + n_s^{\text{eval}} n_t) + n_t(n_t^2 + n_t n_s + n_t^{\text{eval}} n_s^{\text{eval}})$. For the naive evaluation of the spline it is necessary to calculate

$$\left\{\sigma_X^{\text{eval}}\right\}_{i=1}^{n_s^{\text{eval}} n_t^{\text{eval}}} = \left\{\sum_{j=1}^{n_s n_t} \lambda_j K(s_j, s_i^{\text{eval}})\right\}_{i=1}^{n_s^{\text{eval}} n_t^{\text{eval}}} .$$

This amounts to $n_s^{\text{eval}} n_t^{\text{eval}}$ times $n_s n_t$ operations plus the cost of determining $\lambda = (\Sigma_X^{st} + \Sigma_Y^{st})^+ a$ which is dominated by inverting a matrix of dimension $n_s n_t$. The total cost is approximately $n_s n_t (n_s^{\text{eval}} n_t^{\text{eval}} + n_s^2 n_t^2)$ for this implementation and features terms of the sixth power in the number of points compared to only the third power in the tensor spline formulation.

□

Interpretation of the tensorspline approach is relatively straightforward; one gains computational speed by assuming that the task is to extract an element of $\mathcal{H}_X^s \otimes \mathcal{H}_X^t$ from a space of the form $(\mathcal{H}_X^s \oplus \mathcal{H}_Y^s) \otimes (\mathcal{H}_X^t \oplus \mathcal{H}_Y^t)$ which is less general than to extract an element of $\mathcal{H}_X^s \otimes \mathcal{H}_X^t =: \mathcal{H}_X^{\otimes}$ from some $\mathcal{H}_Z^{\otimes}$ which contains $\mathcal{H}_X^{\otimes}$ as a subspace. Figure 3.6 illustrates this.

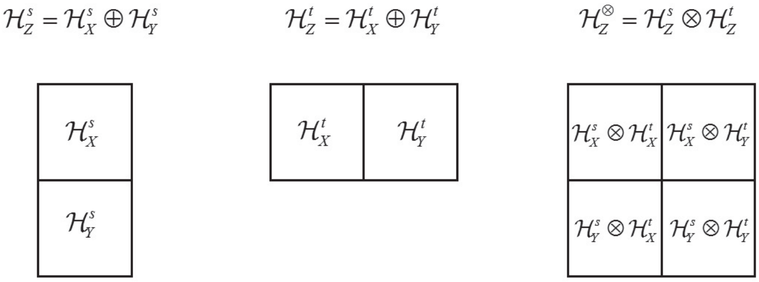

$\mathcal{H}_Z^s = \mathcal{H}_X^s \oplus \mathcal{H}_Y^s$
$\mathcal{H}_Z^t = \mathcal{H}_X^t \oplus \mathcal{H}_Y^t$
$\mathcal{H}_Z^{\otimes} = \mathcal{H}_Z^s \otimes \mathcal{H}_Z^t$
$\mathcal{H}_X^s$
$\mathcal{H}_Y^s$
$\mathcal{H}_X^t$
$\mathcal{H}_Y^t$
$\mathcal{H}_X^s \otimes \mathcal{H}_X^t$
$\mathcal{H}_X^s \otimes \mathcal{H}_Y^t$
$\mathcal{H}_Y^s \otimes \mathcal{H}_X^t$
$\mathcal{H}_Y^s \otimes \mathcal{H}_Y^t$

Figure 3.6: A graphical sketch of the estimation problem underlying tensor splines. The RKHS $\mathcal{H}_X^s \otimes \mathcal{H}_X^t$ consists of functions that have a correlation structure as prescribed by K_X^s in space and K_X^t in time and are to be separated from other constituents of the square $\mathcal{H}_Z^{\otimes} = \mathcal{H}_Z^s \otimes \mathcal{H}_Z^t = (\mathcal{H}_X^s \oplus \mathcal{H}_Y^s) \otimes (\mathcal{H}_X^t \oplus \mathcal{H}_Y^t)$ based on measurements of elements in $\mathcal{H}_Z^{\otimes}$.

It is not possible to reformulate problems of type ”extract an element of $\mathcal{H}_X^{\otimes}$ from $\mathcal{H}_X^{\otimes} \oplus \mathcal{H}_Y^{\otimes}$” as a tensor spline problem; one is limited to ambient spaces with n^2 components and n an integer. Even though this restriction is significant, the time saved during computation makes tensor splines worthwhile entities to investigate and employ when suitable. A simple spatiotemporal separation problem may serve to emphasize that and prepare the stage for later applications of the theory to the separation of atmospheric effects and deformations in terrestrial radar interferometry. The runtimes are documented in table 3.2; some graphical results in figure 3.7.

3.2 Construction of reproducing kernels

In section 3.1 a formalism was introduced that allowed the solution of abstract minimization problems corresponding to interesting real-world applications. The focus was, however, more on how to assemble a complex RKHS from simpler given ones and represent the solution in terms of the kernels of those elementary RKHS. Whereas in previous sections the kernels were assumed given, this section begins a systematic investigation into the properties of positive definite kernels that will allow the reader to guess kernels for a specific problem or infer them naively without any convergence guarantees or error estimates. Regardless of this weakness in rigor, the material presented here is sufficient for applying the theoretical apparatus associated to abstract splines in practice. Finally, a scheme for approximation of kernels by products of simpler kernels will be presented and questions of kernel inference will be posed. The obstacles arising there can only be solved with methods developed later during chapter 4. This section is therefore understood to have an introductory character rather than a definitive one.

3.2.1 Properties of covariance matrices

The solution formula for an abstract spline problem is linear in the data and nontriviality of the estimation is due to the complexity inherent in the covariance matrices describing the correlation structure of the process whose realization is to be inferred. From this point of view covariance matrices in the finite dimensional and covariance operators in the infinite dimensional settings encode all the information available about the processes behaviour in terms of lower order statistical moments and some of these relationships are collected below.

Recall that reproducing kernels $K : T \times T \to \mathbb{R}$ are always positive definite in the sense that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j K(t_i, t_j) \geq 0 \quad \forall n \in \mathbb{N} \ \{t_i\}_{i=1}^{n} \subset T \ \{\alpha_i\}_{i=1}^{n} \subset \mathbb{R}$$

	Size	$5 \times 5 \times 5$	$10 \times 10 \times 10$	$50 \times 50 \times 50$	$100 \times 100 \times 100$	$500 \times 500 \times 500$
Naive	RMSE:	0.42	0.35	——— unsolvable ———		
	Time (s):	$3 * 10^{-3}$	0.27			
2-tensor	RMSE:	0.51	0.42	0.31	——— unsolvable ———	
	Time (s):	$3 * 10^{-4}$	$2 * 10^{-3}$	3.8		
3-tensor	RMSE:	0.58	0.52	0.45	0.42	—— unsolvable ——
	Time (s):	$2 * 10^{-4}$	$4 * 10^{-4}$	$5 * 10^{-2}$	0.94	
Nr. of elements in CM		10^4	10^6	10^{10}	10^{12}	10^{16}

Table 3.2: Average root mean square errors and runtimes over 1000 simulations and successive inferences performed for spatiotemporal data of the size indicated in the top row for different estimation schemes. The 'naive' scheme uses full covariance matrix, the '2-tensor' scheme a splitting of the underlying covariance matrices into spatial and temporal parts and the '3-tensor' scheme factorizes in X and Y direction as well. The last row indicates the number of elements in the full covariance matrix for the differently sized problems. A task is arbitrarily declared as unsolvable if its solution takes more than 5 seconds to compute on an office computer with 3.5 GHz and 32 GB RAM.

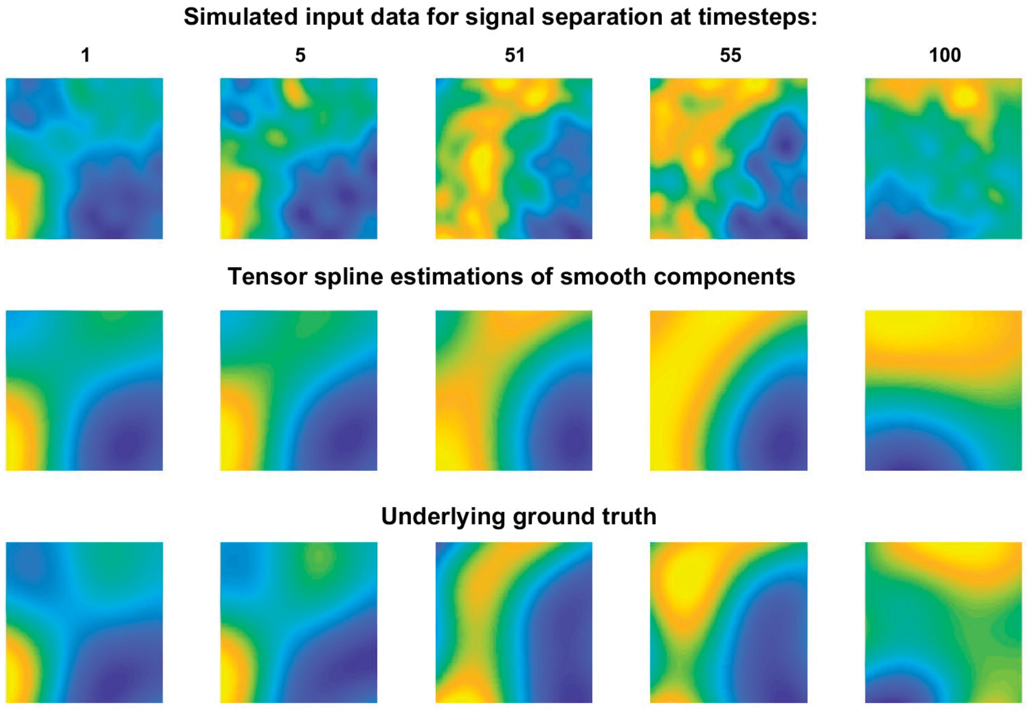


Figure 3.7: The first row of images shows time slices of one realization of a spatiotemporal random field at the times (arbitrary units) indicated by the numbers above the images. The field is a superposition of a spatiotemporal field that is smooth in time and space and one that is rough both in time and space — its covariance function is therefore not decomposable as would be necessary for tensorsplines to be optimal. The second and third row feature the estimations and ground truths for the smooth part of the realization. The colorscale is identical for all images.

as recorded in subsection 2.3.1 and that the RKHS $\mathcal{H}_K$ with RK K is isometrically isomorphic to a Hilbert space of T-indexed mean zero random variables $\{X_t\}_{t\in T}$ whose covariance is given as $E[X_sX_t] = K(s,t)$, see section 3.1. Therefore a kernel can always be interpreted as a covariance function and for any finite n, the matrix with elements $(C)_{ij} = K(t_i,t_j)$ for some choice of $\{t_i\}_{i=1}^n \subset T$ is the covariance matrix of the random vector $\{X_{t_i}\}_{i=1}^n \subset \overline{\mathcal{L}}(X)$. Since C is positive semidefinite and the bounded operators on $\mathbb{R}^n$ form a C^* -algebra, $C : \mathbb{R}^n \ni f \mapsto Cf \in \mathbb{R}^n$ with $(Cf)_i = \sum_{j=1}^n K(t_i,t_j)f_j$ may be written as

$$C = B^*B = A^2 \qquad A, B \in \mathcal{B}(\mathbb{R}^n) \tag{3.33}$$

[48, p. 15] and is seen to be the square of another operator $A : \mathbb{R}^n \to \mathbb{R}^n$. In the limit case, the covariance matrix becomes the covariance operator C_X with

$$(C_Xf)(t) = \int_T K(s,t)f(s)ds. \tag{3.34}$$

It is a positive semidefinite kernel operator that is furthermore compact (subsection 2.1.4) and admits a spectral decomposition $C_X\varphi_i = \lambda_i\varphi_i$ for a sequence $\{\varphi_i\}_{i=1}^\infty$ of eigenfunctions and a sequence $\{\lambda_i\}_{i=1}^\infty$ of nonnegative eigenvalues. This allows a practical implementation of functional calculus and therefore construction of square roots $C_X^{1/2}$, pseudoinverses C_X^+ and arbitrary decompositions $C_X = C_1 + C_2$ when a partition of unity is applied to C_X, see section 2.2. Using the same sequence of eigenvalues λ_i and eigenfunctions φ_i, the kernel K can be written in form of the Mercer decomposition

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(s)\varphi_i(t).$$

For the associated stochastic process $\{X_t\}_{t\in T}$, each of the random variables X_t can be represented as a superposition of deterministic φ_i's with random coefficients to form the Karhunen Loewe expansion

$$X_t = \sum_{j=1}^{\infty} \xi_j \sqrt{\lambda_j}\varphi_j(t) \qquad \text{with } \xi_j \text{ mean zero, unit variance, i.i.d. .}$$

In this way, one can easily generate samples of a stochastic process with covariance function $K(\cdot,\cdot)$ from realizations of white noise if the eigenvalues and eigenfunctions of the covariance operator C_X are known. Note, however, that the sample paths $X_\omega(\cdot) : T \to \mathbb{R}$ interpreted as realizations of random functions do not lie in the associated Hilbert space $\mathcal{H}_K$ with probability one in the case of Karhunen Loewe expansions with infinitely many nonzero expansion coefficients [107].

A straightforward implication of the Karhunen Loewe expansion is that for finite rank C_X, $X_\omega(\cdot)$ is a random superposition of finitely many functions. If for example

$$K(s,t) = \sigma_0^2 st \text{ then } C_X = \sigma_0^2 \varphi(s)\varphi(t)$$

is the spectral decomposition of C_X with $\varphi(s) = s$ and $X_t = \xi\sigma_0 t$, ξ mean zero and unit variance, is just a line passing through the origin chosen at random. This is consistent with $X_t = at$, $a \sim \mathcal{N}(0,\sigma_0^2)$ from which $E[X_s X_t] = \sigma_0^2 st = K(s,t)$ follows. Clearly, the structure of $K(\cdot,\cdot)$ determines important features of both $\mathcal{H}_K$ and $\{X_t\}_{t\in T}$. Recall from page 83 that $\{X_t\}_{t\in T}$ is a second order stationary stochastic process if $E[X_t]$ and $E[X_s X_t] - E[X_s]E[X_t]$ are translation invariant [39, p. 17], i.e.

$$E[X_t] = E[X_{t+\Delta}] \qquad \Delta : t+\Delta \in T \tag{3.35}$$

$$K(s,t) = K(s+\Delta, t+\Delta) \qquad \Delta : (s+\Delta, t+\Delta) \in T \times T. \tag{3.36}$$

This requires T to admit an additive structure. For the random vector $X = [X_1, ..., X_n]^T$, the expected energy $E[\langle X, X\rangle] = E[\sum_{j=1}^n X_j^2]$ is given as $\sum_{j=1}^n \lambda_j = \operatorname{tr}(C_X)$ where C_X is the covariance matrix of X.

In general, constructing a kernel K by forming $K(s,t) = f(s)f(t)$ for any function $f : T \to \mathbb{R}$ leads to the associated stochastic process being a randomly scaled version of $f(\cdot) : T \to \mathbb{R}$ as implied by the Karhunen Loewe expansion. The associated Hilbert space is therefore one-dimensional and does not contain enough different types of functions to be useful in the context of a typical nonparametric estimation. Superpositions of arbitrarily many functions $\{f_j\}_{j=1}^{\infty}$, however, are suited to construct kernels with nontrivial infinite rank structure under certain conditions as recorded in Fortet's theorem [20, p. 22].

Theorem 3.2.1 (Fortets theorem) $K : T \times T \rightarrow \mathbb{C}$ *is a reproducing kernel iff* $\exists \varphi : T \rightarrow \ell^2$ *with* $K(s,t) = \langle \varphi(s), \varphi(t) \rangle_{\ell^2} \quad \forall s, t \in T$.

This allows constructing arbitrarily complex kernels that strongly deviate from the regular ones typically generated from a restricted family of kernels by fixing a small amount of parameters. The associated random fields can feature sharp changes in variance, ridges and locally variable anisotropies. Examples are collected in figure 3.8.

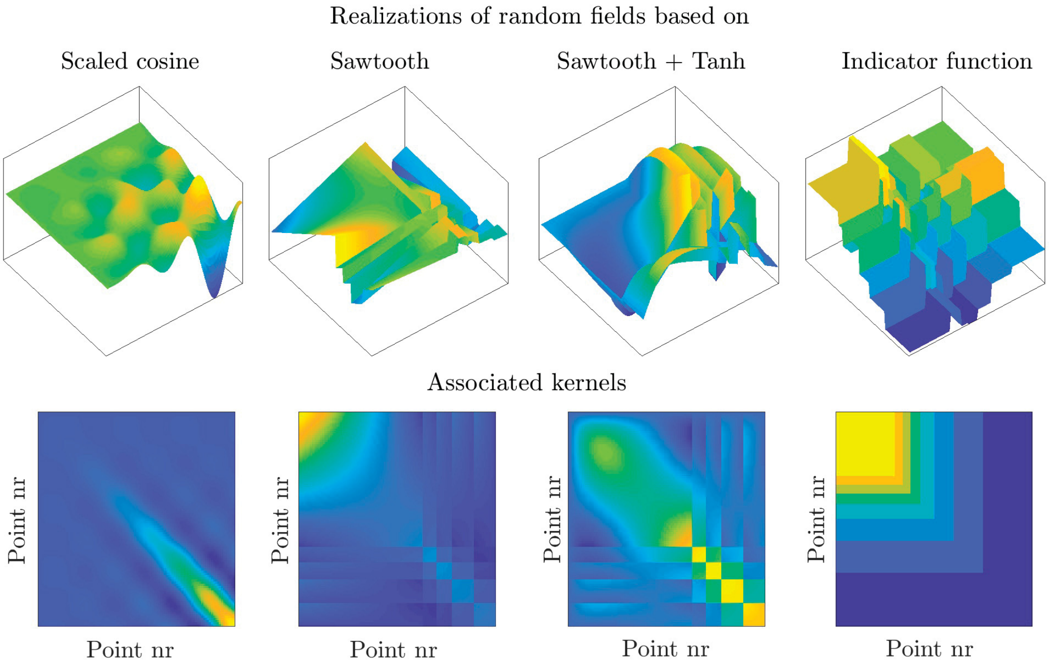


Figure 3.8: Kernels K constructed from specific sets of functions mentioned in the respective headlines. They are plotted only for the case $T \subset \mathbb{R}$ for purposes of clarity; the realizations are generated from the Karhunen-Loewe expansion of tensor products $K \otimes K$ forming kernels on $T \times T \subset \mathbb{R}^2$.

The stochastic interpretation of $K(s,t)$ as a covariance function has direct consequences for the quantification of the estimations reliability. As is recorded in the geostatistical literature, the expected square deviation $E[\epsilon^2] = E[(\hat{X}_t - X_t)^2] = E[(\sigma_X(t) - X_t)^2]$ is explicitly given as [39, p. 169]

$$\text{Var}(\epsilon) = \text{Var}(X_t) - \lambda^T [K(t_1,t_0), ..., K(t_n,t_0)]^T - \mu^T [f_1(t_0), ..., f_m(t_0)]^T \tag{3.37}$$

where $\sigma_X(t)$ is a solution to $\sigma_X(\cdot) = \text{argmin}_{x \in A^{-1}a} \|[x]\|_{\mathcal{H}_K}$. In this formulation A

is just evaluation at $t_1, ..., t_n$, $f_1, ..., f_m$ form a basis of the nullspace of the canonical projection $[\cdot]$ and a is the problem data. The vectors $\lambda \in \mathbb{R}^n$ and $\mu \in \mathbb{R}^m$ are as before solutions of the SLAE documented in theorem 3.1.5. Therefore, confidence intervals can be derived for the estimator $\sigma_X(t)$. Indeed, the aforementioned equation is a special case of the formula to calculate the conditional covariance of a mean-zero Gaussian random vector $Z = [X; Y]$, $X : \Omega \to \mathbb{R}^{n_1}, Y : \Omega \to \mathbb{R}^{n_2}$ for which it reads [160, p. 73]

$$\begin{aligned} \Sigma_{X|Y} &= \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \qquad (3.38) \\ \Sigma_{ZZ} &= \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}. \end{aligned}$$

The conditional mean of X given Y is then $\mu_{X|Y} = \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mu_Y)$ [160, p. 73] where μ_X, μ_Y are the marginal means of X and Y. This allows conditional simulation and the construction of covariance matrices whose associated vectors satisfy certain boundary conditions, see figure 3.9 for examples. This idea will be pursued further in subsection 3.2.2.

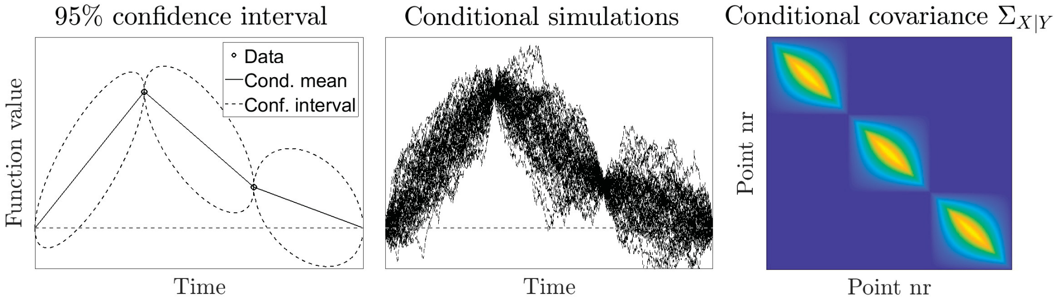


Figure 3.9: Theoretical confidence intervals and conditional simulations of a Brownian bridge process with known values at the locations indicated by circles. The conditional covariance matrix detailing the kernel of the random variables conditioned on the observations is plotted in the last panel.

Lastly, notice that in the finite dimensional case the spectral decomposition of C_X provides a solution to the problem of efficiently representing functions using only a limited number of fixed basis functions and coefficients. This follows from the fact that minimizing $E[\|f - P_{\mathcal{H}_1} f\|_{\mathcal{H}}^2]$ with $\mathcal{H}_1 \overline{\boxslash} \mathcal{H}$ unknown and $\mathcal{H} = \mathbb{R}^n$ is equivalent to minimizing

$$\begin{aligned} E[\|f - P_{\mathcal{H}_1} f\|_{\mathcal{H}}^2 &= E[\langle f - P_{\mathcal{H}_1} f, f - P_{\mathcal{H}_1} f \rangle_{\mathcal{H}}] \\ &= E[\langle f, f \rangle_{\mathcal{H}} - 2\langle P_{\mathcal{H}_1} f, f \rangle_{\mathcal{H}} + \langle P_{\mathcal{H}_1} f, P_{\mathcal{H}_1} f \rangle_{\mathcal{H}}] \\ &= E[\langle f, f \rangle_{\mathcal{H}} - \langle P_{\mathcal{H}_1} f P_{\mathcal{H}_1} f \rangle_{\mathcal{H}}] \\ &= \operatorname{tr} E[f \otimes f^*] = \operatorname{tr} E[(P_{\mathcal{H}_1} f) \otimes (P_{\mathcal{H}_1} f)^*] \\ &= \operatorname{tr}(C_X - P_{\mathcal{H}_1} C_X P_{\mathcal{H}_1}). \end{aligned}$$

If $\dim \mathcal{H}_1 = n_1$, then $P_{\mathcal{H}_1} C_X P_{\mathcal{H}_1}$ is an n_1-rank approximation to C_X. But $\Delta = C_X - P_{\mathcal{H}_1} C_X P_{\mathcal{H}_1}$ is positive semidefinite and therefore $\operatorname{tr}(\Delta)$ is just the trace norm

$\|\cdot\|_{\mathrm{tr}}$ of Δ. It is proven in [136] that the choice of rank-n_1 Q minimizing $\|C_X - Q\|_{\mathrm{tr}}$ is simply $Q = \sum_{j=1}^{n_1} \lambda_j \varphi_j \otimes \varphi_j^*$ where $\{\lambda_j\}_{j=1}^n$ and $\{\varphi_i\}_{i=1}^n$ are the eigenvalues and eigenfunctions of C_X. Therefore the basis vectors $\{\varphi_i\}_{i=1}^n$ provide an efficient system, in which to represent any $f = X_t(\omega)$.

3.2.2 Approximation of kernels

For primarily computational reasons, it is often convenient to consider kernels $K : (S \times T) \times (S \times T) \to \mathbb{R}$ that factorize as the tensor product of two simpler kernels $K^s : S \times S \to \mathbb{R}$ and $K^t : T \times T \to \mathbb{R}$. Inversion and spectral decomposition of the arising kernel matrices can then be done in a fraction of the time needed otherwise and enables efficient simulation and inference. When the kernel does not have a representation as a simple tensor but may be approximated by a sum of simple tensors, the binomial inverse theorem can be of help. Although only situationally applicable and affected by numerical instabilities, this approach reliant on iterative schemes involving spectral decompositions is at times the only possibility to handle large-scale problems.

Recall that the Karhunen-Loewe expansion guaranteed the decomposability of a mean-zero, second order stochastic process $\{X_t\}_{t\in T}$ into a superposition of basis functions as given in theorem 2.3.5. When the expansion is truncated after the n-th term, the resultant expression $\tilde{X}_t^n = \sum_{j=1}^n \xi_j \sqrt{\lambda_j}\varphi_j(t)$ is an approximation to X_t whose error decreases monotonically with n. If as relative error ϵ^n one designates the ratio of expected energy $\|\tilde{X}_t^n - X_t\|_{L^2}^2$ of the error to the expected total energy $\|X_t\|_{L^2}^2$, then one finds for ϵ^n the alternative expression

$$\frac{E[\|\tilde{X}_t^n - X_t\|_{L^2}^2}{E[\|X_t\|_{L^2}^2]} = \frac{\sum_{i=n+1}^{\infty}\sum_{j=n+1}^{\infty} E[\xi_i\xi_j]\sqrt{\lambda_i\lambda_j}\langle\varphi_i,\varphi_j\rangle_{L^2}}{E\left[\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\xi_i\xi_j\sqrt{\lambda_i\lambda_j}\langle\varphi_i,\varphi_j\rangle_{L^2}\right]} = \frac{\sum_{i=n+1}^{\infty}\lambda_i}{\sum_{i=1}^{\infty}\lambda_i}. \tag{3.39}$$

As $\{\lambda_i\}_{i=1}^{\infty} \subset \mathbb{R}_+$ and $\lim_{k\to\infty}\lambda_k = 0$, the relative error ϵ^n converges to zero and if the sequence of eigenvalues decays fast, it does so rapidly. The approximation $\tilde{X}_t^n$ may be interpreted as being drawn from a degenerated stochastic process with covariance operator $\tilde{C}_K$ — an n-term low rank approximation $\sum_{j=1}^n \lambda_j\varphi_j \otimes \varphi_j^*$ to C_K. Constructing a functional calculus on $\tilde{C}_K$, $\psi_{\tilde{C}_K}(x^{-1}) = \sum_{j=1}^n \lambda_j^{-1}\varphi_j \otimes \varphi_j^* = \tilde{C}_K^+$ is readily computable and enables fast, approximate inference.

If the kernel $K : (S \times T) \times (S \times T) \to \mathbb{R}$ decomposes as $K(u,v) = K((s_1,t_1),(s_2,t_2)) = K^s(s_1,s_2)K^t(t_1,t_2)$ and the individual Mercer decompositions are known to be

$$K^s(s_1,s_2) = \sum_{j=1}^{\infty}\lambda_j^s\varphi_j^s(s_1)\varphi_j^s(s_2) \qquad K^t(t_1,t_2) = \sum_{j=1}^{\infty}\lambda_j^t\varphi_j^t(t_1)\varphi_j^t(t_2)$$

then K can be written explicitly by simplifying the product to $K((s_1,t_1),(s_2,t_2)) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\lambda_i^s\lambda_j^t\varphi_i^s(s_1)\varphi_i^s(s_2)\varphi_j^t(t_1)\varphi_j^t(t_2)$. The positive Kernel operator $C_K : f \to$

$\int_{S\times T} K(u,\cdot)f(u)du$ acting on functions on $S \times T$ then has the representation

$$C_K = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \lambda_i^s \lambda_j^t \left(\varphi_i^s(\cdot) \otimes \varphi_j^t(\cdot)\right) \otimes \left(\varphi_i^s(\cdot) \otimes \varphi_j^t(\cdot)\right)^* = \sum_{l=1}^{\infty} \lambda_l^{\otimes} \varphi_l^{\otimes} \otimes (\varphi_l^{\otimes})^* \tag{3.40}$$

where the sequence $\{\lambda_l\}_{l=1}^{\infty}$ contains positive eigenvalues and $\{\varphi_l^{\otimes}\}_{l=1}^{\infty}$ are orthonormal eigenfunctions of C_K. If the sequence of eigenvalues is rearranged into descending order, then a low rank approximation $\tilde{C}_K$ can be found again via truncation. This enables simulation and inference.

Example 17 Consider a spatiotemporal random field F from which a realization is to be drawn. The index set is $S \times T \subset \mathbb{R}^3$ with $S = S_x \times S_y \subset \mathbb{R}^2$ and $T \subset \mathbb{R}^1$. Suppose that S and T are sampled at discrete points leading to realizations of F being sought on an $n \times n \times n$ grid. A naive simulation requires the spectral decomposition of a positive semidefinite covariance matrix with n^6 entries whereas a truncated tensor product approach employs the spectral decomposition of matrices with n^2 entries each. The runtimes for simulating a Gaussian random field with squared exponential covariances in S_x, S_y and T for a relative error $\epsilon < 1\%$ are compared to a standard method of simulation (mvnrnd in Matlab) in table 3.3.

■

If the goal is not simulation but inference, one has to proceed differently. If n denotes the number of observations, the covariance matrix of these observations has $n \times n$ entries and a naive inversion e.g. With Gaussian elimination requires the execution of floating point operations of order approximately n^3. Usually neither inversion nor storage are feasible on normal office computers if n is above 50000. However, this situation occurs routinely when dealing with spatiotemporal problems as those encountered during the investigation of atmospheric effects in terrestrial radar interferometry in section 5.2. If the RK K of $\mathcal{H}_K$ can be written as $K = K^s \otimes K^s$ then a similar statement holds for the covariance matrix $C = C^s \otimes C^t$ and its pseudoinverse $C^+ = (C^s)^+ \otimes (C^t)^+$, $C^s \in \mathbb{R}^{n_s} \otimes \mathbb{R}^{n_s}$, $C^t \in \mathbb{R}^{n_t} \otimes \mathbb{R}^{n_t}$. This reduces storage and inversion of an $n_s n_t \times n_s n_t$ matrix to storage and inversion of both an $n_s \times n_s$ and an $n_t \times n_t$ matrix. The situation gets more complicated when K is a superposition of simple tensors.

Theorem 3.2.2 (Binomial inverse theorem) *Let A, B, U, V be matrices of ap-*

Size	$5 \times 5 \times 5$	$10 \times 10 \times 10$	$50 \times 50 \times 50$	$100 \times 100 \times 100$	$500 \times 500 \times 500$
Runtime mvnrnd	0.0004s	0.170s	unsolvable		
Runtime tensor	0.001s	0.013s	0.331s	2.959s	800s

Table 3.3: Runtimes of different simulation procedures dependent on grid size based on 100 trial runs per scenario. A task is arbitrarily declared as unsolvable if during computation a standard office computer (32 GB RAM) runs out of memory.

propriate dimensions such that the term $(A + UBV)$ is well defined. If A and $B + BVA^{-1}UB$ are invertible, then [160, p. 23]

$$(A + UBV)^{-1} = A^{-1} - A^{-1}UB(B + BVA^{-1}UB)^{-1}BVA^{-1}. \tag{3.41}$$

The binomial inverse theorem is also known as the matrix inversion lemma or as the Woodbury matrix identity. In the simple case where B is the identity and U, V are column and row vectors u and v respectively, the equation specializes to the Sherman-Morrison-Woodbury formula

$$(A + uv)^{-1} = A^{-1} - \frac{(A^{-1}u)(vA^{-1})}{1 + vA^{-1}u}. \tag{3.42}$$

This leads to a cheap update rule for the inverses of perturbed kernel matrices and allows calculation of inverses $(C_1 + \tilde{C}_2^n)^{-1}$, $\tilde{C}_2^n$ a low rank approximation to C_2, that are otherwise intractable. Suppose now that $K_1 = K_1^s K_1^t$ and $K_2 = K_2^s K_2^t$ leading to the kernel matrices $C_1 = C_1^s \otimes C_1^t$ and $C_2 = C_2^s \otimes C_2^t$. When for the sake of inference, $(C_1 + C_2)^{-1}$ is needed, the following theorem provides a computationally efficient solution.

Theorem 3.2.3 *Let C_1 and C_2 be defined as above. Denote by $\tilde{C}_2^n$ the n-term low rank approximation $\tilde{C}_2^n = \sum_{j=1}^n \lambda_j^{\otimes} \varphi_j^{\otimes} \otimes (\varphi_j^{\otimes})^* = \sum_{j=1}^n u_j \otimes u_j^*$. The eigenvalues and $\lambda^{\otimes}$ and the eigenvectors $\varphi_j^{\otimes}$ are constructed as described on page 115. Then with the notation $Q^j = (C_1 + \tilde{C}_2^j)^{-1}$, one finds the recursive equation*

$$Q^0 = (C_1^s)^+ \otimes (C_1^t)^+ \tag{3.43}$$

$$Q^j = Q^{j-1} - \frac{(Q^{j-1}u_j) \otimes (Q^{j-1}u_j)^*}{1 + u_j^T Q^{j-1} u_j} \qquad 1 \leq j. \tag{3.44}$$

The theorem follows directly from a successive application of the binomial inverse theorem as stated in equation 3.42 to the term $(C_1 + \sum_{j=1}^n u_j \otimes u_j^*)$. Theorem 3.2.3 is useful, if $\lambda_j^{\otimes}$ decays fast as in that case only few steps n are necessary to approximate $(C_1 + C_2)$ by $(C_1 + \tilde{C}_2^n$. It is then possible to approximate the inverses of $C = (C_1 + C_2 + ...)$ that are sums of simple tensors $C_j = C_j^s \otimes C_j^t, j = 1, ..., m$ without inverting the sum by reducing the calculation to manipulation of $(C_1^s)^+ \otimes (C_1^t)^+$ and the main spectral components of $C_j, j \geq 2$.

Since sums of simple tensors can be arbitrarily complicated, this allows fast approximate inversion even of anisotropic and instationary kernel matrices C if a decomposition into simple tensors $\sum_{j=1}^n C_j \approx C, C_j = C_j^s \otimes C_j^t$ is available and the spectra of the individual components C_j decay sufficiently fast. As the relationship between the individual spectra of C_1 and C_2 and the spectrum of the sum $C_1 + C_2$ is not straightforward and expensive to compute (see [31]), large scale simulation of random vectors with covariance matrix $C = \sum_{j=1}^n C_j$ is more difficult than inversion and we leave it for future work. In practice, it is sometimes desirable to find a C that satisfies certain boundary conditions. This can be done either by means of solving a constrained variance components estimation problem $C = \sum_{j=1}^{n} \mu_j C_j$ for the coefficient vector $\mu \in \mathbb{R}^n$ as in subsection 4.4.2 or directly in terms of the elements of C. The latter is a special case of the former and will be treated in the succeeding subsection.

3.2.3 Methods of construction and design

A certain amount of standard kernels exist but they are often insufficient in practical applications where a systems behavior is either unknown or satisfies constraints that are not reflected in any of the widely used kernels. The proper choice of kernels based on observations is a task that is tackled in chapter 4. For now the easier problem of constructing a kernel from prior considerations is addressed. It will be shown how kernels can be derived when only their linear transforms are known and and one is given a set of linear constraints. Examples are provided to illustrate the construction rules practical meaning in the context of abstract splines featuring nontrivial measurement and energy operators. Throughout this subsection, an entirely finite-dimensional perspective will be taken and kernels, or rather kernel operators, are replaced by covariance matrices associated to Gaussian random vectors.

§ **Linear relations between spaces**

Let $L_q^2 = \mathbb{R}^{n_q}$ be the Hilbert space of vectors in $\mathbb{R}^{n_q}$ together with the standard inner product $\langle f, g\rangle_{L_q^2} = \sum_{j=1}^{n_q} f_j g_j$ inherited from the identity matrix I. Denote by $\epsilon \sim \mathcal{N}(\mu_\epsilon, I)$ a white noise Gaussian random vector taking values in $\mathbb{R}^{n_\epsilon}$. $\mathcal{H}_X, \mathcal{H}_Y$ and $\mathcal{H}_\epsilon$ are finite dimensional (reproducing kernel) Hilbert spaces associated to Gaussian random vectors

$$X \sim \mathcal{N}(\mu_X, C_X) \qquad Y \sim \mathcal{N}(\mu_Y, C_Y) \qquad \epsilon \sim \mathcal{N}(\mu_\epsilon, I)$$

taking values in $\mathbb{R}^{n_X}, \mathbb{R}^{n_Y}$ and $\mathbb{R}^{n_\epsilon}$ respectively. The corresponding inner products are

$$\langle f, g\rangle_{\mathcal{H}_X} = \langle C_X^+ f, g\rangle_{L_x^2} \qquad \langle f, g\rangle_{\mathcal{H}_Y} = \langle C_Y^+ f, g\rangle_{L_Y^2} \qquad \langle f, g\rangle_{\mathcal{H}_\epsilon} = \langle f, g\rangle_{L_\epsilon^2}.$$

Depending on the exact nature of the known relationships between $\mathcal{H}_X, \mathcal{H}_Y, \mathcal{H}_\epsilon$ some of the covariance matrices are considered known and some are to be deduced. For now, let $\mu_\epsilon = 0$. The following two situations arise regularly in practice and feature closed form solutions under mild assumptions.

a) Case: $L_\epsilon^2 \overset{A}{\to} \mathcal{H}_X$, C_X unknown. ('Variance propagation')

   Let $X = A\epsilon$ with A of full row rank. Then $\mu_X = E[A\epsilon] = 0$ and

$$C_X = E[X \otimes X^*] = E[A\epsilon \otimes \epsilon^* A^*] = AIA^T$$

   implying $X \sim \mathcal{N}(0, AA^T)$. Since A is full row rank, AA^T is invertible and the spaces $\mathcal{H}_X$ and L_ϵ^2 are linked via the operator $A \in \mathbb{R}^{n_X} \otimes \mathbb{R}^{n_\epsilon}$ by the relationship

$$L^2_\epsilon \overset{A}{\rightarrow} \mathcal{H}_X \qquad \mathcal{H}_X = \mathrm{im}(A) \qquad \langle f, g\rangle_{\mathcal{H}_X} = \langle (AA^T)^{-1}f, g\rangle_{L^2_X}.$$

Furthermore, straightforward calculation shows $\langle f, g\rangle_{L^2_X} = \langle AA^T f, g\rangle_{\mathcal{H}_X} = \langle A^T f, A^T g\rangle_{\mathcal{H}_\epsilon}$ and if $\mathcal{H}_X$ is restricted to $\mathrm{im}(A)$, then $\forall\, f, g \in \mathcal{H}_X\ \exists u, v \in \mathcal{H}_\epsilon$ s.t. $f = Au, g = Av$ implying

$$\langle f, g\rangle_{\mathcal{H}_X} = \langle (AA^T)^{-1}f, g\rangle_{L^2_X} = \langle (AA^T)^{-1}Au, Av\rangle_{L^2_X} = \langle A^+Au, v\rangle_{\mathcal{H}_\epsilon}.$$

If A were invertible, then $\langle u, v\rangle_{\mathcal{H}_\epsilon} = \langle (A^T(AA^T)^{-1}A, u\rangle_{\mathcal{H}_\epsilon} = \langle Au, Av\rangle_{\mathcal{H}_X}$.

b) Case: $L^2_\epsilon \overset{B}{\leftarrow} \mathcal{H}_X, C_X$ unknown. ('Energy operator')

Let $\epsilon = BX$ with B full column rank and therefore left invertible. If X is known to be zero mean, then

$$I = C_\epsilon = E[\epsilon \otimes \epsilon^*] = E[BX \otimes X^* B^*] = BC_X B^*$$

implying $C_X = B^+(B^T)^+ = (B^TB)^+$ due to the rank condition. B^TB is invertible, $X \sim \mathcal{N}(0, (B^TB)^{-1})$ and the spaces $\mathcal{H}_X$ and L^2_ϵ are linked via the operator $B \in \mathbb{R}^{n_\epsilon} \otimes \mathbb{R}^{n_X}$ by the relationship

$$L^2_\epsilon \overset{B}{\leftarrow} \mathcal{H}_X \qquad \mathcal{H}_X = B^{-1}(L^2_\epsilon) \qquad \langle f, g\rangle_{\mathcal{H}_X} = \langle B^TBf, g\rangle_{L^2_X}.$$

Furthermore, straightforward calculation shows $\langle f, g\rangle_{\mathcal{H}_X} = \langle Bf, Bg\rangle_{\mathcal{H}_\epsilon}$. If B were invertible, then $\langle u, v\rangle_{\mathcal{H}_\epsilon} = \langle B^TB(B^{-1}u), (B^{-1}v)\rangle_{L^2_X} = \langle B^{-1}u, B^{-1}v\rangle_{\mathcal{H}_X}$.

For both cases, the interpretations are relatively straightforward. If $L^2_\epsilon \overset{A}{\rightarrow} \mathcal{H}_X$, then $\mathcal{H}_X$ contains the responses of A to white noise and for $f \in \mathcal{H}_X$, $\|f\|^2_{\mathcal{H}_X} = \|u\|^2_{L^2_\epsilon}$ where $Au = f$ measures the negative log likelihood of f by reducing it to the underlying noise variable. In contrast, if $L^2_\epsilon \overset{B}{\leftarrow} \mathcal{H}_X$, then Bf contains white noise and penalizing $\|f\|^2_{\mathcal{H}_X} = \|Bf\|^2_{L^2_X}$ means demanding that f should approximately satisfy the equation $Bf = 0$ or at least minimize the energy $\|Bf\|^2_{L^2_X}$ of deviations from 0.

We will continue to use the schematic diagrams of type $\mathcal{H}_X \overset{A}{\rightarrow} \mathcal{H}_Y$ and interpret them as $Y = AX$ on the level of random variables with consequences for the associated RKHS $\mathcal{H}_X$ and $\mathcal{H}_Y$. Before introducing linear constraints on covariance matrices, investigate situations that feature spaces related to white noise in nontrivial ways. Clearly, $L_\epsilon \overset{A}{\rightarrow} \mathcal{H}_X \overset{F}{\rightarrow} \mathcal{H}_Y$ and $L^2_\epsilon \overset{B}{\leftarrow} \mathcal{H}_X \overset{G}{\leftarrow} \mathcal{H}_Y$ are just versions of cases a) and b) with FA replacing A and BG replacing B in such a way that

$$C_Y = FAA^TF^T = FC_XF^T \tag{3.45}$$

$$C_Y = B^+(B^TB)^+(B^T)^+ = (G^+)C_X(G^+)^T \tag{3.46}$$

under the right assumptions for F, A, G, B. However, the following constellation requires special care.

c) Case: $L^2_\epsilon \overset{A}{\to} \mathcal{H}_X \overset{G}{\leftarrow} \mathcal{H}_Y$, C_Y unknown.
From a) and b) one derives $C_X = (AA^T)$ and $C_X = GC_YG^T$. Introduce $Q = C_X^{-1/2}$, the selfadjoint positive semidefinite inverse squareroot of C_X and assume QG is invertible. Then $C_X = GC_YG^T$ is equivalent to

$$\begin{aligned} I &= QGC_YG^TQ^T \\ \Leftrightarrow C_Y &= (QG)^+\left((QG)^T\right)^+ = \left[(QG)^T(QG)\right]^+ &&= \left[G^TQ^TQG\right]^+ \\ & &&= \left[G^T(AA^T)^{-1}G\right]^{-1} \end{aligned}$$

with $C_X = AA^T$ or in a more convenient form $C_Y = G^+C_Y(G^T)^+$ when it applies.

§ Linear constraints

Let $\mathcal{H}_X \overset{A}{\to} \mathcal{H}_\epsilon$, A of full row rank with nullspace $N(A)$ of dimension m. Let L be a linear operator $L : \mathcal{H}_X \to \mathbb{R}^m$ such that $\ker A \cap \ker L = \{0\}$ and $A_c = [A; L] : \mathcal{H}_X \to \mathbb{R}^{n_\epsilon} \oplus \mathbb{R}^m$ invertible. Define as usual

$$A_c \otimes A_cC_X = A_cC_XA_c = \begin{bmatrix} AC_XA^T & AC_XL^T \\ LC_XA^T & LC_XL^T \end{bmatrix}.$$

If $\mathcal{H}_X \overset{A}{\to} \mathcal{H}_\epsilon$, then $AC_XA^T = I$. If AC_XL^T and LC_XL^T can be specified with the help of the boundary conditions introduced by L and given in the form $AC_XL^T = q^T, LC_XL^T = S$, then the system of equations

$$A_cC_XA_c^T = \underbrace{\begin{bmatrix} I & q^T \\ q & S \end{bmatrix}}_{I_c} \tag{3.47}$$

follows. It has solution $C_X = (A_c^{-1})I_c(A_c^{-1})^T$. If slightly more general, the situation is described by the scheme $\mathcal{H}_Y \overset{A}{\to} \mathcal{H}_X$ with C_Y unknown and C_X known, then the SLAE 3.48 ensues.

$$C_Y = (A_c^{-1})F_c(A_c^{-1})^T \tag{3.48}$$

$$A_c = \begin{bmatrix} A \\ L \end{bmatrix} \qquad F_c = \begin{bmatrix} C_X & AC_YL^T \\ LC_YA^T & LC_YL^T \end{bmatrix} \tag{3.49}$$

Example 18 Let $A = \nabla$ be (a discrete version of) the derivative operator and $L = e_0^T$, i.e. $Lf = f(0)$. Then $N(A)$ are the constants and $\dim N(A) = 1$. Since ∇ annihilates constants, C_X can be inferred only up to a constant from the equation

$\nabla A \nabla^T = I$ and L determines that constant. The usual Brownian motion is then constructed from $\mathcal{H}_X \overset{A_c}{\to} \mathcal{H}_\epsilon \oplus \mathbb{R}$ where $A_c = [\nabla; L]$. For the boundary condition $Lf = f(0) = 0$, one explicitly finds

$$A_c = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \qquad F_c = \begin{bmatrix} I & \nabla C_X e_0 \\ e_0^T C_C \nabla^T & e_0^T C_X e_0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

if it is demanded that C_X is zero in the first row and column as forced by $Lf = e_0^T f = 0$. The equation $C_X = (A_c^{-1}) F_c (A_c^{-1})$ gives the covariance matrix for Brownian motion, as desired (see figure 3.10). ■

Example 19 Let $A = \nabla^2 = \Delta$ be (a discrete version of) the second derivative and $L = [e_0^T, e_n^T]$, i.e $Lf = [f(0), f(n)]^T \in \mathbb{R}^2$. Then $N(A)$ are the first order polynomials and $\dim N(A) = 2$. If we choose to let f start and end at zero, then for the underlying kernel K it holds that $K(t,0) = 0, K(t,n) = 0 \ \forall t$ and consequently $\nabla K(t,0) = \nabla K(t,n) = 0$. One explicitly finds

$$A_c = \begin{bmatrix} 1 & -2 & 1 & \cdots & 0 & 0 \\ 0 & 1 & -2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & -2 & 1 \\ 1 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}$$

$$F_c = \begin{bmatrix} I & \nabla C_X e_0 & \nabla C_X e_n \\ e_0^T C_C \nabla^T & e_0^T C_X e_0 & e_0^T C_X e_n \\ e_n^T C_X \nabla^T & e_n^T C_X e_0 & e_n^T C_X e_n \end{bmatrix} = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

The result $C_X = (A_c^{-1}) F_C (A_c^{-1})^T$ is plotted in figure 3.10 and has some similarities to the covariance matrix of a Brownian bridge process but is smoother due to the act of integrating twice to counteract the second derivative. ■

When more constraints are included than strictly necessary, $A_c = [A; L]$ does not have an inverse anymore. For the overdetermined system $A_c C_Y A_c^T = F_c$ a weight matrix $P^\otimes = (\Sigma^\otimes)^{-1}$ could be introduced to weigh the different conditions against each other. Then estimate

$$\begin{aligned} C_Y &= \underset{C_Y \in \mathbb{R}^{n_Y} \otimes \mathbb{R}^{n_Y}}{\operatorname{argmin}} \quad \|A_c \otimes A_c C_Y - F_c\|^2_{\Sigma^\otimes} \\ &= \left((A_c^\otimes)^T P^\otimes A_c^\otimes\right)^+ (A_c^\otimes)^T P^\otimes F_c \qquad\qquad A_c^\otimes = A_c \otimes A_c \end{aligned} \tag{3.50}$$

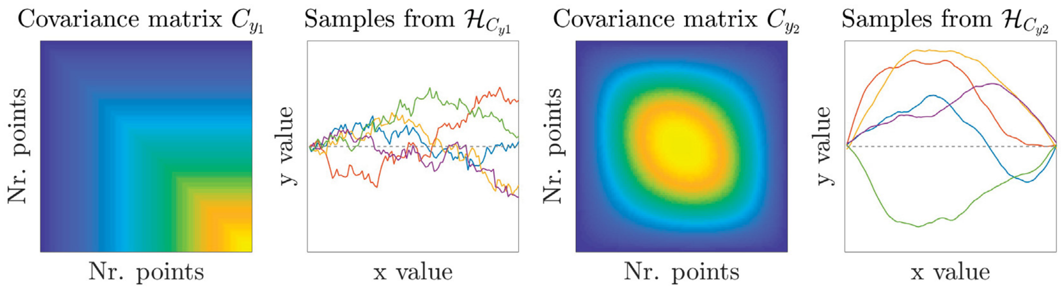


Figure 3.10: A visualization of the results from examples 18 and 19. The constructed C_Y's (panel 1 and 3) are generated using the equation $C_y = (A_c^{-1})(C_X)(A_c^{-1})T$ where C_X is white noise covariance and A_c as in the examples. Panels 2 and 4 contain some realizations of a Gaussian process with the covariance matrices C_{Y1} and C_{Y2}; note that they fulfill the boundary conditions and also that ∇f or Δf respectively look like white noise.

If $\Sigma^{\otimes} = \Sigma \otimes \Sigma$ then $P^{\otimes} = P \otimes P$ and the simplified version of equation 3.50 is

$$C_Y = (A_c^{\psi})F_c(A_c^{\psi})^T \tag{3.51}$$

$$A_c^{\psi} = \left(A_c^T P A_c\right)^+ A_c^T P. \tag{3.52}$$

With this method, one can force for example Brownian motion to be approximately periodic among other things. Note at this point however, that it is always assumed that $C_Y = E[Y \otimes Y^*]$ and $C_X = E[X \otimes X^*]$. If the boundary conditions introduce a non-zero mean, the more complicated terms $C_X = E[X \otimes X^*] - E[X] \otimes E[X]^*$ and $C_Y = E[Y \otimes Y^*] - E[Y] \otimes E[Y]^*$ have to be used.

§ Derivation of the matrix F_c

It is not always straightforward to derive the elements of the matrix F_c from prior knowledge about the function space to be constructed. We will present a scheme to derive a guess for F_c from linear boundary conditions that are in a first step assumed to be stochastic but with mean zero. Represent this as follows.

$$\begin{tikzcd}
& \mathcal{H}_Y \text{ (known)} \arrow[dl, "\text{(known) } A"'] \arrow[dr, "L \text{ (known)}"] & \\
\text{(known)}\mathcal{H}_X & & \mathcal{F}\text{(known)}
\end{tikzcd}$$

$$A_c : \mathcal{H}_Y \mapsto \mathcal{H}_X \oplus \mathcal{F}$$
$$A_c^+ : \mathcal{H}_X \oplus \mathcal{F} \to \mathcal{H}_Y$$

Here $A_c = [A; L]$, $E[X] = E[AY] = 0$ and it is supposed that $\ker A_c = \{0\}$ such that A_c is injective and there exists a left inverse. Then $\exists A_c^+ : A_c^+ A_c = \mathrm{id}_{\mathcal{H}_Y}$ but $A_c A_c^+ \neq \mathrm{id}_{\mathcal{H}_X \oplus \mathcal{F}}$ in general. To use equation 3.51, the terms F, q, S in the following expression need to be determined.

$$A_c \otimes A_c C_Y = \begin{bmatrix} AC_Y A^T & AC_Y L^T \\ LC_Y A^T & LC_Y L^T \end{bmatrix} = \begin{bmatrix} F & q^T \\ q & S \end{bmatrix} = F_c \tag{3.53}$$

The equation $AY = X$ gives information about the covariances of Y by relating them to the known covariances C_X of X. $LY = f$ gives additional information to resolve ambiguities stemming from a nonzero nullspace $N(A)$. It is allowed that f is random but for now let $E[f] = 0$. Writing

$$F_c = \begin{bmatrix} C_X & C_{XF} \\ C_{FX} & C_F \end{bmatrix} \tag{3.54}$$

$$\begin{aligned} C_X &= AC_Y A^T = E[AY \otimes Y^* A^T] = E[X \otimes X^*] && \text{(assumed known)} \\ C_F &= LCYL^T = E[LY \otimes Y^* L^*] = E[f \otimes f^*] && \text{(assumed known)} \\ C_{XF} &= AC_Y L^T = E[AY \otimes Y^* L^*] = E[x \otimes f^*] && \text{(fixed but unknown)} \end{aligned}$$

makes it clear that C_X and C_F are design choices determining the correlation structure of Y and the boundary conditions LY whereas the unknown $C_{XF} = C_{FX}^T$ is determined (not necessarily uniquely) through these choices and needs to be derived. One finds

$$\begin{aligned} C_{FX} = E[f \otimes X^*] = E[LY \otimes (AY)^*] &= E[LA_c^+ \begin{bmatrix} X \\ f \end{bmatrix} \otimes (AA_c^+ \begin{bmatrix} X \\ f \end{bmatrix})^*] \\ &= \underbrace{LA_c^+}_{T} \begin{bmatrix} C_X & C_{XF} \\ C_{FX} & C_F \end{bmatrix} \underbrace{(A_c^+)^T A^T}_{S} \end{aligned} \tag{3.55}$$

In matrix notation, the individual terms have dimensions as listed below. Here by writing $Y[n_Y, 1]$ for example, it is meant that the random vector Y has as realizations elements of $\mathbb{R}^{n_Y}$.

$\dim \mathcal{H}_Y = n_Y$	$\dim \mathcal{H}_X = n_X$	$\dim \mathcal{F} = n_P$	
$Y\ [n_Y, 1]$	$X\ [n_X, 1]$	$f\ [n_P, 1]$	
$C_X\ [n_X, n_X]$	$C_Y\ [n_Y, n_Y]$	$C_F\ [n_P, n_P]$	$C_{FX}\ [n_P, n_X]$
$A\ [n_X, n_Y]$	$A_c\ [n_X + n_P, n_Y]$	$A_c^+\ [n_Y, n_X + n_P]$	$L\ [n_P, n_Y]$
$T\ [n_P, n_X + n_P]$	$T = [T_l, T_r]$	$T_l\ [n_P, n_X]$	$T_r\ [n_P, n_P]$
$S\ [n_X + n_P, n_X]$	$S = [S_u; S_d]$	$S_u\ [n_X, n_X]$	$S_d\ [n_P, n_X]$

Equation 3.55 can be simplified. Rename $C_{FX} = Q$ and let ψ be the linear transposition operator satisfying $\psi M = M^T$ for any matrix M.

$$\begin{aligned} Q &= \begin{bmatrix} T_l & T_r \end{bmatrix} \begin{bmatrix} C_X & Q^T \\ Q & C_F \end{bmatrix} \begin{bmatrix} S_u \\ S_d \end{bmatrix} \\ &= T_l C_X S_u + T_l Q^T S_d + T_r Q S_u + T_r C_F S_d \\ &= T_l \otimes S_u^T C_X + T_r \otimes S_d^T C_F + \left(T_l \otimes S_d^T + T_r \otimes S_u^T\right) Q \end{aligned}$$

This is the case iff

$$(I - T_l \otimes S_d^T \psi - T_r \otimes S_u^T) Q = T_l \otimes S_u^T C_X + T_r \otimes S_d^T C_F$$

and we propose as a rule to infer C_{FX} from C_X and C_F the equation

$$C_{FX} = \left(I - T_l \otimes S_d^T \psi - T_r \otimes S_u^T\right)^+ \left[T_l \otimes S_u^T C_X + T_r \otimes S_d^T C_F\right] \tag{3.56}$$

For the mean zero case with $E[X] = E[Y] = E[f] = 0$, this concludes the inference scheme as C_X, C_F, C_{FX} are then known and one might infer C_Y via $(A_c^+)F_c(A_c^+)^T$. The formulas for this deduction in the presence of a nonzero mean reduce almost completely to what was just derived. This is summarized below.

Define $E[X] = \mu_X \in \mathbb{R}^{nx}, C_X = E[X \otimes X^*] - E[X] \otimes E[X]^*$ and suppose that C_Y, C_F, C_{FX} are defined analogously. In the model

$$\begin{tikzcd} & \mathcal{H}_Y \arrow[dl, "A"'] \arrow[dr, "L"] & \\ \mathcal{H}_X & & \mathcal{F} \end{tikzcd}$$

$$A_c : \mathcal{H}_Y \mapsto \mathcal{H}_X \oplus \mathcal{F}$$
$$A_c^+ : \mathcal{H}_X \oplus \mathcal{F} \to \mathcal{H}_Y$$

with A_c admitting a left inverse assume the following quantities to be known.

$[A, C_X, \mu_X]$ A relates the unknown function space $\mathcal{H}_Y$ to the known space $\mathcal{H}_X$.

$[L, C_F, \mu_F]$ L determines via $LY = f$ the stochastic boundary conditions to be satisfied by elements of $\mathcal{H}_Y$.

Then for the mean of Y one finds

$$\mu_Y = E[Y] = E[A_c^+ \begin{bmatrix} X \\ f \end{bmatrix}] = A_c^+ \begin{bmatrix} \mu_X \\ \mu_F \end{bmatrix}. \tag{3.57}$$

Similarly to before, from $Y = A_c^+[X, f]^T$ it is possible to derive

$$\begin{aligned} C_Y &= (A_c^+) \left(E \left[\begin{bmatrix} X \\ f \end{bmatrix} \otimes \begin{bmatrix} X \\ f \end{bmatrix} \right] - E \begin{bmatrix} X \\ f \end{bmatrix} \otimes E \begin{bmatrix} X \\ f \end{bmatrix}^* \right) (A_c^+)^T \\ &= (A_c^+) F_c (A_c^+)^T \end{aligned} \tag{3.58}$$

$$\begin{aligned} F_c &= E \begin{bmatrix} X \otimes X^* & X \otimes f^* \\ f \otimes X^* & f \otimes f^* \end{bmatrix} - E \begin{bmatrix} X \\ f \end{bmatrix} \otimes \begin{bmatrix} X \\ f \end{bmatrix}^* \\ &= \begin{bmatrix} C_X & C_{FX}^T \\ C_{FX} & C_F \end{bmatrix} \end{aligned} \tag{3.59}$$

Executing the same set of calculations as before and using the notation from equation 3.55, one arrives at

$$\begin{aligned} C_{FX} &= E[f \otimes X^*] - E[f] \otimes E[X]^* \\ &= LA_c^+ E \left[\begin{bmatrix} X \\ f \end{bmatrix} \otimes \begin{bmatrix} X \\ f \end{bmatrix}^* \right] (A_c^+)^T A^T - LA_c^+ E \begin{bmatrix} X \\ f \end{bmatrix} \otimes E \begin{bmatrix} X \\ f \end{bmatrix}^* (A_c^+)^T A^T \end{aligned}$$

$$= \underbrace{LA_c^+}_{T} F_c \underbrace{(A_c^+)^T A^T}_{S} . \tag{3.60}$$

The best guess for C_{FX} emerges as

$$C_{FX} = \left[I - T_l \otimes S_d^T \psi - T_r \otimes S_u^T\right]^+ \left(T_l \otimes S_u^T C_X + T_r \otimes S_d^T C_F\right). \tag{3.61}$$

As the final result, one recovers $\mu_Y = A_c^+[\mu_X, \mu_F]^T$ and $C_Y = (A_c^+)F_c(A_c^+)^T$ where F_c is the matrix from equation 3.59 with C_{FX} as guessed by equation 3.61. Figure 3.11 illustrates some results.

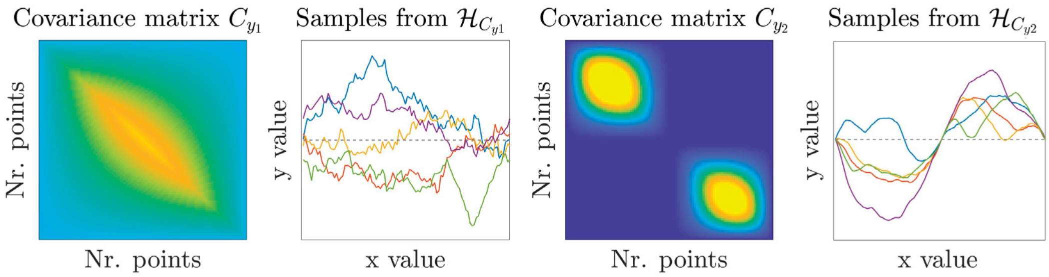


Figure 3.11: Results from the construction method for generating kernels satisfying boundary conditions. The first two panels feature a space of functions which are integrals of white noise that take the same values at the start, middle and ending positions although that function value is random. The second pair of panels features a space of functions the inhabitants of which have derivative $+1$ in the middle position additionally to starting, ending and going through zero. Their second order derivatives behave approximately as white noise.

3.3 Basic geodetic applications

This section is devoted to an exposition of practical applications of stochastic processes and RKHS to problems routinely encountered in geodesy. Its purpose is to provide a hands-on tutorial, expose strengths and weaknesses of RKHS based processing approaches, and to collect a set of test-problems with different flavors. Although mathematical rigor will not constitute one of this sections main concerns, the experiences made and conclusions drawn should be sufficiently representative to allow an intuitive assessment about which type of practical problems can be posed and solved in an RKHS framework and which complications may arise in doing so. The contents of this section are ordered roughly with respect to their difficulty. Therefore stochastic processes $X : T \to L^2(\Omega), \dim T = 1$ are considered first and it is shown how they can provide models for real world processes driven by random inputs. Afterwards the focus will shift to modelling and analysis of random fields, i.e. to assignments $X : T \to L^2(\Omega), \dim T > 1$ before the topic of Hilbert space embeddings of probability distributions is brought up again.

3.3.1 Modelling with stochastic processes

The Wiener process is also known as Brownian motion. Since it is the integral of white noise, it can be considered as the archetypal process representing a system in which measurement noise adds over time or space depending on the definition of the index set T. As such, it also provides a stochastic model for the data gathered during either leveling or total station measurements and the well known rules for distributing errors in leveling emerge as simple consequences of the abstract spline equations. An abstract spline model can be specified for the task of optimally estimating deformations based on total station measurements that are a superposition of deformation, pure noise and atmospheric influences. As soon as measurements to more than one prism are involved, the situation admits nontrivial statements that rest on the measurement operator in the abstract spline formulation being integration along a line rather than evaluation. It is not difficult to extend the calculations slightly and make the estimation procedure applicable to highly dynamic movement processes by introducing a more sophisticated stochastic model. Trajectory estimation of objects based on total station measurements can then be done in virtually the same way as deformation estimation.

Example 20 Recall that the Wiener process $X : T \times \Omega \to \mathbb{R}$ on the interval $T = [0, L] \subset \mathbb{R}$ as integrated white noise has the kernel $\min(s, t)$. A discrete version of X on the set $T = \{1, ..., n\}$ with $X(t) = \sum_{j=1}^{t} \epsilon_j, \ \epsilon_j \sim \mathcal{N}(0, \sigma_0^2)$ was shown to be a reasonable stochastic model for the distribution of errors in leveling on page 84. As a matter of fact, it was even shown in subsection 3.2.3 that the covariance matrices generated by this type of covariance function arise naturally from demanding the leveling error to be zero at the beginning and to be composed of uncorrelated noise that adds up over the course of measuring along the path, i.e. $X(0) = 0$ and $\partial_t X(t) = \epsilon(t)$ with $\epsilon(t)$ white noise.
Now consider the error at loop closure $X(L)$. Since for typical non-sophisticated leveling tasks one starts and ends at the same location, the realization x_L of $X(L)$ is known and one may guess from this the distribution of errors along the whole path T. The interpolating spline problem

$$\sigma_x(\cdot) = \underset{x \in \mathcal{H}_K, e_L x = x_L}{\operatorname{argmin}} \quad \|x\|_{\mathcal{H}_K}$$

for the evaluation functional $e_L : f \mapsto f(L) \ \forall f \in \mathcal{H}_K$ and $\mathcal{H}_K$ the Wiener process RKHS with RK $K(s, t) = \min(s, t)$ provides this best guess in form of σ_x. The explicit solution is

$$\begin{aligned} \sigma_x(\cdot) &= \lambda e_L K(\cdot, \cdot) \qquad\qquad \lambda = K^{-1}(L, L) x_L \\ &= \min(\cdot, L) \frac{x_L}{L} \end{aligned}$$

which is nothing but a function that is linear from $t = 0$ to $t = L$ and then takes on the value of the constant x_L. Comparing to standard references, this is exactly the rule for distribution of leveling errors in practice [209, p. 272]. ■

Example 21 Reconsider the clamped elastic string under load on an interval $T = [0, 1]$ from subsection 2.3.2. Denote by $w(\cdot)$ a smoothly distributed weight function that is supposed to lie in an RKHS $\mathcal{H}_W$ of functions and let $x \in \mathcal{H}_X$ be the displacement induced by w; $\partial_t^2 x = w \in \mathcal{H}_W$. Then the likelihood of x is quantified by

$$\|x\|^2_{\mathcal{H}_X} = \|\partial_t^2 x\|^2_{\mathcal{H}_W}$$

and $B = \partial_t^2, \mathcal{H}_B = \mathcal{H}_W$ are properly physically motivated choices for the energy operator and its associated space. The nullspace of B are the constant and linear polynomials, both of whose coefficients are necessarily zero due to the boundary conditions $x(0) = x(L) = 0$. Assume the displacement has been noisily observed at locations $\{s_i\}_{i=1}^{n_s} \subset T$ where the measurement noise has mean zero and covariance matrix Σ_N. Then the abstract smoothing spline

$$\sigma_x = \underset{x \in \mathcal{H}_X}{\operatorname{argmin}} \ \|Ax - a\|^2_{\Sigma_N} + \|\partial_t^2 x\|^2_{\mathcal{H}_W}$$

with A evaluation at $\{s_i\}_{i=1}^{n_s}$ provides a best guess for the displacement x based on the measurements a and knowledge of the load configurations typical structure. At the same time

$$\sigma_W = \underset{w \in \mathcal{H}_W}{\operatorname{argmin}} \ \|AL_G w - a\|^2_{\Sigma_N} + \|w\|^2_{\mathcal{H}_W}$$

with L_G being integration against Greens function , $(L_G f)(t) = \int_0^1 G(s,t) f(s) ds$, provides a best guess for the underlying load distribution. For specific choices of parameters, results can be seen in figure 3.12

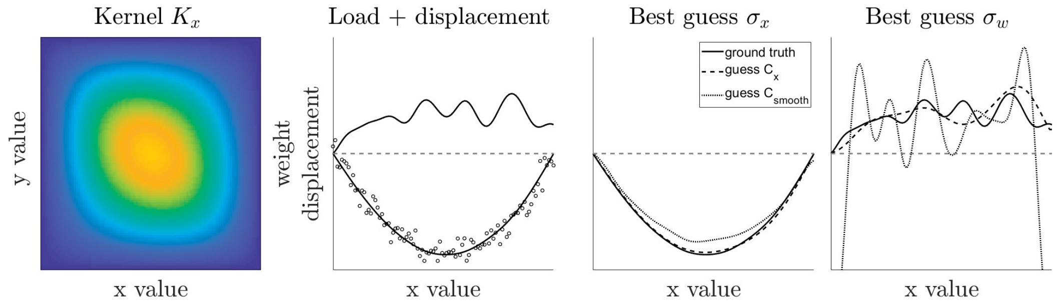


Figure 3.12: The kernel K_X of $\mathcal{H}_X$ such that $\partial_t^2 x = w \in \mathcal{H}_W$ for some smooth load w is featured in panel 1. True underlying load and displacement together with noisy observations serving as inputs to the estimation is plotted in panel 2 whereas the last two images show the best guesses for displacement and load based on our physical model and a generic smooth squared exponential kernel. Ground truth is plotted for comparison.

The noisily observed loaded string is only a toy problem without any direct implications for geodesy but the general approach of optimally estimating a physical quantity based on unreliable measurements and understanding of the underlying differential equations touches upon the interaction between measurements and simulations of physical systems. It has received heightened attention since the advent of highly performant FEM-programs and measurement instruments which are capable of generating geometrically dense samples, see for example [177] who considers this in the context of terrestrial laserscanning. ■

Example 22 A total station measures distances and angles by employing electromagnetic waves with wavelengths approximately in the spectrum of visible light. They are emitted by the instrument and reflected by prisms mounted on the object whose coordinates are to be determined. Presuming the total station's position to be known, a sequence of measurements to an object leads to a noisy sequence of measurements of three dimensional positions of that object. Before considering in more detail the nature of the noise, the following two abstract spline problems arise naturally when dealing with time series of coordinates.

I) Suppose that the measurements have been processed to yield a sequence of x coordinates over which white noise of known variance is superimposed leading to the data vector $a = \{x_t + \epsilon_{t_j}\}_{j=1}^n$. If the true behavior of the x coordinates is smooth, that is $x \in \mathcal{H}_{K_X}$ with K_X some smooth kernel, one may split signal and noise by calculating the smoothing spline

$$\sigma_x = \underset{x \in \mathcal{H}_{K_X}}{\operatorname{argmin}} \ \|Ax - a\|_\Sigma^2 + \|[x]\|_{\mathcal{H}_{K_X}}^2$$

where A is the operator of evaluation at $t_1, ..., t_n$, $[\cdot]$ is the canonical projection that annihilates constants, $a \in \mathbb{R}^n$ is the data and Σ is the noise covariance matrix with entries $(\Sigma)_{ij} = \sigma_0^2 \delta_{ij}$. Depending on $\mathcal{H}_{K_X}$, the classical mean estimator $\sigma_x = n^{-1} \sum_{j=1}^n a_j$ can be recovered for degeneratedly smooth kernel choices K_X. In all other cases, one arrives at more complex, time varying estimators. This is illustrated in figure 3.13.

II) Suppose one has measured at certain times the three-dimensional trajectory of an object that is rapidly moving to such an extent that the influence of the noise is negligible. If the goal is to estimate the full trajectory $x(\cdot) : T \to \mathbb{R}^3$, this can be formulated as the abstract interpolating vector spline.

$$\sigma_x = \underset{x \in A^{-1}a}{\operatorname{argmin}} \ \|[x]\|_{\mathcal{H}_{K_X}}^2$$

A : The measurement operator consisting of linear evaluation functionals at times $t_1, ..., t_n$ for all three coordinates.

a : The data containing x, y, z coordinates as a vector in $\mathbb{R}^{3n}$.

$[\cdot]$: The canonical projection annihilating constants and linear functions.

$\mathcal{H}_{K_X}$: A Hilbert space of vector-valued functions $T \to \mathbb{R}^3$ assembled as the direct sum of the three component Hilbert spaces $\mathcal{H}_{X_c} \oplus_e \mathcal{H}_{Y_c} \oplus_e \mathcal{H}_{Z_c}$.

The canonical projection allows for a drift in expected value of the objects position. If noise would be incorporated into the model, the behavior would be similar as in a). In the limit, infinitely strong regularity requirements on $\mathcal{H}_{K_X}$ and high noise variances lead to the best guess for the trajectory being a straight line; see figure 3.13.

There are more realistic models for the noise in total station measurements than white noise. During the propagation through the atmosphere, the transmitted signal is delayed compared to its propagation through vacuum. That delay is highly variable in time and related to the refractive index of the medium, which in case of the

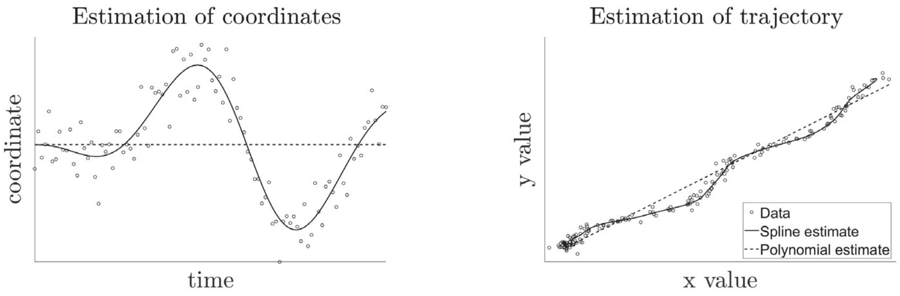


Figure 3.13: The left panel shows data from a total station and two associated smoothing splines; one which presupposes the x-coordinates to vary smoothly and one that presupposes that they do not change at all. Similar cases are also plotted in the right panel which exhibits the result of a smoothing spline estimation of trajectory based on a sequence of three dimensional coordinates. The data plotted on the right come from a real world experiment, in which trajectories of a skier were recorded, see [23] .

atmosphere itself mainly depends on temperature, pressure and water vapor content. Then one invokes an argument similar to the one put forward during the analysis of leveling noise and claims that the incremental delays experienced by the wave on a short part of its path are second order stationary and their integral along the whole propagation path forms the actual total delay observed in the measurements.
The overall variance is then composed of an instrument specific pure noise part that subsumes e.g. the relatively distance invariant quantization errors and electronic crosstalk and an atmospheric part whose variance increases with the length of the propagation path. A reasonable stochastic model is then that the measurements m of coordinate changes are superpositions of uncorrelated noise n, atmospheric effects q and real deformations d. Associate to each of these terms an RKHS and set

$$\mathcal{H}_M = \mathcal{H}_S \oplus \mathcal{H}_Q \oplus \mathcal{H}_N.$$

$\mathcal{H}_Q$ consists of functions that are line integrals through the field of refraction changes r which is itself supposed to be a spatial random field associated to the RKHS $\mathcal{H}_R$ such that $\mathcal{H}_Q = \{Lr(\cdot) : r(\cdot) \in \mathcal{H}_R\}$ where $L : \mathcal{H}_R \ni r \mapsto \int_{s_0}^{\cdot} r(s)ds \in \mathcal{H}_Q$ is the operator of line integration from the instruments position s_0 to an arbitrary position. Clearly, this model is too simple to capture all of the intricacies and interdependencies encountered in total station measurements and therefore more a toy example than state of the art. The following situations are interesting from a theoretical perspective.

III) If total station measurements have been made towards a set of stable prisms located at positions $\{s_j\}_{j=1}^{n_s}$ then $\mathcal{H}_M = \mathcal{H}_Q \oplus \mathcal{H}_N$ and one may try to tackle the tomography problem of inferring the spatial distribution of refractive index changes $r(\cdot) : \mathbb{R}^3 \to \mathbb{R}$ by solving the abstract smoothing spline problem

$$\sigma_r = \underset{r\in\mathcal{H}_R}{\operatorname{argmin}} \ \|ALr - a\|^2_{A\mathcal{H}_N} + \|r\|^2_{\mathcal{H}_R}$$

$$A: \text{ Linear operator of evaluation at locations } \{s_j\}_{j=1}^{n_s}$$

L : Linear operator of line integration from s_0, $(Lr)(\cdot) = \int_{s_0}^{\cdot} r(s)ds$

a : Data of the n_s measured distance changes

$A\mathcal{H}_N$: The Hilbert space of n_s-vectors with the inner product

$$\langle n_1, n_2 \rangle_{A\mathcal{H}_N} = \sigma_0^{-1} \sum_{j=1}^{n_s} (n_1)_j (n_2)_j$$

$\mathcal{H}_R$: A Hilbert space of smooth spatial functions.

Some results of this procedure for a simulated dataset are plotted in figure 3.14. Note that obviously the quality of the estimation depends strongly on a physically justifiable choice of the space of functions $\mathcal{H}_R$. More information about this topic can be found in the comments discussing stochastic models for atmospheric effects observed in terrestrial radar interferometry in subsection 5.2.1.

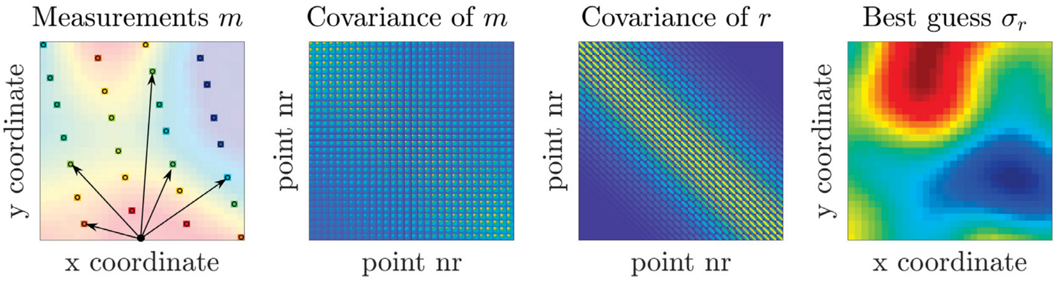


Figure 3.14: The solution of the abstract smoothing spline problem is an estimator σ_r for the underlying field of refractive indices r. The black arrows in the left panel signify that each measurement $m(s_j)$ is a line integral from the instrument position to the prism at s_j through r. There is therefore a noticeable difference between the correlation structure of the refraction indices and the correlation structure of the atmospheric effects induced by them.

IV) if n_s prisms at positions $\{s_j\}_{j=1}^{n_s}$ have been measured over the time $T = \{t_1, ..., t_{n_t}\}$ by the same total station, then the measurements m are vector valued with

$$\mathcal{H}_M = \mathcal{H}_D \oplus \mathcal{H}_Q \oplus \mathcal{H}_N$$

$$\mathcal{H}_D = \bigoplus_{j=1}^{n_s} {}_i\mathcal{H}_{D_j} \qquad \mathcal{H}_Q = \bigoplus_{j=1}^{n_s} {}_i\mathcal{H}_{Q_j} \qquad \mathcal{H}_N = \bigoplus_{j=1}^{n_s} {}_i\mathcal{H}_{N_j}.$$

Each of the component spaces $\mathcal{H}_{D_j}, \mathcal{H}_{Q_j}, \mathcal{H}_{N_j}$ represents deformations, atmospheric effects or noise at location s_j and an element of e.g. $\mathcal{H}_{D_j}$ is an n_t vector containing a time series of deformations for that location. The non-orthogonal direct sum was chosen to reflect that the quantities may well be nontrivially correlated between different points.

The problem of jointly estimating the time series of deformations at each location seems to be not immediately falling into the abstract spline framework but it can be cast as a simple smoothing spline, a vector spline or as a tensor spline. To see this, consider the new indexset $T_s := \sqcup_{j=1}^{n_s} T$ formed by the disjoint union (see e.g. [2, p. 14] for a definition) of n_s copies of T. A measurement m is then a function on

T_s and the kernels for the spaces $\mathcal{H}_X, X = D, Q, N$ are spatiotemporal covariance functions $K_X((s_i,t_i),(s_j,t_j)) = E[X(s_i,t_i)X(s_j,t_j)]$. The best guess for d is the smoothing spline

$$\sigma_d = \underset{d\in\mathcal{H}_D}{\operatorname{argmin}} \ \|Ad - a\|^2_{\mathcal{H}_Q \oplus_i \mathcal{H}_N} + \|d\|^2_{\mathcal{H}_D}$$

with as usual A being evaluation and a the spatiotemporal dataset. If one would consider a whole spatiotemporal field of (potential) measurements and factorize $\mathcal{H}_M$ into $\mathcal{H}^s_M \otimes \mathcal{H}^t_M$ which is evaluated only at the times T and locations $\{s_j\}_{j=1}^{n_s}$, then one would recover a tensor spline. Similarly, if one interpreted the setup as producing n_s vectors with n_t entries then it is possible to formulate a vector spline problem to the same effect.

In any case, all three problems can easily be solved with the usual equations. The solutions for the case of three points and a specific set of data and assumptions are recorded in figure 3.15. Note that the results would have been different if the three time series were not processed jointly. ■

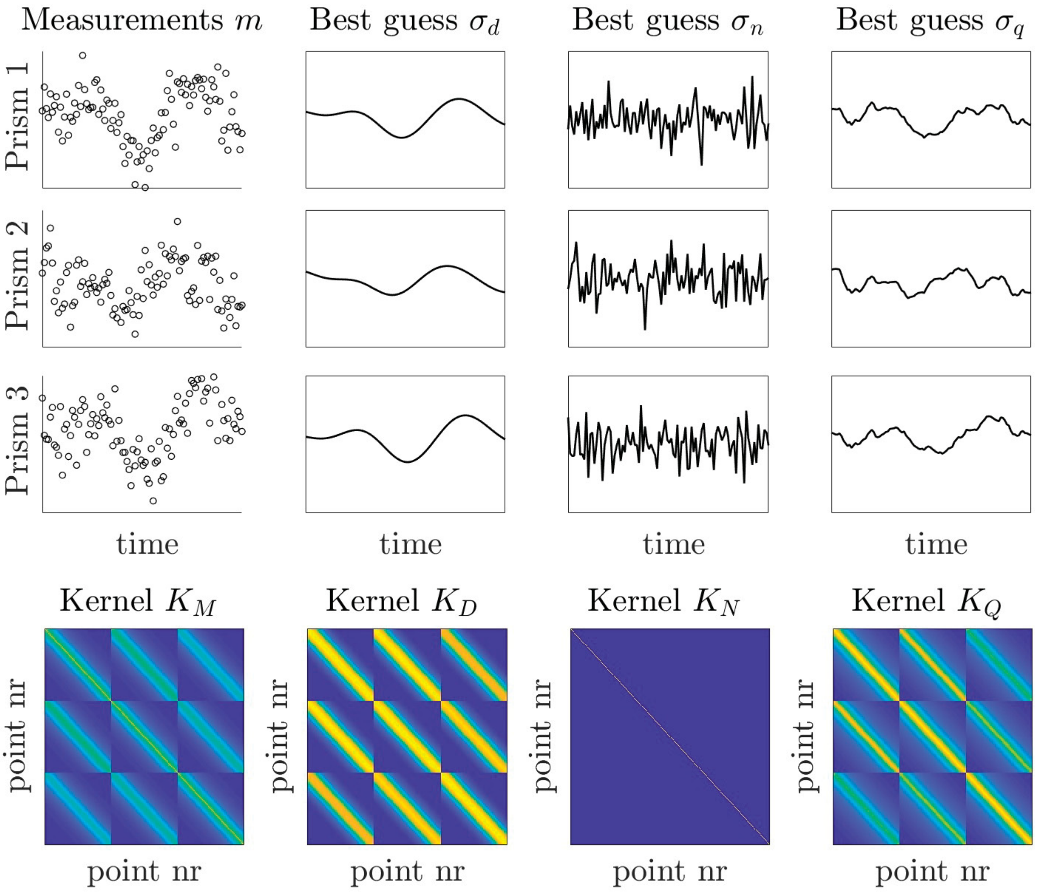


Figure 3.15: A signal separation problem in which several mutually correlated time series are to be decomposed into different parts. The top collection of panels collects the results whereas the lower part showcases the nontrivial correlations between the measurements at different times and at the three different locations.

Interpolation and smoothing of vector fields $f : T \to \mathbb{R}^d$ is possible in exactly the same way. Just consider an RKHS $\mathcal{H}_K$ of functions on $\sqcup_{j=1}^d T$ and solve

$$\sigma_x = \underset{x \in \mathcal{H}_K}{\operatorname{argmin}} \ \|Ax - a\|_{\mathcal{H}_A}^2 + \|Bx\|_{\mathcal{H}_B}^2$$

or the vector spline problem

$$\sigma_x = \underset{(x_1,x_2) \in \mathcal{H}_{X_1} \oplus \mathcal{H}_{x_2}}{\operatorname{argmin}} \ \|A(x_1, x_2) - a\|_{\mathcal{H}_A}^2 + |x_1|_{\mathcal{H}_{X_1}}^2 + |x_2|_{\mathcal{H}_{X_2}}^2$$

for appropriate choices of measurement operators, energy operators or corresponding seminorms and data. An exemplary result can be seen in figure 3.16 and could represent a field of deformations interpolated for example from GNSS or total station measurements.

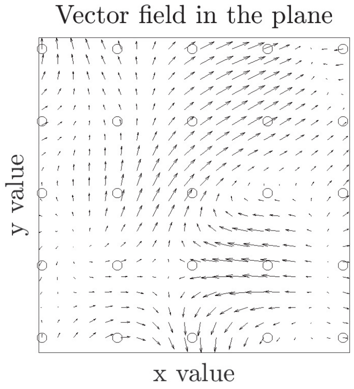


Figure 3.16: Illustration of vector field interpolation. Black circles mark observed vectors.

It is possible to include special relationships as for example discontinuities or boundary conditions via the kernel construction method outlined in subsection 3.2.3. In principle, it is also possible to perform estimation on the sphere, the torus or other arbitrary manifolds $\mathcal{M}$ by constructing kernels satisfying certain boundary conditions (see figure 3.17).

Alternatively it is possible to consider spaces of functions $\mathcal{H}$ on $\mathbb{R}^d \supset \mathcal{M}$ and work in the larger auxiliary space $\mathbb{R}^d$ before restricting any functions $f \in \mathcal{H}$ to $f \ |_{\mathcal{M}}$ to get the trace of a spline on a manifold [21, pp. 135-145]. However, both of these methods are computationally inconvenient and if, like for the sphere, orthonormal systems and kernels constructed on them are available, these are used instead.

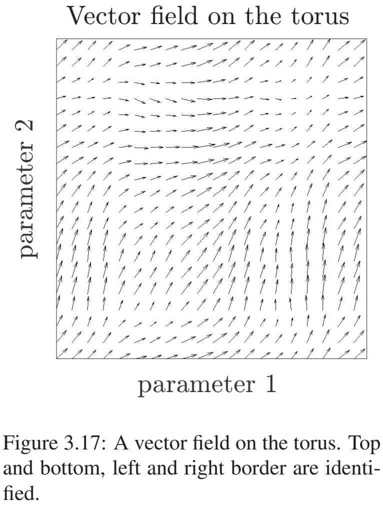


Figure 3.17: A vector field on the torus. Top and bottom, left and right border are identified.

This was investigated for example by Wahba [202] and Swarztrauber [147] who came up with solutions to estimating vorticity and divergence of vector fields on the sphere. They also provide interpolating and smoothing spline procedures that are related to the Laplacian — an energy operator associated to the Poisson equation governing the gravity fields behavior [203, pp. 28-30].

The model for atmospheric influences on total station measurements contains already first elements of a tomography problem. The connection between atmospheric influences and refractive index could in principle be used to solve the inverse problem of optimally estimating the three-dimensional spatial distribution of refractive indices based on total station measurements. The estimation of spatial quantities for which measurements are available only in the form of aggregated values such as line integrals or averages is quite recurrent in geodesy and another example for this type of problem is given by GPS based estimation of atmospheric quantities.

Simple spatial estimation can also be used to create smooth maps for fingerprinting-based WLAN positioning systems. Figure 3.18 showcases the effect of kernel-smoothing on received signal strengths at different locations in an indoor area on Hoenggerberg-campus, ETH. For more details, consult [213].

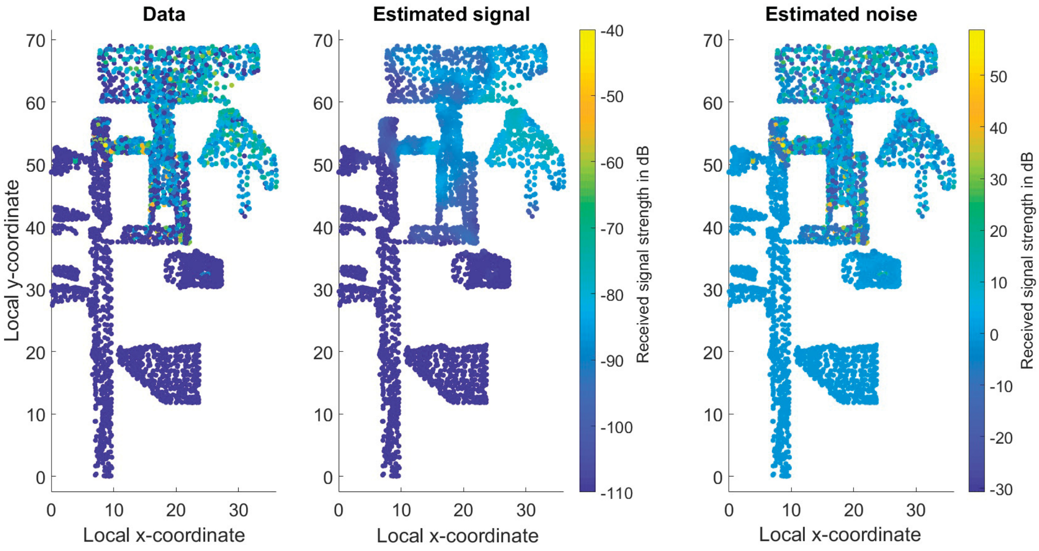


Figure 3.18: The left panel shows the received signal strength of a W-LAN access point. Several outliers and implausible values are visible. If kernel smoothing (squared exponential kernel for the signal, white noise kernel for the noise) is applied to the dataset, the best estimates of the underlying smooth signal and noise seem intuitively reasonable.

3.3.2 Statistical testing

Some geodetic problems like the comparison of different samples from the same instrument are of a low dimensional nature but still interesting. Due to the prevalent focus on linearity and correlation in RKHS-based approaches, problems requiring nonlinear measures of dependence between random variables and manipulation of non-Gaussian probability density functions have been neglected so far. This is rectified by using the RKHS embedding of probability distributions. The Hilbert-Schmidt independence criterion (HSIC) is compared to the correlation coefficient by application to a simple example. The HSIC is employed to answer geodetically motivated questions pertaining to the dependency of measurements on auxiliary variables. Other than in this subsection, however, the view of RKHS as function spaces with physical meaning will stay the predominant perspective.

Suppose two different instruments A and B of the same type are used to produce a series of measurements. Denote by $X : \Omega \to \mathbb{R}, Y : \Omega \to \mathbb{R}$ the random variables with probability distributions P, Q associated to measuring with A and B respectively. If two samples $x = \{x_1, ..., x_m\}$ and $y = \{y_1, ..., y_n\}$ of realizations of X and Y are given, the question may arise if A and B have been manufactured alike or if one of the instruments is suffering from a malfunction detectable in the statistical distribution of the measurements. In short, the task is to infer from $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ if $P \neq Q$. This task has been analyzed in [82] where a two sample test is derived on the basis of RKHS embeddings of distributions.

Suppose $K(\cdot, \cdot)$ is a positive definite kernel and denote by $\mathcal{H}_K$ the corresponding RKHS. One may try to find a smooth function $f \in \mathcal{H}_K$ that distinguishes the sample x and y maximally and designate the mean of $f(x) - f(y)$ as an indicator of distance between x and y. This leads to the maximum mean discrepancy

$$\begin{aligned} MMD = \sup_{\|f\|_{\mathcal{H}_K} \leq 1} E_P[f(X)] - E_Q[f(Y)] &= \sup_{\|f\|_{\mathcal{H}_K} \leq 1} \langle \mu_P, f \rangle_{\mathcal{H}_K} - \langle \mu_Q, f \rangle_{\mathcal{H}_K} \\ &= \|\mu_P - \mu_Q\|_{\mathcal{H}_K} \end{aligned} \tag{3.62}$$

where $\mu_P = E_P[K(X, \cdot)], \mu_Q = E_Q[K(Y, \cdot)]$ are the kernel embeddings of P and Q. The element $f \in \mathcal{H}_K$ leading to the maximum discrepancy is $f = \|\mu_P - u_Q\|_{\mathcal{H}_K}^{-1}(\mu_P - \mu_Q)$ and an unbiased estimate of the MMD is given by [82]

$$\begin{aligned} \widehat{MMD} = & \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} K(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K(y_i, y_j) \\ & - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} K(x_i, y_j) \end{aligned} \tag{3.63}$$

Let $m = n$. Then for the hypotheses $H_0 : P = Q$ and $H_1 : P \neq Q$ the distributions of the test statistic $\widehat{MMD}$ can be derived. Demanding an acceptance region leading to a type I error of α leads to the thresholding test

$$\widehat{MMD}^2 < \left(4 \sup_{x_i, x_j} K(x_i, x_j) \, [m]^{-1/2} \sqrt{\log(\alpha^{-1})} \right) \tag{3.64}$$

where the nullhypothesis is rejected if statement 3.64 is untrue. See [82] for the details. This test is illustrated schematically in figure 3.19 thereby concluding the example.

Suppose now that radar interferometric measurements are made over a precisely known stable length l with wavelength $\lambda \approx 17.6\text{mm}$. As is shown later in chapter 5, the difference between received and sent phase is

$$\Delta\varphi = \mathrm{mod}\left(\frac{4\pi}{\lambda} l_{\text{optic}}, 2\pi \right) \qquad \Delta\varphi \in [0, 2\pi]$$

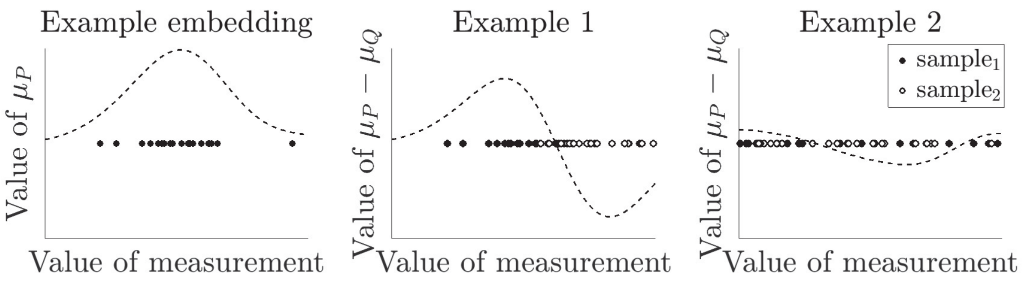


Figure 3.19: An example of the empirical kernel embedding — since the smooth squared exponential kernel was used, the result looks like a kernel density estimate but should not be confused with one. The two panels on the right side showcase in dashed lines the elements $\mu_P - \mu_Q$ which are used to assemble the test statistic $MMD = \|\mu_P - \mu_Q\|_{\mathcal{H}_K}$. The more $\mu_P - \mu_Q$ deviates from zero, the bigger the test statistic implying a higher tendency towards accepting $P \neq Q$ - this coincides with the ground truth underlying the simulations (Gaussian vs uniform with different means in middle panel, two identical uniform distributions in right panel).

and the optical path length l_{optic} depends on meteorological parameters as recorded on page 231. For the sake of simplicity, we consider here only the partial pressure e of water vapor the temperature T. Given synthetic noisy observations of $\Delta\varphi, T, e$ that were generated using this functional relationship, we want to use the Hilbert Schmidt independence criterion to diagnose if $\Delta\varphi, T, e$ are statistically independent. For reasons of contrast, also a dummy variable z completely independent of the others has been included, see figure 3.20.

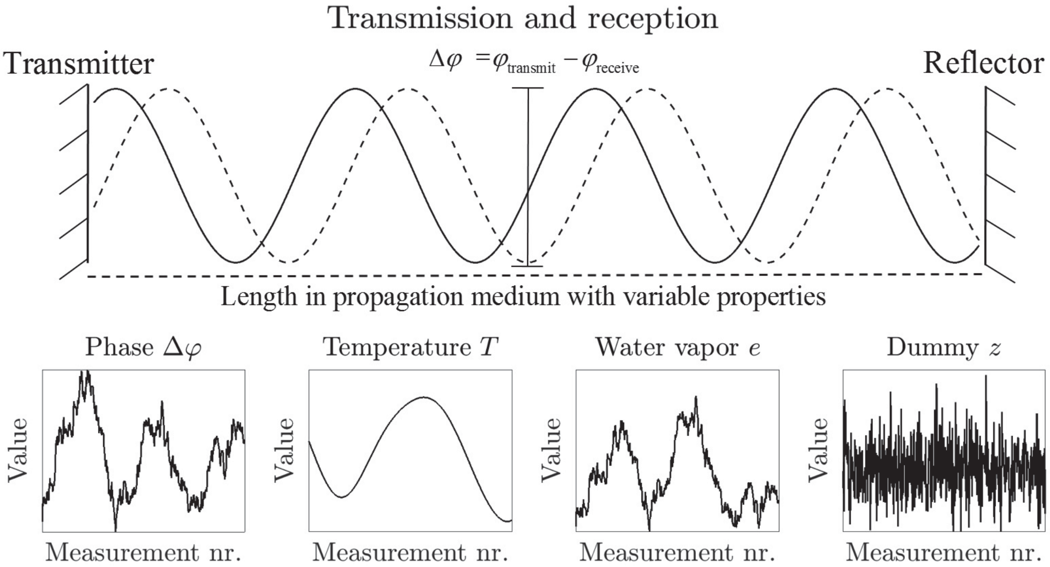


Figure 3.20: An illustration of the radarinterferometric toy example. Given the measurements for the phases, temperatures, water vapor content and an unrelated dummy variable, HSIC is used to diagnose dependencies between the four sets of measurements plotted in the second row. The results are reported in figure 3.21.

Given $\{\Delta\varphi_i\}_{i=1}^n, \{T_i\}_{i=1}^n, \{e_i\}_{i=1}^n, \{z_i\}_{i=1}^n$ and kernels $K_X(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}, X \in \{\Delta\varphi, T, E, z\}$, calculating HSIC between for example $\Delta\varphi$ and T is akin to estimating

$$\|C_{\Delta\varphi T}\|_{HS} = \sqrt{\sum_{i=1}^{\infty} \sigma_i^2(C_{\Delta\varphi T})} \qquad C_{\Delta\varphi T} = \sum_{i=1}^{n} K_{\Delta\varphi}(\Delta\varphi_i, \cdot) \otimes K_T(t_i, \cdot)^*$$

for example via [83]

$$HSIC = \left[\frac{1}{n-1}\right]^2 \operatorname{tr}\left(K_{L_{\Delta\varphi}} H L_{K_T} H\right) \qquad (H)_{ij} = \delta_{ij} - m^{-1}$$

which is essentially a comparison metric evaluating how strongly $(L_{K_{\Delta\varphi}})_{ij} = K_{\Delta\varphi}(\Delta\varphi_i, \Delta\varphi_j)$ and $(L_{K_T})_{ij} = K_T(t_i, t_j)$ are correlated as vectors in $\mathbb{R}^n \otimes \mathbb{R}^n$. Upper bounding the type I error by α and setting $H_0 : \Delta\varphi \coprod T, H_1 : \Delta\varphi \not\coprod T$, a threshold can be determined [83]. The results are documented in figure 3.21.

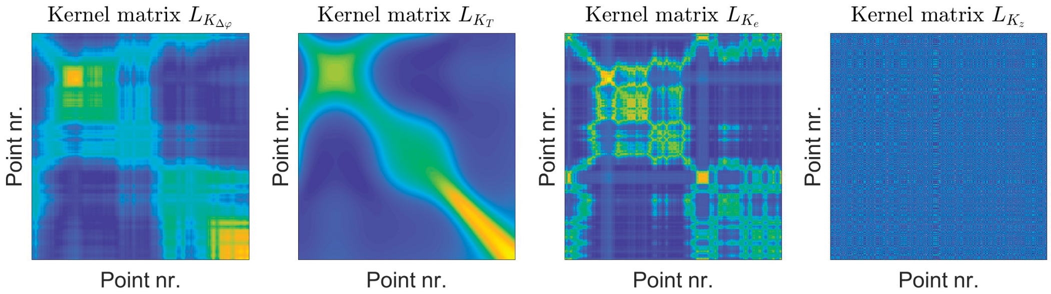


Figure 3.21: The different kernel matrices of the observations. Notice that L_{K_z} seems to be rather unrelated to the other three as confirmed via a small inner product $\langle HL_{K_z}, HL_{K_X}\rangle_F \propto HSIC, X \in \{\Delta\varphi, T, e\}$.

For the HSIC values of the pairs $(\Delta\varphi, T), (\Delta\varphi, e), (\Delta\varphi, z)$ one finds after 100 simulations the empirical mean values

$$HSIC(\Delta\varphi, T) = 0.04 \qquad HSIC(\Delta\varphi, e) = 0.03 \qquad HSIC(\Delta\varphi, z) = 0.0001$$

of which the last one is significantly below the other two. We reject $\Delta\varphi \coprod T$ and $\Delta\varphi \coprod e$ and thereby come to the conclusion that the dependence of $\Delta\varphi$ on temperature and content of water vapor is much stronger than explainable by random chance. This is in accordance to what the data generation process suggests.

Chapter 4

Theory of kernel inference

This is the last chapter in which new theoretical content will be presented. We will focus primarily on deriving a systematic approach to kernel inference, i.e. the determination of covariance functions and their corresponding RKHS from data. For the sake of a more thorough comprehension, positive definite functions and positive definite kernels K are first investigated in their own right independently of their potential relation to elements of an RKHS $\mathcal{H}_K$. The theory of group (C^*)-algebras is one of the settings in which positive definiteness admits an easy characterization in terms of group algebraic multiplication and a $*$-structure. Bochner's theorem provides necessary and sufficient conditions for a function to be the covariance function of a second order stationary stochastic process. It makes use of the abstract notion of a Fourier transform on locally compact abelian groups and relates positive definiteness of a function to positivity of its Fourier transform therefore opening up routes to a formulation of kernel inference as an optimization problem with constraints on its Fourier transform. The choice of correct distance measure to minimize is as difficult as the derivation of an actual algorithm capable of minimizing it within the feasible region determined by the constraints. Both questions receive some attention and their resolution comes in the form of a type of optimization problem that has some resemblance to semidefinite programs, or SDP for short. While a brief description of their numerical implementation is unavoidable, the main interest of this chapter is SDP-based kernel inference and the analysis and solution of other geodetically motivated problems admitting a formulation in terms of spectral quantities.

4.1 Group algebras

During the definition of second order stationary stochastic processes as those possessing a translation invariant covariance function $K(\cdot,\cdot), K(t_i+\tau, t_j+\tau) = K(t_i,t_j) \ \forall t_i, t_j, \tau \in T$, use was made of an additional group structure on T that allowed addition of its elements. The implications of a group structure on T for the vector space of functions $T \to \mathbb{C}$ have so far been neglected together with the existence of a natural multiplication of functions given either by pointwise multiplication or convolution. Both aspects are interrelated and function algebras with an involution on locally compact abelian groups have irreducible one dimensional representations in terms of special elements called group characters. The set of characters can be endowed with a structure that mimics the original group. It is termed the dual group and is a necessary ingredient for the abstract Fourier transform, for which instructive applications from signals and systems theory will be collected.

4.1.1 (Locally) Compact groups and their algebras

A group T is called compact if it has a topology and every open cover of T includes a finite open subcover. It is called locally compact and abelian if its group operation is commutative and every group element has a closed compact neighborhood w.r.t. the topology τ. Complete spaces $L^1(T)$ of functions $f : T \to \mathbb{C}$ that satisfy an L^1-norm finiteness condition and are endowed with the usual convolution of functions $f * g(t) = \int_{\tau \in T} f(\tau)g(t-\tau)d\tau$ as multiplication go by the name of group algebras. They are of interest because the additional structure of a multiplication of functions on T is necessary to algebraically distinguish positive definite functions from generic ones by e.g. relating positive definite functions to squares $f * f^* \in L^1(T)$. Ideals of an algebra $L^1(T)$ are subspaces that are also absorbing with respect to multiplication and the set of maximal ideals forms a particularly convenient decomposition if $L^1(T)$ into one dimensional subspaces of T is a locally compact abelian group.

Definition 4.1.1 Let T be a set and a binary operation $T \times T \to T$ be given. This operation is typically denoted by $*$ or $+$ depending on its properties. However, often no special symbol is used at all and the operation is implicitly assumed to be applied to two elements $t_1, t_2 \in T$ if they are written in juxtaposition as $t_1 t_2$.

I The set T together with the binary operation is called a group if [156, p.25] $\forall s, t, u \in T$ it holds that i) $(st)u = s(tu)$, ii) $\exists e \in T : te = et = t$, iii) $\exists\, t^{-1} \in T : t^{-1}t = tt^{-1} = e$.

II If the group operation is commutative and therefore satisfies iv) $st = ts \;\; \forall s, t \in T$ then T is called an abelian group [156, p. 28].

III The group T is called topological if a topology $\mathcal{O}$ of open sets is defined on T with respect to which the group operation $(s,t) \mapsto st$ and inversion $t \mapsto t^{-1}$ are continuous for all $s, t \in T$ [95, p.16] and T as a topological space is Hausdorff, that is $\forall s, t \in T \;\exists\, O_s, O_t \in \mathcal{O} : s \in O_s, t \in O_t$ and $O_s \cap O_t = \emptyset$ [208, p. 85].

IV A topological group T is called compact if T is compact as a topological space, i.e. if $\bigcup_{j \in \mathcal{J}} O_j \supset Y \Rightarrow \exists \mathcal{J}' \subset \mathcal{J}, |\mathcal{J}'|$ finite, such that $\bigcup_{j \in \mathcal{J}'} O_j \supset T$ where in all of the above the O_j are elements of the topology $\mathcal{O}$. T is called locally compact if $\forall t \in T \;\exists$ a neighborhood U of t such that $\overline{U}$ is compact [95, p. 11].

For a locally compact abelian group T that is also Hausdorff, the string of adjectives will routinely be shortened by saying that T is LCA. Examples of compact groups include the complex numbers of modulus 1 with multiplication, 3×3 orthogonal matrices with determinant 1 and rotations in the plane or 3-dimensional euclidean space [74, p. 45] [63, p. 135]. The algebraic first pair of examples is obviously related to the second pair which details groups of transformations. The real numbers with addition as well as their finite products $(\mathbb{R}^n, +)$ are examples of locally compact abelian groups [113, p. 213] .

A function f on a group T is completely determined when its values $f(t)$ at the group members $t \in T$ are known. In the case of countably many elements this suggests a decomposition

$$f(t) = \sum_{s \in T} \alpha_s \delta_s(t) \qquad \alpha_s \in \mathbb{C} \;\; \forall s \in T$$

where $\delta_s(t)$ is one for $s = t$ and zero otherwise. A natural condition to impose on the multiplication $* : (f, g) \mapsto f*g$ of two functions f and g is that it is compatible with the group operation in the sense of $\delta_s(\cdot) * \delta_t(\cdot) = \delta_{st}(\cdot)$ as then the delta functions under multiplication mirror the groups behavior under the group operation. In this way the basis of the space of functions on T may be identified with T itself. This type of structure preserving multiplication is termed the (discrete) convolution and extends to functions f, g on T via

$$(f * g)(\cdot) = \left(\sum_{s \in T} \alpha_s \delta_s(\cdot) \right) * \left(\sum_{t \in T} \beta_t \delta_t(\cdot) \right) = \sum_{s \in T} \sum_{t \in T} \alpha_s \beta_t \delta_{st}(\cdot)$$

and renaming $st = u \in T$ this leads to

$$(f * g)(\cdot) = \sum_{s \in T} \sum_{t \in T} \alpha_s \beta_t \delta_{st}(\cdot) = \sum_{u \in T} \left(\sum_{s \in T} \alpha_s \beta_{s^{-1}u} \right) \delta_u(\cdot)$$

implying $(f * g)(\cdot) = h(\cdot) = \sum_{s \in T} \gamma_s \delta_s(\cdot)$ with coefficients $\gamma_s = \sum_{v \in T} \alpha_v \beta_{v^{-1}s}$. One then notices that $f(t) = \alpha_t$, $g(t) = \beta_t$ and therefore

$$(f * g)(t) = \sum_{s \in T} \alpha_s \beta_{s^{-1}t} = \sum_{s \in T} f(s) g(s^{-1}t)$$

which gives an explicit formula for $f * g$ in terms of values of $f(\cdot)$ and $g(\cdot)$. To guarantee existence of the convolution, it is enough to demand the coefficients of $f(\cdot)$ and $g(\cdot)$ to be absolutely summable, i.e. $f = \sum_{s \in T} \alpha_s \delta_s(\cdot)$ should satisfy $\sum_{s \in T} |\alpha_s| < \infty$. In the algebra $\ell^1(T)$ of absolutely summable functions on a countable group T defined like this, it holds that $(\ell^1(T), *)$ is a Banach algebra with continuous involution $f(s) \to \overline{f}(s^{-1})$ under norm $\| \cdot \|_{\ell^1}$ w.r.t. the counting measure [128, pp. 120-122]. Furthermore $\ell^1(T)$ is commutative iff T is abelian because

$$(f * g)(t) - (g * f)(t) = \sum_{s \in T} f(s) g(s^{-1}t) - \sum_{s \in T} g(s) f(s^{-1}t) \tag{4.1}$$

$$\overset{s'=s^{-1}t}{=} \sum_{s\in T} f(s)g(s^{-1}t) - \sum_{s'\in T} g(ts'^{-1})f(s') \tag{4.2}$$

$$= \sum_{s\in T} f(s)\left[g(s^{-1}t) - g(ts^{-1})\right] \tag{4.3}$$

is zero iff $s^{-1}t = ts^{-1}\ \ \forall s,t \in T$. Any representation $\pi : T \to U(\mathcal{H}_\pi)$ of T by means of unitary operators $U_t \in U(\mathcal{H}_\pi)$ on some Hilbert space $\mathcal{H}_\pi$ can be lifted to a bounded $*$-representation on $\ell^1(T)$ by $\pi(f) = \sum_{t\in T} f(t)U_t\ \ \forall f \in \ell^1(T)$ [128, p. 127]. This construction parallels the usual Fourier transform as for $f, g \in \ell^1(T)$

$$\begin{aligned} U_{f*g} = \sum_{t\in T}(f*g)(t)U_t &= \sum_{t\in T}\sum_{s\in T} f(s)g(s^{-1}t)U_t \\ &\overset{t'=s^{-1}t}{=} \sum_{s\in T} f(s)\sum_{t'\in T} g(t')U_{st'} \\ &= \sum_{s\in T} f(s)U_s\sum_{t\in T} g(t)U_t \qquad = U_fU_g \end{aligned} \tag{4.4}$$

and $\pi[f*g] = \pi(f)\pi(g)$ maps convolution to regular products. Since for any $q \in \mathcal{H}_\pi$ the function $f(s^{-1}t) = \langle U_{s^{-1}t}q, q\rangle_{\mathcal{H}_\pi}$ is positive definite by

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\overline{\alpha_j}f(t_j^{-1}t_i) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\overline{\alpha_j}\langle U_{t_i}q, U_{t_j}q\rangle_{\mathcal{H}_\pi} = \|\sum_{i=1}^{n}\alpha_iU_{t_i}q\|^2_{\mathcal{H}_\pi} \geq 0$$

for all choices of α_i and evaluation points t_i, the algebra $\ell^1(T)$ and its unitary representations are related to positive definite functions that correspond to covariance functions of second order stationary stochastic processes. Extending the construction to cover uncountable groups while at the same time carrying over the convenient properties requires introduction of an essentially unique invariant measure — called the Haar measure — on T. The next theorem only sketches its properties; more details, a proof and rigorous definitions of the involved quantities can be found in [63, p. 41].

Theorem 4.1.2 *For any locally compact group T there exists a nonzero Radon measure $\mu(\cdot) : B \mapsto \mu(B) \in \mathbb{C}$ such that $\mu(tB) = \mu(\{tb : b \in B\}) = \mu(B)$ for all $t \in T$ and Borel sets $B \subset T$. This so called Haar measure is unique up to a positive scaling constant.*

Employing this theorem, integration against the Haar measure on T is invariant against translation of functions $f(t)$ by $s \in T$ to $f(s^{-1}t)$ as

$$\begin{aligned} \int_{t\in T} f(t)d\mu(t) = \int_{t\in T} f(t)d\mu(st) &= \int_{t\in s^{-1}T} f(t)d\mu(st) \\ &= \int_{u\in T} f(s^{-1}u)d\mu(u)\ \ \forall s\in T \end{aligned} \tag{4.5}$$

and essentially the calculations made during the exemplary study of countable groups T carry over unperturbedly. One can then formulate the following theorem listing properties of $L^1(T) = \{f : T \to \mathbb{C} : \int_T |f(t)|d\mu(t) < \infty\}$.

Theorem 4.1.3 *Let T be an LCA group and μ its Haar measure. Then for $L^1(T)$ with convolution $(f * g)(t) = \int_{s\in T} f(s)g(t-s)d\mu(s)$ as multiplication and involution $* : f(s) \mapsto \overline{f}(s^{-1})$*

I *$L^1(T)$ is a commutative Banach $*$-algebra. The closure of $L^1(T)$ under the norm $\|f\|_* = \sup_\pi \|\pi(f)\|_{\mathcal{H}_\pi}$, π any continuous unitary representation of T on $\mathcal{H}_\pi$, is the group C^*-algebra $C^*(T)$ [48, p. 303].*

II *Functions $f \in L^1(T)$ that are squares in the sense of $f = g^* * g$ or $f = g * g^*$ for some $g \in L^1(T)$ are positive definite. This follows from*

$$\begin{aligned}\sum_{i,j=1}^{n} \alpha_i\overline{\alpha_j} f(t_i - t_j) &= \sum_{i,j=1}^{n} \alpha_i\overline{\alpha_j} \int_T g(s-t_j)\overline{g}(s-t_i)d\mu(s) \\ &= \int_T h(s)\overline{h}(s)d\mu(s) \qquad\qquad \geq 0 \qquad (4.6)\end{aligned}$$

for $h(s) = \sum_{i=1}^n \overline{\alpha_i} g(s-t_i)$ and implies a systematic way to construct covariance functions for second order stationary processes. Another method of construction using more elementary functions is given in item IV. Note that $\|f\|_1 = \|g\|_1^2$ automatically by the Banach algebra property.

III *Unitary representations $\pi : T \ni t \mapsto \pi(t) = U_t \in \mathcal{B}(\mathcal{H}_\pi)$ that are continuous as maps $T \ni t \mapsto \pi(t)q \in \mathcal{H}_\pi$ for $q \in \mathcal{H}_\pi$ induce $*$-representations of $L^1(T)$ on $\mathcal{B}(\mathcal{H}_\pi)$ via $\pi(f) = \int_{t\in T} f(t)U_t d\mu(t)$ [128, p. 127].*

IV *The functions $f(t-s) = \langle U_{t-s}q, q\rangle_{\mathcal{H}_\pi}$ for some nonzero $q \in \mathcal{H}_\pi$ are then bounded, continuous and positive definite with $\|f\|_\infty = f(1) = \|q\|^2_{\mathcal{H}_\pi}$ [63, pp.83,92].*

We illustrate briefly for $T = (\mathbb{R}, +)$ and $L^1(T) := \{f : T \to \mathbb{C} : \|f\|_1 = \int_{\mathbb{R}} |f(t)|dt < \infty\}$ where we interpret $f * h = L_h f$ as the response of a linear time invariant system to the input f.

We find by I that $L^1(\mathbb{R})$ is commutative with $f * h = h * f$ and $(f*h)^* = h^* * f^*$ $\forall f, h \in L^1(\mathbb{R})$ where $(f^*)(s) = \overline{f}(-s)$ and $\|f*h\|_1 \leq \|f\|_1\|h\|_1$. The group C^*-algebra is the closure of $L^1(\mathbb{R})$ under the norm $\|f\|_* = \|\hat{f}\|_{\sup}$ and $C^*(\mathbb{R}) \cong C_0(\mathbb{R})$, the continuous functions on $\mathbb{R}$ vanishing at infinity [63, p.225]. Since for every $\omega \in \mathbb{R}$, $\pi_\omega : T \ni t \mapsto \pi_\omega(t) = \exp(i\omega t) \in \mathbb{C}$ creates a unitary operator on $\mathbb{C}$ for every $t \in T$ by III $\pi_\omega(f) = \int_{\mathbb{R}} f(t)e^{i\omega t}dt$ $\forall f \in L^1(\mathbb{R})$ which as a function of ω is just the usual Fourier transform. It holds that $\pi_\omega(f * h) = \pi_\omega(f)\pi_\omega(h)$ which already hints at $\pi_\omega(f) = \pi_\omega(h)^{-1}\pi_\omega(f * h)$ and solvability of the inverse problem in case when the transfer function has nonvanishing Fourier transform.

Finally note that $f_\omega(t-s) = \langle \pi_\omega(t-s)q, q\rangle_{\mathbb{C}} = |c|^2 e^{i\omega(t-s)}$ is a continuous bounded positive definite function on $\mathbb{R}$ for all $\omega \in \mathbb{R}$. The complex constant $c \in \mathbb{C}$ is arbitrary and $\|f_\omega\|_{\text{sup}} = f_\omega(0) = |c|^2$. The complex exponentials take a central role in the next subsections as they turn out to be the extreme points of the convex set of positive definite functions, which can consequently be written as integrals of complex exponentials against a probability measure. The situation is clarified in figure 4.1

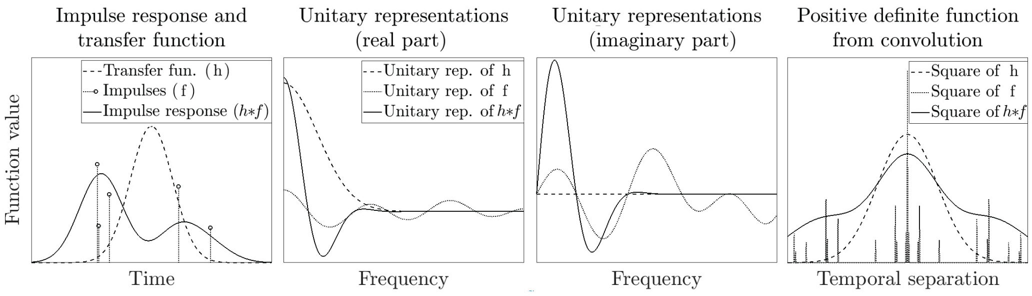


Figure 4.1: Convolution as a way of superimposing signals has a simple relationship to the family of unitary representations $\pi_\omega(t) = \exp(i\omega t)$, that form elementary building blocks of positive definite functions.. On the left, an input signal f, the transfer function h and its convolution can be seen, whereas the middle plots show their complex unitary representations for a range of parameters ω. The right panel finally exhibits the positive definite convolution squares $f * f^*$ and $h * h^*$ and $(h * f) * (h * f)^*$.

The role of complex exponentials as constituting simple positive definite functions is no coincidence. They arise naturally when one tries to decompose the algebra $L^1(T)$ into minimal parts to derive a structure theorem for algebras that generalizes theorems about orthogonal decomposability of separable Hilbert spaces into direct sums of one dimensional subspaces.
A subalgebra $\mathcal{J}$ of a commutative Banach $*$-algebra $\mathcal{A}$ is called an ideal if $\forall f \in \mathcal{J}$and $g \in \mathcal{A}$ one has $f * g \in \mathcal{J}$. It is termed proper if $\mathcal{J} \neq \mathcal{A}$ and maximal if $\nexists$ a proper ideal $\mathcal{J}' \subsetneq \mathcal{J}$. Let as before $\Gamma : \mathcal{A} \to C(\sigma(\mathcal{A}))$ be the Gelfand transform and $\sigma(\mathcal{A})$ be the set of multiplicative linear functionals from $\mathcal{A}$ to the complex numbers.

Theorem 4.1.4 *For a complex commutative Banach algebra $\mathcal{A}$ with ideal $\mathcal{J}$, the following statements hold*

I *Every maximal ideal $\mathcal{J}$ is the kernel of an $h \in \sigma(\mathcal{A})$. If $\mathcal{A}$ is unital, $\mathcal{A}/\mathcal{J}$ has dimension* 1 *[128, p. 69].*

II *If $\mathcal{J}$ is maximal, the quotient space $\mathcal{A}/\mathcal{J}$ can be made into a Banach algebra by defining multiplication on representations of equivalence classes. If $\mathcal{A}$ is unital and every nonzero element of $\mathcal{A}/\mathcal{J}$ is invertible, then $\mathcal{A}/\mathcal{J} \cong \mathbb{C}$ [95, p. 473].*

III *If $\mathcal{A}$ is unital, $A \in \mathcal{A}$ has an inverse if and only if $A \notin \mathcal{J}$ for any proper ideal $\mathcal{J}$ [128, p. 64].*

IV *If $\mathcal{A} = L^1(T)$ for a commutative locally compact group T, then each homomorphism $h \in \sigma(\mathcal{A})$ can be written as $h(f) = \int_T f(t)\overline{\chi_\omega}(t)d\mu(t)$ for some*

> $\chi_\omega(\cdot) \in L^\infty(T)$. *The so called character* $\chi_\omega(\cdot)$ *is unique and a homomorphism from* T *to* $\mathbb{C}$ *[128, p. 135].*

Let us again illustrate the theorems implications by investigating an instructive special case. Let $\mathcal{A}$ be the algebra of selfadjoint linear bounded operators on $\mathcal{H} = \mathbb{C}^n$ which are generated from a selfadjoint $A \in \mathcal{B}(\mathcal{H})$ with simple spectrum by applying bounded Borel functions f to it. $\mathcal{A}$ is unital and commutative since $A^0 = I$ and $f(A)g(A) = g(A)f(A)\ \ \forall f, g, \in \mathcal{B}(\sigma(A))$. They can be simultaneously diagonalized by some unitary matrix $U = [u_1, ..., n_n] \in \mathbb{R}^n \otimes \mathbb{R}^n$, see spectral theorem.

One can then construct explicitly $h \in \sigma(\mathcal{A})$ such that h is a multiplicative functional on $\mathcal{A}$ by setting — now for $B, C \in \mathcal{A}$ arbitrary —

$$h_k(B) = \langle U^* B U e_k, e_k \rangle_{\mathcal{H}} = d_k(U^* B U)$$

where e_k is the k-th Euclidean unit vector and d_k denotes the k-th diagonal element. Since $B = f(A)$ and $C = g(A)$, $U^*BU = \Lambda_B$ and $U^*CU = \Lambda_C$ are diagonal which implies the homomorphism properties

$$\begin{aligned} h_k(B+C) = \langle (\Lambda_B + \Lambda_C)e_k, e_k\rangle_{\mathcal{H}} \quad &= d_k(\Lambda_B + \Lambda_C) \\ &= h_k(B) + h_K(C) \\ h_k(BC) = \langle (\Lambda_B \Lambda_C)e_k, e_k\rangle_{\mathcal{H}} \quad &= d_k(\Lambda_B \Lambda_C) \\ &= h_k(A)h_k(B) \end{aligned}$$

The map associating to $B \in \mathcal{A}$ its k-th eigenvalue (apart from some reordering) is therefore a homomorphism for all $k = 1, ..., n$. By theorem 4.1.4.I the corresponding maximal ideals $\mathcal{J}_k$ are the kernels of h_k, i.e. are all those $B \in \mathcal{A}$ such that their k-th diagonal element $h_k(B) = d_k(\Lambda_B)$ is zero. This property is conserved under addition and under multiplication by elements of $\mathcal{A}$. $\mathcal{J}_k$ consists only of noninvertible elements by theorem 4.1.4.III. For each $k = 1, ..., n$ an element of $\mathcal{A}/\mathcal{J}_k$ is an equivalence class containing linear operators in $\mathcal{A}$ whose k-th diagonal element coincides. They form a one dimensional subalgebra of $\mathcal{A}$ in which every nonzero element $[B] \in \mathcal{A}/\mathcal{J}_k$ can be inverted by multiplying with some $[C]$ for which $h_k(C) = h_k(B)^{-1}$. By theorem 4.1.4.II this allows to conclude $\mathcal{A}/\mathcal{J}_k \cong \mathbb{C}$. Notice that $\mathcal{A}/\mathcal{J}_k \cong \mathbb{C}$ and $\mathbb{C}^n = \bigoplus_{k=1}^n \mathcal{A}/\mathcal{J}_k$. Since bounded operators satisfy $\|B^*B\|_{op} = \|BB^*\|_{op} = \|B^2\|_{op}\ \forall B \in \mathcal{B}(\mathcal{H})$, $\mathcal{A}$ is also a C^*-algebra, $\mathcal{A} \cong C(\sigma(\mathcal{A})) \cong C(\text{range}\,\hat{A})$ by Gelfand Naimark theorem and [63, p. 7]. Since the functions on an n-element domain (A was assumed to have simple spectrum) can be represented as a vector in $\mathbb{C}^n$, we find $\mathcal{A} \cong \bigoplus_{k=1}^n \mathcal{A}/\mathcal{J}_k$ and multiplication on $\mathcal{A}$ may equivalently be carried out via the trivial pointwise multiplication of elements $\lambda = \{\lambda_k\}_{k=1}^n \in \bigoplus_{k=1}^n \mathcal{A}/\mathcal{J}_k \cong \mathbb{C}^n$. All of this is relatively close to the discussion of the spectral theorem in subsection 2.2.3 but now with an explicit focus on the algebraic aspects of the operator algebra generated by a covariance operator.

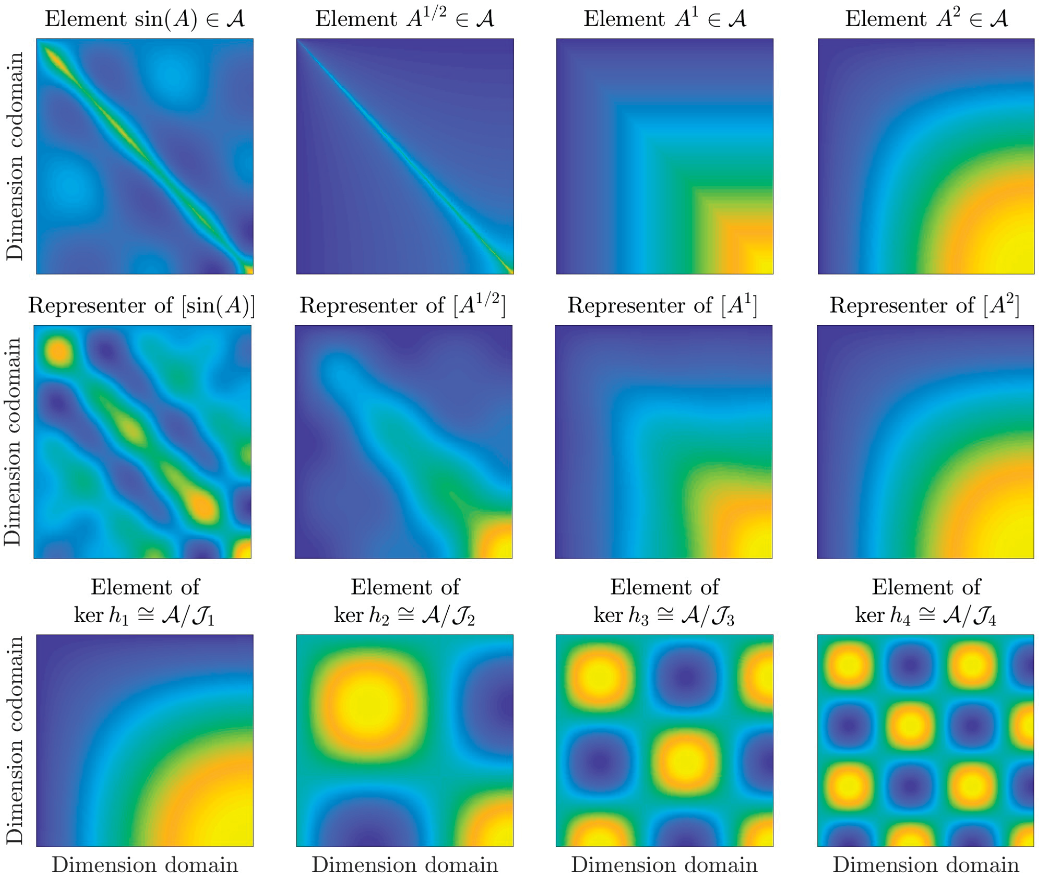


Figure 4.2: In the first row the elements $\sin(A), A^{1/2}, A^1$, and A^2 of the algebra $\mathcal{A}$ generated by A are plotted. In the middle row representing elements $[\sin(A)], [A^{1/2}], [A^1]$, and $[A^2]$ of the quotient space $\bigoplus_{k=1}^{4} \mathcal{A}/\mathcal{J}_k$ are exhibited. The representers $B \in [B]$ were chosen such that $h_k(B) = 0$ for $k \geq 5$. The complements of $\ker(h_k)$ in $\mathcal{A}$ for $k = 1, ..., 4$ are shown in the last row and correspond to a representing element of the one dimensional spaces not annihilated by h_k.

4.1.2 Group characters and the Fourier transform

Every maximal ideal $\mathcal{J}$ of $L^1(T)$ is the kernel of a unique algebra homomorphism $\varphi_{\mathcal{J}}$: $L^1(T) \ni f \mapsto \hat{f}_{\mathcal{J}} \in \mathbb{C}, \varphi_{\mathcal{J}} \in \mathrm{Hom}_{\mathrm{Alg}}(L^1(T), \mathbb{C})$. Since $\varphi_{\mathcal{J}}$ as a linear multiplicative functional is also an element of the dual space $(L^1(T))^* \cong L^\infty(T)$, it can be written in the form $\varphi_{\mathcal{J}}(f) = \int_T \overline{\chi_{\mathcal{J}}}(\tau) f(\tau) d\tau$ where the $\chi_{\mathcal{J}}(\cdot) \in L^\infty$ are called group characters. They map T to $\mathbb{S}^1$ in a structure preserving way and their relevance lies in the fact that there is a one to one correspondence between multiplicative linear functionals, maximal ideals and group characters. Once they have been determined for certain groups T_1 and T_2 they are readily computable for product and quotient groups involving T_1 and T_2. Group characters form themselves an abelian group $(\hat{T}, \cdot)$ under pointwise multiplication. Products as well as powers of functions $f, g \in L^1(T)$ are straightforward to analyze with the help of $\chi_{\mathcal{J}} \in \hat{T}$ since $\varphi_{\mathcal{J}}(f * g) = \varphi_{\mathcal{J}}(f) \cdot \varphi_{\mathcal{J}}(g)$ and f is approximable as a weighted sum over group characters $\chi_{\mathcal{J}}$ indexed by the set of all maximal ideals. Convolutions of functions are encountered in the theory of linear time invariant systems and when calculating the probability density of sums of random variables.

Theorem 4.1.4 from the previous subsection suggests a one to one relationship between maximal ideals of an algebra $\mathcal{A}$, the set of homomorphisms from $\mathcal{A}$ to $\mathbb{C}$ and certain unitary representations that act as homomorphisms from T to $\mathbb{C}$

if T is LCA and $\mathcal{A} = L^1(T)$. These functions χ have important properties that stem from the fact that, for LCA groups T, their structure preserving properties as maps $T \to \mathbb{S} = \{z \in \mathbb{C} : |z|^2 = 1\}$ lift to structure preserving properties as maps $L^1(T) \to \mathbb{C}$ when they are identified with linear functionals $l_\chi(f) = \int_T f(t)\overline{\chi}(t)d\mu(t)$. In this case one finds among other relations

$$\begin{aligned}
l_\chi(f * g) &= \int_T (f * g)(t)\overline{\chi}(t)d\mu(t) \\
&= \int_T \int_T f(s)g(t-s)\overline{\chi}(t)d\mu(s)s\mu(t) \\
&\overset{t'=t-s}{=} \int_T \int_T f(s)g(t')\overline{\chi}(t')\overline{\chi}(s)d\mu(s)d\mu(t) \\
&= \int_T f(s)\overline{\chi}(s)d\mu(s) \int_T g(t)\overline{\chi}(t)d\mu(t) \qquad = l_\chi(f)l_\chi(g)
\end{aligned} \tag{4.7}$$

in which $l_\chi(\cdot) : L^1(T) \to \mathbb{C}$ reveals itself as an algebra homomorphism from the group algebra to the complex numbers and a one-dimensional analogue of the Fourier transform but obviously more general. The exact nature of these $\chi(\cdot)$ is fixed in the next definition [128, pp. 135-138].

Definition 4.1.5 Let T be LCA and $\mathbb{S}$ the set of all complex numbers of modulus 1.

I If $\chi(\cdot) : T \ni t \mapsto \chi(t) \in \mathbb{S}$ is a continuous homomorphism into the multiplicative group of complex numbers in the sense that $\chi(t+s) = \chi(t)\chi(s)$ then χ is called a character of T.

II The set $\hat{T} := \{\chi : T \to \mathbb{S}, \chi \text{ a character}\}$ together with pointwise multiplication $\chi_1\chi_2 = \chi$ with $\chi(t) = \chi_1(t)\chi_2(t) \;\; \forall t \in T, \chi_1, \chi_2 \in \hat{T}$ is called the dual group of T.

It is standard [128, pp. 137-138] to show that the set $\hat{T}$ with multiplication as defined above is actually again a locally compact group with $\chi^{-1} = \overline{\chi}$. From $\chi \in \hat{T}$ by $l_\chi(f+g) = l_\chi(f) + l_\chi(g)$ and $l_\chi(f*g) = l_\chi(f)l_\chi(g) \;\; \forall f, g \in L^1(T) =: \mathcal{A}$ as shown in equation 4.7 it follows that $l_\chi(\cdot) : L^1(T) \to \mathbb{C}$ is a multiplicative functional and therefore $l_\chi \in \sigma(\mathcal{A})$. That also the converse is true and every $l \in \sigma(\mathcal{A})$ arises as l_χ for some $\chi \in \hat{T}$ was the topic of theorem 4.1.4.IV. As under these conditions the topology of compact convergence on $\hat{T}$ coincides with the weak* topology on $\{l_\chi : \chi \in \hat{T}\} \subset L^\infty$ [63, p. 97], the map

$$l : \hat{T} \ni \chi \mapsto l_\chi \in \sigma(\mathcal{A}), \qquad\qquad l_\chi(f) = \int_T f(t)\overline{\chi}(t)d\mu(t)$$

is a homomorphism and $\hat{T} \overset{l}{\cong} \sigma(L^1(T))$, so that the spectrum of the algebra $L^1(T)$ may be identified with the dual group $\hat{T}$ in the future. If T is even compact with $\mu(T) = 1$ then $\hat{T}$ is an orthonormal basis for $L^2(T) := \{f : T \to \mathbb{C} :$

$\int_T |f(t)|^2 d\mu(t0 < \infty\}$, i.e. [63, p. 109]

$$\langle \chi_i, \chi_j \rangle_{L^2(T)} = \int_T \chi_i(t)\overline{\chi_j}(t) d\mu(t) = \delta_{ij}.$$

This allows orthogonal decomposition of functions for example on the circle $\mathbb{S}$ without reference to the exterior space $\mathbb{R}^2$ into which one might embed it. Originally $\chi(\cdot)$ came from a unitary representation of the group T, and most of its properties can be derived from this connection. It implies an interesting relationship between the dual group $\hat{T}$ and unitary representations of T by means of one dimensional complex numbers or their Cartesian products.

Theorem 4.1.6 *Let T be LCA. Then every irreducible unitary representation, i.e. every $\pi : T \to \mathcal{B}(\mathcal{H})$ which cannot be further decomposed into $\pi = \pi_1 \oplus \pi_2$ with $\pi_k : T \to \mathcal{B}(\mathcal{H}_k), \mathcal{H}_k \overline{\boxslash} \mathcal{H}$ unitary and $\mathcal{H}_k \neq \{0\}$ for $k = 1, 2$, is one dimensional [63, p. 72].*

Furthermore for every irreducible $\pi, \mathcal{H}_\pi \cong \mathbb{C}$ and $\pi(t) \in \mathcal{B}(\mathbb{C}) \cong \mathbb{C}$ can be identified with a complex number $\chi(t) \in \mathbb{C}\ \forall t \in T$. The $\chi(t)$ constructed in this way is a character and all characters have this form so that the dual group $\hat{T}$ is in one to one correspondence with all irreducible unitary representations [63, p. 96]. Consequently $\pi(f) = \int_T f(t)\overline{\chi}(t) d\mu(t)\ \ \forall f \in L^1(T)$ is a multiplicative functional whose kernel is a maximal ideal in $L^1(T)$ [128, p. 135].

As a corollary, notice that if for $k = 1, 2\ \pi_k : T_k \to \mathcal{B}(\mathcal{H}_k) \cong \mathbb{C}$ is an irreducible representation of of LCA T_k's on $\mathcal{H}_k$, then $\pi_1(s)\pi_2(t), s \in T_1, t \in T_2$ is a complex number of modulus 1 and

$$\pi = \pi_1 \otimes \pi_2 : T_1 \times T_2 \ni (s, t) \mapsto \pi_1(s)\pi_2(t) \in \mathbb{S}$$

is an irreducible unitary representation of $T_1 \times T_2$ on $\mathcal{H}_\pi \cong \mathbb{C}$ [128, p. 139]. It follows that $\widehat{T_1 \times T_2} \cong \hat{T}_1 \times \hat{T}_2$. In a similar vein, if one decomposes

$$T \cong T/S \times S \qquad\qquad \hat{T} \cong (\widehat{T/S}) \times \hat{S} \qquad\qquad S \text{ subgroup of } T$$

one finds that the unitary representation

$$\pi(t) = \pi([t])\pi(s)$$

for $t \in T, [t] \in T/S, s \in S$ implies that the characters of T/S are those characters of T which are constant on S [128, p. 139]. In fact, if $[\cdot]$ is the canonical projection $T \to T/S$, then $\pi \circ [\cdot]$ is a character on T/S. Typical examples of groups T for which we will need the characters to construct positive definite functions include [95, p. 366-388]

$$\begin{array}{llll} i)\ T = (\mathbb{R}, +) & \hat{T} \cong T & \chi_\omega(t) = \exp(2\pi i \omega t) & \omega \in \mathbb{R} \\ ii)\ T = (\mathbb{R}^n, +) & \hat{T} \cong T & \chi_\omega(t) = \exp(2\pi i \langle w, t \rangle_{\ell^2}) & \omega \in \mathbb{R}^n \end{array}$$

$iii)\ T = (\mathbb{S}, *)$	$\hat{T} \cong \mathbb{Z}$	$\chi_\omega(t) = \exp(2\pi i \omega t)$	$\omega \in \mathbb{Z}$
$iv)\ T = (\mathbb{R}_+, *)$	$\hat{T} \cong \mathbb{R}$	$\chi_\omega(t) = \exp(2\pi i \omega \log t)$	$\omega \in \mathbb{R}$

where the properties of the positive reals under multiplication follow from the first statement and the fact that $(\mathbb{R}_+, *)$ and $(\mathbb{R}, +)$ are isomorphic as groups via the exponential map [74, p. 15].

The relevance of examples i) and ii) is clear: when dealing with functions on $\mathbb{R}^n$ from a stochastic perspective, the characters are elementary building blocks of translation invariant positive definite kernels which one might identify with covariance functions of second order stationary processes. They will be shown to provide something akin to a basis for covariance functions enabling inference of a processes correlation structure in a non-parametric way. For iii) and iv) note that it might be interesting to look at functions on $\mathbb{R}^2$ that are rotation invariant. We might then decompose $\mathbb{R}^2$ in some way with $T = (\mathbb{R}_+, *) \times \mathbb{S}$ via polar coordinates, focus on the quotient space T/S and investigate functions and kernels on it.

As was just seen, $\varphi_{\mathcal{J}}$ maps convolutions to pointwise multiplication and its central ingredient $\chi_{\mathcal{J}} : T \to \mathbb{S}^1$ sets up a norm preserving isomorphism $\mathcal{F} : L^2(T) \ni f(t) \mapsto (\mathcal{F}f)_{(\omega)} = \int_{\tau \in T} f(\tau)\chi_\omega d\tau \in L^2(\hat{T})$ so that analysis of $f \in L^2(T)$ is interchangeable with analysis of $\hat{f} = \mathcal{F}f \in L^2(\hat{T})$. The construction of $\mathcal{F}f$ is analogous to the usual Fourier transform $\hat{f}$ when $T = \mathbb{R}$ and carries its advantages over to function spaces on arbitrary locally compact abelian groups.

Definition 4.1.7 let T be LCA and denote again by $\hat{T}$ its dual group. Then for $f \in L^1(T)$ the function $\hat{f} : \hat{T} \to \mathbb{C}$ defined as

$$\hat{f}(\omega) := \int_T \overline{\chi_\omega}(t) f(t) d\mu(t) \tag{4.8}$$

is called the Fourier transform of f. The map $\mathcal{F} : L^1(T) \ni f \mapsto \hat{f} \in C_0(\hat{T})$ is called Fourier transform on T [63, p. 102].

The Fourier transform on T is a $*$-homomorphism between $L^1(T)$ and $C_0(\hat{T})$ and many of the convenient properties listed in the theorem below are direct consequences of $(\mathcal{F}f)(\omega)$ being effectively $\pi_\omega(f)$ for a family of unitary representations indexed by $\chi_\omega \in \hat{T}$. It is standard to identify the $\hat{T}$ and $\{\omega : \chi_\omega \in \hat{T}\}$.

Theorem 4.1.8 *Let $\mathcal{F}$ be the Fourier transform on some LCA T with dual group $\hat{T}$. Then the following assertions hold [63, pp. 101-114].*

I If $\hat{f} \in L^1(\hat{T})$ then the Fourier transform can be inverted to yield

$$f(t) = \int_{\hat{T}} \chi_\omega(t) \hat{f}(\omega) d\nu(\omega)$$

where equality holds almost everywhere and ν is the dual measure of the Haar measure μ on T. The Fourier transform can be extended to a unitary isomorphism between $L^2(T)$ and $L^2(\hat{T})$.

II *For $f, g \in L^1(T)$, the Fourier transform diagonalizes the convolution operator $L_f g = f * g$ which is unitarily equivalent to $M_{\hat{f}}$ via*

$$\mathcal{F} L_f g = \mathcal{F}(f * g) = \hat{f}\hat{g} = M_{\hat{f}} \mathcal{F} g$$

by the homomorphism property implying $L_f = \mathcal{F}^* M_{\hat{f}} \mathcal{F}$.

III *The covariance operator C with kernel $K(s,t) = f(t-s)$ given as $Cg = \int_T f(t-s)g(s)d\mu(s)$ for some continuous positive definite $f(\cdot)$ absolutely integrable on T is diagonalizable via Fourier transform. This is implied by II via $Cg = f * g = L_f g = \mathcal{F}^* M_{\hat{f}} \mathcal{F} g$. Particularly, C's positivity means that $\hat{f}$ takes on only nonnegative values as $\sigma(L_f) = \sigma(M_{\hat{f}}) = \operatorname{range} \hat{f}$ where the last equality holds because for a continuous function $\hat{f}$, essential range and range coincide [168, p. 369] [196, p. 80].*

IV *(Poisson summation) If S is a closed subgroup of T, $f \in L^1(T)$ with compact support and $\hat{f}|_{S^\perp} \in L^1(S^\perp)$ then*

$$\int_S f(st)d\mu(s) = \int_{S^\perp} \hat{f}(\omega)\chi_\omega(t)d\nu(\omega)$$

where equality again holds almost everywhere, μ (and ν) a suitably normalized (dual) Haar measure and $S^\perp := \{\chi_\omega \in \hat{T} : \chi_\omega(s) = 0 \; \forall s \in S\} \cong \widehat{T/S}$.

On a less formal level, theorem 4.1.8.I implies that one may investigate $\hat{f}$ instead of f without losses in explanatory power. Statement II furnished with additional surroundings provides a method for fast inversion of covariance operators whose kernel is a positive definite function. This speeds up one of the steps whose computational cost is most prohibitive to practical applications of spline formulas $\sigma_f = \sum_{j=1}^n \lambda_j K(t_j, \cdot), \lambda = C^{-1}a$ to large scale estimation problems whereas III assures that the action of applying C is unitarily equivalent to multiplication by some $\hat{f}$, whose nonnegativity is necessary for C's positive definiteness. Finally the Poisson summation formula exhibited in IV makes possible an extension of Shannon's sampling theorem to functions on LCA groups.

4.1.3 Bochners theorem

The set of elementary positive definite functions $\chi_\omega, \omega \in \hat{T}$ forms an important subset of positive definite functions f with $\|f\|_\infty = 1$ for which they play a role similar to a basis in ordinary vector space theory. The representation of f in terms of weighted sums of $\chi_\omega, \omega \in \hat{T}$ has characteristic positivity properties that are stated formally in Bochner's theorem and provide a canonical link between correlation functions and different functions occurring in probability theory. In its most simplistic form, Bochner's theorem implies the possibility of inferring positive definite functions f via classical least squares estimation of type $\|Af - a\|_2 \to \min$ augmented with a positivity constraint on the Fourier transform $\mathcal{F}f$.

Recall that a function f on an LCA group T was called positive definite if

$$\sum_{i,j=1}^{n} \alpha_i \overline{\alpha_j} f(t_i - t_j) \geq 0 \quad \forall \text{ choices of } \alpha_i \in \mathbb{C}, t_i \in T \tag{4.9}$$

where n is an arbitrary natural number. If equation 4.9 holds for a function $K : T \times T \to \mathbb{C}$ of two variables with $f(t_i - t_j)$ replaced by $K(t_i, t_j)$ then K is called a positive definite kernel [19, pp. 67,87]. Denote by $\mathfrak{P}$ the set of continuous positive definite functions on some LCA T and let $\mathfrak{P}_1$ be the subset consisting of those functions f for which $|f(0)| = 1$ [63, p. 86]. Similarly let $\mathfrak{K}$ be the set of positive definite kernels without any continuity conditions imposed on them [19, p. 80]. To investigate the structure of $\mathfrak{P}, \mathfrak{P}_1$ and $\mathfrak{K}$, some definitions taken from convex geometry are in order.

Definition 4.1.9 I A set $\mathfrak{C}$ for which it holds that $\forall c_1, c_2 \in \mathfrak{C}$, $\lambda(c_1, c_2) := \lambda c_1 + (1-\lambda)c_2 \in \mathfrak{C}$ whenever $\lambda \in [0,1]$ is called convex and if furthermore for all $\lambda > 0$, $\lambda\mathfrak{C} \subset \mathfrak{C}$ then $\mathfrak{C}$ is a convex cone [50, p. 135].

II $\mathfrak{A}$ is termed an extreme subset of the convex cone $\mathfrak{C}$ if $\forall c_1, c_2 \in \mathfrak{C}$ and $\lambda \in (0,1), \lambda(c_1, c_2) \in \mathfrak{A}$ implies $c_1, c_2 \in \mathfrak{A}$. $\mathfrak{F} \subset \mathfrak{C}$ is an extreme ray if $\mathfrak{F} = \{\lambda c : \lambda > 0\}$ for some fixed $c \in \mathfrak{C}$ and for any $f \in \mathfrak{F}$, $f \neq c_1 + c_2$ for any $c_1, c_2 \in \mathfrak{C}/\mathfrak{F}$. Single points $f \in \mathfrak{C}$ are termed extreme points if $\{f\}$ is an extreme subset of $\mathfrak{C}$ [112, p. 239], [19, p 55].

As an example of a convex cone, take the set of positive semidefinite matrices S_+^n acting on $\mathbb{R}^n$. Obvious extreme rays are sets of the form $\{\lambda f \otimes f^* : \lambda > 0\}$ for some $f \in \mathbb{R}^n$ [112, p. 239]. Since for any nonzero $\Sigma \in S_+^n$, $-\Sigma \notin S_+^n$ we find that $\nexists$ nonzero $\Sigma_1, \Sigma_2 \in S_+^n$ with $\Sigma_1 + \Sigma_2 = 0$ and the zero matrix forms an extreme point $\{0\}$ of S_+^n.

In general for $K_1, K_2 \in \mathfrak{K}$ and $\lambda_1, \lambda_2 \geq 0$

$$\begin{aligned}\sum_{i,j=1}^{n} \alpha_i \overline{\alpha_j} \left[\lambda_1 K_1(t_i, t_j) + \lambda_2 K_2(t_i, t_j)\right] = \lambda_1 &\left[\sum_{i,j=1}^{n} \alpha_i \overline{\alpha_j} K_1(t_i, t_j)\right] \\ &+ \lambda_2 \left[\sum_{ij=1}^{n} \alpha_i \overline{\alpha_j} K_2(t_i, t_j)\right] \geq 0\end{aligned}$$

and $\mathfrak{P}$ as well as $\mathfrak{K}$ are also closed under scaling by a positive scalar λ. Therefore $\mathfrak{P}$ and $\mathfrak{K}$ are convex cones. The extreme rays of $\mathfrak{K}$ and the extreme points of $\mathfrak{P}_1$ are given explicitly in the next theorem.

Theorem 4.1.10 *Let $\mathfrak{P}_1$ be the set of continuous positive definite functions $f \in \mathfrak{P}$ for which also $|f(0)| = 1$ and denote by $\mathfrak{K}$ the set of all positive definite kernels.*

I Any positive definite function $f \in \mathfrak{P}$ can be written as $f(t) = \langle \pi_f(t)q, q\rangle_{\mathcal{H}_f}$ for the unitary representation π_f given by

$$\pi_f(t)[g] = [g(\cdot - t)]$$

for some $g \in \mathcal{H}_f$. Here $[\cdot]$ is the canonical projection from $L^1(T)$ to $\mathcal{H}_f$, the Hilbert space of equivalence classes $\overline{L^1(T)/N}$ where N is the nullspace of the seminorm induced by the semi-inner product $\langle g, h\rangle_{\mathcal{H}_f} = \int_T \int_T g(s)\overline{h}(t)f(s-t)d\mu(s)d\mu(t)$ [63, p. 84].

II The extreme points of $\mathfrak{P}_1$ are those $f \in \mathfrak{P}$ for which the unitary representation π_f as determined in I is irreducible [63, p. 86]. These are also often called elementary positive definite functions.

III The set $\{\lambda K : \lambda > 0\}$ is an extreme ray of $\mathfrak{K}$ if and only if $K(s,t) = g(s)\overline{g}(t)$ for some nonzero $g : T \to \mathbb{C}$. The RKHS $\mathcal{H}_k$ associated to $k(\cdot,\cdot)$ via the Moore-Aronszajn theorem has dimension 1 [19, p. 83].

Theorem 4.1.10.II provides a systematic way to list all the extreme points of normalized positive definite functions $f \in \mathfrak{P}_1$ by enumerating the characters of T. Item III does essentially the same for $\mathfrak{K}$ and comes as no surprise due to it being simply a further specification of properties for the Mercer decomposition of a kernel. Since convex cones are the closure of the set of convex combinations of its extreme points by the Krein-Milman theorem [119], both parts suggest that a) one may deconstruct a given $f \in \mathfrak{P}_1$ or $K \in \mathfrak{K}$ into linear combinations of elementary positive definite functions and b) synthesize $f \in \mathfrak{P}_1, K \in \mathfrak{K}$ by combining either complex exponentials or tensor products of functions according to certain rules. At least for the case of $\mathfrak{P}$ and $\mathfrak{P}_1$ this is made formal in Bochners theorem [63, pp. 103-104].

Theorem 4.1.11 (Bochners theorem) *Let T be LCA. Then $f \in \mathfrak{P}$ on T iff*

$$f(t) = \int_{\hat{T}} \chi_\omega(t)d\nu(\omega) \quad \forall t \in T \tag{4.10}$$

for some unique positive measure ν on $\hat{T}$. If $\|f\|_\infty = 1$ or $f(0) = 1$ and therefore even $f \in \mathfrak{P}_1$, then $\nu(\cdot)$ is a probability measure, i.e. $\nu(\hat{T}) = 1$.

For not necessarily translation invariant positive definite kernels $K(\cdot,\cdot)$ a similar but weaker equivalence result is provided by Fortet's theorem 3.2.1 on page 112. The general situation is less accessible to systematic treatment and yields no easily implementable constraints expressible via linear transforms of $K(\cdot,\cdot)$. This cannot be sidestepped and will make necessary the development of optimization procedures that operate under cone constraints in the following sections.

Now four different options are available to construct positive definite kernels K on $T\times T$ or positive definite functions f on T. In the translation invariant case, one may either form square functions $f*f^*$ or integrate against a probability measure on $\hat{T}$ to get $f=\int_T \chi_\omega(\cdot)d\nu(\omega)$. When constructing positive definite kernels in general there are two options as well: choose an arbitrary but normalized sequence of functions $\{g_k\}_{k=1}^\infty$ and employ a converse of the Mercer decomposition to the effect of creating a kernel $K(s,t)=\sum_{j=1}^n \lambda_j g_j(s)\overline{g_j}(t)$ for some summable sequence of positive λ_j or freely specify the right-hand side of the equation in Fortet's theorem. According to Bochner's and Fortet's theorems as well as Mercers decomposition, the last three constructions establish a one-to-one correspondence. This opens up the possibility to approximate and estimate positive definite functions or kernels via linear combinations of complex exponentials or squares $g_j\overline{g}_j$ on which the complicated constraint of positive (semi)definiteness is replaced by the easy one of positivity of the expansion coefficients — a condition that is not only straightforward to check but can also be included as a constraint during optimization procedures.

A similar but weaker statement can be made for the square construction $f=g*g^*$. Suppose $f\in\mathfrak{P}\cap L^1(T)$ with $\hat{f}\in L^1(\hat{T})$. By theorem 4.1.8.III $\hat{f}\geq 0$ and if $\hat{g}=\sqrt{\hat{f}}\in L^1(\hat{T})$ as well by the Fourier inversion theorem 4.1.8.I $\exists g\in L^1(T)$ with

$$g(t)=\int_{\hat{T}}\hat{g}(\omega)\chi_\omega(t)d\nu(\omega)$$
$$\mathcal{F}(g*g^*)=\mathcal{F}(g)\mathcal{F}(g^*)=\hat{g}\overline{\hat{g}}=|\hat{f}|=\mathcal{F}f$$

which implies that $g*g^*(t)=\int_{\hat{T}}\hat{f}(\omega)\chi_\omega(t)d\nu(\omega)=f(t)$. Therefore under stronger conditions on the positive definite function f, approximating or estimating them via sums of type $\sum_{j=1}^n \lambda_j g_j * g_j^*$ for $\lambda_j\geq 0$ is a reasonable approach as well.

4.2 Kernel inference

In this section different methods for estimating reproducing kernels from data are evaluated. The goal is to create inference procedures based on nonparametric representations of kernels K in terms of orthonormal bases in some Hilbert space $\mathcal{H}$. The theory outlined in section 4.1 suggests the existence of a decomposition into a weighted sum of complex exponentials when the reproducing kernel is translation invariant. This observation serves as a suitable starting point for constrained ℓ^2 norm minimization based kernel inference. It is easily implementable but the performance of the inferred kernels for estimation purposes can be disappointing necessitating further investigations into appropriate measures of closeness between kernels and the possibility of regularization.

4.2.1 Naive kernel inference

The optimal decomposition of a normalized translation invariant kernel into a weighted superposition of group characters requires solution algorithms for constrained least squares problems. One of these is the sequential coordinatewise algorithm for nonnegative least squares. Its performance for estimation of positive definite functions and kernels is assessed. While the approximative qualities of this approach are satisfying at first, serious drawbacks emerge as soon as the p.d. functions and kernels are interpreted as covariances and used for statistical inference. Further deficiencies become apparent when the approach is employed to estimate instationary kernels with unknown eigenfunctions and unknown linear relations to stationary kernels; i.e. when more is subject to incomplete information than simply the spectrum of the covariance operator C_K. The investigations leading up to these realizations are interesting in their own right as they pave the way for more sophisticated approaches of kernel inference and are presented in what follows.

From Bochner's theorem it is known that for each probability measure ν on $\hat{T}$, the formula $f(s-t) = \int_{\hat{T}} \chi_\omega(s-t)d\nu(\omega)$ defines a p.d. function $f(\cdot)$ on the LCA T. Particularly, this implies that for ν a weighted counting measure on some subset $W \subset \hat{T}$, the equation

$$K(s,t) = f(s-t) = \sum_{\omega \in W} \alpha_\omega \chi_\omega(s-t)$$

defines both a translation invariant kernel $k(\cdot,\cdot)$ and a p.d. function $f(\cdot)$ if $\alpha_\omega \geq 0 \ \forall \omega \in W$. This relation can be used for purposes of estimating non-parametrically a p.d. function from data. Let the LCA T be given and again introduce the $\mathbb{R}^n$-valued linear measurement operator A with domain either p.d. functions, kernels or as needed on a case-to-case basis. $K_{\text{emp}} \in \mathbb{R}^n$ will denote any empirical data used to narrow down the search for an appropriate $K(\cdot,\cdot)$, which satisfies $AK \approx K_{\text{emp}}$. The following three inference strategies suggest themselves immediately.

I If $W \subset \hat{T}$ is a finite subset with $|W| = m$, then set the estimator $K_{\text{est}}(s,t) \ \forall s,t \in T$ to

$$K_{\text{est}}(s,t) = f(s-t) = \sum_{\omega \in W} \alpha_\omega \chi_\omega(s-t)$$

where the parametervector $\alpha = \{\alpha_\omega\}_{\omega \in W} \in \mathbb{R}^m_+$ is determined as the solution to the optimization problem

$$\alpha^{\text{opt}} = \underset{\alpha \in \mathbb{R}^m_+}{\operatorname{argmin}} \ \|\Psi\alpha - K_{\text{emp}}\|_2^2 \tag{4.11}$$

Here $\Psi = [\psi_1, ..., \psi_m] \in \mathbb{C}^n \otimes \mathbb{C}^m$ with $\psi_j = A\chi_{\omega_j}(\cdot - \cdot)$ and ω_j is the j-th element of W.

II If $\{\varphi_j\}_{j=1}^m$ is a sequence of functions in $L^1(T)$, then set the estimator $K_{\text{est}}(s,t) \ \forall s,t \in T$ to

$$K_{\text{est}}(s,t) = f(s-t) = (g * g^*)(s-t) \quad g(\cdot) = \sum_{j=1}^m \beta_j \varphi_j(\cdot)$$

where the parametervector $\beta = \{\beta_j\}_{j=1}^m \in \mathbb{C}^m$ is determined as the solution to the optimization problem

$$\beta^{\text{opt}} = \operatorname*{argmin}_{\beta \in \mathbb{C}^m} \|\Psi v(\beta \otimes \overline{\beta}) - K_{\text{emp}}\|_2^2 \tag{4.12}$$

Here $v(\cdot) : \mathbb{C}^m \otimes \mathbb{C}^m \to \mathbb{C}^{m \cdot m}$ is the vectorization operator and $\Psi = [\psi_{11}, ..., \psi_{mm}] \in \mathbb{C}^n \otimes \mathbb{C}^{m \cdot m}$ with $\psi_{jk} = A \int_T \varphi_j(\cdot - \cdot + t)\varphi_k(t)d\mu(t)$.

III If $\{\varphi_j\}_{j=1}^\infty$ is an ONB of $L^2(T)$, then set the estimator $K_{\text{est}}(s,t)\ \forall s,t \in T$ to

$$K_{\text{est}}(s,t) = \sum_{j=1}^m \alpha_j \varphi_j(s)\varphi_j(t)$$

where the parameter vector $\alpha = \{\alpha_j\}_{j=1}^m \in \mathbb{R}_+^m$, m finite , is determined as the solution to the optimization problem

$$\alpha^{\text{opt}} = \operatorname*{argmin}_{\alpha \in \mathbb{R}_+^m} \|\Psi\alpha - K_{\text{emp}}\|_2^2 \tag{4.13}$$

Here $\Psi = [\psi_1, ..., \psi_m] \in \mathbb{C}^n \otimes \mathbb{C}^m$ with $\psi_j = A\varphi_j(\cdot)\varphi_j(\cdot)$.

The optimization problem posed in strategy II is obviously very complicated since the unknown parameter vector enters the objective function in a nonlinear fashion already before norms are evaluated and we will not pursue this approach further. Note, however, that if in II $\{\varphi_j\}_{j=1}^m$ is an orthonormal system w.r.t. $L^2(T,\mu)$, e.g. if $\varphi_j = \chi_{\omega_j}$ for T compact and abelian by Peter-Weyl theorem [63, p. 143], then

$$\begin{aligned} f(s) = (g * g^*)(s) &= \sum_{k=1}^m \sum_{l=1}^m \beta_k \overline{\beta_l} \int_T \chi_{\omega_k}(s-t)\overline{\chi_{\omega_l}}(-t)d\mu(t) \\ &= \sum_{k=1}^m \sum_{l=1}^m \beta_k \overline{\beta_l} \chi_{\omega_k}(s) \langle \chi_{\omega_l}, \chi_{\omega_k} \rangle_{L^2(T,\mu)} \\ &= \sum_{j=1}^m |\beta_j|^2 \chi_{\omega_j}(s) \end{aligned}$$

and processing strategy II reduces to I. The theoretical drawback of I and II is that only translation invariant kernels can be inferred. The same limitation does not hold for III which is slightly more general in this respect as setting φ_j to χ_{ω_j} recovers I. However, III is not a suitable strategy if the exact nature of K is not known beforehand as the family constructible via positive linear combinations of elements of $\cup_{j=1}^m \{\varphi_j \otimes \varphi_j\}$ differs only in terms of the kernel operator C_K's spectrum whereas the spectral family stays the same for all the families members. This is due to the fact that whatever $\alpha_j \in \mathbb{R}_+$ are chosen in the expansion

$K(s,t) = \sum_{j=1}^m \alpha_j \varphi_j(s)\varphi_j(t)$,

$$(C_K \varphi_i)(s) = \langle K(s,\cdot), \varphi_i(\cdot)\rangle_{L^2(T)} = \langle \sum_{j=1}^m \alpha_j \varphi_j(s)\varphi_j(\cdot), \varphi_i(\cdot)\rangle_{L^2(T)} = \alpha_i \varphi_i(s).$$

This is particularly obvious in the real finite-dimensional matrix case, where we can set φ_j to the canonical euclidean basis vector $e_j \in \mathbb{R}^n$ which implies that $C_K = \sum_{j=1}^n e_k \otimes e_k^*$ could only ever be a diagonal matrix. The number of freely choosable parameters $\alpha_j \in \mathbb{R}_+$ then amounts to n whereas the set of $n \times n$ covariance matrices has dimension at least $(n^2 - n)/2$ because the manifold of symmetric matrices has dimension $n(n+1)/2$ [45, p. 70] and only n further linear positivity constraints are needed to enforce positive semidefiniteness.

Regarding the practical implementation of I and III there are at least two options: Solve the constrained optimization problem for the parameters numerically with an algorithm for nonnegative least-squares; e.g. the sequential coordinatewise algorithm [67]. Alternatively solve a relaxation of the problem with $\alpha^{\text{opt}} \in \mathbb{R}_+^m$ replaced by $\alpha_2^{\text{opt}} \in \mathbb{R}^m$ via pseudoinverse methods and project the resulting α_2^{opt} onto $\mathbb{R}_+^m$ by discarding any of its negative entries. We briefly describe both for the practically relevant case where kernel and measurements are explicitly demanded to be real-valued. Since α has trivial imaginary part as well by Bochner's theorem, it holds that

$$\| \operatorname{Re}\left(AK_{\text{est}}(\cdot,\cdot) - K_{\text{emp}}\right) \|_2 = \| \operatorname{Re}\left(\Psi\alpha - K_{\text{emp}}\right) \|_2 = \| \operatorname{Re}(\Psi)\alpha - K_{\text{emp}}\|_2$$

and one might write the objective functions from equations 4.11-4.13 explicitly with Ψ replaced by $\operatorname{Re}(\Psi) = \Psi_r$.

The sequential coordinatewise algorithm [67] minimizes $(1/2)\|\Psi_r\alpha - K_{\text{emp}}\|_2^2, \alpha \in \mathbb{R}_+^m$ by first swapping it for the equivalent quadratic program

$$\begin{aligned} \min \ & \frac{1}{2}\langle \alpha, \Psi_r^* \Psi_r \alpha\rangle_{\mathbb{R}^m} + \langle \alpha, -2\Psi_r^* K_{\text{emp}}\rangle_{\mathbb{R}^m} \\ \text{subject to } \ & \alpha \succeq_{\mathfrak{C}} \ \text{ with } \mathfrak{C} \text{ the nonnegative orthant in } \mathbb{R}^m \end{aligned}$$

and then solving it by once initializing the algorithm in step i) and then applying repeatedly step ii) until convergence.

i) Set $\alpha^0 = 0, \mu^0 = -2\Psi_r^* K_{\text{emp}}$ and $H = \Psi_r^*\Psi_r = [h_1, ..., h_m]$.

ii) Loop through the index set $J = \{1, ..., m\}$ and update α_j^k to α_j^{k+1} where k denotes the k-th iteration by setting

$$\begin{aligned} \alpha_j^{k+1} &= \max\left(0, \alpha_j^k - \frac{\mu_j^k}{H_{jj}}\right) \\ \mu^{k+1} &= \mu^k + \left(\alpha_j^{k+1} - \alpha_j^k\right) h_j \end{aligned}$$

i.e. the vector μ is modified after each update of an $\alpha_j\ \forall j \in J$. One typically lets j run through J several times until a certain stopping criterion concerning closeness between the parameter vectors α^k and α^{k+1} as measured by the norm of their differences is reached.

The algorithm is known to always converge for nonnegative least-squares problems and typically does so rapidly, even for high-dimensional decision variables α [67].

The pseudoinverse method:

i) Solve $\alpha_2^{\text{opt}} = \underset{\alpha \in \mathbb{R}^m}{\operatorname{argmin}}\ \|\Psi_r \alpha - K_{\text{emp}}\|_2^2 = \Psi_r^+ K_{\text{emp}}$.

ii) Project α_2^{opt} onto its positive part via $\Pi_+ \alpha_2^{\text{opt}}$ where $\Pi_+ : \mathbb{R}^m \to \mathbb{R}^m$ is the orthogonal projection $\Pi_+ : v \mapsto \Pi_+ v = \{[v_k]_+\}_{k=1}^m$ that maps every entry v_k of a vector $v \in \mathbb{R}^m$ to 0 if $(\alpha_2^{\text{opt}})_k < 0$ and leaves it unperturbed otherwise.

iii) The vector $\Pi_+ \alpha_2^{\text{opt}}$ is a vector of coefficients such that $f = \sum_{j=1}^m \left(\Pi_+ \alpha_2^{\text{opt}}\right)_j \chi_{\omega_j}$ is positive definite and close to K_{emp}.

Since $\alpha_2^{\text{opt}} \in R^m$ is a solution to a relaxed problem and consequently the search space for the minimum is bigger ($\mathbb{R}^m \supset \mathbb{R}_+^m$), $\Psi_r \alpha_2^{\text{opt}}$ is closer to K_{emp} than $\Psi_r \alpha^{\text{opt}}$ (or equally far away). Furthermore $\Pi_+ \alpha_2^{\text{opt}}$ is an element of $\mathbb{R}_+^m$ but not necessarily the actually optimal α^{opt} implying $\Psi_r \alpha^{\text{opt}}$ to be closer to K_{emp} than $\Psi_r \Pi_+ \alpha_2^{\text{opt}}$ (or equally far away). This is summarizable as

$$\|\Psi_r \alpha_2^{\text{opt}} - K_{\text{emp}}\|_2 \leq \|\Psi_r \alpha^{\text{opt}} - K_{\text{emp}}\|_2 \leq \|\Psi_r \Pi_+ \alpha_2^{\text{opt}} - K_{\text{emp}}\|_2$$

and leads via triangle inequality to an upper bound for the difference between true optimum and approximation.

$$\begin{aligned} \|\Psi_r \alpha^{\text{opt}} - \Psi_r \Pi_+ \alpha_2^{\text{opt}}\|_2 &\leq \|\Psi_r \alpha^{\text{opt}} - K_{\text{emp}}\|_2 + \|\Psi_r \Pi_+ \alpha_2^{\text{opt}} - K_{\text{emp}}\|_2 \\ &\leq 2\|\Psi_r \Pi_+ \alpha_2^{\text{opt}} - K_{\text{emp}}\|_2 \end{aligned} \tag{4.14}$$

Strategies I and III are illustrated in figure 4.3 by applying them to exemplary data.

While strategy I's performance looks reasonable for the translation invariant squared exponential covariance matrix, it fails as soon as instationary CM's are encountered. Strategy III demands prior fixation of an ONB $\{\varphi_k\}_{k=1}^\infty$ of $L^2(T)$. If the right one is chosen, it succeeds (see figure 4.3 column 4, row 2) even for instationary CM — here the CM of Brownian motion on $T = [0, 1]$ with $\varphi_k(t) = \sqrt{2}\sin\left(\frac{\pi}{2}(2k+1)t\right)$. However, it fails for the example in figure 4.3 (column 4, row 3) because $\{\varphi_k\}_{k=1}^m$ is ”too far away” from the correct spectral family of the underlying CM.

Even though the approximations might seem sensible, the predictions resulting from estimated kernels $K_{\text{est}}(\cdot, \cdot)$ deviating from the true kernel only marginally in the mean-square sense can be almost arbitrarily bad. This is illustrated in figure 4.4 and the reasons for the observed shortcomings are twofold:

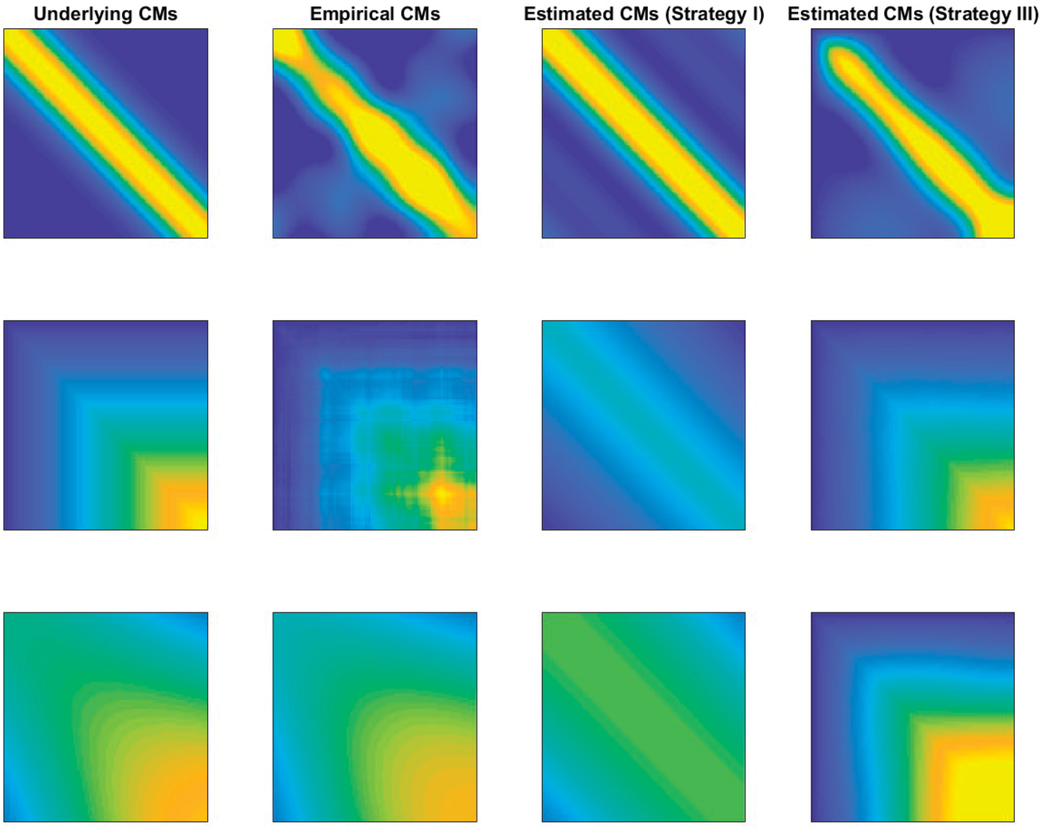


Figure 4.3: An illustration of strategy I and III's performance in estimating different types of covariance matrices from samples $AK(\cdot,\cdot) = K_{\text{emp}}$ where A is here simple evaluation and the data to be approximated are the full empirical covariance matrices. In the third column, strategy I and in the last column strategy III was employed to first estimate a covariance function $K(\cdot,\cdot)$ nonparametrically via the sequential coordinatewise minimization and then form a covariance matrix. To approximate the covariance function, 50 expansion terms were used. The colorscale is identical for all images.

i) Whereas the ℓ^2-norm is a good measure of the distance between functions, kernels mostly take on the role of key ingredients of linear operators, for which different measures of closeness are appropriate.

ii) When choosing the optimization objective to be $\|A_\otimes K - a\|_2^2$, $A_\otimes$ measurement operator, a data, K decision variable, no regularization term $\|K\|_{\mathcal{H}}$ related to the energy of K or any other indicator of its 'likelihood' is included in the problem formulation.

Insufficient reliability arising due to overfitting and lack of robustness are the consequence. One might conclude then, that it is necessary to investigate more appropriate measures of closeness between kernels than the simple ℓ^2-norm. Recall for example that the differentiability of $K(s,t) = f(t-s)$ around 0 is primarily responsible for the global smoothness properties of a stochastic process with covariance function $K(\cdot,\cdot)$. However, $\|\cdot\|_{\ell^2}$ does not distinguish between deviations around 0 and any other element of the parameter set T and therefore fails as a distance measure for kernels.

One of the reasons is that the Frobenius norm $\|A\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$ for matrices

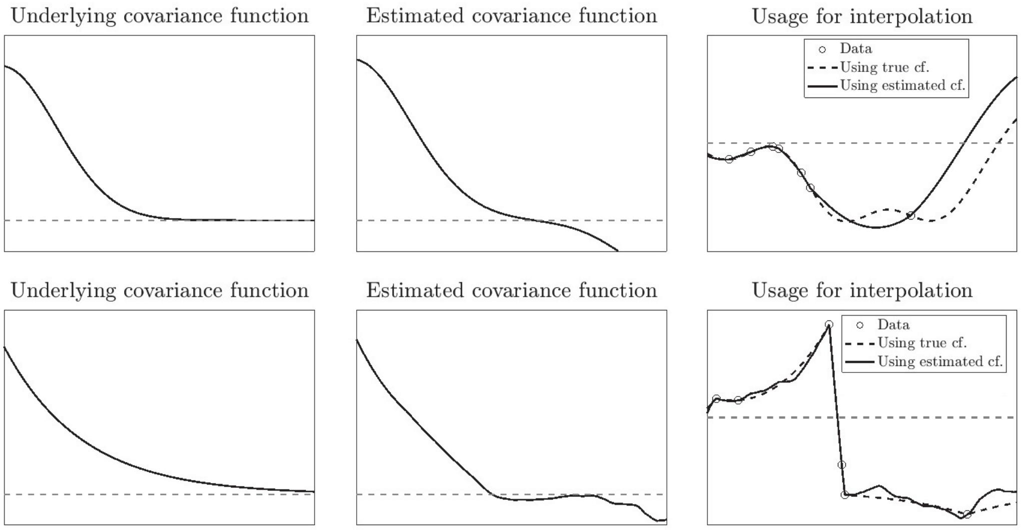


Figure 4.4: In the first column a squared exponential (smooth, infinitely often differentiable sample paths) and an exponential (nowhere differentiable, everywhere continuous sample paths) [162, pp. 83-85] covariance function are plotted. From some sample values the positive definite functions in the middle columns are inferred as estimates for the underlying covariance functions. Time series predictions using true and estimated covariances are plotted in the final two panels. Notice the large deviations. Y values of zero are marked by the dashed grey line.

$A \in \mathcal{B}(\mathbb{R}^n, \mathbb{R}^n)$ is not an $\ell_p - \ell_q$ operator norm (if it were $\|I\|_F$ would have to take the value 1) and as such fails to uphold inequalities of the type $\|Ax\|_p \leq \|A\|_{\text{op}}\|x\|_q$ that relate the norm of an operator $A : \mathcal{H} \to \mathcal{H}$ to its action on $\mathcal{H}$. Therefore the proper interpretation of $\|A\|_F$ is simply that of the Euclidean length of a vector in $\mathbb{R}^n \otimes \mathbb{R}^n$ equipped with the standard inner product and norm of $\mathbb{R}^{n^2}$. This is inappropriate both in the sense that $\|A\|_F$ tells very little of A seen as an operator and in the sense that the size of A is measured by considering the entries a_{ij} to be white noise (then $\|A\|_F \leftrightarrow$ probability density of A). A proper choice of norm should be related to a more suitable choice of probability distribution of the elements. The Wishart distribution as a generalization of the χ^2 distribution for matrix valued random variables is a natural candidate and regularization terms based on prior assumptions of this type will be investigated in the next section.

4.2.2 Regularization terms

If one wants to answer questions regarding which one of a choice of two kernels K_1, K_2 is the more reasonable option given some apriori preconceptions of the kernels behavior, a probabilistic perspective is helpful. We pursue ideas similar to those in section 3.1, where the link between norms and Gaussian processes helped reformulating statistical inference as an optimization problem in a quadratic form. If functions f_j are realizations of Gaussian processes then in analogy kernels K_j are realizations of a kernel valued process that must be a multivariate generalization of the χ^2-distribution due to the K_j being sums of squares according to the Mercer decomposition.

Likelihood based regularization terms require knowledge of the underlying probability distribution and may therefore be hard to justify. They can be swapped for functional analytic terms that express the intuitive belief that kernels K as correlation functions representing the laws governing a physical system are more likely to change smoothly than abruptly. Depending on the physical system whose state f is modelled as an element of $\mathcal{H}_K$ this assumption is incorrect to different degrees. However, this does not change the fact that in doing so one derives one of the few kernel inference formulations, for which a solution can be found. High likelihood, smoothness and robustness to inversion are all properties that make a kernel more useful for statistical inference and their relative importance depends on the task at hand and the preferences of the user.

Recall that covariance matrices K are positive elements of the C*-algebra of bounded operators in $\mathbb{R}^n$ and have therefore always a decomposition $K = A^*A$ for some $A \in \mathcal{B}(\mathbb{R}^n)$ [48, p. 15]. This decomposition may be used to incorporate prior knowledge about K's nature while prescribing only a stochastic and flexible model for K by

i) demanding A to be a random matrix drawn from a distribution related to prior assumptions on K as expressed by a guess C_p for K.

ii) setting $K = A^*A$ which is automatically positive semidefinite, has expected value $E[K] = C_p$ and a probability density suitable for inference.

Specific parts of this strategy and its properties are collected in the next theorem.

Theorem 4.2.1 *Let a prior covariance matrix $C_p = \sum_{i=1}^n \lambda_i \varphi_i \otimes \varphi_i^*$ be prescribed and then set*

$$A = \frac{1}{\sqrt{\operatorname{tr} C_p}} \sum_{i=1}^n \sum_{j=1}^n \xi_{ij} \sqrt{\lambda_i \lambda_j} \varphi_i \otimes \varphi_j^* \qquad \xi_{ij} \sim \mathcal{N}(0,1). \tag{4.15}$$

The matrix A can be interpreted as both a random field on $T \times T$ and a linear operator indexed by elements of $T \times T$ where T is the indexset of the process for which inference is to be performed. When denoting by $C(x)$ the function $E[x \otimes x^*] - E[x] \otimes E[x]^*$ and extending the trace to the usual tensor contraction, one finds

I $E[A] = 0$, *i.e.* A *has mean* 0.

II $C(A) = E[A \otimes A^*] = \frac{1}{\operatorname{tr} C_p} (C_p \otimes C_p)^{T_{24}}$, *i.e. A apart from some reordering has covariance proportional to the tensor product of the prior guess C_p. Here T_{24} denotes higher order transposition which interchanges the second and fourth dimension of a tensor.*

III $E[A^*A] = C_p$ *,i.e. the expected correlation for the model* $K = A^*A$ *is* C_p.

Proof: The statements are straightforward to verify. For I, note simply that $E[A] = \frac{1}{\sqrt{\operatorname{tr} C_p}} \sum_{i=1}^n \sum_{j=1}^n E[\xi_{ij}] \sqrt{\lambda_i \lambda_j} \varphi_i \otimes \varphi_j^* = 0$. Similarly for II, we find

$$\begin{aligned}
C(A) &= E[A \otimes A^*] - E[A] \otimes E[A]^* \\
&= \frac{1}{\operatorname{tr} C_p} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n E[\xi_{ij}\xi_{kl}] \sqrt{\lambda_i \lambda_j \lambda_k \lambda_l} \varphi_i \otimes \varphi_j^* \otimes \varphi_l \otimes \varphi_k^* \\
&= \frac{1}{\operatorname{tr} C_p} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \varphi_i \otimes \varphi_j^* \otimes \varphi_j \otimes \varphi_i^* \\
&= \frac{1}{\operatorname{tr} C_p} \left(C_p \otimes C_p\right)^{T_{24}}
\end{aligned}$$

and III follows from

$$\begin{aligned}
E[A^*A] &= \frac{1}{\operatorname{tr} C_p} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n E[\xi_{ij}\xi_{kl}] \sqrt{\lambda_i \lambda_j \lambda_k \lambda_l} \varphi_j \otimes \varphi_l^* \langle \varphi_i, \varphi_k \rangle \\
&= \frac{1}{\operatorname{tr} C_p} \left(\sum_{i=1}^n \lambda_i \right) \sum_{j=1}^n \lambda_j \varphi_j \otimes \varphi_j^* \\
&= C_p
\end{aligned}$$

□

It is necessary to derive information that will help to evaluate the probability density of A^*A. A simple calculation results in

$$\begin{aligned}
A^*A &= \frac{1}{\operatorname{tr} C_p} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n \xi_{ij}\xi_{kl} \sqrt{\lambda_i \lambda_j \lambda_k \lambda_l} (\varphi_j \otimes \varphi_i^*)(\varphi_k \otimes \varphi_l^*) \\
&= \frac{1}{\operatorname{tr} C_p} \sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^n \xi_{ij}\xi_{il} \lambda_i \sqrt{\lambda_j \lambda_l} \varphi_j \otimes \varphi_l^* \\
&= \frac{1}{\operatorname{tr} C_p} \sum_{j=1}^n \sum_{l=1}^n \varphi_j \otimes \varphi_l^* \underbrace{\left(\sum_{i=1}^n \xi_{ij}\xi_{il} \lambda_i \sqrt{\lambda_j \lambda_l} \right)}_{\gamma_{jl}} \\
&= \frac{1}{\operatorname{tr} C_p} \sum_{j=1}^n \sum_{l=1}^n \gamma_{jl} \varphi_j \otimes \varphi_l^* \qquad (4.16)
\end{aligned}$$

It is possible to approximate the distribution of the matrix γ with $(\gamma)_{ij} = \gamma_{ij}$ for $i, j = 1, ..., n$ as having a Wishart probability density. The goal is to prove this. To this end, start noticing that

$$\gamma = \sqrt{\Lambda}\ \ \Xi\ \ \Lambda\ \ \Xi^*\ \ \sqrt{\Lambda}^* \tag{4.17}$$

where Λ is an $n \times n$ diagonal matrix with entries $(\Lambda)_{ii} = \lambda_i$ and Ξ is an $n \times n$ random matrix with $(\Xi)_{ij} = \xi_{ji} \sim \mathcal{N}(0,1)$. To see this, note that if $\psi = \sqrt{\Lambda}\Xi = [\psi_1, ..., \psi_n]$ one has $\gamma = \psi\Lambda\psi^* = \sum_{i=1}^n \lambda_i \psi_i \otimes \psi_i^*$ and

$$\sum_{i=1}^n \lambda_i (\psi_i \otimes \psi_i^*)_{jl} = \sum_{i=1}^n \lambda_i \sqrt{\lambda_j \lambda_l} \xi_{ij} \xi_{il} = \gamma_{jl}$$

as defined in equation 4.16 and claimed by equation 4.17. This obviously implies the decomposition

$$\gamma = \sum_{i=1}^n \lambda_i \psi_i \otimes \psi_i^* = \sum_{i=1}^n \lambda_i S_i$$

where $\psi_i = [\sqrt{\lambda_1}\xi_{i1}, ..., \sqrt{\lambda_n}\xi_{in}]^T \sim \mathcal{N}(0, \Lambda)$ from which it follows that $S_i \sim \mathcal{W}_n(1, \Lambda)$ for $i = 1, ..., n$. Typically the sum $c_1 M_1 + c_2 M_2, c_1, c_2 > 0$ of two Wishart distributed matrices M_1, M_2 is not Wishart distributed anymore and the sums probability density is given by a rather complicated term involving confluent hypergeometric functions of the first kind [121]; optimization and maximum likelihood estimation become seemingly intractable.

Adopting an approximation approach proposed in [158] and slightly extending it to cover sums of Wishart matrices whose expected values are not simply multiples of the identity, we exchange the term

$$\gamma = \sum_{i=1}^n \lambda_i S_i \qquad\qquad S_i \sim \mathcal{W}_n(1, \Lambda)$$

for an approximation $\tilde{\gamma}$ defined as

$$\tilde{\gamma} = \frac{\operatorname{tr} C_p}{q} S \qquad\qquad S \sim \mathcal{W}_n(q, \Lambda) \tag{4.18}$$

$$q = \lceil [\operatorname{tr}(C_p^2)]^{-1} [\operatorname{tr}(C_p)^2] \rfloor \tag{4.19}$$

where $\lceil \cdot \rfloor$ is the operator taking a real number to the nearest integer. The approximation is motivated by the fact that the expected value of $\tilde{\gamma}$ equals $E[\gamma]$ and the second moments of γ and $\tilde{\gamma}$ coincide approximately on the diagonal [158]:

$$E[\gamma] = \sum_{i=1}^n \lambda_i E[S_i] = \sum_{i=1}^n \lambda_i \Lambda = \frac{\operatorname{tr} C_p}{q}(q\Lambda) = E[\tilde{\gamma}]$$

$$\operatorname{Var}[(\gamma)_{kk}] = \operatorname{Cov}\left(\sum_{i=1}^n \lambda_i S_i^{kk}, \sum_{j=1}^n \lambda_j S_j^{kk}\right)$$

$$\begin{aligned}
&\overset{S_i \coprod S_j, i\neq j}{=} \sum_{i=1}^{n} \lambda_i^2 \operatorname{Cov}\left(S_i^{kk}, S_j^{kk}\right) = 2\lambda_k^2 \operatorname{tr}(C_p^2) \\
\operatorname{Var}[(\tilde{\gamma})_{kk}] &= \frac{\operatorname{tr}(C_p)^2}{q^2} \operatorname{Var}[S_k k] \\
&= \frac{\operatorname{tr}(C_p)^2}{q} 2\lambda_k^2 = 2\lambda_k^2 \left\lceil \frac{\operatorname{tr}(C_p^2)}{\operatorname{tr}(C_p)^2} \right\rceil \operatorname{tr}(C_p)^2
\end{aligned}$$

When $[\operatorname{tr}(C_p^2)]^{-1}[\operatorname{tr}(C_p)^2]$ is approximately integer, then $\operatorname{Var}[(\gamma)_{kk}] \approx \operatorname{Var}[(\tilde{\gamma})_{kk}]$. For the calculations we employed [160, p. 115] to derive the variance of diagonal elements of Wishart distributed matrices.

Therefore γ is approximately distributed like $\tilde{\gamma} \sim \mathcal{W}_n(q, \Lambda)$; equivalently in terms of the probability density function associated to the Wishart distribution [160, p. 108] $p_{\tilde{\gamma}}(V)$ for any positive definite $V \in S_{++}^n$ with entries $(V)_{ij} = v_{ij}$

$$\begin{aligned}
p_{\tilde{\gamma}}(V) &= \frac{1}{2^{\frac{nq}{2}} |\Lambda|^{\frac{q}{2}} \Gamma_n\left(\frac{q}{2}\right)} |V|^{\left(\frac{q-n-1}{2}\right)} \exp\left(-\frac{1}{2} \operatorname{tr}(\Lambda^{-1} V)\right) \\
&= \left[2^{\frac{nq}{2}} \Gamma_n\left(\frac{q}{2}\right) \prod_{j-1}^{n} \lambda_j^{q/2} \right]^{-1} |V|^{\left(\frac{q-n-1}{2}\right)} \exp\left(-\frac{1}{2} \sum_{j=1}^{n} \frac{v_{jj}}{\lambda_j}\right) \qquad (4.20)
\end{aligned}$$

in which $\Gamma(\cdot)$ is the gamma function. As routinely encountered before, $-\log p_{\tilde{\gamma}}(V)$ is more suitable for inference. By denoting the appropriate term constant in V by c_0, one finds

$$-\log p_{\tilde{\gamma}}(V) = c_0 - \left(\frac{q-n-1}{2}\right) \log\det V + \frac{1}{2} \sum_{j=1}^{n} \frac{v_{jj}}{\lambda_j} \qquad (4.21)$$

This formula allows expressing the likelihood of a positive definite random $A^*A = \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} \varphi_i \otimes \varphi_j^*$ with $E[A^*A] = C_p$ for some prespecified prior C_p. In successive chapters equation 4.21 will be employed to numerically determine the maximum aposteriori (MAP) estimate for covariance matrices and kernels by supposing the model $K(\gamma) = \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij} \varphi_i \otimes \varphi_j^*$ with parameters $(\gamma)_{ij} \;\; i, j = 1, ..., n$ and interchanging $p_{\gamma|f}(\gamma|f) \propto p_{f|\gamma}(f|\gamma) p_{\gamma}(\gamma)$ with the more easily optimizable $p_{f|\gamma}(f|\gamma) p_{\tilde{\gamma}}(\gamma)$ for some observations f.

Note that the mode of the Wishart distribution $\mathcal{W}_n(q, \Lambda)$ lies at $(q-n-1)\Lambda$ [59, p. 653]. Therefore in specifying q we have the alternative choice of setting $q = n+2$ which guarantees that the mode of the distribution is exactly Λ but induces a deviation in variance of the diagonal elements compared to the original model stemming from $A^*A = \sum_{i,j=1}^{m} \varphi_i \otimes \varphi_j^*$. Compared to the classical Wishart assumption $K \sim \mathcal{W}_n(q, C_p)$, expressing K as a random variable A^*A results in more flexible class of models, see figure 4.5.

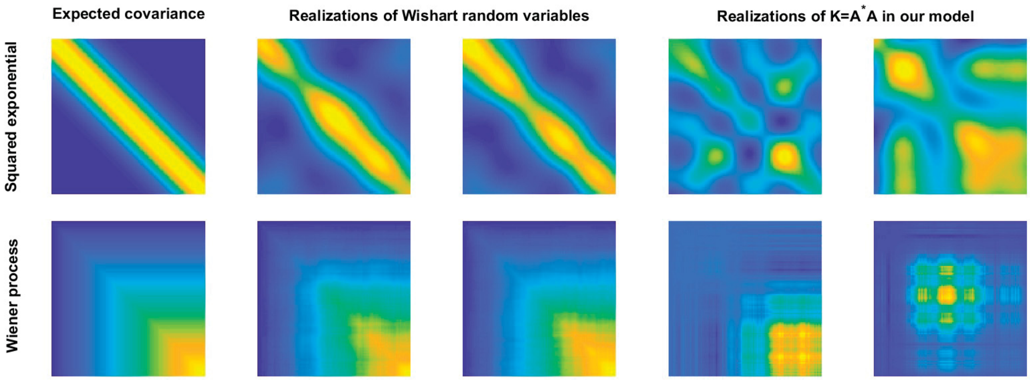


Figure 4.5: Comparison between two different probability distributions on covariance matrices. The expected values C_p of the random variables (first row: squared exponential; second row Wiener covariance matrix) are shown in column one. Columns two and three show realizations of Wishart random matrices $K \sim \mathcal{W}_n(n, C_p)$ while the two rightmost columns show realizations of $K = A^*A$ for A a Gaussian random field with $E[A^*A] = C_p$. Note that the random field model is able to generate highly flexible covariance matrices unlike the model based on directly sampling from a Wishart distribution.

4.2.3 Formulation of kernel inference tasks

When kernel inference is formulated as an optimization problem with a mixture of terms measuring smoothness and plausibility given observations, the interpretation is similar as in the abstract spline framework. In the previous subsection a probabilistic perspective was developed that allowed to concretely express the prior likelihood of a kernel. An analysis of the multivariate normal distribution extends this line of approach to derive equations relating the likelihood of a kernel to observation data and the resulting optimization problem features simple linear terms as well as complexity penalties involving determinants. The optimization problem has similarities to a generalization of semidefinite programs called maxdet problems which have received attention in the past due to their links to multivariate statistics and experiment design. We relate our formulation of kernel inference to this problem class and investigate certain trivial subcases. The theoretical analysis is aided by results provided by implementations of numerical procedures recorded in the literature.

For K an $n \times n$ positive definite matrix, the probability density function $p_f(\cdot)$ of a multivariate Gaussian $f \sim \mathcal{N}(\mu, K)$ is given by [160, p. 68]

$$p_f(f_\omega) = \left[\sqrt{(2\pi)^n \det K}\right]^{-1} \exp\left(-\frac{1}{2}(f_\omega - \mu)^T K^{-1}(f_\omega - \mu)\right) \tag{4.22}$$

$$= (2\pi)^{-n/2}(\det K)^{-1/2} \exp\left(-\frac{1}{2}\|f_\omega - \mu\|^2_{\mathcal{H}_k}\right) \tag{4.23}$$

where we employ the custom of denoting a specific realization of the random variable $f : \Omega \ni \omega \mapsto f_\omega \in \mathbb{R}$ by f_ω where ω is an element of the probability space Ω. Suppose now that a certain fixed f_ω was observed and the task was to infer the most probable K from a θ-parametrized family of covariance matrices explaining the data. The Bayes-law for probability densities suggests investigation of the term

$$p_{\theta|f}(\theta|f_\omega) = \frac{p_{f|\theta}(f_\omega|\theta)p_\theta(\theta)}{p_f(f_\omega)}$$

which is the posterior distribution of the parameters given the observations and an obvious objective function to be optimized w.r.t. θ to arrive at an estimate of K. To

derive this so called maximum aposteriori estimator, the normalization constant in the denominator can be discarded as it does not depend on θ. Furthermore, instead of maximizing $p_{\theta|f}(\theta|f_\omega)$ one might instead minimize $-\log p_{\theta|f}(\theta|f_\omega)$.
Finally one arrives at

$$\begin{aligned}\hat{\theta}_{MAP} &= \operatorname*{argmin}_{\theta\in\Theta} \quad -\log p_{f|\theta}(f_\omega|\theta) - \log p_\theta(\theta)\\ &= \operatorname*{argmin}_{\theta\in\Theta} \quad \frac{n}{2}\log 2\pi + \frac{1}{2}\log\det K(\theta) + \frac{1}{2}\|f_\omega - \mu\|^2_{\mathcal{H}_{K(\theta)}} - \log p_\theta(\theta)\end{aligned}$$

with Θ the set of permissible values for θ. The term $p_{f|\theta}(f_\omega|\theta) = \mathcal{L}(\theta; f_\omega)$ is also called the likelihood of θ. In absence of prior knowledge on the distribution of θ, it is a common practice to either assume a uniform prior on Θ or to just ignore $p_\theta(\theta)$ altogether leading in both cases to a maximum likelihood (ML) estimate for K [162, pp. 112-114] whose properties and construction we briefly sketch. More details are found in a dedicated chapter on inference of hyperparameters for Gaussian processes in the book by Rasmussen and Williams [162, pp. 105-125].

$$\begin{aligned}\hat{\theta}_{ML} &= \operatorname*{argmin}_{\theta\in\Theta} \quad \log\det K(\theta) + \|f_\omega - \mu\|^2_{\mathcal{H}_{K(\theta)}}\\ &= \operatorname*{argmin}_{\theta\in\Theta} \quad \sum_{\lambda_j\in\sigma(K(\theta))} \lambda_j + \frac{\langle(f_\omega-\mu),\varphi_j\rangle^2_{\mathbb{R}^n}}{\lambda_j}\end{aligned} \tag{4.24}$$

where $\sigma(K)$ denotes the spectrum of $K(\theta)$ and φ_j is the eigenfunction of $K(\theta)$ corresponding to eigenvalue λ_j. Seen purely from a perspective of RKHS based estimation, the term $\|f_\omega - \mu\|^2_{\mathcal{H}_{K(\theta)}}$ quantifies the likelihood of the residuals $f_\omega - \mu$ whereas terms of type $\log\det K$ not expressible via norms have not been encountered before. They are often referred to as complexity penalties [162, p. 113] and are vital for kernel estimation as the following train of thought suggests, in which we compare the results of estimating a covariance matrix in absence (situation **A**) or presence (situation **B**) of the complexity penalty. For simplicity's sake, we suppose $\mu = 0$.

A) If we ignored the complexity penalty and would instead try to find a covariance matrix K such that the negative log likelihood of f_ω as represented by $\|f_\omega\|^2_{\mathcal{H}_K}$ was minimal for some prespecified f_ω, there would be a simple solution. As the maximizer K_{opt} for the likelihood of f_ω satisfies

$$K_{opt} = \operatorname*{argmin}_{K\in S^n_+} \ \|f_\omega\|^2_{\mathcal{H}_K} = \operatorname*{argmin}_{K\in S^n_+} \ f_\omega^T K^{-1} f_\omega$$

it is clear that by choosing an arbitrary p.d. K and scaling it by some large $\alpha \in \mathbb{R}$ the term $f_\omega^T(\alpha K)^{-1}f_\omega = \frac{1}{\alpha}f_\omega^T K^{-1} f_\omega$ can be driven arbitrarily close to 0. Without complexity penalty, the best guess for K making f_ω most likely would therefore be any p.d. matrix scaled to large proportions as to keep $\|K^{-1}\|_{op}$ small — this is obviously degenerated behavior to be avoided. ■

B) Now in contrast to item A, investigate the same situation with the complexity penalty present. Given the true K, K_{true}, suppose the goal is to find $\alpha \in \mathbb{R}_+$ such that $E[\log\det(\alpha K_{\text{true}}) + \|f\|^2_{\mathcal{H}(\alpha K_{true})}]$ is minimal. One has

$$\begin{aligned} E[\log\det(\alpha K_{\text{true}}) + f^T(\alpha K_{\text{true}})^{-1} f] &= \log\det(\alpha K_{\text{true}}) + \frac{1}{\alpha} E[\operatorname{tr}(f \otimes f^T K_{\text{true}})] \\ &= n\alpha + \log\det K_{\text{true}} + \frac{1}{\alpha}\operatorname{tr} I \\ &= n\left(\alpha + \frac{1}{\alpha}\right) + \log\det K_{\text{true}} \end{aligned}$$

To find the α minimizing the last expression in item B, solve equivalently

$$\alpha_{opt} = \operatorname*{argmin}_{\alpha \in \mathbb{R}_+} \quad \alpha + \frac{1}{\alpha}$$

where the first part of the optimization objective comes from the complexity penalty and drives α towards 0. The second term originates from the data-fit penalty and would again, analogously to the case discussed in A, drive α to ∞. Solving this problem is possible analytically. Notice that

$$\frac{\partial}{\partial\alpha}\left(\alpha + \frac{1}{\alpha}\right) = 1 - \frac{1}{\alpha^2} = 0$$

if and only if $\alpha^2 = 1$; i.e. $\alpha \in \{1, -1\}$ with $\alpha = -1$ being ruled out by reasons of positive semidefiniteness. Furthermore the Hessian is given by

$$\frac{\partial^2}{\partial\alpha^2}\left(\alpha + \frac{1}{\alpha}\right) = \frac{1}{\alpha^3} \geq 0 \quad \text{for} \quad \alpha \in \mathbb{R}_+$$

implying $\alpha = 1$ to be indeed a minimum and $K_{opt} = \alpha_{opt} K_{\text{true}} = K_{\text{true}}$. It is therefore demonstrated that the addition of the log det term can successfully avoid the degeneracies encountered in the naive formulation $K_{opt} = \operatorname{argmin}_{K \in S^n_+} \|f_\omega\|^2_{\mathcal{H}_K}$ that find their expression in preference of unboundedly scaled versions of K_{true}.

■

A system similar to equation 4.24 has been investigated by Mardia and Marshal [133] in the context of spatial statistics. They assume the mean function $\mu(t), t \in T = \{1, ..., n\}$ to be unknown as well but of type $\mu(t) = \sum_{j=1}^{n_\beta} \beta_j g_j(t)$ for some given family $\{g_j\}_{j=1}^{n_\beta}$ and suggest for the joint estimation of $\beta \in \mathbb{R}^{n_\beta}$ and $\theta \in \mathbb{R}^{n_\theta}$ the following Levenberg-Marquard type of algorithm.

1. Start with an initial, potentially randomized guess $\theta_0 \in \Theta$ of the parameters θ. Compute $\beta_0 = (G^T K_0^{-1} G)^{-1} G^T K_0^{-1} f_\omega$ where $K_0 = K(\theta_0)$ and G is an $n \times n_\beta$ regressor matrix with $(G)_{ij} = g_j(t_i)$. The initial estimate is $\phi_0 = (\beta_0, \theta_0)$.

2. Calculate the gradients ∇_β and ∇_θ of the log likelihood w.r.t the mean and

covariance function parameters β and θ. They are

$$\nabla_\beta = -G^T K_0^{-1} G\beta_0 + G^T K_0^{-1} f_\omega$$
$$(\nabla_\theta)_j = -\frac{1}{2}\operatorname{tr}\left(K_0^{-1}\frac{\partial K}{\partial\theta_j}\right) + \frac{1}{2}\left(f_\omega - G\beta_0\right)^T K_0^{-1}\frac{\partial K}{\partial\theta_j} K_0^{-1}\left(f_\omega - G\beta_0\right)$$

3. Update $\phi_1 = \phi_0 + B^{-1}(\nabla_\beta, \nabla_\theta)^T$ where B is block diagonal with $B{=}B_\beta \oplus B_\theta$, $B_\beta = G^T K_0^{-1} G$ and the individual elements of B_θ given as

$$(B_\theta)_{ij} = \frac{1}{2}\operatorname{tr}\left(K_0^{-1}\frac{\partial K}{\partial\theta_i}K_0^{-1}\frac{\partial K}{\partial\theta_j}\right).$$

4. Set ϕ_1 as the initial estimate in step 1. and repeat until convergence.

Note that in a practical implementation, it is recommended to work with the logarithm of typical parameters θ representing variances or ranges in covariance function models to avoid volatile behavior near the origin. The problem is not convex and local optima may exist [162, p. 115] which suggests to start either with good initial values for β and θ or to solve a sequence of problems with β and θ initialized randomly and then pick the set of parameters with the highest likelihood as given by equation 4.22. The implementation is straightforward otherwise and the maximum likelihood estimator of covariance functions as described above seems to behave favorably compared to the character-based estimation, see figure 4.6 and compare to the predictions in figure 4.4. The supposition of already known parametric type of covariance function is often approximately justified but can be a drawback.

If in model 4.22 one refrains from using covariance functions that are nonlinear in the parameters, one recovers a convex optimization problem known as maxdet and closely related to semidefinite programming.

Theorem 4.2.2 *Fix an ONB $\{\varphi_j\}_{j=1}^\infty$ of the real (separable, countably infinite) Hilbert space $L^2(T)$ for $T \subset \mathbb{R}^p$ and truncate the sequence at $n \in \mathbb{N}$ giving the orthonormal sequence $\{\varphi_j\}_{j=1}^n$. A linear superposition $K(\cdot,\cdot) = \sum_{i=1}^n \alpha_i\varphi_i(\cdot)\varphi(\cdot)$ with nonnegative α_i is obviously positive (semi)definite; the more flexible*

$$K(\cdot,\cdot) = \sum_{i=1}^n\sum_{j=1}^n \alpha_{ij}\varphi_i(\cdot)\varphi_j(\cdot)$$

is positive definite if and only if $A \in \mathbb{R}^n \otimes \mathbb{R}^n$, $(A)_{ij} = \alpha_{ij}$ is positive semidefinite as a matrix.

Proof: $\forall m \in \mathbb{N}$ and $\forall \{t_k\}_{k=1}^n \subset T, \{\beta\}_{k=1}^m \subset \mathbb{R}$, one has

$$\sum_{k=1}^m\sum_{l=1}^m \beta_k\beta_l K(t_k,t_l) = \sum_{k=1}^m\sum_{l=1}^m\sum_{i=1}^n\sum_{j=1}^n \alpha_{ij}\beta_k\varphi_i(t_k)\beta_l\varphi_j(t_l)$$

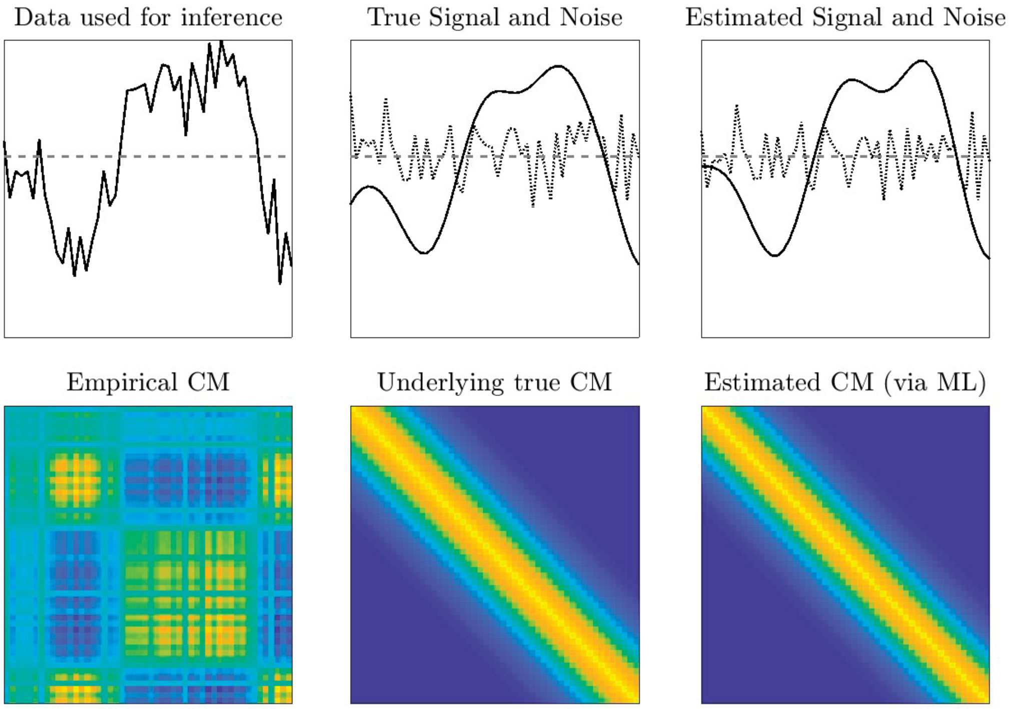


Figure 4.6: The result of estimating a covariance function (squared exponential + white noise) based on observation data and a correctly specified parametric model, the exact parameters of which are unknown. The inferred kernel exhibits convincing performance when used to derive solutions to a signal separation problem.

$$
\begin{aligned}
&= \sum_{i=1}^{n} \underbrace{\sum_{k=1}^{m} \beta_k \varphi_i(t_k)}_{\gamma_i} \alpha_{ij} \sum_{j=1}^{n} \underbrace{\sum_{l=1}^{m} \beta_l \varphi_j(t_l)}_{\gamma_j} \\
&= \gamma^T A \gamma
\end{aligned}
$$

where $\gamma \in \mathbb{R}^n$ is a vector with components $(\gamma)_i = \gamma_i$. Now one may proceed to show necessity and sufficiency of $\gamma^T A\gamma \geq 0 \quad \forall \gamma \in \mathbb{R}^n$ for K to be a positive (semi)definite kernel. Sufficiency is simple; if $\gamma^T A \gamma \geq 0 \quad \forall\gamma \in \mathbb{R}^n$, then also $\sum_{k=1}^m \sum_{l=1}^m \beta_k \beta_l K(t_k, t_l) \geq 0$.

Necessity follow from the fact that if $\exists\gamma$ with $\gamma^T A\gamma < 0$ then one may associate a corresponding $\beta \in \mathbb{R}^m$ and a sequence $\{t_k\}_{k=1}^m \subset T$ such that $\gamma_i = \sum_{k=1}^m \beta_k \varphi_i(t_k)$ for all $i = 1, ..., n$. Constructability is guaranteed by solvability of $\gamma = \phi\beta$ for β where $\phi \in \mathbb{R}^n \otimes \mathbb{R}^m$ with $\phi_{ik} = \varphi_i(t_k)$ which in turn depends on the existence of a solution β with $\phi\beta = \gamma$. Therefore if $\exists$ a right inverse ϕ^+ with $\phi\phi^+ = I$, $\beta = \phi^+\gamma$ satisfies

$$\sum_{k=1}^{m} \sum_{l=1}^{m} \beta_k \beta_l K(t_k, t_l) = \gamma^T A \gamma < 0$$

establishing $K(\cdot, \cdot)$ as not positive semidefinite. Existence of ϕ^+ is asserted by the following lemma. □

Lemma 4.2.3 *Let $\{\varphi_i\}_{i=1}^n$ be an orthonormal sequence in a separable Hilbert space $\mathcal{H}$ of functions on T. Then $\exists m \in \mathbb{N}$ and $\{t_k\}_{k=1}^m \subset T$ such that the matrix with entries $(\phi)_{ik} = \varphi_i(t_k)$ has a right inverse.*

Proof: The matrix ϕ has a right inverse iff it is of full row rank. Note that the statement of ϕ having full row rank is trivial for $n = 1$ as the orthonormality condition ensures that $\exists t_1 \in T : \varphi_1(t_1) \neq 0$. Then proceed by complete induction through dimensions q up to n. Suppose that the matrix

$$\phi_q = \begin{bmatrix} \varphi_1(t_1) & \cdots & \varphi_1(t_q) \\ \vdots & \ddots & \vdots \\ \varphi_q(t_1) & \cdots & \varphi_q(t_q) \end{bmatrix} = \begin{bmatrix} \text{—}(\varphi_1^q)\text{—} \\ \vdots \\ \text{—}(\varphi_q^q)\text{—} \end{bmatrix} \in \mathbb{R}^q \otimes \mathbb{R}^q$$

has full row rank of q. For $q + 1 \leq n$, if ϕ_q is augmented with the row vector $\varphi_{q+1}^q = [\varphi_{q+1}(t_1), ...\varphi_{q+1}(t_q)]$ to form a matrix of dimension $(q+1) \times q$, then that matrix has row rank q since row rank equals column rank equals q. Then for the unique $c \in \mathbb{R}^q$ such that $\varphi_{q+1}^q = \sum_{j=1}^q c_j \varphi_j^q$, there exists at least one $t_{q+1} \in T$ with $\varphi_{q+1}(t_{q+1}) \neq \sum_{j=1}^q c_j \varphi_j(t_{q+1})$ as otherwise $\langle \varphi_{q+1}, \sum_{j=1}^q c_j \varphi_j \rangle \neq 0$ in violation to the presupposition that $\{\varphi_j\}_{j=1}^n$ is an ONS. Consequently, for any sequence of coefficients $c_j, j = 1, ..., q$ s.t. $\sum_{j=1}^q c_j \varphi_j^q = \varphi_{q+1}^q$, one finds $\sum_{j=1}^q c_j \varphi_j(t_{q+1}) \neq \varphi_{q+1}(t_{1+1})$ and the matrix ϕ_{q+1} has full row rank admitting a right inverse. □

Now suppose the task is to find the sequence of coefficients $\alpha = \{\alpha_{ij}\}_{i \leq j=1}^m$ such that the function

$$K(s,t) = \sum_{i \leq j=1}^m \alpha_{ij} \left(1 - \frac{1}{2}\delta_{ij}\right) [\varphi_i(s)\varphi_j(t) + \varphi_j(s)\varphi_i(t)]$$

best explains the data $Af = f_\omega$ for $f \in \mathcal{H}(T)$, $f_\omega \in \mathbb{R}^n$ and A evaluation at $\{t_k\}_{k=1}^n$ in the maximum likelihood sense. If one denotes the covariance matrix $A \otimes AK(\cdot, \cdot)$ induced by the kernel function $K(\cdot, \cdot)$ by K and sets

$$F(\alpha) = \sum_{i \leq j=1}^m \alpha_{ij}\phi_{ij} \quad , (\phi_{ij})_{kl} = \left(1 - \frac{1}{2}\delta_{ij}\right) [\varphi_i(t_k)\varphi_j(t_l) + \varphi_j(t_k)\varphi_i(t_l)],$$

optimizing α_{ij} with respect to the likelihood $\mathcal{L}(\alpha; f_\omega) = p_{f|K}(f_\omega | F(\alpha))$ leads to

$$\begin{aligned} \text{minimize} \quad & \log\det F(\alpha) + f_\omega^T F(\alpha)^{-1} f_\omega \\ \text{subject to} \quad & F(\alpha) \succeq 0. \end{aligned}$$

However, $\log\det F$ is concave in F [28, p. 73] precluding minimization [28, p. 356]. Reformulating the estimation problem in terms not of the kernel $K(\cdot, \cdot)$ but of a function $Q(\cdot, \cdot)$ such that the matrix Q with $(Q)_{kl} = Q(t_k, t_l)$ is the inverse covariance matrix changes it into a convex optimization problem [199]. When setting

$$Q(s,t) = \sum_{i\leq j=1}^{m} \alpha_{ij}\left(1-\frac{1}{2}\delta_{ij}\right)[\varphi_i(s)\varphi_j(t)+\varphi_j(s)\varphi_i(t)]$$
$$Q(\alpha) = \sum_{i\leq j=1}^{n} \alpha_{ij}\phi_{ij} \quad ,(\phi_{ij})_{kl} = \left(1-\frac{1}{2}\delta_{ij}\right)[\varphi_i(t_k)\varphi_j(t_l)+\varphi_i(t_l)\varphi_j(t_k)]$$

and interpreting $Q(\alpha)$ as an estimator for K_{true}^{-1}, maximization of the likelihood finds its expression in the optimization problem

$$\begin{aligned} \text{maximize} &\quad -\log\det Q(\alpha)^{-1} - f_\omega^T Q(\alpha) f_\omega \\ \text{subject to} &\quad Q(\alpha) \succeq 0 \end{aligned}$$

which can be slightly simplified to

$$\begin{aligned} \text{minimize} &\quad \log\det Q(\alpha)^{-1} + \operatorname{tr}(f_\omega \otimes f_\omega^T Q(\alpha)) \\ \text{subject to} &\quad Q(\alpha) \succeq 0. \end{aligned}$$

This is a maxdet problem with decision variables $\{\alpha_{ij}\}_{i\leq j=1}^{m}$ and linear term $\langle c, \alpha\rangle$ for c a vector with $(c)_{ij} = 2(1-\frac{1}{2}\delta_{ij})\langle f_\omega, A\varphi_i\rangle\langle f_\omega, A\varphi_j\rangle$ since, denoting $A\varphi_i = \psi_i$ for all $i \in \mathbb{N}$

$$\begin{aligned} \operatorname{tr}(f_\omega \otimes f_\omega^T Q(\alpha)) &= \langle f_\omega \otimes f_\omega^T, Q(\alpha)\rangle_F \\ &= \sum_{i\leq j=1}^{m} \alpha_{ij}\langle f_\omega \otimes f_\omega^T, \left(1-\frac{1}{2}\delta_{ij}\right)[\psi_i \otimes \psi_j^T + \psi_j \otimes \psi_i^T]\rangle_F \\ &= \sum_{i\leq j=1}^{m} \alpha_{ij} 2\left(1-\frac{1}{2}\delta_{ij}\right)\langle f_\omega, \psi_i\rangle\langle f_\omega, \psi_j\rangle. \end{aligned}$$

This follows from the Frobenius inner products behavior on simple tensors:

$$\begin{aligned} \langle a\otimes b^T, c\otimes d^T\rangle_F = \operatorname{tr}((a\otimes b^T)^T c\otimes d^T) &= \operatorname{tr}(b\otimes a^T c\otimes d^T) \\ &= a^T c b^T d \end{aligned}$$

The formula is readily extendable to the case where more than one f_ω was observed, say a sequence $\{f_{\omega_i}\}_{i=1}^{n_{\text{obs}}}$ in which case maximum likelihood estimation of the inverse covariance matrix K^{-1} is given by

$$\begin{aligned} \text{minimize} &\quad \log\det Q(\alpha)^{-1} + \langle K_{emp}, Q(\alpha)\rangle_F \\ \text{subject to} &\quad Q(\alpha) \succeq 0. \end{aligned} \tag{4.25}$$

If Q is subjugated to no constraint whatsoever, $m = n$, and the $\{\psi_i\}_{i=1}^{n}$ provide the full canonical euclidean basis for $\mathbb{R}^n$, the solution to 4.25 is the inverse of the empirical covariance matrix $K_{emp} = \frac{1}{n_{\text{obs}}} \sum_{j=1}^{n_{\text{obs}}} f_{\omega_j} \otimes f_{\omega_j}^T$ (provided it exists) [199][211].

Remark Rename the vector of logarithms of eigenvalues to $\{\log \lambda_j\}_{j=1}^n = l_\lambda \in \mathbb{R}^n$ and notice the formal similarity between the log det complexity penalty $cp(K)$, the Riemannian distance between positive definite matrices and linear/bilinear forms. One has

$$\begin{aligned} d^2(K, I) = \| \log K^{-1/2} I K^{-1/2} \|_F^2 \qquad &= \sum_{j=1}^n |\log \lambda_j|^2 \qquad &&= \langle l_\lambda, l_\lambda \rangle_{\mathbb{R}^n} \\ cp(K) = \log \det K \qquad &= \sum_{j=1}^n \log \lambda_j \qquad &&= \langle \mathbf{1}, l_\lambda \rangle_{\mathbb{R}^n} \end{aligned}$$

with $\mathbf{1}$ a vector of ones in $\mathbb{R}^n$. To introduce a prior on K whose purpose it is to prefer matrices K close to some prior guess $C_p \in S_{++}^n$, one might set $p_\theta(\theta) = c \exp\left(-(1/2) d^2(K(\theta), C_p)\right)$. Assume that $K(\theta)$ lies in the von-Neumann-Algebra generated by C_p, i.e. it commutes with C_p and their spectral families coincide. Then the MAP is

$$\begin{aligned} \hat{\theta}_{MAP} &= \underset{\theta \in \Theta}{\text{argmin}} \quad (f_\omega - \mu)^T K(\theta)^{-1} (f_\omega - \mu) + cp(K) + d^2(K, C_p) \\ &= \underset{\theta \in \Theta}{\text{argmin}} \quad \sum_{j=1}^n |\langle f_\omega - \mu, \varphi_j \rangle_{\mathbb{R}^n}|^2 e^{-(l_\lambda^K)_j} + \langle \mathbf{1}, l_\lambda^K \rangle_{\mathbb{R}^n} + \| l_\lambda^K - l_\lambda^{C_p} \|_{\mathbb{R}^n}^2 \end{aligned}$$

which is a quadratic program in l_λ^K with an additional data term of the form $\| f_\omega - \mu \|_{K(\theta)}^2$.

Alternatively, and this turns out to be easier, it is possible to directly include a prior on the coefficients α by specifying a prior covariance matrix $C_p = \sum_{i=1}^n \lambda_i \varphi_i \otimes \varphi_i^*$ and employing the induced probability distribution on the coefficients α derived in previous subsections. Several approximations and assumptions are integral to this formulation — the whole procedure including problem specification, mathematical setup, and the associated optimization task, is outlined below. In the future, the author will refer to this and similar procedures as 'kernel inference'.

Problem: Given measurements $Lf = \{f(t_k)\}_{k=1}^n = f_\omega$ of some function f element of the RKHS $\mathcal{H}_K$ on T, $L \in \mathcal{B}(\mathcal{H}_K, \mathbb{R}^n)$ being evaluation, infer optimally the reproducing kernel $K(\cdot, \cdot)$ employing some prior guess K_p. Interpreted as a stochastic process, the functions f chosen at random from the RKHS have mean μ equal to zero.
Given: $Lf \in \mathbb{R}^n$, $K_p(s,t) = \sum_{i=1}^n \lambda_i^p \varphi_i^p(s) \otimes \varphi_i^p(t)$, the evaluation operator L.
Assume: The true underlying kernel $K(\cdot, \cdot)$ is actually a realization of a degenerate, finite rank kernel valued random variable that is at the same time a stochastic process on $T \times T$. One may write $K(\cdot, \cdot) = K^\omega(\cdot, \cdot)$ for $\omega \in \Omega, \Omega$ some probability

space and

$$K^\omega(\cdot,\cdot) = \frac{1}{\operatorname{tr} K_p}\sum_{i=1}^m\sum_{j=1}^m \alpha_{ij}\varphi_i(\cdot)\otimes\varphi_j(\cdot)^* \quad \alpha\succeq 0 \quad \alpha_{ij} = \sum_{k=1}^n \xi_{ki}\xi_{kj}\lambda_k.$$

The coefficient matrix α is to be inferred and can be assembled to provide a good estimator for $K^\omega(\cdot,\cdot)$ even when only the analogue finite dimensional problem with covariance matrices replacing kernels is solved.

Setup: Replace $K_p(\cdot,\cdot)$ by $C_p = L\otimes LK_p = \{K_p(t_i,t_j)\}_{i,j=1}^n$ and write $C_p = \sum_{i=1}^n \lambda_i\varphi_i\otimes\varphi_i^*$ where λ_i and φ_i are the eigenvalues and eigenvectors of C_p. If $C = L\otimes LK^\omega = \{K(t_i,t_j)\}_{i,j=1}^n$ is the true underlying covariance matrix for f_ω based on evaluating the kernel K^ω; $E[f_\omega\otimes f_\omega^*] = C$; then estimate C via

$$\hat{C} = \frac{1}{\operatorname{tr} C_p}\sum_{i=1}^m\sum_{j=1}^m \alpha_{ij}\varphi_i\otimes\varphi_j^*$$
$$\alpha\sim\mathcal{W}_m(q,\Lambda), \quad q = \left\lceil[\operatorname{tr}(C_p^2)]^{-1}[(\operatorname{tr} C_p)^2]\right\rceil$$

where Λ is the $m\times m$ diagonal matrix containing $\lambda_1,...,\lambda_m$. Solve for α, the positive semidefinite matrix of coefficients α_{ij} by solving an optimization problem.

Optimize: The maximum aposteriori optimization problem for the determination of the $\alpha\in S_+^m$ explaining best the data $f_\omega\in\mathbb{R}^n$ while generating a covariance matrix close to C_p is

$$\hat{\alpha}_{MAP} = \underset{\alpha\in S_+^m}{\operatorname{argmax}}\ p_{f|\alpha}(f_\omega|\alpha)p_\alpha(\alpha) \tag{4.26}$$

which we replace by

$$\hat{\alpha} = \underset{\alpha\in S_+^m}{\operatorname{argmax}}\ p_{f|\alpha}(f_\omega|\alpha)p_{\tilde{\alpha}}(\alpha) \tag{4.27}$$

where $p_{f|\alpha}$ is multivariate Gaussian and $p_{\tilde{\alpha}}$ is Wishart according to equation 4.20. Consequently one may equivalently minimize $-\log p_{f|\alpha}(f_\omega|\alpha)p_{\tilde{\alpha}}(\alpha)$ where $\alpha\succeq 0$ are the decision variables. ■

The rest of this section deals with the simplification of equation 4.27 towards an optimization problem solvable with standard numerical methods. Denoting $p_{f|\alpha}(f_\omega|\alpha)p_{\tilde{\alpha}}(\alpha)$ by $\tilde{p}(\alpha,f_\omega)$ and using $\tilde{l}(\alpha,f_\omega)$ for $-\log\tilde{p}(\alpha,f_\omega)$, then for some constant c the usual formulas for the probability density functions [160, p. 108,p. 68] imply

$$\tilde{p}(\alpha,f_\omega) = \left[(2\pi)^{n/2}2^{nq/2}\Gamma_n\left(\frac{q}{2}\right)\prod_{j=1}^m \lambda_j^{q/2}\sqrt{\det F(\alpha)}\right]^{-1}[\det G(\alpha)]^{\left(\frac{q-n-1}{2}\right)}$$

$$\cdot \exp\left(-\frac{1}{2}\left[\operatorname{tr}(\Lambda^{-1}G(\alpha)) + f_\omega^T F(\alpha)^{-1} f_\omega\right]\right). \tag{4.28}$$

Consequently the negative log density is found — apart from some constants irrelevant to the optimization performed later — as

$$\begin{aligned}\tilde{l}(\alpha, f_\omega) = \frac{1}{2}&\left[\log\det F(\alpha) + f_\omega^T F(\alpha)^{-1}\right] \\ &+ r\left[-\left(\frac{q-n-1}{2}\right)\log\det G(\alpha) + \frac{1}{2}\operatorname{tr}(\Lambda^{-1}G(\alpha))\right] \end{aligned} \tag{4.29}$$

where $r = 1$ is a parameter that can be changed to adjust the weight of the prior and

$$F(\alpha) = \frac{1}{\sqrt{\operatorname{tr} C_p}} \sum_{i \le j = 1}^{m} \alpha_{ij}[1 - (1/2)\delta_{ij}]\left[\varphi_i \otimes \varphi_j^* + \varphi_j \otimes \varphi_i^*\right] \in \mathbb{R}^n \otimes \mathbb{R}^n$$

$$G(\alpha) = \frac{1}{\sqrt{\operatorname{tr} C_p}} \sum_{i \le j = 1}^{m} \alpha_{ij}[1 - (1/2)\delta_{ij}]\left[e_i \otimes e_j^* + e_j \otimes e_i^*\right] \in \mathbb{R}^m \otimes \mathbb{R}^m$$

with $e_k, k = 1, ..., m$ being the canonical euclidean basis in $\mathbb{R}^m$. By splitting $\tilde{l}(\alpha, f_\omega)$ into the two terms

$$\tilde{l}_1(\alpha, f_\omega) = \frac{1}{2}\left[\log\det F(\alpha) + f_\omega^T F(\alpha)^{-1} f_\omega\right] \tag{4.30}$$

$$\tilde{l}_2(\alpha, f_\omega) = r\left[-\left(\frac{q-n-1}{2}\right)\log\det G(\alpha) + \frac{1}{2}\operatorname{tr}(\Lambda^{-1}G(\alpha))\right] \tag{4.31}$$

and subsequently applying the substitution $F(\alpha)^{-1} = Q(\alpha)$ while using $-\log\det A = \log(\det A)^{-1} = \log\det A^{-1}$ one may realize that both $\tilde{l}_1(\alpha, f_\omega)$ and $\tilde{l}_2(\alpha, f_\omega)$ can in principle be interpreted as constituents of two maxdet problems that have been joined and need to be solved simultaneously. [204] reports solvability of a closely related but more complicated problem via Block-Coordinate-Descent; implementation details will be given later. One might extend the framework to include only indirect measurements based on arbitrary bounded linear operators A more general than just pointwise evaluation and incorporate a joint estimation of mean and covariance functions similar to what was done in [133]. We leave this for future work.

The roles of $\tilde{l}_1$ and $\tilde{l}_2$ are clear. The solution to the pure $\tilde{l}_1$ problem

$$X^1_{\text{opt}} = \underset{X \in S^n_+}{\operatorname{argmin}}\ \frac{1}{2}\log|X| + \frac{1}{2}\operatorname{tr}\left(SX^{-1}\right) \tag{4.32}$$

is the empirical covariance matrix $X^1_{\text{opt}} = S = [n_{\text{obs}}]^{-1}\sum_{j=1}^{n_{\text{obs}}} f_{\omega_j} \otimes f^*_{\omega_j}$; the first term drives X to zero and the second term prefers large X. In direct comparison, the solution to the pure $\tilde{l}_2$ problem

$$X^2_{\text{opt}} = \underset{X \in S^n_+}{\operatorname{argmin}} \quad -\frac{1}{2}\log|X| + \frac{1}{2}\operatorname{tr}\left(\Lambda^{-1}X\right) \tag{4.33}$$

with $q = n+2$ is $X^2_{\text{opt}} = \Lambda$. The first term prefers large X while the second one drives X to zero. Only solving the $\tilde{l}_1$ problem amounts to absolute data fidelity while solving the $\tilde{l}_2$ problem amounts to ignoring the data and choosing the prior covariance. A weighted mixture of both objectives is therefore a natural target for optimization. As before, the derivation of $\tilde{l}(\alpha, f_\omega)$ is extendable to cover more than one observation f_ω. If the coefficient vector α is swapped for the coefficient matrix γ and the more straightforward model $C(\gamma) = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}\varphi_i \otimes \varphi_j^*, \gamma \succeq_{S^n_+} 0$ is employed, the calculations become less cluttered. Assume n_{obs} independent observations $f_{\omega_1}, ..., f_{\omega_{n_{\text{obs}}}} \in \mathbb{R}^n$ are given. Then

$$p_{\gamma|f}\left(\gamma|f_{\omega_1}, ..., f_{\omega_{n_{\text{obs}}}}\right) = \prod_{j=1}^{n_{\text{obs}}} p_{f|\gamma}\left(f_{\omega_j}|\gamma\right) p_\gamma(\gamma) \approx \prod_{j=1}^{n_{\text{obs}}} p_{f|\gamma}(f_{\omega_j}|\gamma)\tilde{p}(\gamma)$$

and the last term can be simplified to yields for $\prod_{j=1}^{n_{\text{obs}}} p_{f|\gamma}(f_{\omega_j}|\gamma)\tilde{p}(\gamma)$ the term

$$\begin{aligned}
&\prod_{j=1}^{n_{\text{obs}}} \frac{1}{\sqrt{2\pi}^n |C(\gamma)|^{1/2}} \exp\left(-\frac{1}{2} f^*_{\omega_j} C^{-1}(\gamma) f_{\omega_j}\right) \\
&\qquad\qquad \left[\frac{1}{\sqrt{2}^{nq}\Gamma_n\left(q/2\right)|\Lambda|^{n/2}} |\gamma|^{\frac{q-n-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\Lambda^{-1}\gamma\right)\right)\right] \\
&= \tilde{c}|C(\gamma)|^{\frac{-n_{\text{obs}}}{2}} |\gamma|^{\frac{q-n-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\Lambda^{-1}\gamma\right) - \frac{1}{2}\sum_{j=1}^{n_{\text{obs}}} f^*_{\omega_j} C^{-1}(\gamma) f_{\omega_j}\right) \\
&=: \tilde{p}(\gamma, f_{\omega_1}, ..., f_{\omega_{n_{\text{obs}}}})
\end{aligned}$$

for some normalization constant $\tilde{c}$. Writing

$$\sum_{j=1}^{n_{\text{obs}}} f_{\omega_j} C^{-1}(\gamma) f_{\omega_j} = \sum_{j=1}^{n_{\text{obs}}} \operatorname{tr}\left(f_{\omega_j} \otimes f^*_{\omega_j} C^{-1}\right) = n_{\text{obs}} \operatorname{tr}\left(SC^{-1}\right)$$

with S being the empirical covariance matrix, and taking the negative logarithm $\tilde{l} = -\log \tilde{p}(\gamma, f_{\omega_1}, ..., f_{\omega_{n_{\text{obs}}}})$ one finds

$$\begin{aligned}
\tilde{l}(\gamma, S) = -\log(\tilde{c}) + n_{\text{obs}} &\left[\frac{1}{2}\log|C(\gamma)| + \frac{1}{2}\operatorname{tr}\left(SC^{-1}(\gamma)\right)\right] \\
&+ \left[-\left(\frac{q-n-1}{2}\right)\log|\gamma| + \frac{1}{2}\operatorname{tr}\left(\Lambda^{-1}\gamma\right)\right]
\end{aligned} \tag{4.34}$$

Discarding the constant $-\log(\tilde{c})$, minimizing $\tilde{l}(\gamma, S)$ with respect to γ is equivalent to minimizing $L(\gamma)$,

$$L(\gamma) = \underbrace{\left[\log|C(\gamma)| + \operatorname{tr}\left(SC^{-1}(\gamma)\right)\right]}_{L^1(\gamma,S)} + \underbrace{\frac{1}{n_{\text{obs}}}\left[-c\log|\gamma| + \operatorname{tr}\left(\Lambda^{-1}\gamma\right)\right]}_{rL^2(\gamma,\Lambda)} \tag{4.35}$$

for $r = n_{\text{obs}}^{-1}, c = (q - n - 1)$. More observations therefore simply affect the optimization objective via a decrease of the regularization parameter r.

4.3 Practical implementation

The kernel inference formulations encountered in subsection 4.2.3 exhibit constraints on the spectrum of a symmetric matrix and amount to demanding that the decision variables can be assembled to lie in the convex cone of positive semidefinite matrices. Convex optimization problems with covariance matrix valued decision variables are called semidefinite programs (SDP) and include as subclasses linear, quadratic and second order cone programs. Many tasks in geodesy such as robust estimation, ℓ^p-norm minimization, variance components estimation and campaign design problems can be stated and solved in the SDP framework although the SDP-embeddings are often nonobvious and the solution algorithms are less reliable and less performant than those for linear or quadratic programs. Typical implementations are based on interior point methods: A self concordant barrier function is added to the objective function and acts to enforce the constraints, then a sequential minimization of incrementally modified problems generates a curve of solutions that converges to the solution of the original problem in the limiting case. This procedure will be modified to cope with the task of approximately minimizing expression 4.35 in order to derive a finite dimensional matrix representation of the positive definite kernel most compatible with prior knowledge and observed data. The algorithm can be extended in several directions; the most important one is the inclusion of affine constraints.

4.3.1 Semidefinite programming

> *Semidefinite programs are optimization problems in a matrix X with a linear objective function expressible as $\langle C, X\rangle_F$ subject to linear equalities $\langle A_i, X\rangle_F = b_i$ and the cone constraint $X \succeq_{S^n_+} 0$ demanding the decision variable to lie in the convex cone S^n_+ of $n \times n$ positive semidefinite matrices . The class of semidefinite programs properly includes as subclasses linear (LP) quadratic (QP) and second order cone programs (SOCP) and thus also generalizes least-squares based estimation procedures. Consequently, they arise naturally in statistics as soon as robustness (as measured by the ℓ^1-norm), maximum-likelihood estimation under a joint Gaussian assumption (as measured by the ℓ^2-norm) or the properties of covariance matrices (as measured by the distribution of their spectra) are of interest.*

The positive definiteness constraint is not easy to enforce and requires algorithms significantly different from the ones known to work for LP, QP and SOCP. Some of the most successful ones trace a continuous path through the interior of the feasible set by solving a sequence of problems in which the constraint $X \succeq_{S^n_+} 0$ is replaced by adding the self-concordant barrier function $-\mu \det X$ to the objective. The steps undertaken to derive these interior point algorithms can be modified slightly to construct a practically implementable algorithm for kernel inference.

Definition 4.3.1 A function $f : \mathbb{R}^n \to \mathbb{R}$ is called convex if it satisfies

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2)$$

for all $x_1, x_2 \in \mathbb{R}^n$ and $\alpha \in [0,1]$ [149, p. 112].

The definition implies that for a convex function f, every line connecting two codomain elements $f(x_1)$ and $f(x_2)$ lies above the function values $f(\alpha x_2 + (1-\alpha)x_1), \alpha \in [0,1]$ generated by applying f to the line connecting two domain elements x_1 and x_2. For these functions, any local optimum is globally optimal [28, p. 138] and they automatically possess a weak analogue of differentiability. Subgradients may then be defined to approximate f's local behavior and lower bound its minimum value, so that a generalized type of gradient descent is at least in theory always available to perform minimization [149, pp. 141-143].

Many convex optimization problems can be written as ones with linear objective and conic constraints [5]; the generality of the order induced by the cones $\mathfrak{C}$ through $X \succeq_{\mathfrak{C}} Y \Leftrightarrow X - Y \in \mathfrak{C}$ corresponds to the generality of the convex optimization problem. A simple cone is the nonnegative orthant $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x_k \geq 0 \;\; k = 1...n\}$. Cone constraints of this type appear in linear programming whereas the so, called Lorenz cone $\mathfrak{C}_L := \{(x,t) : \|x\|_2 \leq t\}$ allows a linear formulation of quadratically constrained quadratic programs which include least squares problems as trivial cases. A more general partial order containing $\succeq_{\mathbb{R}^n_+}$ and $\succeq_{\mathfrak{C}_L}$ is induced by the cone S^n_+ of positive semidefinite $n \times n$ matrices. If constraints of the type $\sum_j x_j A_j \in S^n_+$ are imposed, one speaks of a linear matrix inequality (LMI) and the whole problem of finding $X \succeq_{S^n_+} 0$ to minimize $\langle C, X\rangle_F$ subject to linear equalities is a semidefinite program (SDP). See figure 4.7 for an illustration of different cones involved and notice the heavy nonlinearities implicit in the cone constraints as visible in the boundaries of the feasible sets.

Definition 4.3.2 (SDP) A semidefinite program in standard form [210, p. 113] is an optimization problem

$$\begin{aligned} \text{minimize} \;\; & \langle C, X\rangle_F \\ \text{subject to} \;\; & \langle A_i, X\rangle_F = b_i \qquad i = 1, ..., n_c \\ & X \succeq_{S^n_+} 0 \end{aligned} \tag{4.36}$$

where $\langle C, X\rangle_F = \operatorname{tr}(C^* X)$ is called the objective function, $\langle A_i, X\rangle_F = b_i$ is a set of n_c affine linear constraints, $X \succeq_{S^n_+} 0$ is the cone constraint and the positive semidefinite matrix $X \in \mathbb{R}^n \otimes \mathbb{R}^n$ is the decision variable to be chosen as to

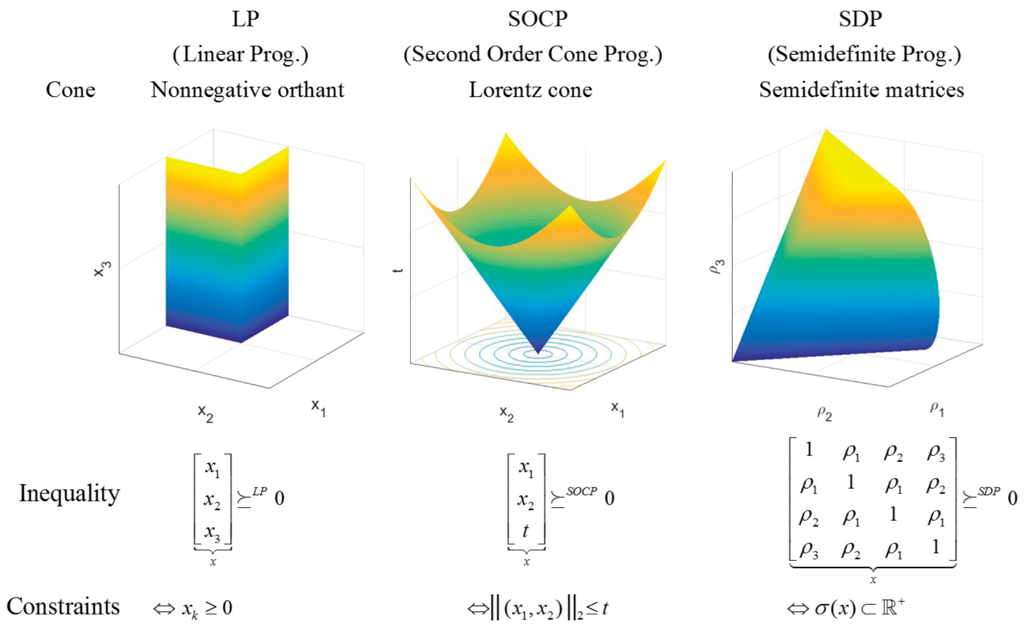

LP
(Linear Prog.)
Nonnegative orthant
SOCP
(Second Order Cone Prog.)
Lorentz cone
SDP
(Semidefinite Prog.)
Semidefinite matrices
Cone
x3
x2
x1
t
x2
x1
ρ3
ρ2
ρ1
Inequality
$\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{x} \succeq^{LP} 0$
$\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ t \end{bmatrix}}_{x} \succeq^{SOCP} 0$
$\underbrace{\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}}_{x} \succeq^{SDP} 0$
Constraints
$\Leftrightarrow x_k \geq 0$
$\Leftrightarrow \|(x_1, x_2)\|_2 \leq t$
$\Leftrightarrow \sigma(x) \subset \mathbb{R}^+$

Figure 4.7: An illustration of the positive orthant, the Lorenz cone and the semidefinite cone in $\mathbb{R}^3$. The latter is sometimes also called a spectrahedron; its specific form depends on the exact type of inequality. The one plotted here is constructed on the basis of a parametrization taken from [200].

minimize the objective function.

Alternatively, SDPs can be provided in inequality form as a minimization problem with vector valued decision variable and a linear matrix inequality [28, p. 169]. They are then written as

$$\begin{aligned} &\text{minimize } \langle C, X\rangle_F \\ &\text{subject to } F_0 + \sum_{i=1}^{m} x_i F_i \succeq_{S_+^n} 0 \end{aligned} \tag{4.37}$$

with the F_i symmetric $n \times n$ matrices for $i = 0, ..., m$. This form is especially suitable to demonstrate embeddability of LPs and QPs into SDPs when one furthermore employs the following lemma.

Lemma 4.3.3 (Schur complement) *The* $(n+m) \times (n+m)$ *dimensional symmetric matrix*

$$M = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \qquad A \in S_{++}^n, C \in S_{++}^m$$

is positive semidefinite if and only iff $C \succeq_{S_+^n} B^* A^+ B$ *[210, p. 21].*

Example 23 (LP) Employing the relationship between positive definiteness of a matrix and the positive definiteness of its submatrices, it immediately follows that a linear program in standard form

$$\begin{aligned} \text{minimize } & \langle c, x\rangle_{\mathbb{R}^n} \\ \text{subject to } & Ax = b \\ & x \geq 0 \end{aligned}$$

is nothing more than an SDP in standard form with non-diagonal elements of the matrix X in equation 4.36 constrained to be zero [112, p. 3]. This constraint is obviously linear and LPs are therefore also often called diagonal SDPs. ■

Example 24 (QP) Similarly, minimizing the quadratic form $\|Ax - a\|_{\ell^2}^2 = (Ax - a)^*(Ax - a)$ can be embedded into the SDP

$$\begin{aligned} \text{minimize } & t \\ \text{subject to } & \begin{bmatrix} tI & Ax - a \\ (Ax - a)^* & 1 \end{bmatrix} \succeq_{S_+^n} 0 \end{aligned}$$

where the decision variable is the $n + 1$-dimensional vector $[t; x]$. By the Schur complement lemma, the constraint is satisfied, iff $tI \succeq_{S_+^n} 0$ and $1 - t^{-1}(Ax - a)^*(Ax - a) > 0 \Leftrightarrow t > \|Ax - a\|_{\ell^2}^2$ therefore forcing $x \in \mathbb{R}^n$ to take the values minimizing $\|Ax - a\|_{\ell^2}^2$. As the matrix appearing in the semidefiniteness constraint can also be written as

$$\begin{bmatrix} tI & Ax - a \\ (Ax - a)^* & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & -a \\ -a^* & 1 \end{bmatrix}}_{F_0} + t\underbrace{\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}}_{F_t} + x_1\underbrace{\begin{bmatrix} 0 & a_1 \\ a_1^* & 0 \end{bmatrix}}_{F_1} + ... + x_n\underbrace{\begin{bmatrix} 0 & a_n \\ a_n^* & 0 \end{bmatrix}}_{F_n}$$

with $A = [a_1, ..., a_n]$ and all the matrices F symmetric, the problem is clearly an SDP in inequality form. ■

Adherence to a cone constraint of type $X \succeq_{S_+^n} 0$ is implicitly guaranteed by adding the barrier-term $-\mu \log |X|, \mu \geq 0$, to the minimization objective $\langle C, X\rangle_F$. Roughly explained, $-\mu \log |X|$ grows towards infinity on the border of the semidefinite cone ($|X| = 0$ iff X is singular) enforcing a strictly positive definite solution when performing unconstrained minimization of

$$l^\mu := \langle C, X\rangle_F - \mu \log |X|$$

with respect to X. Therefore for strictly positive μ this produces a minimizer X_μ that lies in the interior of the semidefinite cone; solving a sequence of minimizations of terms $\{l^{\mu_j}\}_{j=1}^\infty$ with $\lim \mu_j = 0$ and the sequence $\{\mu_j\}_{j=1}^\infty$ of barrier weights decreasing monotonically traces out a curve through the interior of the semidefinite cone that converges to the correct solution of problem 4.36 and is known as the primal path.

More detailed information on these interior point algorithms including convergence guarantees and theoretical justifications can be found in Nesterov's book on convex optimization [149, pp. 192-210] and in [112, p. 77], where specific schemes of type

$$\begin{aligned} X^{j+1} &= \text{argmin } \; l^{\mu_j}(X), & X^{\text{initial}} &= X^j \\ \mu_{j+1} &= \alpha\mu_j, & \alpha &\in [0,1) \end{aligned} \tag{4.38}$$

are outlined which construct and follow the primal path even for cones $\mathfrak{C}$ different from S^n_+ if $\log|X|$ is replaced by a self-concordant barrier function for $\mathfrak{C}$. Solving the subproblems of minimizing l^{μ_j} appearing in algorithm 4.38 is efficiently possible with Newtons method, because the self-concordant barrier function $-\mu\log|X|$ has gradients and Hessians that can be calculated explicitly as $\nabla l^\mu = C - \mu X^+$ and $\nabla^2 l^\mu = \mu X^+ \otimes X^+$ respectively [112, p. 79].

To every convex optimization problem $\mathcal{P}$ there exists a dual problem $\mathcal{D}$. In the SDP case a problem is given as $\min_{X\succeq 0}\langle C, X\rangle_F$ subject to $\langle A_i, X\rangle_F = b_i, i = 1, ..., m$ and its Lagrangian dual is given as $\max_{S\succeq 0, y\in\mathbb{R}^m}\langle b, y\rangle_{\mathbb{R}^m}$ subject to $\sum_{i=1}^n y_i A_i + S = C$. Notice the similarity of $\mathcal{D}$ and the LMI formulation of an SDP. In a very coarse sense the constraints in the primal problem are stated in terms of membership to a cone $\mathfrak{C}$ whereas its dual makes use of the dual cone $\mathfrak{C}^* := \{X \in \mathbb{R}^{n^2} : \langle X, Y\rangle_{\mathbb{R}^{n^2}} = 0 \;\; \forall Y \in \mathfrak{C}\}$. The cone S^n_+ of positive semidefinite $n \times n$ matrices is one of only three self-dual cones over the reals, the other ones being $\mathbb{R}^n_+$ and the Lorentz cone $\mathfrak{C}_L$ ([112], p. 21).

The main information a dual offers about its corresponding primal are lower bounds on the latter's optimal value. For any feasible point (y, S) of the dual, $\langle y, b\rangle_{\mathbb{R}^m} \leq \langle C, X\rangle_F \;\; \forall X$ feasible for the primal problem and the difference of these two values is called the duality gap. If $\mathcal{P}$ and $\mathcal{D}$ fulfill the Slater constraint qualification, the primal's optimal value coincides with the dual's optimal value and it is reasonable to manipulate both decision variables of $\mathcal{P}$ and $\mathcal{D}$ to minimize the duality gap. In the limit the result is a smooth curve through feasible primal and dual variables — the central path — that as in the purely primal case converges to the solutions of $\mathcal{P}$ and $\mathcal{D}$.

One of these primal-dual path following methods is Mehrotras predictor corrector method [138]. It is comprising of a sequence of steps that roughly follow the idea outlined above and are implemented in existing openly accessible software packages SeDuMi by Sturm [192] and SDPT3 by Toh et al.[195]. Both software implementations do not require feasible starting points and are as such suitable to find positive semidefinite solutions X to linear equations by solving the standard SDP

$$\begin{aligned} \text{minimize } & 0 \\ \text{subject to } & \langle A_i, X\rangle_F = b_i \\ & X \succeq_{S^n_+} 0. \end{aligned}$$

When this is handled via primal path following and therefore effectively maximization of $|X|$, the result is called the analytic center of the inequalities [199]. This way to construct the analytic center provides a feasible initial iterate as necessary in the kernel inference algorithms devised later.

If not for the problem that the Frobenius norm is unsuitable to measure distances between covariance matrices in any meaningful way one could now solve an SDP constructed as described below as a means for kernel inference. Recall the SDP embedding of a least squares problem and notice that $K = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}\varphi_i \otimes \varphi_j^*$ is positive semidefinite if γ_{ij} is positive semidefinite due to the fact that $\gamma \succeq_{S_+^n} 0$ implies $v^*Kv = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}\langle v, \varphi_i\rangle\langle v, \varphi_j\rangle = w^*\gamma w \geq 0$. Write $K \in \mathbb{R}^n \otimes \mathbb{R}^n$ as the superposition of matrices $\Psi_{ij} \in \mathbb{R}^n \otimes \mathbb{R}^n$;

$$K = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}\varphi_i \otimes \varphi_j^* = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}\Psi_{ij}, \qquad (\Psi_{ij})_{mn} = (\varphi_i \otimes \varphi_j^*)_{mn}$$

and identify K with its vectorized version inhabiting $\mathbb{R}^{n^2}$. Each element of K_{mn} is again a weighted sum of elements of the matrices $\Phi_{mn} \in \mathbb{R}^{n_{\text{exp}}} \otimes \mathbb{R}^{n_{\text{exp}}}$.

$$K_{mn} = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}(\Psi_{ij})_{mn} = \langle \gamma, \Phi_{mn}\rangle_{\mathbb{R}^{n^2}}, \qquad (\Phi_{mn})_{ij} = (\varphi_i)_m(\varphi_j)_n$$

$$\underbrace{\begin{bmatrix} K_{11} \\ \vdots \\ K_{nn} \end{bmatrix}}_{K} = \underbrace{\begin{bmatrix} (\Psi_{11})_{11} & \cdots & (\Psi_{n_{\text{exp}}n_{\text{exp}}})_{11} \\ \vdots & \ddots & \vdots \\ (\Psi_{11})_{nn} & \cdots & (\Psi_{n_{\text{exp}}n_{\text{exp}}})_{nn} \end{bmatrix}}_{\Psi} \underbrace{\begin{bmatrix} \gamma_{11} \\ \vdots \\ \gamma_{n_{\text{exp}}n_{\text{exp}}} \end{bmatrix}}_{\gamma}.$$

Given an empirical covariance matrix K_{emp} one could now search for that positive semidefinite γ for which $\|\Psi\gamma - K_{\text{emp}}\|_F^2 \to \min$. The associated SDP in inequality form is

$$\begin{aligned} &\text{minimize } \langle c, \tilde{\gamma}\rangle_{R^{n_{\text{exp}}^2+1}} \\ &\text{subject to } F_0 + tF_t + \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}F_{ij} \succeq_{S_+^n} 0 \end{aligned} \tag{4.39}$$

$$F_0 = \begin{bmatrix} 0 & -K_{\text{emp}} & 0 \\ -K_{\text{emp}}^* & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad F_t = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad F_{ij} = \begin{bmatrix} 0 & \Psi_{ij} & 0 \\ \Psi_{ij}^* & 0 & 0 \\ 0 & 0 & Q_{ij} \end{bmatrix}$$

with $Q_{ij} = e_i \otimes e_j^* \in \mathbb{R}^{n_{\text{exp}}} \otimes \mathbb{R}^{n_{\text{exp}}}$, the tensor product of the i-th and j-the canonical euclidean basis vectors in $\mathbb{R}^{n_{\text{exp}}}$. In the above, $c = [1, 0, ..., 0] \in \mathbb{R}^{n_{\text{exp}}^2+1}$ and $\tilde{\gamma} = [t, \gamma_{11}, ..., \gamma_{n_{\text{exp}}n_{\text{exp}}}] \in \mathbb{R}^{n_{\text{exp}}^2+1}$. In the context of what was said before, it should be clear that the block diagonal structure of the linear matrix inequality implies its equivalence to the two separate LMIs

$$\begin{bmatrix} tI & (\Psi\gamma - K_{\text{emp}}) \\ (\Psi\gamma - K_{\text{emp}})^* & 1 \end{bmatrix} \succeq_{S_+^n} 0 \quad \text{and} \quad \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1n_{\text{exp}}} \\ \vdots & \ddots & \vdots \\ \gamma_{n_{\text{exp}}1} & \cdots & \gamma_{n_{\text{exp}}n_{\text{exp}}} \end{bmatrix} \succeq_{S_+^n} 0$$

which in turn means problem 4.39 has the simple interpretation

$$\begin{aligned} &\text{minimize} \;\; \|\Psi\gamma - K_{\text{emp}}\|_F^2 \\ &\text{subject to} \;\; \gamma \succeq_{S_+^n} 0; \end{aligned}$$

i.e. γ is the positive semidefinite least squares solution to the approximation problem $K = \Psi\gamma \approx K_{\text{emp}}$. We will not pursue this approach farther due to the suboptimality of the Frobenius norm for this task and the fact that solving SDP 4.39 poses a severe computational challenge.

4.3.2 Unconstrained kernel inference

Maximum likelihood and maximum aposteriori estimation of covariance matrices $C \in \mathbb{R}^n \otimes \mathbb{R}^n$ under the model

$$C = \sum_{j=1}^{n_{exp}} \gamma_{ij}\varphi_i \otimes \varphi_j^*$$

is possible. Inclusion of prior knowledge in form of a probability distribution on the coefficients $\gamma \sim \mathcal{W}_{n_{exp}}(q, \Lambda)$ for $(\Lambda)_{ij} = \delta_{ij}\lambda_i, \varphi_i \in \mathbb{R}^n$ is straightforward for λ_i, φ_i eigenvalues and eigenfunctions originating from some prior covariance matrix C_p's spectral decomposition $C_p = \Phi\Lambda\Phi^*$. We record in the following the calculations necessary to formulate a Newton-type algorithm based on Fisher scoring to approximately minimize the regularized discrepancy measure $L = L_1 + rL_2$ from equation4.35 with respect to the positive semidefinite coefficient matrix $\gamma \in S_+^{n_{exp}}$.

Bearing in mind the general setup of the optimization problem whose goal is inference of a reproducing kernel, task is to find

$$\begin{aligned} \gamma_{\text{opt}} &= \underset{\gamma \in S_+^{n_{\text{exp}}}}{\text{argmin}} \;\; L(\gamma) \\ L(\gamma) &= \underbrace{\log|C(\gamma)| + \text{tr}(SC^+(\gamma))}_{C(\gamma)=\sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}\varphi_i \otimes \varphi_j^*} + r\underbrace{\left[-c\log|F(\gamma)| + \text{tr}(\Lambda^{-1}F(\gamma))\right]}_{F(\gamma)=\sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}e_i \otimes e_j^*} \end{aligned} \tag{4.40}$$

Here $\{e_i\}_{i=1}^n$ is the canonical Euclidean basis of $\mathbb{R}^n$, $S = (n_{\text{obs}})^{-1}\sum_{j=1}^{n_{\text{obs}}} f_{\omega_j} \otimes f_{\omega_j}^*$ is the empirical covariance matrix calculated from realizations $f_{\omega_j} \in \mathbb{R}^n$ of the random vector f whose covariance matrix is to be inferred and r is a regularization parameter. The value of r may be set to n_{obs}^{-1} to recover the objective function corresponding to the statistically motivated MAP estimation derived in section 4.2.3 equation 4.35, the constant $c = q - n - 1$ could play a similar role as r though we will typically fix it to be 1.

The goal is to find the gradient L_γ and the information matrix B_γ such that the numerical scheme

$$\gamma^{k+1} = \gamma^k - B_\gamma^+ L_\gamma \tag{4.41}$$

produces a minimizer of equation 4.40. As it holds that $\log |C| = \operatorname{tr} \log C$, one has $\partial/\partial\gamma_{ij} \log |C| = \operatorname{tr}(C^+ C_{ij})$ [133] where for now the sub- and superscripts indicate differentiation; i.e.

$$\frac{\partial}{\partial \gamma_{ij}} C = C_{ij} \qquad \text{and} \qquad \frac{\partial}{\partial \gamma_{ij}} C^+ = C^{ij}$$

Lemma 4.3.4 *Let $Q_{ij} = \varphi_i \otimes \varphi_j^*$ and Λ be diagonal with elements λ_i. For every ONB $\{\varphi_j\}_{j=1}^n$ of $\mathbb{R}^n$ and any matrix $C = \sum_{i,j=1}^{n_{exp}} \gamma_{ij} Q_{ij}$ with $n_{exp} \le n$ it holds that*

I $C_{ij} = Q_{ij}$

II $Q_{kl} Q_{ij} = \delta_{il} Q_{kj}$

III $C^+ = \sum_{i,j=1}^{n_{exp}} \gamma^{ij} Q_{ij}$ *for* $\gamma^{ij} = (\gamma^+)_{ij}$

IV $C^+ C_{ij} = \sum_{k=1}^{n_{exp}} \gamma^{ki} Q_{kj}$

V $C^+ C_{ij} C^+ = \sum_{k,l=1}^{n_{exp}} \gamma^{ki} \gamma^{jl} Q_{kl}$

VI $\operatorname{tr}(C^+ C_{ij}) = \gamma^{ji}$

VII $\operatorname{tr}(C^+ C_{ij} C^+ C_{kl}) = \gamma^{il} \gamma^{jk}$

Proof: It is clear that $C_{ij} = \partial/\partial\gamma_{ij} \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{k,l} \varphi_k \otimes \varphi_l^* = \varphi_i \otimes \varphi_j^* = Q_{ij}$ thus proving I. As $Q_{kl} Q_{ij} = (\varphi_k \otimes \varphi_l^*)(\varphi_i \otimes \varphi_j^*) = \langle \varphi_i, \varphi_l \rangle_{\mathbb{R}^n} \varphi_k \otimes \varphi_j^* = \delta_{il} Q_{kj}$, II follows. For III, note that if $n_{\text{exp}} \le n$ finite, the matrix C may be written as

$$C = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij} \varphi_i \otimes \varphi_j^* = \begin{bmatrix} \varphi_1, ..., \varphi_{n_{\text{exp}}} \end{bmatrix} \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1 n_{\text{exp}}} \\ \vdots & \ddots & \vdots \\ \gamma_{n_{\text{exp}} 1} & \cdots & \gamma_{n_{\text{exp}} n_{\text{exp}}} \end{bmatrix} \begin{bmatrix} \varphi_1^* \\ \vdots \\ \varphi_{n_{\text{exp}}}^* \end{bmatrix} = \Phi \gamma \Phi^*.$$

It is then trivial to show that the matrix $\tilde{C} = \sum_{i,j=1}^{n_{\text{exp}}} \gamma^{ij} Q_{ij}$ with $\gamma^{ij} = (\gamma^+)_{ij}$ satisfies the four Moore-Penrose conditions [18, p. 40]

$$\tilde{C} C \tilde{C} = \tilde{C} \qquad C \tilde{C} C = C \qquad (C\tilde{C})^* = C\tilde{C} \qquad (\tilde{C} C)^* = \tilde{C} C$$

implying $\tilde{C}$ to be the pseudoinverse C^+. The matrix multiplication of C^+ with C_{ij} yields

$$C^+ C_{ij} = \sum_{k,l=1}^{n_{\text{exp}}} \gamma^{kl} Q_{kl} Q_{ij} = \sum_{k,l=1}^{n_{\text{exp}}} \delta_{il} Q_{kj} \sum_{k=1}^{n_{\text{exp}}} \gamma^{ki} Q_{kj}$$

and ultimately leads to the trace formula

$$\operatorname{tr}(C^+ C_{ij}) = \sum_{k=1}^{n_{\text{exp}}} \gamma^{ki} \underbrace{\operatorname{tr}(Q_{kj})}_{\operatorname{tr}(\varphi_k \otimes \varphi_j^*) = \langle \varphi_k, \varphi_j \rangle_{\mathbb{R}^n}} = \sum_{k=1}^{n_{\text{exp}}} \gamma^{ki} \delta_{kj} = \gamma^{ji}.$$

This proves IV and VI; equations V and VII are a result of straightforward computation as well.

$$C^+C_{ij}C^+ = \sum_{k,l=1}^{n_{\exp}} \gamma^{kl} Q_{kl} Q_{ij} \sum_{p,q=1}^{n_{\exp}} \gamma^{pq} Q_{pq} = \sum_{k,l=1}^{n_{\exp}} \sum_{p,q=1}^{n_{\exp}} \gamma^{kl}\gamma^{pq}\delta_{il}\delta_{jp}Q_{kp}$$
$$= \sum_{k,l=1}^{n_{\exp}} \gamma^{ki}\gamma^{jl} Q_{kl}$$

Equation VII can be proven be reformulating the term $\operatorname{tr}(C^+C_{ij}C^+C_{kl})$ as

$$\operatorname{tr}\left(\sum_{p,q=1}^{n_{\exp}} \sum_{r,s=1}^{n_{\exp}} \gamma^{pq}\gamma^{rs} Q_{pq}Q_{ij}Q_{rs}Q_{kl}\right) = \sum_{p,q=1}^{n_{\exp}} \sum_{r,s=1}^{n_{\exp}} \gamma^{pq}\gamma^{rs}\delta_{qi}\delta_{sk}\delta_{jr}\operatorname{tr}(Q_{pl})$$
$$= \sum_{p=1}^{n_{\exp}} \gamma^{pi}\gamma^{jk} \underbrace{\operatorname{tr}(Q_{pl})}_{\operatorname{tr}(\varphi_p \otimes \varphi_l^*) = \langle \varphi_p, \varphi_l \rangle_{\mathbb{R}^n}}$$
$$= \gamma^{li}\gamma^{jk}$$

□

With the help of lemma 4.3.4 it is possible to calculate gradients and (expected) Hessians of L^1 and L^2 from equation 4.35 with $L(\gamma) = L^1(\gamma) + rL^2(\gamma)$.

Theorem 4.3.5 *With the same notation as before, the gradients L^1_γ and L^2_γ with respect to the parameters $\gamma_{ij}, i, j = 1, ..., n_{exp}$ are given by*

$$L^1_\gamma = \left[\gamma^+ - \gamma^+ S_\varphi \gamma^+\right] \tag{4.42}$$
$$L^2_\gamma = r\left[-c\gamma^+ + \Lambda^{-1}\right] \tag{4.43}$$

where $r = n_{obs}^{-1}, c = q - n - 1$ and $S_\varphi = n_{obs}^{-1} \sum_{j=1}^{n_{obs}} f^j_\varphi \otimes (f^j_\varphi)^*$ with $f^j_\varphi \in \mathbb{R}^{n_{exp}}$ a vector with elements $(f^j_\varphi)_i = \langle f_{\omega_j}, \varphi_i \rangle_{\mathbb{R}^n}$. The information matrices B^1_γ and B^2_γ act on symmetric matrices $A \in S^{n_{exp}}$ as

$$B^1_\gamma A = \gamma^+ A \gamma^+ \tag{4.44}$$
$$B^2_\gamma A = rc\gamma^+ A \gamma^+ \tag{4.45}$$

Proof: (Equation 4.42) From $L^1 = \log|C| + \operatorname{tr}(SC^+)$ it follows that

$$(L^1_\gamma)_{ij} = \operatorname{tr}(C^+C_{ij}) + \operatorname{tr}(SC^{ij}) = \gamma^{ji} - \sum_{k,l=1}^{n_{\exp}} \gamma^{ki}\gamma^{jl}\operatorname{tr}(SQ_{kl}).$$

Since γ is symmetric, $\gamma^{ij} = \gamma^{ji}$ and one also finds that $\mathrm{tr}(SQ_{kl})$ satisfies

$$\mathrm{tr}(SQ_{kl}) = \frac{1}{n_{\text{obs}}} \sum_{j=1}^{n_{\text{obs}}} \mathrm{tr}\left(f_{\omega_j} \otimes f^*_{\omega_j} Q_{kl} \right) = \frac{1}{n_{\text{obs}}} \sum_{j=1}^{n_{\text{obs}}} \langle f_{\omega_j}, \varphi_k \rangle_{\mathbb{R}^n} \langle f_{\omega_j}, \varphi_l \rangle_{\mathbb{R}^n}.$$

Employing the notation $f^k_j = \langle f_{\omega_j}, \varphi_k \rangle_{\mathbb{R}^n}$ for the coordinates of f_{ω_j} in the ONS spanned by $\{\varphi_i\}_{i=1}^{n_{\text{exp}}}$ one arrives at

$$\sum_{k,l=1}^{n_{\text{exp}}} \gamma^{ki} \gamma^{jl} \, \mathrm{tr}(SQ_{kl}) = \frac{1}{n_{\text{obs}}} \sum_{p=1}^{n_{\text{obs}}} \sum_{k=1}^{n_{\text{exp}}} \gamma^{ki} f^k_p \sum_{l=1}^{n_{\text{exp}}} \gamma^{jl} f^l_p = \frac{1}{n_{\text{obs}}} \sum_{p=1}^{n_{\text{obs}}} \left(\gamma^+ f^p_\varphi \right)_i \left(\gamma^+ f^p_\varphi \right)_j$$

where $(f^p_\varphi)_k = \langle f_{\omega_p}, \varphi_k \rangle_{\mathbb{R}^n}$. To get the whole matrix $\{(L^1_\gamma)_{ij}\}_{i,j=1}^{n_{\text{exp}}}$ we may simply form $\left(\gamma^+ f^p_\varphi\right) \otimes \left(\gamma^+ f^p_\varphi\right)^* = \gamma^+(f^p_\varphi) \otimes (f^p_\varphi)^* \gamma^+$ and collect terms yielding $\frac{1}{n_{\text{obs}}} \sum_{p=1}^{n_{\text{obs}}} \gamma^+(f^p_\varphi) \otimes (f^p_\varphi)^* \gamma^+ = \gamma^+ S_\varphi \gamma^+$ and finally

$$L^1_\gamma = \gamma^+ - \gamma^+ S_\varphi \gamma^+$$

for $S_\varphi = \frac{1}{n_{\text{obs}}} \sum_{j=1}^{n_{\text{obs}}} (f^j_\varphi) \otimes (f^j_\varphi)^*$ the empirical coefficients γ_{emp}.

(Equation 4.43) From $L^2 = r\left[-c \log|F(\gamma)| + \mathrm{tr}\left(\Lambda^{-1} F(\gamma)\right)\right]$ it can be derived that

$$(L^2_\gamma)_{ij} = r\left[-c \, \mathrm{tr}\left(F^+ F_{ij}\right) + \mathrm{tr}\left(\Lambda^{-1} F_{ij}\right)\right] = r\left[-c\gamma^{ij} + \delta_{ij} \lambda_i^{-1}\right]$$

By the same arguments as for the preceding equation, we may explicitly rewrite L^2_γ in matrix form as

$$L^2_\gamma = r\left[-c\gamma^+ + \Lambda^{-1}\right].$$

(Equation 4.44). Mardia and Marshal [133] prove that under the simplifying assumption $E[f \otimes f^*] = C$, the Fisher information matrix $B^1_\gamma = E[L^1_{\gamma\gamma}]$ for $L^1_{\gamma\gamma}$ the Hessian of L^1 has elements

$$(B^1_\gamma)_{(i,j)(k,l)} = \mathrm{tr}\left(C^+ C_{ij} C^+ C_{kl}\right) = \gamma^{jk} \gamma^{li}$$

Then for $A \in S^n_+$, $B^1_\gamma A$ can be written as

$$\begin{aligned} B^1_\gamma A &= \sum_{i,j=1}^{n_{\text{exp}}} \sum_{k,l=1}^{n_{\text{exp}}} \gamma^{jk} \gamma^{li} e_i \otimes e^*_j \otimes e_k \otimes e^*_l A \\ &= \sum_{i,j=1}^{n_{\text{exp}}} \sum_{k,l=1}^{n_{\text{exp}}} \gamma^{il} a_{lk} \gamma^{kj} e_i \otimes e^*_j \\ &= \sum_{i,l=1}^{n_{\text{exp}}} \gamma^{il} e_i \otimes e^*_l A \sum_{k,j=1}^{n_{\text{exp}}} \gamma^{kj} e_k \otimes e^*_j \qquad = \gamma^+ \otimes \gamma^+ A \end{aligned}$$

Similarly, it is straightforward to prove equation 4.45. The information matrix $B^2_\gamma = L^2_{\gamma\gamma}$ (since the random variables f_{ω_j} does not appear here, taking expectations is not required) has elements $(B^2_\gamma)_{(i,j)(k,l)}$ computable as the ij-th element of $-rc(\partial/\partial\gamma_{kl})\gamma^+ = rcF^+F_{kl}F^+$. Since $F_{kl} = e_k \otimes e_l^* =: P_{kl}$, one finds

$$-rc\left(\frac{\partial}{\partial\gamma_{kl}}\gamma^+\right)_{ij} = rc\langle F^+P_{kl}F^+, P_{ij}\rangle_F \quad = rc\,\mathrm{tr}\left(F^+P_{kl}^TF^+P_{ij}\right) \quad = rc\gamma^{lj}\gamma^{ik}$$

The direct calculation $B^2_\gamma A = rc\sum_{ij=1}^{n_{\mathrm{exp}}}\sum_{k,l=1}^{n_{\mathrm{exp}}}\gamma^{lj}a_{ji}\gamma^{ik}e_l \otimes e_k^* = rc\sum_{k,l=1}^{n_{\mathrm{exp}}}\sum_{i,j=1}^{n_{\mathrm{exp}}}\gamma^{il}a_{lk}\gamma^{kj}$, where we exchanged l and k, then commuted the scalar factors, establishes $B^2_\gamma A = rc\gamma^+A\gamma^+$ for symmetric A. Therefore B^2_γ acts on $A \in S^n_+$ as $rc\gamma^+ \otimes \gamma^+$. □

To minimize $L = L^1 + L^2$, find the zeros of the gradient

$$L_\gamma = L^1_\gamma + L^2_\gamma = \gamma^+ - \gamma^+S_\varphi\gamma^+ + r\left[-c\gamma^+ + \Lambda^{-1}\right] \overset{!}{=} 0 \tag{4.46}$$

This is a continuous time algebraic Riccati equation [126] for which solutions γ^+ can be constructed from the spectral decomposition of a certain symplectic matrix. However, in our experiments these solutions were typically not positive definite and instead of the closed form solutions, we employ a sequence of Newton steps.

$$\gamma^{k+1} = \gamma^k - B^+_\gamma L_\gamma \tag{4.47}$$

$$\begin{aligned} B^+_\gamma L_\gamma &= \left(B^1_\gamma + B^2_\gamma\right)^+ \left(L^1_\gamma + L^2_\gamma\right) \\ &= \frac{1}{1+rc}\gamma \otimes \gamma\left[(1-rc)\gamma^+ - \gamma^+S_\varphi\gamma^+ + r\Lambda^{-1}\right] \\ &= \left(\frac{1-rc}{1+rc}\right)\gamma - \frac{1}{1+rc}S_\varphi + \frac{r}{1+rc}\gamma\Lambda^{-1}\gamma \end{aligned} \tag{4.48}$$

This implies the following explicit formula for γ^{k+1} dependent on the previous iteration γ^k, the prior Λ and the matrix $S_\varphi = \Phi^TS\Phi$ representing the empirical γ-values (see remarks later).

$$\begin{aligned} \gamma^{k+1} &= \gamma^k - \left(\frac{1-rc}{1+rc}\right)\gamma^k + \frac{1}{1+rc}S_\varphi - \frac{r}{1+rc}\gamma^k\Lambda^{-1}\gamma^k \\ &= (1+rc)^{-1}\left[2rc\gamma^k + s_\varphi - r\gamma^k\Lambda^{-1}\gamma^k\right] \end{aligned} \tag{4.49}$$

Recalling $r = n_{\mathrm{obs}}^{-1}$ and using $rc(1+rc)^{-1} = c(c+n_{\mathrm{obs}})^{-1}$ one may alternatively write

$$\gamma^{k+1} = \left(\frac{2c}{c+n_{\mathrm{obs}}}\right)\gamma^k + \left(\frac{n_{\mathrm{obs}}}{c+n_{\mathrm{obs}}}\right)S_\varphi - \left(\frac{1}{c+n_{\mathrm{obs}}}\right)\gamma^k\Lambda^{-1}\gamma^k \tag{4.50}$$

§ Detailed investigation of the kernel inference procedure

When initializing $\gamma^0 = \Lambda$, equation 4.50 leads to fast and reliable guesses for γ; see figures 4.8 to 4.11. When no observations are available, equation 4.50 specializes to

$$\gamma^{k+1} = 2\gamma^k - \frac{1}{c}\gamma^k \Lambda^{-1} \gamma^k. \tag{4.51}$$

Similarly, if no prior knowledge about γ is specified and $c = 0$, then

$$\gamma^{k+1} = S_\varphi. \tag{4.52}$$

The matrix S_φ can be interpreted as a low-dimensional approximation to the matrix $\tilde{\gamma}$ that forms the empirical covariance matrix S via $S = \Phi\tilde{\gamma}\Phi^*$. Notice that

$$S = \frac{1}{n_{\text{obs}}} \sum_{k=1}^{n_{\text{obs}}} f_{\omega_k} \otimes f_{\omega_k}^* = \sum_{i,j=1}^{n} \tilde{\gamma}_{ij} \varphi_i \otimes \varphi_j^* = \Phi\tilde{\gamma}\Phi^*. \tag{4.53}$$

If n_{exp} were n, one would have $\tilde{\gamma} = \Phi^* S \Phi$ and $(\tilde{\gamma})_{ij} = \frac{1}{n_{\text{obs}}} \sum_{k=1}^{n_{\text{obs}}} \langle \varphi_i, f_{\omega_k} \rangle \langle \varphi_j, f_{\omega_k} \rangle$. Since $n_{\text{exp}} \leq n$, $S_\varphi = \Phi^* S \Phi$ is only an approximation to $\tilde{\gamma}$ derived from the norm minimization problem

$$\|\Phi S_\varphi \Phi^* - S\|_F^2 \to \min \Leftrightarrow S_\varphi = (\Phi)^+ \otimes (\Phi^*)^+ S = \Phi^* S \Phi.$$

Equation 4.52 determines therefore a reasonable guess for the coefficient matrix γ in the absence of prior knowledge. Equation 4.51 has a specific interpretation as well. Suppose the task would be to only minimize the prior term $L^2(\gamma)$. The obvious solution is to investigate the gradient $L^2_\gamma = r\left[-c\gamma^+ + \Lambda^{-1}\right]$ which equals zero for $\gamma^+ = c^{-1}\Lambda^{-1}$ and $\gamma = c\Lambda$, the mode of the Wishart distribution. If these exact solutions were unknown, one could try to optimize $L^2(\gamma)$ (and $L^1(\gamma)$) numerically via semidefinite programming. For $L^2(\gamma)$, the projected Newton direction $\Delta\gamma$ can be calculated directly [112, p. 77-80] as

$$\begin{aligned} \Delta\gamma = \underset{\Delta\gamma \in S^n}{\operatorname{argmin}} \ \langle L^2_\gamma, \Delta\gamma \rangle_F + \frac{1}{2} \langle L^2_{\gamma\gamma} \Delta\gamma, \Delta\gamma \rangle_F &= -\gamma^{1/2} \left[\frac{\gamma^{1/2}\Lambda^{-1}\gamma^{1/2}}{c} - I \right] \gamma^{1/2} \\ &= \gamma - \frac{\gamma\Lambda^{-1}\gamma}{c} \end{aligned} \tag{4.54}$$

and leads to the update rule $\gamma^{k+1} = 2\gamma^k - c^{-1}\gamma^k\Lambda^{-1}\gamma^k$ exhibited in equation 4.51. Optimizing only the prior probability is therefore expressable as semidefinite optimization leading to an iterative update rule for the coefficient matrix γ. A similar investigation is possible for equation 4.52. Suppose for the sake of illustration that $n_{\text{exp}} = n$ and γ is invertible. Then $|C(\gamma)| = |\Phi\gamma\Phi^*| = |\gamma|$ by the properties of determinants. If one writes $L^1(\gamma)$ as

$$L^1(\gamma) = \log|C(\gamma)| + \operatorname{tr}\left(SC^{-1}(\gamma)\right) = \log|\gamma| + \operatorname{tr}\left(S_\varphi \gamma^+\right)$$

$$= -\log|\gamma^{+}| + \operatorname{tr}\left(S_{\varphi}\gamma^{+}\right) \tag{4.55}$$

the result is a term in $Z = \gamma^{+}$ that is again amenable to semidefinite programming and has projected Newton direction [112, p. 77-80]

$$\Delta Z = Z - ZS_{\varphi}Z \tag{4.56}$$

where ΔZ is an update for the inverse γ^{+} of γ. In conclusion, investigation of the special cases inherent to the iteration scheme proposed in equation 4.50 reveals it to be a mixture of weighted simultaneous manipulations of the coefficient matrix γ and its pseudoinverse γ^{+} via the action of regularization and data terms. We suggest to set $c = 1$ to avoid issues evoked by the mode of the Wishart distribution scaling with c. The pseudocode below summarizes our approach.

Nonparametric unconstrained kernel inference

```
    Input
        $S = \frac{1}{n_{\text{obs}}}\sum_{j=1}^{n_{\text{obs}}} f_{\omega_j} \otimes f_{\omega_j}^{*}$, the empirical $n \times n$ covariance matrix
        $K_p$, the prior kernel with decomposition $K_p(s,t) = \sum_{i=1}^{n} \lambda_i \varphi_i(s)\varphi_i(t)$
Parameters
        $n_{\text{exp}} \leq n$, the expansion depth determining the dimension of $\gamma$
        $r \in [0,\infty)$, a regularization parameter which can be set to $r = n_{\text{obs}}^{-1}$
    Begin
        $\gamma^{0} = \Lambda$, with $\Lambda \in S_{+}^{n_{\text{exp}}}$ and $(\Lambda)_{ij} = \delta_{ij}\lambda_i$
        Do until convergence:
        $\gamma = \frac{1}{1+r}\left[2r\gamma + S_{\varphi} - r\gamma\Lambda^{+}\gamma\right]$
                    where $S_{\varphi} \in S_{+}^{n_{\text{exp}}}$    $S_{\varphi} = (\Phi^{+})S(\Phi^{*})^{+}$, $\Phi = [\varphi_1, ..., \varphi_{n_{\text{exp}}}]$
     End
   Output
        $\gamma$, a $n_{\text{exp}} \times n_{\text{exp}}$ coefficient matrix such that $C(\gamma) = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}\varphi_i \otimes \varphi_j^{*}$ is both likely
        under the observations and probable under the prior assumptions.
```

Since all operations act only on the finite dimensional coefficient matrices, there is no inherent reason, why the algorithm should not work for infinite dimensional covariance matrices — or the underlying kernel functions if one swaps $\mathbb{R}^{n}$ for L^{2}. The algorithm is illustrated on subsequent pages by plotting in the figures 4.8 to 4.11 some results showing performance for inference and simulation tasks and by comparing it to other methods.

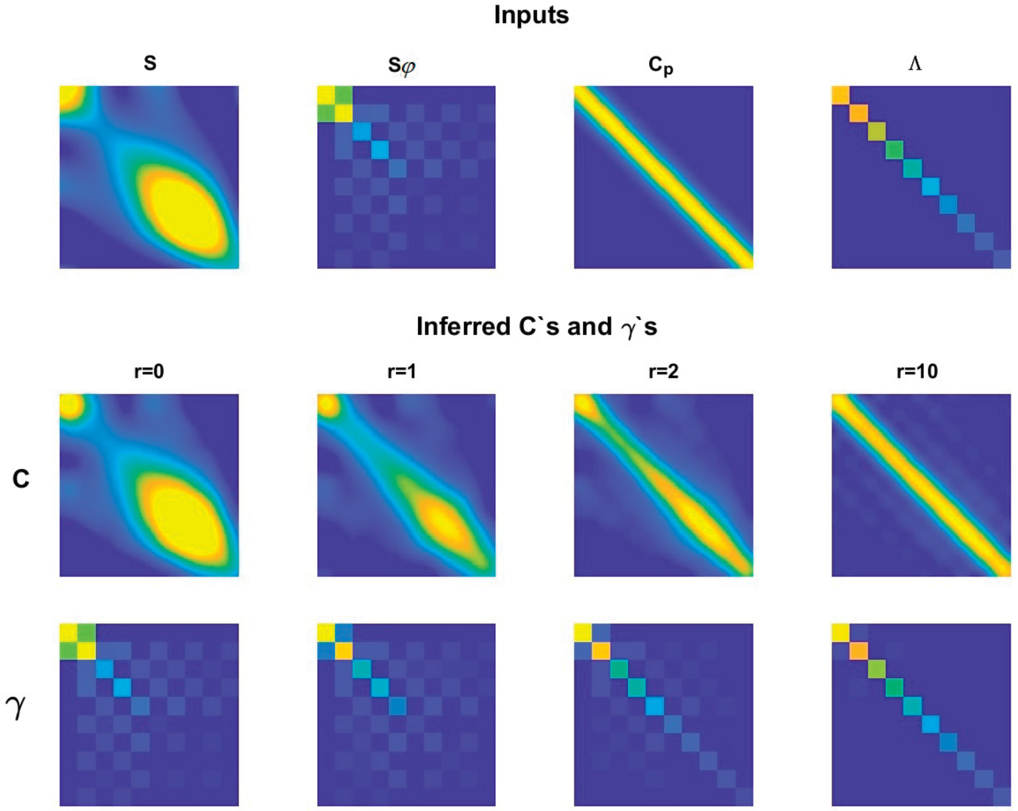


Figure 4.8: Inferred coefficient matrices γ and covariance matrices $C = \Phi\gamma\Phi^*$ for different regularization parameters r. The expansion depth n_{exp} is 10, the dimension n of the observed process is 100 whereas the number n_{obs} of observations that contributed to the empirical covariance matrix S is assumed unknown precluding any informed choice of r. The influence of the regularization parameter r on the result of the inference procedure is well visible. The algorithm terminated in approximately 0.01 seconds. The colorscale is identical for all plots of the same type. In all cases the true underlying covariance function was the squared exponential kernel.

4.3.3 Extensions

It is possible to solve the kernel inference problem under affine constraints on the coefficient matrix γ albeit at the cost of further computational effort and reliability of the solutions. The resultant algorithm can be used for example to formulate a procedure for variance components estimation centered around the likelihood as optimization objective. The calculation of the closed forms of gradients, Hessians and derived information matrices provided by theorem 4.3.5 exploited the orthonormality of the family $\{\varphi_i\}_{i=1}^{n_{exp}}$. Especially in the infinite dimensional case, where the Mercer decomposition of the prior $K_p(s,t)$ is $\sum_{i=1}^{\infty}\lambda_i\varphi_i(s)\varphi_i(t)$ with $\{\varphi_i\}_{i=1}^{\infty}$ an ONS of functions in $L^2(T)$, one typically has access and works with only a finite vector of point evaluations $[\varphi_i(t_1),...,\varphi_i(t_n)]^T \in \mathbb{R}^n$ or more generally with $A\varphi_i = \psi_i$ for some linear operator $A : L^2(T) \to \mathbb{R}^n$. Then the ψ_i are not necessarily orthogonal anymore; however if the sequence $\{\psi_i\}_{i=1}^{n} \subset \mathbb{R}^n$ is linearly independent, the results derived for the ONB-case mostly still hold as will be shown in what follows.

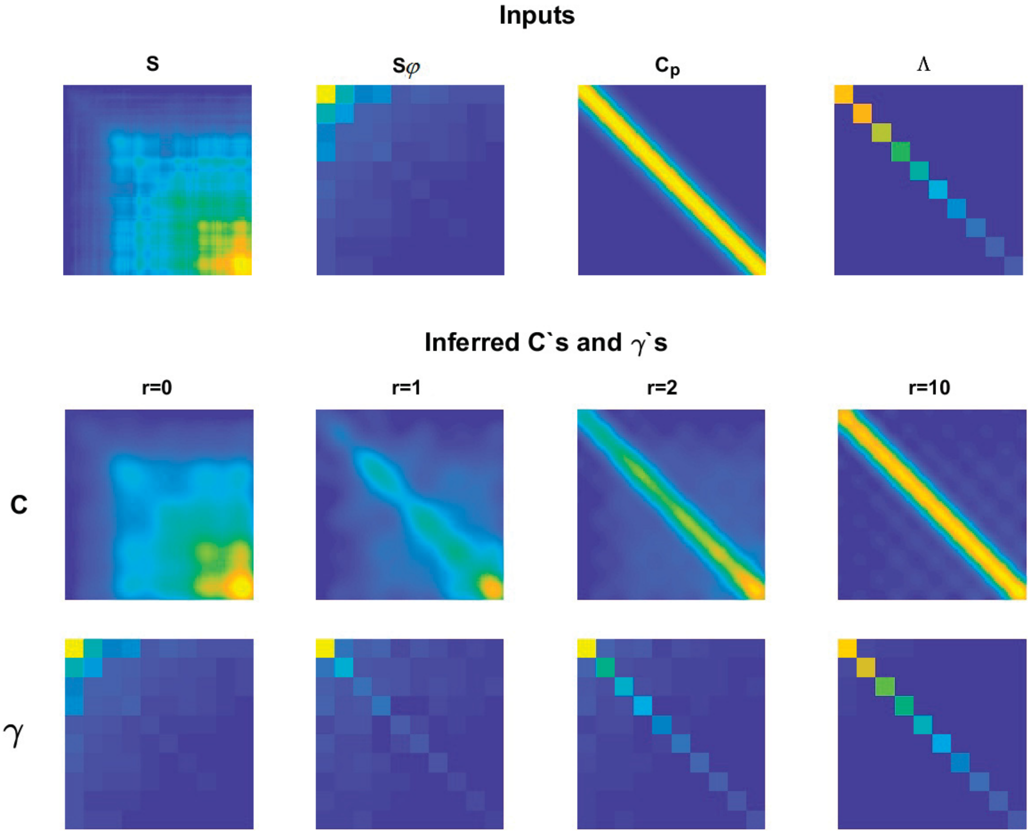


Figure 4.9: The same setup as in the previous figure but with a massively misspecified prior covariance. The true underlying covariance function is the $\min(s,t)$-kernel of the Wiener process. If r is set to $n_{\text{obs}}^{-1} = 0.05$ (20 observations), the prior is overridden as more data comes in and the result of inference is a covariance matrix that lies between the one plotted in the first and second column.

§ Inclusion of affine constraints

Recall from theorem 4.3.5 that the gradient and Hessian of the approximate maximum a posteriori objective function

$$L = \log|C(\gamma)| + \operatorname{tr}\left(SC^{+}(\gamma)\right) + r\left[-\log|\gamma| + \operatorname{tr}\left(\Lambda^{+}\gamma\right)\right] \tag{4.57}$$

were given by the expressions

$$L_{\gamma} = (1-r)\gamma^{+} + r\Lambda^{+} - \gamma^{+}S_{\varphi}\gamma^{+} \tag{4.58}$$

$$L_{\gamma\gamma} = (1+r)\gamma^{+} \otimes \gamma^{+}. \tag{4.59}$$

If it is now demanded that $\gamma \in \mathbb{R}^{n_{\text{exp}}} \otimes \mathbb{R}^{n_{\text{exp}}}$ is the solution to some linear equation $A\gamma = b$ where $b \in \mathbb{R}^{n_c}$, $A : \mathbb{R}^{n_{\text{exp}}} \otimes \mathbb{R}^{n_{\text{exp}}} \to \mathbb{R}^{n_c}$ and n_c is the number of constraints, then one may try to optimize L with respect to $\gamma \in A^{-1}b$, i.e. enforce γ to lie in the preimage of b under A. Given some initial guess γ^0 lying in the feasible set $A^{-1}b$, the projected Newton direction $\Delta\gamma \in S^{n_{\text{exp}}}$ is that symmetric matrix, which minimizes a Taylor expansion based approximation to L at γ^0 while

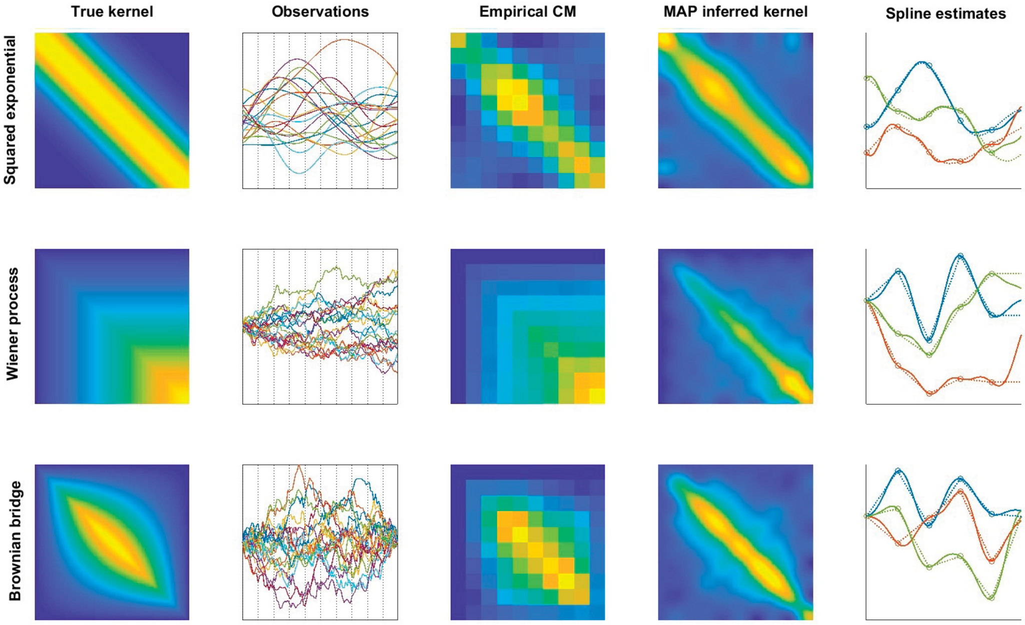


Figure 4.10: Kernel estimations of the kernels exhibited in the first column based on several observations whose locations are marked by the dashed black lines in the second column. The empirical covariance matrices are illustrated in column 3, the MAP estimates in column 4. The regularization parameter was chosen according to the rule $r = n_{\text{obs}}^{-1}$. As a prior, the smooth squared exponential kernel was used. Finally column 5 reports the results of spline estimation for a simple interpolation problem using the true kernels (dashed lines) and the inferred ones (unbroken lines). Compare to figure 4.4.

additionally satisfying $A\Delta\gamma = 0$. Then, if the initial iterate is chosen correctly, $\gamma^1 = \gamma^0 + \Delta\gamma$ has a smaller objective value $L(\gamma^1)$ than γ^0 and still satisfies $A\gamma^1 = A\gamma^0 + A\Delta\gamma = b$; see [112, p. 78] for more details.

Theorem 4.3.6 *With notation as above, the projected Newton direction*

$$\Delta\gamma = \underset{\Delta\gamma\in S^{n_{exp}}}{\operatorname{argmin}} \ \langle L_\gamma, \Delta\gamma\rangle_F + \frac{1}{2}\langle L_{\gamma\gamma}\Delta\gamma, \Delta\gamma\rangle_F \tag{4.60}$$

$$\textit{subject to} \quad A\Delta\gamma = 0$$

is — presupposing invertibility of γ and $A\gamma A^*$ — given by the expression

$$\Delta\gamma = \frac{1}{1+r}\left[\gamma A^*\left(A\gamma A^*\right)^+ A - I\right]\left[(1-r)\gamma + r\gamma\Lambda^+\gamma - S_\varphi\right]. \tag{4.61}$$

Proof. Write $A : \mathbb{R}^{n_{\text{exp}}} \otimes \mathbb{R}^{n_{\text{exp}}} \to \mathbb{R}^{n_c}$ as the matrix $A^* = [A_1, ..., A_{n_c}]$ where each $A_i \in \mathbb{R}^{n_{\text{exp}}^2}$ is the vectorization of the matrix A_i such that $\operatorname{tr}(A_i\gamma) = \langle A_i, \gamma\rangle_F = (A\gamma)_i$. When convenient, we will identify $A_i, \gamma, \Delta\gamma$ either as elements of $\mathbb{R}^{n_{\text{exp}}^2}$ or of $\mathbb{R}^{n_{\text{exp}}} \otimes \mathbb{R}^{n_{\text{exp}}}$. The Lagrangian of the problem is

$$\mathcal{L} = \langle L_\gamma, \Delta\gamma\rangle_F + \frac{1}{2}\langle L_{\gamma\gamma}\Delta\gamma, \Delta\gamma\rangle_F + \sum_{i=1}^{n_c}\mu_i\langle A_i, \Delta\gamma\rangle_F$$

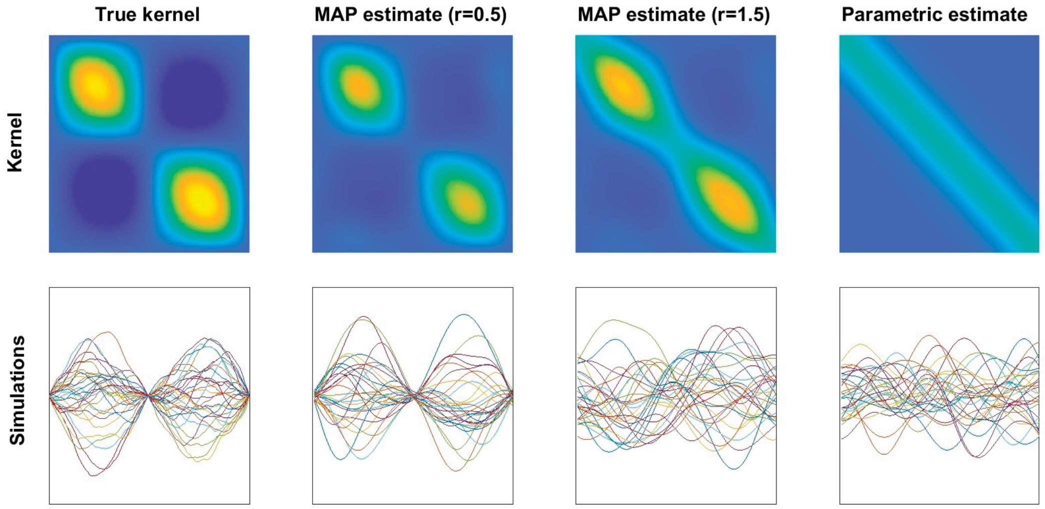


Figure 4.11: Column one shows a certain kernel K and exemplary members of $\mathcal{H}_K$. These samples are used to infer K via MAP and parametrically. Simulations illustrate that the elements of $\mathcal{H}_{K_{\text{MAP}}}$ capture more of the underlying global structure than an arbitrarily chosen parametric model (squared exponential). Notice that the input is an empirical covariance matrix whereas the outputs of the inference procedures are kernel functions that can be employed in spline constructions.

and consequently the KKT conditions [25, p. 161] are

$$\begin{aligned}
\nabla_{\Delta\gamma}\mathcal{L} &= L_\gamma + L_{\gamma\gamma}\Delta\gamma + \sum_{i=1}^{n_c}\mu_i A_i && = 0 \\
\nabla_{\mu_i}\mathcal{L} &= \langle A_i, \Delta\gamma\rangle_F && = 0.
\end{aligned}$$

This can be written as a system of linear equations for $\Delta\gamma$. With $\mu \in \mathbb{R}^{n_c}$ such that $(\mu)_i = \mu_i$, the SLAE has the form

$$\begin{aligned}
&L_{\gamma\gamma}\Delta\gamma + A^*\mu = -L_\gamma \\
&A\Delta\gamma \qquad\qquad = 0
\end{aligned}$$

One might denote as $\widetilde{A}$ the matrix $[L_{\gamma\gamma}, A^*; A, 0]$, set $\widetilde{\Delta\gamma} = [\Delta\gamma; \mu]$ and $\widetilde{L_\gamma} = [L_\gamma; 0]$. This would allow to write the equation as

$$\widetilde{A}\widetilde{\Delta\gamma} = -\widetilde{L_\gamma}$$

and solve it by means of simple matrix inversion, albeit very inefficiently. Instead, under assumption of invertibility of γ and full row rank of A, a sequence of substitutions produces

$$\begin{aligned}
&I && \Delta\gamma && = -L_{\gamma\gamma}^{+}(L_\gamma + A^*\mu) \\
&II && A\Delta\gamma && = 0 \\
&\quad\Leftrightarrow && \mu && = -(AL_{\gamma\gamma}^{+}A^*)^{+}AL_{\gamma\gamma}^{+}L_\gamma \\
&III && L_{\gamma\gamma}\Delta\gamma + A^*\mu && = -L_\gamma
\end{aligned}$$

$$\begin{aligned}\Leftrightarrow \qquad \Delta\gamma \qquad &= -L^+_{\gamma\gamma}\left(L_\gamma - A^*(AL^+_{\gamma\gamma}A^*)^+AL^+_{\gamma\gamma}L_\gamma\right)\\ &= L^+_{\gamma\gamma}\left(A^*(AL^+_{\gamma\gamma}A^*)^+AL^+_{\gamma\gamma} - I\right)L_\gamma\end{aligned}$$

Since $L_{\gamma\gamma}$ and $L^+_{\gamma\gamma}$ only act on symmetric matrices in the above expression, they can be simplified by using $L_{\gamma\gamma} = (1+r)\gamma^+ \otimes \gamma^+$ on $S^{n_{\text{exp}}}$. Now investigate separately

$$\Delta\gamma = \underbrace{-L^+_{\gamma\gamma}L_\gamma}_{\Delta\gamma_1} \quad + \quad \underbrace{L^+_{\gamma\gamma}A^*\left(AL^+_{\gamma\gamma}A^*\right)^+AL^+_{\gamma\gamma}L_\gamma}_{\Delta\gamma_2}. \tag{4.62}$$

For $\Delta\gamma_1$, one arrives in analogy to the unconstrained case at

$$\begin{aligned}\Delta\gamma_1 = -L^+_{\gamma\gamma}L_\gamma \qquad &= -\frac{1}{1+r}\gamma\otimes\gamma\left[(1-r)\gamma^+ + r\Lambda^+ - \gamma^+S_\varphi\gamma^+\right]\\ &= \frac{1}{1+r}\left[(r-1)\gamma + S_\varphi - r\gamma\Lambda^+\gamma\right] \qquad (4.63)\end{aligned}$$

For $\Delta\gamma_2$, notice that $\Delta\gamma_2 = (1+r)^{-2}\gamma A^*\left(AL^+_{\gamma\gamma}A^*\right)^+ A\gamma L_\gamma\gamma\gamma$ and that since $AL^+_{\gamma\gamma}A^*z = (1+r)^{-1}A\gamma A^*z\gamma$, the inverse $\left(AL^+_{\gamma\gamma}A^*\right)^+$ acts on some $b \in \mathbb{R}^{n_c}$ as $(1+r)\left(A\gamma A^*\right)^+ b\gamma^+$. Since we assume γ to be invertible, we may write

$$\begin{aligned}\Delta\gamma_2 = \frac{1}{1+r}\gamma A^*\left(A\gamma A^*\right)^+\left[A\gamma L_\gamma\gamma\gamma\right]\gamma^+ &= \gamma A^*\left(A\gamma A^*\right)^+A\left(\frac{1}{1+r}\gamma L_\gamma\gamma\right)\\ &= -\gamma A^*\left(A\gamma A^*\right)^+A\Delta\gamma_1 \qquad (4.64)\end{aligned}$$

Adding $\Delta\gamma_1$ and $\Delta\gamma_2$ and collecting terms proves the theorem. □

We can then summarize the result of the projected Newton steps as

$$\gamma + \Delta\gamma = \gamma + \Delta\gamma_1 - \gamma A^*\left(A\gamma A^*\right)^+A\Delta\gamma_1 \tag{4.65}$$
$$\Delta\gamma_1 = (1+r)^{-1}\left[(r-1)\gamma + S_\varphi - r\gamma\Lambda^+\gamma\right]. \tag{4.66}$$

The term $(A\gamma A^*)^+$ has dimension $[n_c, n_c]$ and is easy to invert. It can be constructed as the pseudoinverse to the operator $AM_\gamma A^*$ where $M_\gamma = I\otimes\gamma$ is matrix multiplication by γ but slightly modified to act on vectors of dimension n^2_{exp} in an analogous way.

§ Nonorthogonal basis vectors

Suppose an arbitrary ONB $\{\varphi_i\}_{i=1}^n \subset \mathbb{R}^n$ is given and express the covariance matrix in terms of a matrix η of coefficients for the elementary matrices $\psi_i\otimes\psi_j^*$ formed by tensoring together the elements of a linearly independent but otherwise arbitrary sequence of vectors $\{\psi_i\}_{i=1}^n$.

$$C_\eta = \sum_{i,j=1}^n \eta_{ij}\psi_i\otimes\psi_j^* \qquad \{\psi_i\}_{i=1}^n \text{ linearly independent}. \tag{4.67}$$

Let the matrix η have prior Λ. In the loss function $L(\eta) = L_1(\eta) + L_2(\eta) = \log|C_\eta| + \operatorname{tr}\left(SC_\eta^+\right) + r\left[-\log|\eta| + \operatorname{tr}\left(\Lambda^+\eta\right)\right]$, the first term $L_1(\eta)$ is related to the likelihood of C_η given the data and one might express C_η using the elementary matrices $Q_{ij} = \varphi_i \otimes \varphi_j^*$. Naming the coefficients of this expansion γ; i.e.

$$C_\eta = \sum_{i,j=1}^{n} \eta_{ij}\psi_i \otimes \psi_j^* = \sum_{i,j=1}^{n} \gamma_{ij}\varphi_i \otimes \varphi_j^* = C_\gamma$$

one has $L_1(\eta) = \log|C_\eta| + \operatorname{tr}\left(SC_\eta^+\right) = \log|C_\gamma| + \operatorname{tr}\left(SC_\gamma^+\right) = L_1(\gamma)$, where the expansion in terms of γ uses the ONB $\{\varphi_i\}_{i=1}^n$. Gradients and Hessians for this case are provided in theorem 4.3.5. If a prior Λ is given only for η and not for γ, then the second term $L_2(\eta)$ has to be rewritten slightly. Denote by D the matrix of expansion coefficients $(D)_{ij} = d_{ij}$ where $\psi_i = \sum_{j=1}^n d_{ij}\varphi_j$, or $\Psi = \Phi D^T$ for short with $\Psi = [\psi_1, ..., \psi_n]$ and $\Phi = [\varphi_i, ..., \varphi_n]$. Then rewrite

$$\begin{aligned}
C_\eta = \sum_{i,j=1}^{n} \eta_{ij}\psi_i \otimes \psi_j^* \qquad &= \sum_{i,j=1}^{n} \eta_{ij} \left(\sum_{k=1}^{n} d_{ik}\varphi_k\right) \otimes \left(\sum_{l=1}^{n} d_{jl}\varphi_l\right)^* \\
&= \sum_{k,l=1}^{n} \underbrace{\sum_{i,j=1}^{n} d_{ik}\eta_{ij}d_{jl}}_{\gamma_{kl}} \varphi_k \otimes \varphi_l^* \\
&= \sum_{k,l=1}^{n} \gamma_{kl}\varphi_k \otimes \varphi_l^* \qquad\qquad =C_\gamma
\end{aligned}$$

The relationship between the coefficient matrices η and γ is therefore seen to be

$$\gamma = D^T\eta D \tag{4.68}$$

as $\gamma_{ij} = \sum_{k,l=1}^n d_{ki}\eta_{kl}d_{lj} = d_i^*\eta d_j$ for $D = [d_1, ..., d_n]$. As $D = (\Phi^+\Psi)^* = \Psi^*\Phi$ with both Ψ^* and Φ quadratic and full rank, it holds that D has also full rank and is therefore invertible [100, p. 13] with inverse $F = D^+$.
This allows to reformulate $L_2(\eta)$ in terms of γ by replacing η with $F^*\gamma F$ leading to the loss function

$$\begin{aligned}
L_\eta &= \log|C_\eta| + \operatorname{tr}\left(SC_\eta^+\right) + r\left[-\log|\eta| + \operatorname{tr}\left(\Lambda^+\eta\right)\right] &(4.69)\\
&= \log|C_\gamma| + \operatorname{tr}\left(SC_\gamma^+\right) + r\left[-\log|F^*\gamma F| + \operatorname{tr}\left(\Lambda^+F^*\gamma F\right)\right] &(4.70)
\end{aligned}$$

completely described by the coefficients γ of an expansion in terms of the ONB $\{\varphi_i\}_{i=1}^n$. Now note that $-\log|F^*\gamma F| = -\log|F^*||F||\gamma| = -2\log|F| - \log|\gamma|$ with $-2\log|F|$ being a finite constant due to F being invertible. It is independent of γ and therefore ignorable for purposes of minimization w.r.t γ. One may therefore equivalently optimize

$$\widetilde{L}_2 = \log|C_\gamma| + \operatorname{tr}\left(SC_\gamma^+\right) + r\left[-\log|\gamma| + \operatorname{tr}(G\gamma F)\right] \tag{4.71}$$

with $G = \Lambda^+ F^*$. Apart from the term $\mathrm{tr}\,(G\gamma F)$ all first and second derivatives have already been calculated before. It is obvious that the second derivatives all vanish and that its first derivatives are given as $F\Lambda^+ F^*$ because

$$\begin{aligned}\frac{\partial}{\partial\gamma_{ij}}\,\mathrm{tr}\,(G\gamma F) = \frac{\partial}{\partial\gamma_{ij}}\sum_{k,l=1}^{n} G_{mk}\gamma_{kl}F_{lm} &= \frac{\partial}{\partial\gamma_{ij}}\sum_{k,l=1}^{n}\gamma_{kl}\sum_{m=1}^{n} G_{mk}F_{lm}\\ &= \sum_{m=1}^{n} F_{jm}G_{mi}\\ &= (FG)_{ji}\end{aligned}$$

and $(F\Lambda^+ F^*)_{ji} = (F\Lambda^+ F^*)_{ij}$ by symmetry. Then it is possible to employ the same Newton-type algorithm as before to minimize $\widetilde{L}_2$ by iterating $\gamma^{k+1} = \gamma^k - B_\gamma^+ \nabla_\gamma \widetilde{L}_2$ where

$$\begin{aligned}B_\gamma^+\nabla_\gamma\widetilde{L}_2 &= \left(\frac{1}{1+r}\right)\gamma\otimes\gamma\left[(1-r)\gamma^+ - \gamma^+ S_\varphi\gamma^+ + rF\Lambda^+F^*\right]\\ &= \left(\frac{1-r}{1+r}\right)\gamma - \left(\frac{1}{1+r}\right)S_\varphi + \left(\frac{r}{1+r}\right)\gamma F\Lambda^+F^*\gamma^*.\end{aligned}$$

The expression for γ^{k+1} based on the $k-$th iterate γ is then

$$\gamma^{k+1} = \left(\frac{1}{1+r}\right)\left[2r\gamma + S_\varphi - r\gamma F\Lambda^+F^*\gamma^*\right] \tag{4.72}$$

where after convergence the relationship $\eta = F^*\gamma F$ can be used to find the optimal η. Alternatively one might directly write down an iteration scheme for η by exploiting $\gamma = D^*\eta D$ via

$$\begin{aligned}\eta^{k+1} &= (1+r)^{-1}(D^*)^+\left[2rD^*\eta D + S_\varphi - rD^*\eta DF\Lambda^+F^*D^*\eta D\right]D^+\\ &= (1+r)^{-1}\big[2r\eta + \underbrace{F^*S_\varphi F}_{S_\psi} - r\eta\Lambda^+\eta\big]\end{aligned} \tag{4.73}$$

Effectively, the scheme is the same as the one presented previously for γ but with S_φ replaced by $F^*S_\varphi F =: S_\psi$. The interpretation is as follows.

$$S_\psi = F^*S_\varphi F = F^*\Phi^+S\Phi F = (\Phi D^*)^+S[(\Phi D^*)^+]^* = (\Psi^+)S(\Psi^+)^*$$

At the same time one finds

$$\underset{\tilde{\eta}\in\mathbb{R}^n\otimes\mathbb{R}^n}{\mathrm{argmin}}\ \|\Psi\tilde{\eta}\Psi^* - S\|_F^2 = \Psi^+\otimes\Psi^+S = (\Psi^+)S(\Psi^+)^* = S_\psi$$

i.e. S_ψ is the coefficient matrix such that $\sum_{i,j=1}^{n}(S_\psi)_{ij}\psi_i\otimes\psi_j^*$ is closest to S. The matrix contains the expansion coefficients of S in the linearly independent system given by $\{\psi_i\}_{i=1}^n$. When $n_{\text{exp}} < n$ expansion coefficients are used, the least squares interpretation still holds. The surprising conclusion is that the algorithm as sketched on page 185 still holds even if only a system of linearly independent vectors spanning $\mathbb{R}^n$ is used but not an ONB.

Remark The reader may notice that this situation was already encountered during the computations for figure 4.10. There an $n_p \times n_p$ covariance matrix C with prior $C_p = \sum_{i=1}^{n_p} \lambda_i \varphi_i \otimes \varphi_i^* \in \mathbb{R}^{n_p} \otimes \mathbb{R}^{n_p}$ had to be inferred from samples $f_{\omega_j} \in \mathbb{R}^n$ with $n < n_p$.

We summarize the results in the following general algorithm for constrained kernel inference for which some geodetic applications are provided in the final section of this chapter.

Nonparametric kernel inference with affine constraints

Problem

The kernel $K(s,t)$ on an indexset $T \times T$ is to be inferred based on n_{obs} observation vectors $f_{\omega_l} = [f_{\omega_l}(t_1), ..., f_{\omega_l}(t_n)]^T \in \mathbb{R}^n$ that provide point evaluations of the functions $f_{\omega_l}(\cdot)$ drawn from $\mathcal{H}_K$ for $l = 1, ..., n_{\text{obs}}$.

A prior guess $K_p(\cdot,\cdot)$ for the kernel with Mercer decomposition $\sum_{i=1}^{\infty} \lambda_i \varphi_i(\cdot)\varphi_i(\cdot)$ is to be respected and the result of inference has to satisfy the affine constraints $\langle A_m, \gamma \rangle = b_m, m = 1, ..., n_c$ where γ is a coefficient matrix determining the guess for K.

Model

The best guess for $K(\cdot,\cdot)$ is determined by that $\gamma \in S_+^{n_{\text{exp}}}$ for which the log probability is maximal; i.e. which minimizes

$$L(\gamma) = \log |C_\gamma| + \operatorname{tr}\left(SC_\gamma^+\right) + r\left[-\log|\gamma| + \operatorname{tr}\left(\Lambda^+\gamma\right)\right]$$

where $C_\gamma = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij} \psi_i \otimes \psi_j^*$ with $\psi_i = [\varphi_i(t_1), ..., \varphi_i(t_n)]^T \in \mathbb{R}^n$ is the covariance matrix induced by the kernel guess $K(s,t) = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}\varphi_i(s)\varphi_i j(t)$ and other quantities are as defined below.

This probabilistic model corresponds to the assumption of Gaussianity for $f_{\omega_l} \in \mathbb{R}^n$ and approximately a Wishart distribution on $\gamma \in S_+^{n_{\text{exp}}}$.

Input

$S = \frac{1}{n_{\text{obs}}} \sum_{j=1}^{n_{\text{obs}}} f_{\omega_j} \otimes f_{\omega_j}^*$, the empirical $n \times n$ covariance matrix

K_p, the prior kernel with known Mercer decomposition $K_p(s,t) = \sum_{i=1}^{n_\infty} \lambda_i \varphi_i(s)\varphi_i(t)$ where n_∞ is allowed to be finite or infinite and the sequence $\{\varphi_i(\cdot)\}_{i=1}^{n_\infty} \subset L^2(T)$ constitutes an ONB.

$A \in \mathbb{R}^{n_c} \otimes \mathbb{R}^{n_{\text{exp}}^2}$, the operator determining the n_c affine constraints $A\gamma = b = \{\langle A_m, \gamma \rangle_F\}_{m=1}^{n_c}$.

$\gamma^0 \in S_+^{n_{\text{exp}}} \cap A^{-1}b$, a feasible initial guess for γ used as a starting point for iteration.

Parameters

n_{exp}, the expansion depth determining the dimension of γ

$r \in [0, \infty)$, a regularization parameter which can be set to $r = n_{\text{obs}}^{-1}$

```
    Begin
        Set $\Lambda \in S_+^{n_{\text{exp}}}, (\Lambda)_{ij} = \delta_{ij}\lambda_i$
        Do until convergence:
        $\gamma = \gamma + \Delta\gamma_1 - \gamma A^*(A\gamma A^*)^+ A\Delta\gamma_1$
                        where $\Delta\gamma_1 = (1+r)^{-1}[(r-1)\gamma + S_\psi - r\gamma\Lambda^+\gamma] \in S_+^{n_{\text{exp}}}$
                        and $S_\psi = (\Psi^+)S(\Psi^*)^+ \in S_+^{n_{\text{exp}}}$, $\Psi = [\psi_1, ..., \psi_{n_{\text{exp}}}]$
     End
  Output
        $\gamma$, a $n_{\text{exp}} \times n_{\text{exp}}$ coefficient matrix subject to linear constraints such that $C(\gamma) = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}\varphi_i \otimes \varphi_j^*$ is both likely under the observations and probable under the prior assumptions. A kernel estimate $K(\cdot,\cdot) = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}\varphi_i(\cdot)\varphi_j(\cdot)$ having those same properties and generalizing $C(\gamma)$.
```

4.4 Selected applications

In this last section of chapter 4, the theoretical apparatus developed in sections 4.1 to 4.3 will be given practical meaning by applying it to geodetically motivated problems. These include but are not limited to different versions of variance and covariance estimation tasks for which kernel inference methodologies are an obvious choice due to the central role played by positivity in the characterizations of variances, covariance matrices, and kernels. Less expected might be the application of linear programming to robust estimation, in which primarily ℓ^1-norm based minimization problems are employed to handle sparsity constraints and introduce robustness into geodetic standard procedures such as the otherwise ℓ^2-norm based estimation of Helmert transformations. The use of LP methods raises the question, which additional geodetic problems might admit a formulation in terms of quadratic programs or the more general semidefinite programs and first nonobvious applications of kernel inference are investigated.

4.4.1 Robust estimation

The choice of parameters x based on ℓ^2 norm minimization of residuals $r = Ax - b$, with A, b problem data, results usually in nonsparse residuals even if an x exists with $(Ax)_j = b_j$ for all but one $j \in J$. Furthermore, due to the square r_j^2 in $\|r\|_2^2 = \sum_{j=1}^n r_j^2$ large residuals are strongly penalized and during the minimization of $\|r\|_2^2$ it regularly happens that the decrease of a large residual r_l by Δr_l is bought by an increase of several small r_j by $\Delta r_j, j \in J/\{l\}$ s.t. $\sum_{j\in J} |r_j + \Delta r_j|^2 < \sum_{j \in J} |r_j|^2$ but the absolute value of the residuals has grown: $\sum_{j \in J} |r_j + \Delta r_j| > \sum_{j\in J} |r_j|$. This tendency to overemphasize the largest residuals can lead to inconvenient behavior when outliers are encountered. Sparsity and robustness to outliers are linked to the ℓ^1-norm rather than the ℓ^2-norm and intuitive explanations for that fact are presented in this subsection. As ℓ^1, ℓ^2 and ℓ^∞-norm minimization can be cast in an LP respectively QP framework, SDP embeddings exist and solutions to $\sigma = \operatorname{argmin}_{x\in\mathbb{R}^n} \|Ax - b\|_p, p \in \{1, 2, \infty\}$ can be calculated numerically. These then help to solve both the robust estimation of parameters in a linear model and the more challenging nonlinear problem of inferring the parameters of a Helmert transformation $H : \mathbb{R}^2 \to \mathbb{R}^2$ given identical points in different coordinate systems.

As shown previously, norm minimization tasks of the type

$$\sigma_f = \underset{f\in\mathcal{H}_K}{\operatorname{argmin}} \|Af - y\|_{\ell^2}^2 + \|f\|_{\mathcal{H}_K}^2 \tag{4.74}$$

correspond to optimal estimation in presence of white noise on the measurements Af of a stochastic process f with covariance function $K(\cdot,\cdot)$. This interpretation is also valid for adjustment problems $A\lambda - y = v$ with white noise v on the measurements and a prior that favors small lengths of the coefficient vector λ although A is a design matrix rather than a measurement operator in this case. Even though a prior on coefficient vectors λ with $(A\lambda)_k = \left(\sum_{j=1}^m \lambda_j g_j(x_k)\right)_k \approx y_k$ seems, at least from this perspective, puzzling at first, it enters naturally if one assumes that the linear combination $\sum_{j=1}^m \lambda_j g_j(\cdot)$ is itself chosen randomly with the λ_j's distributed as multivariate Gaussian.

This establishes interpretations of further machine learning methods that are similar in flavor to the abstract spline problem 4.74. Consider for example

$$\begin{aligned}
\text{Ridge regression}: \ \sigma_f &= \underset{f\in\mathbb{R}^m}{\operatorname{argmin}} \|Af - y\|_{\ell^2}^2 + \alpha\|Bf\|_{\ell^2}^2 \\
\text{LASSO}: \ \sigma_f &= \underset{f\in\mathbb{R}^m}{\operatorname{argmin}} \|Af - y\|_{\ell^2}^2 + \alpha\|f\|_{\ell^1} \\
\text{Elastic net}: \ \sigma_f &= \underset{f\in\mathbb{R}^m}{\operatorname{argmin}} \|Af - y\|_{\ell^2}^2 + \alpha_1\|f\|_{\ell^2}^2 + \alpha_2\|f\|_{\ell^1}
\end{aligned}$$

[92, pp. 61,68,118] where the α's are some positive constants that determine if faithfulness to the data or regularity of the estimator are prioritized and B is some linear operator. In the above, $\|\cdot\|_{\ell^p}$ denotes the classical ℓ^p norms, i.e.

$$\|f\|_{\ell^p} = \sqrt[p]{\sum_{k=1}^m |f_k|^p}.$$

Note that for some nonnegative function $q(f) \geq 0 \ \forall f \in \mathbb{R}^m$ satisfying additional constraints $\exp(-q(f))$ is normalizable with $c^{-1} = \int_{\mathbb{R}^m} e^{-q(f)} df < \infty$ implying that $c\exp(-q(f))$ is a valid probability density function (pdf). Therefore to each norm type there corresponds a unique probability density function: to the ℓ^2 norm one may associate the multivariate Gaussian and to the ℓ^1 norm a multivariate version of the Laplacian distribution. See Fig. 4.12 below for some sketches of the respective norms and densities in the instructive 1 dimensional case.

The Gaussian pdf's derivative at its mean is zero; the pdf's value converges to zero extraordinarily fast. The Laplacian pdf in contrast has heavy tails but its derivative at the mean is undefined. We extract the following from our discussion and the images in Fig. 4.12:

I When minimizing the ℓ^2-norm or equivalently maximizing the likelihood under a Gaussian pdf, small residuals are considered almost irrelevant since the

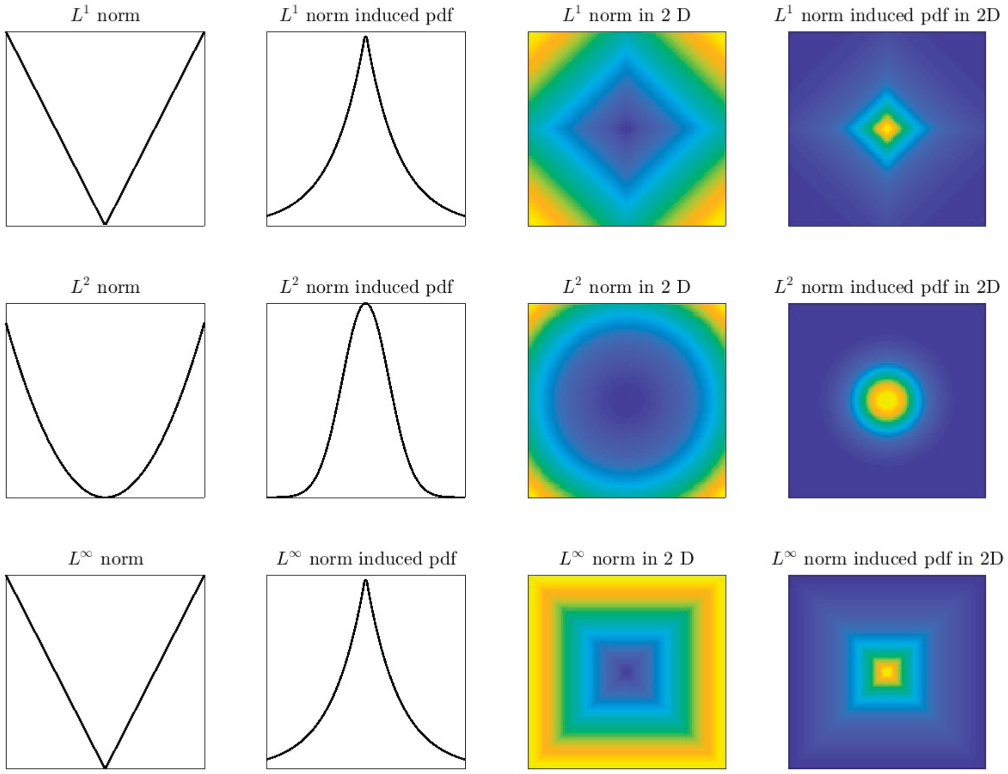


Figure 4.12: The ℓ^1, ℓ^2 and ℓ^∞ norms and their corresponding probability densities associated with the Gaussian and Laplacian distribution. Note the Laplacian's heavier tails. In the 2-D images, brighter color corresponds to higher density values. To the best of the authors knowledge, the exact nature of the pdfs associated to ℓ^∞-norms in dimensions greater or equal than two has not yet been investigated.

gradient of $\|\cdot\|_{\ell^2}^2$ around 0 is zero. Large deviations are punished disproportionately strong: During minimization decreasing a big residual is considered more favorable than decreasing several small ones by the same amount.

II When minimizing the ℓ^1-norm or equivalently maximizing the likelihood under a Laplacian pdf, small residuals are punished less than big ones but still severely as the gradient of $\|\cdot\|_{\ell^1}$ around 0 is constant and positive driving either f to sparsity (if $\|f\|_{\ell^1} \to \min$) or leading to sparse residuals (if $\|Af - y\|_{\ell^1} \to \min$). Big residuals are penalized proportionally: decreasing a big residual is as good as decreasing an already small residual by the same amount.

Combining I and II explains why ℓ^1-norm minimization leads to sparse and robust estimators that can systematically outperform ℓ^2-norm based least squares solutions. Therefore ridge regression might be seen as adjustment with a prior on the length of $B\lambda$, LASSO has a sparsity prior on the parameter vector λ and elastic net regularization balances both. To obtain the usual interpretations, swap f for λ in the above and assume the stochastic process f to be determined by a multivariate Gaussian on Bf, sparse or a combination of both.

The behavior of ℓ^1, ℓ^2 and ℓ^∞-norm based estimation procedures differs significantly. For an illustration of the results of fitting optimally a line and a parabola to data subject to noisy measurements and outliers see figure 4.13. Minimizing $\|A\lambda - y\|_p$ can be cast as a linear program for $p \in \{1, \infty\}$ and as a quadratic program for $p = 2$, which establishes the optimization task as a subproblem of semidefinite programming. The exact SDP-embeddings are reproduced below for the reader's convenience, with more details available in [28, pp. 293-294,].

- The ℓ^1-norm based estimation problem is a diagonal SDP. It can be written as

  $$\lambda^* = \operatorname*{argmin}_{\lambda=[\lambda_1,\ldots,\lambda_m]\in\mathbb{R}^m} \|A\lambda - y\|_1 = \operatorname*{argmin}_{\lambda=[\lambda_1,\ldots,\lambda_m]\in\mathbb{R}^m} \sum_{k=1}^{n} |\langle a_k, \lambda\rangle_{\ell^2} - y_k|$$

  This is equivalent to the following formulation.

  $$\begin{aligned} \min_{q\in\mathbb{R}^s} \quad & \langle c, q\rangle_{\ell^2} \\ \text{subject to} \quad & Gq \le h \end{aligned}$$

  $$c = \begin{bmatrix} \vec{0} \\ \vec{1} \end{bmatrix}; \quad q = \begin{bmatrix} \vec{\lambda} \\ \vec{v} \end{bmatrix}; \quad G = \begin{bmatrix} -A & -I \\ A & -I \end{bmatrix}; \quad h = \begin{bmatrix} -\vec{y} \\ \vec{y} \end{bmatrix}$$

  where λ is the set of m parameters to be inferred, y contains the n observations, and A is the design matrix with rows a_k.

- The ℓ^2-norm based estimation problem has an SDP-embedding as well; its explicit form was derived earlier in example 24. As there are closed-form solutions to the problem that do not rely on numerical optimization, those will be given instead.

  $$\lambda^* = \operatorname*{argmin}_{\lambda=[\lambda_1,\ldots,\lambda_m]\in\mathbb{R}^m} \|A\lambda - y\|_2^2 = \operatorname*{argmin}_{\lambda=[\lambda_1,\ldots,\lambda_m]\in\mathbb{R}^m} \sum_{k=1}^{n} |\langle a_k, \lambda\rangle_{\ell^2} - y_k|^2$$

  The solution is

  $$\lambda = A^+ y$$

  where the notation is as in the ℓ^1-norm minimization case and A^+ denotes the pseudoinverse of A.

- The ℓ^∞-norm based estimation problem is a diagonal SDP. It can be written as

  $$\lambda^* = \operatorname*{argmin}_{\lambda=[\lambda_1,\ldots,\lambda_m]\in\mathbb{R}^m} \|A\lambda - y\|_\infty = \operatorname*{argmin}_{\lambda=[\lambda_1,\ldots,\lambda_m]\in\mathbb{R}^m} \sup_{k\in\{1,\ldots,n\}} |\langle a_k, \lambda\rangle_{\ell^2} - y_k|$$

  This is equivalent to the following formulation.

  $$\begin{aligned} \min_{q\in\mathbb{R}^s} \quad & \langle c, q\rangle_{\ell^2} \\ \text{subject to} \quad & Gq \le h \end{aligned}$$

$$c = \begin{bmatrix} \vec{0} \\ \vec{1} \end{bmatrix}; \quad q = \begin{bmatrix} \vec{\lambda} \\ \vec{v} \end{bmatrix}; \qquad G = \begin{bmatrix} -A & -\mathbf{1} \\ A & \mathbf{1} \end{bmatrix}; \quad h = \begin{bmatrix} -\vec{y} \\ \vec{y} \end{bmatrix}$$

where the notation is as in the ℓ^1-norm minimization case and $\mathbf{1}$ is a vector of ones.

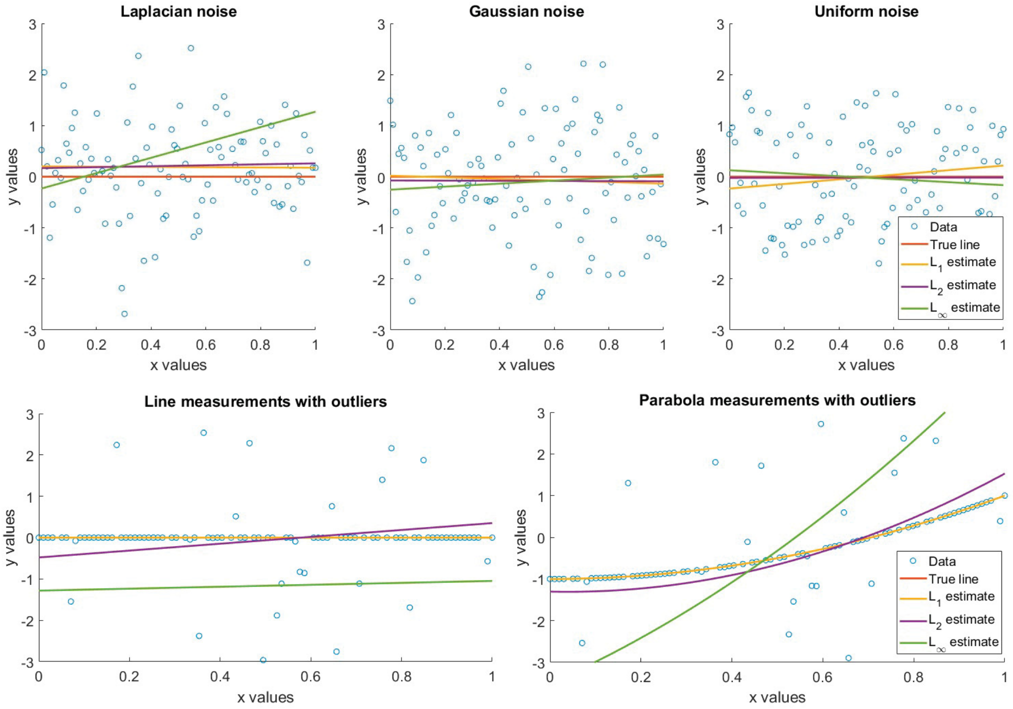


Figure 4.13: The best fitting lines and parabolas based on data contaminated by noise with different probability distributions where "best" is meant as the lowest ℓ^p-norm, $p \in \{1, 2, \infty\}$ of the residuals. Both of the lower images show the estimation's sensitivity with respect to outliers. Where the red line (ground truth) is not visible, it is covered by the coinciding ℓ^1-norm based estimation.

This performance difference of ℓ^2 and ℓ^1-norm based estimation in the presence of outliers carries over to the nonlinear but nonetheless typical geodetic task of inferring a Helmert transformation with fixed scale from coordinate measurements, see Fig. 4.14. We briefly sketch the algorithm used to find the optimal transformation $A(\lambda^*)$ with

$$\lambda^* = \underset{\lambda=[x_A, y_A, \varphi_A] \in \mathbb{R}^3}{\operatorname{argmin}} \|A(\lambda)x - y\|_{\ell^p} \qquad p = 1, 2 \tag{4.75}$$

that maps the coordinates x in system 1 onto the coordinates y in system 2:

1. Get initial solution: $\lambda^0 \in \mathbb{R}^3$.
2. Set up problem: $y^k = A(\lambda^k)x, \Delta y^k = y^k - y$.
3. Estimation step: $\Delta\lambda^* = \underset{\Delta\lambda \in \mathbb{R}^3}{\operatorname{argmin}} \|DA_{[\lambda^k]}\Delta\lambda - \Delta y^k\|_{\ell^p}$.
4. Update step: $\lambda^{k+1} = \lambda^k + \Delta\lambda^*$. Repeat steps 2-4 until convergence.

In the above, D denotes the differential with respect to the parameters. The initial solution can be guessed via an initial least squares step or by solving a subproblem which is neither over- nor underdetermined. The minimization problem in step 3 is either solved analytically (ℓ^2-norm) or via linear programming (ℓ^1-norm) [28, p. 294].

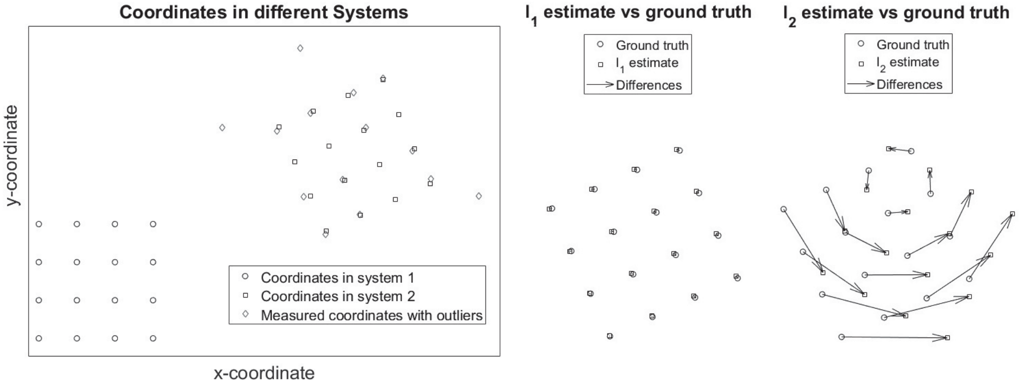


Figure 4.14: The left panel shows three sets of coordinates. They represent the same set of material, real-world points in two different coordinate systems in the presence of noise and outliers. The noisy dataset was used to infer translation and orientation parameters of a Helmert transformation using the two different ℓ^p-based schemes as outlined on page 198. As is visible in the figure, ℓ^1-norm based estimation is much more robust than the ℓ^2-norm based estimation. The scale in the pair of images placed on the righthand side is identical.

4.4.2 Kernel estimation

The estimation of variance components has a long standing history in geodesy. It is used most widely to provide a test of the validity of a hypothetical stochastic model based on real data in conjunction with an updated guess on the diagonal covariance matrix characterizing the precision ratios of the observations. Kernel inference includes as a special case the estimation of a positive diagonal matrix and for purely illustrative purposes the two methods are compared w.r.t the results they offer when applied to instructive numerical examples. As soon as nonzero nondiagonal terms $E[X_iX_j] = \sigma_{ij}, i \neq j$ arise, the simple variance components estimation associated to the covariance matrix model $\Sigma_X = \sum_{k=1}^n \lambda_k e_k \otimes e_k \succeq 0, e_k$ standard basis of $\mathbb{R}^n$, is not sufficient anymore. The covariance model $K_X = \sum_{i,j=1}^n \lambda_{ij}\varphi_i \otimes \varphi_j \succeq 0$ includes nonzero crossterms and by a proper choice of φ_k it is possible to estimate covariance functions and covariance matrices as superpositions of orthogonal parametric or nonparametric basis functions depending on auxiliary information.

Approximation via sequence expansions can be modified to cover decompositions of covariance matrices $\Sigma_X \succeq 0$ into two covariance matrices $\Sigma_{X_1}, \Sigma_{X_2} \succeq 0$ by projection $\sum_{i,j=1}^n \lambda_{ij}\varphi_i \otimes \varphi_j$ onto subspaces spanned by subsets $\{\varphi_k\}_{k\in S}$. This is illustrated by splitting the empirical covariance matrix of exemplary total station measurements into parts corresponding to pure noise and correlated influences explainable by atmospheric effects or smooth movements of the measured object.

Nonparametric kernel inference with affine constraints can be used for what is called variance components estimation in the geodetic literature [114, p. 265]. There, a combination of linearly independent elementary positive semidefinite matrices $C_j, j = 1, ..., n_{\text{exp}}$ into a single covariance matrix C_μ is sought in such a way that C_μ and the observations are minimally contradictory.

A covariance matrix $C_\mu = \sum_{j=1}^{n_{\text{exp}}} \mu_j C_j$ assembled from the $C_j \in S_+^n$ is interpretable as the joint covariance matrix resulting from having several independent observation processes each with covariance matrix C_j. Suppose for example that $x = [x_1, x_2] = [\epsilon_1, \epsilon_1 + \epsilon_2]$ for two independent white noise variables ϵ_1, ϵ_2 with unknown variances μ_1 and μ_2. Then

$$C_\mu = E[x \otimes x^*] = \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_1 + \mu_2 \end{bmatrix} = \mu_1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \mu_2 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \sum_{j=1}^{2} \mu_j C_j$$

motivating this approach and giving the matrices C_j the interpretation of being covariance matrices of blocks of observations. One can come up with similarly sensible explanations for nondiagonal C_j, see for example [151, p. 320]. Given a prior Wishart distribution with scale matrix Λ for μ (considered as a diagonal matrix), finding the maximum aposteriori parameter matrix μ corresponds to minimizing

$$L(\mu) = \log |C_\mu| + \operatorname{tr}\left(SC_\mu^+\right) + r\left[-\log|\mu| + \operatorname{tr}\left(\Lambda^+ \mu\right)\right] \tag{4.76}$$

where $C_\mu = \sum_{j=1}^{n_{\exp}} C_j$ and S is an empirical covariance matrix derived from observations. One might swap $C_\mu = \sum_{j=1}^{n_{\exp}} \mu_j C_j$ for the more general model $C_\gamma = \sum_{i,j=1}^{n} \gamma_{ij} \varphi_i \otimes \varphi_j^*$ with $\{\varphi_i\}_{i=1}^n$ some ONB of $\mathbb{R}^n$ and as a compensation impose an affine constraint on $\gamma \in S_+^n$ to ensure that $\exists \mu \in \mathbb{R}^{n_{\exp}}$ such that $C_\mu = C_\gamma$. Since converting a prior on μ into a prior on γ is nontrivial in the case where the transformation is not of the form $\mu = \Phi\gamma\Phi^*$, attention will be restricted to the special case of $r = 0$, thereby effectively neglecting regularization and only performing maximum likelihood estimation.

Note that demanding $C_\gamma \in \operatorname{span}\{C_1, ..., C_{n_{\exp}}\}$ is the same as demanding $PC_\gamma = C_\gamma$ where $P : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ is the orthogonal projection onto the column space of $C = [C_1, ..., C_{n_{\exp}}] \in \mathbb{R}^{n^2} \otimes \mathbb{R}^{n_{\exp}}$. The appropriate linear constraint on γ is therefore

$$(I - P)\Phi\gamma\Phi^* = 0 \qquad\qquad P = C(C^*C)^+C^* \tag{4.77}$$

and by renaming $A = (I - P)\Phi \otimes \Phi$, the problem of maximum likelihood estimation can be encapsulated as the optimization problem

$$\begin{aligned} &\text{minimize} && \log|C_\gamma| + \operatorname{tr}\left(SC_\gamma^+\right) \\ &\text{subject to} && A\gamma = 0. \end{aligned} \tag{4.78}$$

But this is just constrained kernel inference with regularization parameter $r = 0$ and can be solved by employing the algorithm outlined on page 193. The steps are briefly listed below for convenience.

Maximum likelihood variance components estimation

```
Input
    $S = \frac{1}{n_{\text{obs}}} \sum_{j=1}^{n_{\text{obs}}} f_{\omega_j} \otimes f_{\omega_j}^*$, the empirical $n \times n$ covariance matrix
    $\{\varphi_i\}_{i=1}^n$, an ONB of $\mathbb{R}^n$
    $\{C_j\}_{j=1}^{n_{\text{exp}}}$ a linearly independent sequence of positive semidefinite matrices
Begin
    $\gamma^0 = \Phi^* \widetilde{C} \Phi$ with $\Phi = [\varphi_1, ..., \varphi_n], \widetilde{C} = \sum_{j=1}^{n_{\text{exp}}} C_j$
    Do until convergence:
    $\gamma = S_\varphi - \gamma A^*(A\gamma A^*)^+ A S_\varphi + \gamma A^*(A\gamma A^*)^+ A\gamma$
                where $S_\varphi = \Phi^* S \Phi$ and $A = (I - C(C^*C)^+ C^*)\Phi \otimes \Phi$
                and $C = [C_1, ..., C_{n_{\text{exp}}}] \in \mathbb{R}^{n^2} \otimes \mathbb{R}^{n_{\text{exp}}}$
End
Output
    $\gamma$, a $n \times n$ coefficient matrix such that $C(\gamma) = \sum_{i,j=1}^n \gamma_{ij} \varphi_i \otimes \varphi_j^*$ is the matrix of the
    form $C_\gamma = \sum_{j=1}^{n_{\text{exp}}} \mu_j C_j$ with the highest likelihood given the data.
```

Remark If γ or $A\gamma A^*$ are not invertible, numerical problems during formation of $\Delta\gamma$ can induce nonzero $A\Delta\gamma$ implying that later iterations of $\Phi\gamma\Phi^*$ are not expressible anymore as $\sum_{j=1}^{n_{\text{exp}}} \mu_j C_j$. Typically noninvertible γ are encountered, when the initial construction $C^{\text{init}} = \sum_{j=1}^{n_{\text{exp}}} C_j$ does not have full rank as then $\operatorname{rank} \gamma^0 = \operatorname{rank}(\Phi C^{\text{init}} \Phi) = \operatorname{rank} C^{\text{init}}$ [100, p. 13] for Φ constructed from an ONB $\{\varphi_i\}_{i=1}^n$.

§ Comparison with least squares

Estimation of variance components for geodetic applications has received some attention particularly during the advent of affordable compute power during the late 1970's. Iterative schemes were published during that era, see [64] for an overview. They generalize the well known formulas for empirical variances and covariances and are of type BIQUE (best invariant quadratic unbiased estimators). Estimation of variances or coefficients in a functional model describing variances is still an important problem lacking a universally agreed upon success criterion and closed form solutions. We will briefly sketch the construction of a least squares estimator that is fast and relatively reliable but can lead to negative guesses for variances and under unfortunate circumstances results in covariance matrices that are not positive semidefinite. In what follows, the outputs of kernel inference and the the least squares based algorithm are compared with respect to the computational cost associated to deriving them, several error metrics and their performance for estimation purposes where applicable.

Estimation of coefficients μ_j may be solved within a least squares framework by solving the problem

$$\mu_{opt} = \underset{\mu\in\mathbb{R}^{n_{\text{exp}}}}{\text{argmin}} \; \|\sum_{j=1}^{n_{\text{exp}}} \mu_j C_j - S\|_F^2 = C^+ S. \tag{4.79}$$

Although a direct generalization of the straightforward scalar estimators for empirical variances and covariances, least squares estimators can lead to model covariance matrices $\sum_{j=1}^{n_{\text{exp}}} \mu_j C_j$ that are not positive semidefinite, violate the conditions sensibly imposed on covariance matrices and perform insufficiently when themselves used for estimation. Three problems will illustrate the behavior exhibited by the least squares solution and the Maximum Likelihood estimator derived in the kernel-inference framework. The summary statistics provided for each of the problems are generated using simulated data based on synthetic ground truth and 1000 trial runs, in which both kernel inference and the classical method are employed to derive a solution. The entries $\|\hat{x}_{C_{\text{est}}} - \hat{x}_{C_{\text{true}}}\|_2$ are the average ℓ^2-norms quantifying the lengths of the deviations between solutions of the problems

$$\hat{x}_{C_{\text{est}}} = \underset{x\in\mathbb{R}^{n_{\text{exp}}}}{\text{argmin}} \; \|Ax-b\|_{C_{\text{est}}}^2 \qquad \text{and} \qquad \hat{x}_{C_{\text{true}}} = \underset{x\in\mathbb{R}^{n_{\text{exp}}}}{\text{argmin}} \; \|Ax-b\|_{C_{\text{true}}}^2;$$

i.e they quantify how far estimation with C_{est} deviates from estimation with C_{true}. In the above, A and b are white noise variables of the appropriate sizes randomly drawn for each of the trial runs and C_{est} and C_{true} are, of course, the estimated and the underlying true covariance matrices of the problem.

Example 25 (Problem I) Estimation of basis variances μ_1 and μ_2 in total station measurements with observation variances $\sigma_{obs} = \mu_1 + \mu_2 d_{obs}$ where d_{obs} is the distance to reflective targets in m. The goal is to find μ_1 and $\mu_2 \geq 0$ such that $\mu_1 I + \mu_2 D$ is most likely given an empirical covariance matrix S. D is a diagonal matrix with $(D)_{ij} = \delta_{ij} d_{obs_i}$ with i,j ranging from 1 to n_{obs}, the number of observations. In the specific instance of problem I treated below, $n_{obs} = 5, D = 100\,\text{diag}[1,2,3,4,5]$, the true μ_1 and μ_2 are $\mu_1^{\text{true}} = 1$ m and $\mu_2^{\text{true}} = 10^{-3}$, and the empirical covariance matrix S comes from 50 simulations drawn from the true underlying distribution. The results of this and the following examples are summarily interpreted after example 27.

	Runtime in s	Absolute error	RMSE	Likelihood	$\|\hat{x}_{C_{\text{est}}} - \hat{x}_{C_{\text{true}}}\|_2$
LS	$4.3 * 10^{-5}$	$\mu_1 : 0.2042$ $\mu_2 : 0.0007$	$\mu_1 : 0.2556$ $\mu_2 : 0.0008$	7.7140	0.0409
KI	0.0449	$\mu_1 : 0.2022$ $\mu_2 : 0.0007$	$\mu_1 : 0.2533$ $\mu_2 : 0.0008$	7.7136	0.0406

■

Example 26 (Problem II) Estimation of variance of leveling instruments. Let a simple leveling setup with known point heights and lengths $L_1, L_2, L_3 = L_1 + L_2$

be given as specified in figure 4.15.

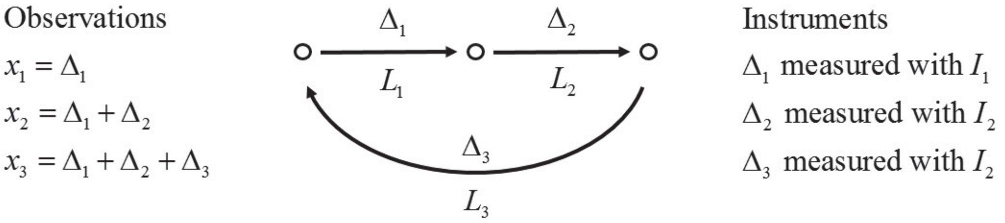

Observations
$x_1 = \Delta_1$
$x_2 = \Delta_1 + \Delta_2$
$x_3 = \Delta_1 + \Delta_2 + \Delta_3$
Δ_1
Δ_2
L_1
L_2
Δ_3
L_3
Instruments
Δ_1 measured with I_1
Δ_2 measured with I_2
Δ_3 measured with I_2

Figure 4.15: The hypothetical leveling setup for purposes of instrument testing investigated in Problem II. The Δ's are measured height differences between the points of known height marked as empty circles and can therefore be directly converted to residuals.

Presupposing the usual model claiming leveling variances to be proportional to leveling lengths, it is trivial to check that

$$\begin{aligned} \sigma_{x_1}^2 &= \mu_1 L_1 \quad & \sigma_{x_2}^2 &= \mu_1 L_1 + \mu_2 L_2 \quad & \sigma_{x_3}^2 &= \mu_1 L_1 + \mu_2 L_2 + \mu_2 (L_1 + L_2) \\ \sigma_{x_1 x_2}^2 &= \mu_1 L_1 \quad & \sigma_{x_1 x_3}^2 &= \mu_1 L_1 \quad & \sigma_{x_2 x_3}^2 &= \mu_1 L_1 + \mu_2 L_2 \end{aligned}$$

where $\sigma_{x_k}^2$ denotes the variance of x_k and $\sigma_{x_i x_j}^2$ denotes the covariance between x_i and x_j. This implies that the model $\hat{S}$ for the covariance matrix of the observations depending on the variance factors μ_1, μ_2 is

$$\hat{S} = \mu_1 \begin{bmatrix} L_1 & L_1 & L_1 \\ L_1 & L_1 & L_1 \\ L_1 & L_1 & L_1 \end{bmatrix} + \mu_2 \begin{bmatrix} 0 & 0 & 0 \\ 0 & L_2 & L_2 \\ 0 & L_2 & L_1 + 2L_2 \end{bmatrix}$$

Now infer from 50 realizations of x and the associated empirical covariance matrix S the coefficients μ_1 and μ_2. They correspond to $2\sigma_{I_1}^2$ and $2\sigma_{I_2}^2$, multiples of the two instrument variances per 1 km double run. The synthetic ground truth is $\mu_1^{\text{true}} = 1$ and $\mu_2^{\text{true}} = 2$. Note that an imaginary average likelihood implies, that invalid covariance matrices have been produced by the LS procedure.

	Runtime in s	Absolute error	RMSE	Likelihood	$\lVert \hat{x}_{C_{\text{est}}} - \hat{x}_{C_{\text{true}}} \rVert_2$
LS	$3.5 * 10^{-5}$	$\mu_1 : 0.3055$ $\mu_2 : 0.3258$	$\mu_1 : 0.3867$ $\mu_2 : 0.4067$	$5.873 + 0.003i$	0.083
KI	0.0610	$\mu_1 : 0.2354$ $\mu_2 : 0.2834$	$\mu_1 : 0.3010$ $\mu_2 : 0.3539$	5.8484	0.063

■

Example 27 (Problem III) Approximately decompose an empirical covariance matrix S derived from time series into a superpositions of two parts S_l, S_h that correspond to low and high frequency components. The goal is to use S_l and S_h for signal separation and filter out the high-frequency 'noise' from a time series. To this end, set the ground truth covariance function to be a superposition $K(\cdot, \cdot)$ of a squared exponential covariance and a pure white noise covariance, both with coefficients 1. This ground truth is unknown to the algorithm. Afterwards sample 50 realizations of a stochastic process with covariance K, then form the empirical covariance matrix S. Decompose S into the weighted sum $C_0 + \sum_{j=1}^{n_{\text{exp}}} \mu_j C_j$ where C_0 is the identity matrix and the $C_j, j = 1, ..., n_{\text{exp}}$ are covariance matrices of smooth processes. Figure 4.16 illustrates this procedure and the associated table again compares performances of the classical solution and kernel inference. Surprisingly, Least squares and kernel inference are identical up to machine precision in this example.

	Runtime in s	Absolute error	RMSE	Likelihood	Prediction RMSE
LS	$1.1 * 10^{-4}$	$\mu_0 : 0.0284$ $\mu_1 : 1.0399$	$\mu_0 : 0.036$ $\mu_1 : 1.3081$	20.3766	1.3945
KI	0.2134	$\mu_0 : 0.0284$ $\mu_1 : 1.0399$	$\mu_0 : 0.036$ $\mu_1 : 1.3081$	20.3766	1.3945

Remark Note that in these tests, we are restricting kernel inference to solve a task similar to the one solved by variance components estimation. It should be clear, that the least squares algorithm can be outperformed significantly w.r.t. prediction RMSE by employing the more general model $S \approx \hat{S} = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij} \varphi_i \otimes \varphi_j^* + \gamma_0 I$ for $\{\varphi_i\}_{i=1}^{n_{\text{exp}}}$ some ONB coming from the spectral decomposition of a smooth covariance matrix. However, inference with this model is impossible with least squares as it is unable to assure a choice of γ that leads to positive (semi-)definite $\hat{S}$.

■

Note that our algorithms runtime scales unfavorably with the dimension $n \times n$ of the matrices involved because the full $n \times n$ matrix γ is is formed and inference is performed by iterating over γ subject to affine constraints. The least squares algorithm manipulates the coefficients directly and only once; it is therefore much faster. By extrapolation from problems I to III, one might conclude that least squares estimation finds itself at an advantage over kernel inference in situations where numerical simplicity and lower runtimes are of the essence whereas the kernel inference shows better results in terms of those parameters likelihood, which can be imaginary owing to violations of positive definiteness constraints in the least squares approach, and prediction performance.

§ Inference of instationary covariance functions

The inference of covariance functions of random fields is included as a subcase in the kernel inference algorithm. It is mentioned here explicitly only because the practical application may seem nonobvious. Suppose an instationary random field N has been observed at several locations $\{s_k\}_{k=1}^{n} \subset \mathbb{R}^2$ where s_k are spatial coordinates and for purposes of illustration $N_.(\cdot)$ is the distribution of a modified refractivity $N = 10^6(n_{\text{air}} - 1)$ where n_{air} is the refractive index of air. For each $\omega \in \Omega$, $N_\omega(\cdot) : \mathbb{R}^2 \to \mathbb{R}$ is a specific spatial distribution of N in x and z, i.e. a profile drawn at random and indexed by an element in the probability space Ω.

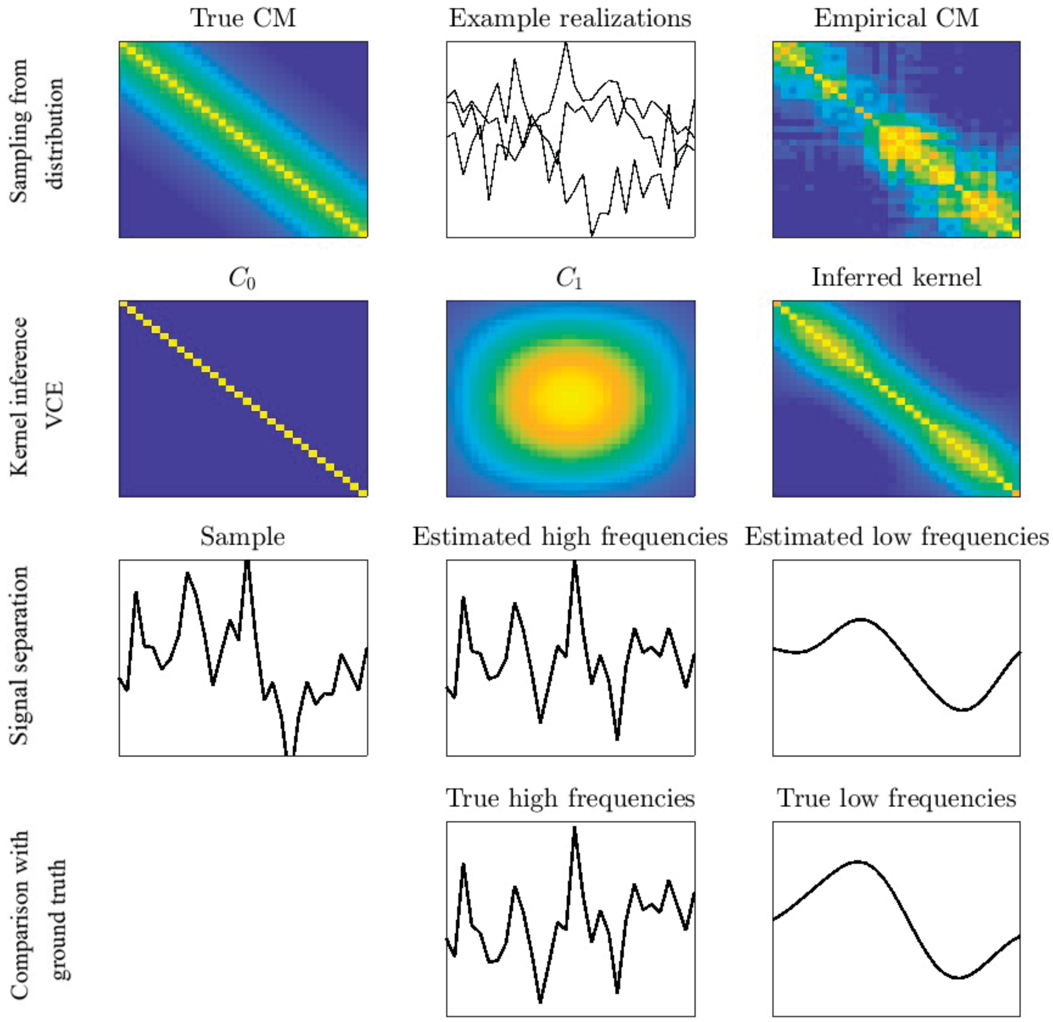


Figure 4.16: In Problem III, variance components estimation (VCE) is used to approximately decompose an empirical covariance matrix into a rough part (C_0) and a superposition of smooth parts $(e.g.\ C_1)$. Later on, this decomposition of covariance matrices is used for signal separation.

Given the $n_{\text{meas}} = 40$ measurements of the of the $n_{\text{obs}} = 6$ realizations $N_{\omega 1}, ..., N_{\omega 6}$ at the locations $s_1, ..., n_{\text{meas}}$ shown in figure 4.17, the task is to infer the kernel $K_N(\cdot, \cdot) : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ thereby enabling inference of the whole of N.

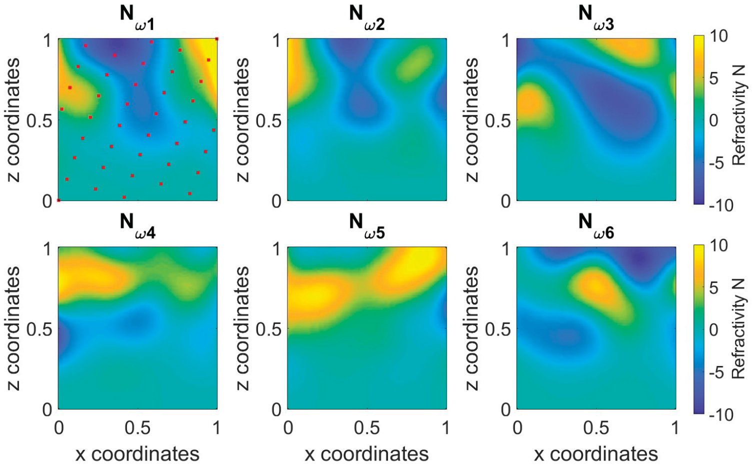


Figure 4.17: Six realizations of the random field $N.(\cdot) : \Omega \times \mathbb{R}^2 \to \mathbb{R}$ representing a modified refractivity. For the simulations a stochastic model was used; it features smooth changes in $N.(\cdot)$ and increasing variability for high altitudes. The red boxes mark the locations of measurements to be used for kernel inference.

To perform inference, use an exponential kernel K_N^x and K_N^z in x and z direction respectively. Choose the range parameters reasonably and construct the prior

$$K_N^{\text{prior}} = K_N^x K_N^z.$$

From the Mercer decompositions $K_N^x = \sum_{i=1}^{\infty} \lambda_i^x \varphi_i^x \otimes \varphi_i^x$ and $K_N^z = \sum_{i=1}^{\infty} \lambda_i^z \varphi_i^z \otimes \varphi_i^z$ construct via tensorization the Mercer decomposition of K_N^{prior} as

$$K_N^{\text{prior}}(\cdot, \cdot) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(\cdot) \otimes \varphi_i(\cdot)$$

where λ_i is the i-th element of the sequence $\{\lambda_k^x \lambda_l^z\}_{k,l=1}^{\infty}$ sorted in descending order and $\varphi_i(\cdot) = \varphi_k^x(\cdot) \otimes \varphi_l^z(\cdot)$ is the tensor product of the eigenfunctions corresponding to the eigenvalues λ_k^x and λ_l^z. Each $\varphi_i(\cdot)$ is therefore a function from $\mathbb{R}^2$ to $\mathbb{R}$. There are no affine constraints. To minimize the objective function

$$L(\gamma) = \log |C_\gamma| + \operatorname{tr}(SC^+) + r \left[-\log |\gamma| + \operatorname{tr}(\Lambda^+ \gamma) \right]$$

$$C_\gamma = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij} \Psi_i \otimes \Psi_i^* \qquad \Psi_i \in \mathbb{R}^n, (\Psi_i)_k = \varphi_i(s_k)$$

with $n_{\text{exp}} = 20, \Lambda \in \mathbb{R}^{n_{\text{exp}}} \otimes \mathbb{R}^{n_{\text{exp}}}, \lambda_{ij} = \delta_{ij} \lambda$ and S the empirical covariance matrix, execute the iterative scheme proposed for unconstrained kernel inference.

This means, repeat

$$\gamma^{k+1} = \frac{1}{1+r}\left[2r\gamma + S_\psi - r\gamma\Lambda^+\gamma\right]$$

until convergence, where the $n_{\text{exp}} \times n_{\text{exp}}$ symmetric matrix S_ψ is defined as $S_\psi = (\Psi)^+ S(\Psi^*)^+$. The result is a symmetric positive semidefinite matrix γ such that $K_N(\cdot,\cdot) = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij}\varphi_i(\cdot) \otimes \varphi_j(\cdot)$ is a reasonable kernel given the observations. Using the classical Kriging formulas, interpolation can then be performed, yielding the results plotted in figure 4.18. The figure suggests that, compared to other approaches, interpolation with covariance functions determined by kernel inference potentially frees the user from the negative impact that misspecified covariance models can have on the estimation performance. This is especially the case when correlation data is reliable and has been gathered densely; then the advantage of additional flexibility of a non-parametric model outweigh the disadvantage of worse regularity properties.

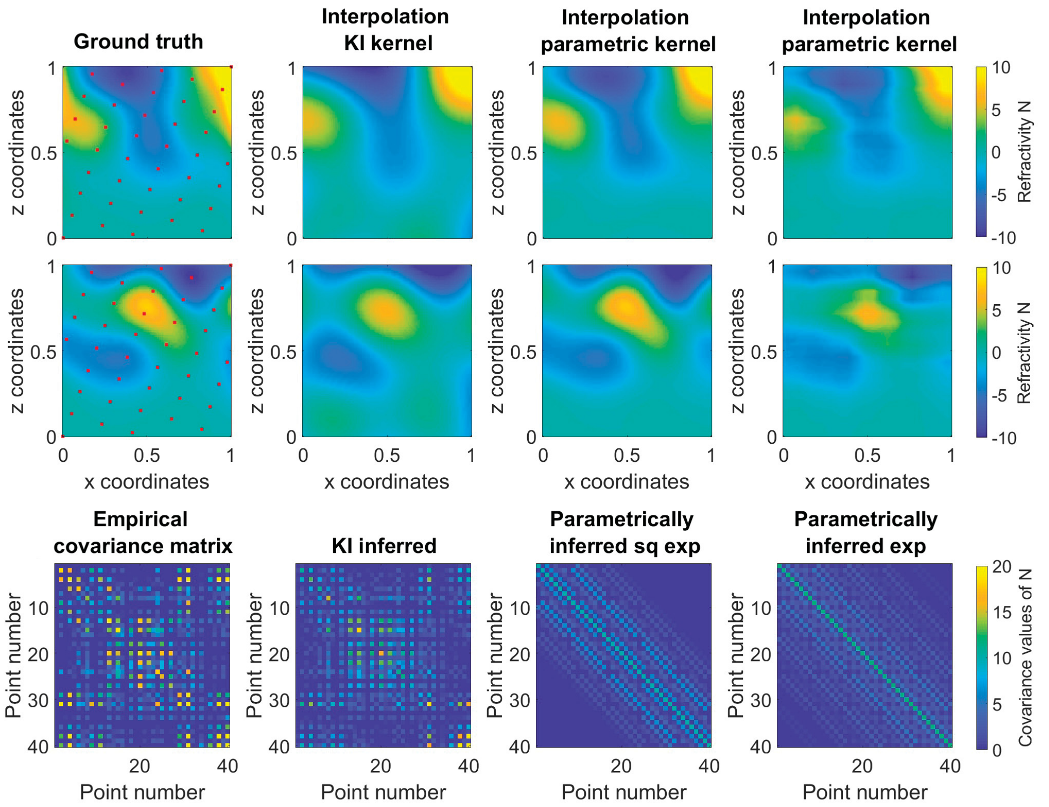


Figure 4.18: The interpolations performed with kernels extracted from the data plotted in the previous figure. The first column shows the ground truth, two different realizations and an empirical covariance matrix. The three columns right of that show the guesses of the ground truth based on the measurements marked with the red squares. Estimation was performed using the kernel from kernel inference (2nd column), a squared exponential kernel close to the true underlying kernel (3rd column) and an exponential kernel (4th column) that one may actually count as a gross misspecification. Notice that the KI-kernel is nonstationary and exhibits a truthful representation of the variation structure even though the prior was the poorly performing covariance function in the fourth column.

Chapter 5

Applications to terrestrial radar interferometry

In this last chapter, the concepts introduced in chapters 2 to 4 — Hilbert spaces of functions, probability distributions on them, and their choice — will be employed in the context of radar interferometry. In order to do this, it is necessary to review the basic principles of terrestrial radar interferometry (TRI) first and explain the relationship between objects surveyed, data produced, and the role of the transmission medium. It will not take long to identify the atmosphere as one of the main contributors to the artifacts contaminating TRI-data whereupon we focus on extracting the deformations from the noisy data in an RKHS framework. The way the data is normally represented as a complex number encoding phase and amplitude leads us to consider complex covariance matrices whose interpretation is followed by practical strategies for estimating them in the framework of kernel inference. The standard approaches for solving TRI signal separation tasks turn out to be special cases of RKHS-based processing and we provide the appropriate Hilbert space embeddings that establish this link.

5.1 Basics of terrestrial radar interferometry

Terrestrial radar interferometry (TRI) is a relatively young technology devoted to employing electromagnetic waves to derive coarse digital elevation models or highly precise deformation estimates for surveying or monitoring purposes. TRI as a remote sensing technology also used by geodesists has been applied successfully in practice to derive safety-relevant deformation estimates for landslides, glaciers, volcanoes, and man-made infrastructure among other objects [194, 166, 201, 167, 47]; more comprehensive overviews can be found in [35] and [143]. However, the technology is still only partially understood from a scientific standpoint as testified by still ongoing discussions as to what is actually measured and how the data are to be interpreted exactly. Instruments have been designed, developed, completed and made available to the market before the real-world implications of their wave-theoretical underpinnings had fully been worked out.

This precludes a completely satisfactory and closed treatment of TRI from a mathematical perspective, and topics like scattering theory and phasor arithmetics will only be touched upon briefly herein. Instead, a rough survey of principles and instruments will be provided and serves as an intuitive guide to the functional relations between surveyed objects and data. We will allow ourselves to be guided by empiricism and use mathematically sound reasoning based on formulae only in a supporting role. In the same spirit, the standard processing chain producing unwrapped interferograms from SLCs will be presented; uses and motivations behind the products generated in this way are emphasized. A detailed look at some applications will reveal current limitations that are inherent to the technology and can only be approached via more sophisticated data analysis.

5.1.1 Survey of principles and instruments

TRI is as an active remote sensing technology that uses radar waves to derive deformations by evaluating changes of the backscattering behavior of natural surfaces. Even though the physical and electrotechnical realization of the instrument and explanations of its characteristics rely on quantum mechanical arguments, simply relating changes of a surface's mean-distance (to the instrument) to differences in measured phase and focusing almost exclusively on these works well enough as a mental model and first order approximation to reality.

Spatial resolution in range direction is generated via frequency modulation whereas the way in which the resolution in angular (cross-range) direction is formed depends on the type of system used. One mainly differentiates between systems with real or synthetic aperture and it makes sense to treat these separately. The geometric configuration inherent to the measurement process has immediate consequences in terms of certain effects on the distribution of noise, radar shadows, and classes of movements that are indistinguishable from the perspective of TRI.

Before any interferometric post processing happens, a terrestrial radar interferometer provides amplitudes A and phase values φ for a discrete set of positions $(x_{kl}, y_{kl}), (k,l) \in \mathbb{Z}_m \times \mathbb{Z}_n$ corresponding to real-world surface patches. These sets of values $(A_{kl}, \varphi_{kl}), A_{kl} \geq 0, \varphi_{kl} \in [-\pi, \pi]$, are typically stored in complex form as a matrix $z \in \mathbb{C}^m \otimes \mathbb{C}^n$ representing the totality of waves scattered back by surfaces associated to the positions $\{(x_{kl}, y_{kl})\}_{(k,l)\in\mathbb{Z}_m\times\mathbb{Z}_n} \subset \mathbb{R}^2$. More on the exact interpretation of these complex values can be found in the subsequent subsection and the general scheme is illustrated in figure 5.1.

The matrix z is also called a Single-Look Complex image, or SLC for short. Whereas the phases $\{\varphi_{kl}\}_{(k,l)=(1,1)}^{(m,n)}$ are basically realizations of spatial uniform noise on the interval $[-\pi, \pi]$ [91, p. 89] they should stay constant in time in absence of any changes of instrument conditions, propagation medium or backscattering surfaces. Conversely for different times t_1 and t_2, non-zero phase differences $\Delta\varphi_{kl} = \varphi_{kl}^{t_2} - \varphi_{kl}^{t_1}$ correspond to changes in topography or atmosphere and often have a strongly autocorrelated and highly intricate structure that reflects underlying meteorological phenomena or deformation processes; see figure 5.2 for interferograms featuring both. Note that it is common to use the word 'phase' also for the phase difference $\Delta\varphi$ of the interferogram. This introduces a slight ambiguity and terminological overlap as the same nomenclature is used for the phase of an SLC.

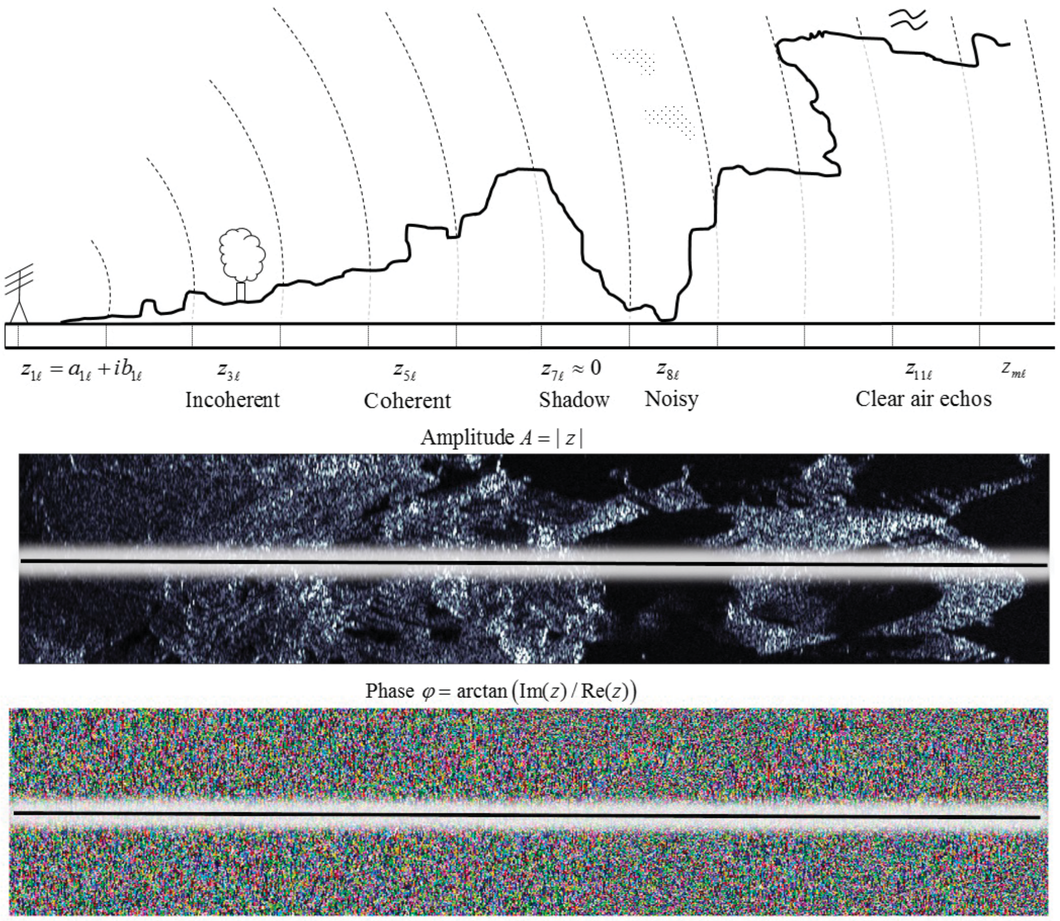


Figure 5.1: The top panel illustrates the measurement process schematically. Waves are emitted, scattered back by terrain and objects and then received by the radar. Due to occlusion, certain regions should theoretically have a backscattering intensity of 0, however, noise, small scale scatterers and meteorological effects preclude this. Different systems patch these one-dimensional slices together differently. The result is nonetheless always a set of two two-dimensional images in polar coordinates that detail amplitude and phase values (bottom panels). If, as will be the case in later practical applications, an instrument with a real aperture is used, then each line of constant polar angle in the resultant images is associated to the signal backscattered by a one-dimensional section of the topography. In this way, the measurement situation depicted in the top panel may lead to the single lines marked in the bottom two panels, see comments in the text.

However, since the original phases φ of the SLCs are rarely the subject of discussion directly, the potential for confusion is limited. The same holds for the symbol 'φ' which may either denote a phase or a phase difference depending on context.

Of the four quantities $A_{kl}, \varphi_{kl}, x_{kl}, y_{kl}$ provided by a TRI measurement, the way the positions as given by tuples (x_{kl}, y_{kl}) are formed is indicative of the type of instrument used. All of them employ either frequency modulated continuous waves or stepped frequency continuous waves to derive distances d between instrument and backscattering surface by comparing the difference in phases of transmitted and received signals. These differences are time-dependent and by analyzing the mixture of transmitted and received signals one finds the formula

$$d = \frac{c}{2}\frac{f_{\text{beat}}}{\mu} \tag{5.1}$$

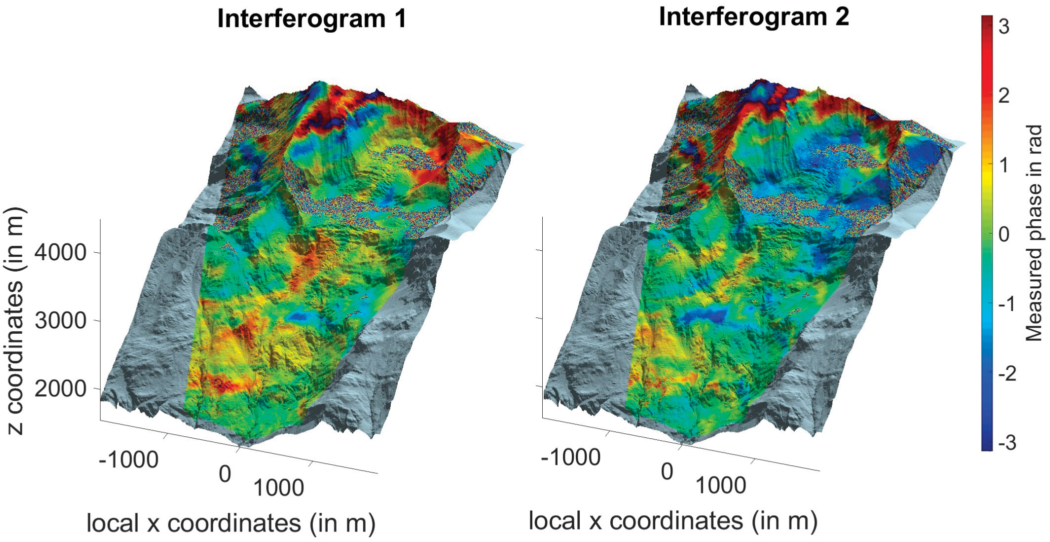


Figure 5.2: The phases of two subsequent interferograms formed by differencing three SLCs that were acquired two minutes apart. In these interferograms, one part of the phase is dynamic and changes even in the course of two minutes time difference and the other one is more slowly varying. The data are taken from the dataset described in subsection 5.2.2.

where c is the speed of light, μ the rate of change in frequency, f_{beat} the time derivative of the mixtures phase and d is the distance to be inferred. More details can be found for example in [189], [109, p. 14], and [111, pp. 90-92] . In this way, a one-dimensional slice of distance measurements is generated from the superposition of received signals by low-pass filtering their mixtures with the transmitted signal and performing Fourier analysis to decompose them into elementary waves with different frequencies (distances) and complex coefficients (amplitudes and phases).

This however is not yet sufficient to derive a spatial distribution of amplitudes and phases, which is why additional measurements have to be made to guarantee sufficient cross-range resolution. One may distinguish between radars with real aperture (RAR), which are limited in terms of cross-range resolution by the antenna's gain pattern and radars with synthetic aperture (SAR). For the latter ones, the width at half maximum is typically not a limiting factor as they acquire the set of range-profiles necessary for generating a spatial distribution of amplitudes and phases not via reorientation of their antenna but via movement of the whole instrument. Both types of radars are in practical operation. A brief sketch of the TRI instruments available on the market follows. Although care was taken to compile all information found on that topic in scientific papers, no guarantee can be given as towards the list's completeness. Where detailed information regarding an instrument's specifications is otherwise unavailable from scientific sources, webpresences of manufacturing companies are referenced.

At the time of writing there exist at least six commercially available systems for measuring medium- to large-scale surface deformations with TRI [143]. Three types of essentially different working principles may be distinguished. Four systems that have reached technological and commercial maturity rely on the SAR principle and use a multitude of acquisitions from slightly different perspectives to synthesize a nontrivial cross-range resolution purely computationally. To this end, the TRIs ’FastGBSAR’ of Metasensing BV, ’IBIS-FM’ manufactured by IDS, and ’LiSALab’ by Ellegi S.r.l. translate their antennas along a rail of approximately 2 m length whereas ’IBIS-ArcSAR’ by IDS employs a circular motion. The instruments going by the name of ’GPRI’ and ’SSR-XT’ produced by Gamma Remote Sensing AG and GroundProbe ltd. respectively are real aperture radars. All systems apart from SSR-XT and LiSALab operate in a frequency spectrum ranging from 17.1 to 17.3 GHz.

FastGBSAR: With acquisition times of less than 5 seconds per image, this currently fastest commercially available system is part of the products offered by the Dutch company Metasensing BV [167]. It is operable in two different modes (either as SAR or RAR) which, according to the manufacturer, enables its usage for the purposes of both structural health monitoring as well as for risk assessment of geotechnical constructions and widespread geological phenomena. The instrument has to be fixed on a tripod for RAR-usage. This mode of operation allows measurement of approximately 4000 range profiles per second and enables the analysis of vibrations and rapid movements of structures or industrial objects requiring supervision.

Figure 5.3: The FastGBSAR system (Image: [167]).

IBIS-FM: This system by the Italian company IDS is widely used in the mining industry. With an acquisition time of approximately 3 minutes, it is still suitable for continuous monitoring [102]. It has already been used for risk assessment in different scenarios related to geohazards; usage for structural health monitoring is possible as well in a slightly modified version that goes by the name of IBIS-FS. The mode of operation is analogous to that of FastGBSAR but offers only a temporal resolution of 200 Hz in RAR mode [103].

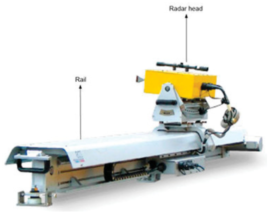


Figure 5.4: The IBIS-FL system (Image: IDS).

LiSALab: This instrument by company Ellegi S.r.l is based on research done in cooperation with the Institute for the Protection and Security of the Citizen which is a Joint Research Center of the European Commission. It is very similar to the other two ground-based SAR systems mentioned previously but differs slightly with regard to the frequency spectrum, that for LiSALab spans the interval from 17.0 to 17.2 GHz [187, 47, 193]. Neither the homepage nor the papers available to the author mention any possibility of operating it in a stationary mode enabling temporally highly frequent measurements.

The three instruments mentioned above are able to operate within ranges of up to 4 km. The precision of deformations derived from SAR measurements is approximately 0.1 mm under laboratory conditions whereas the operation in RAR-mode leads to results an order of magnitude better. For SAR measurements, the instruments are typically installed at a fixed location that stays constant for the measurement campaign's duration. This 'zero baseline condition' guarantees negligibility of topography-induced phases and allows for easy extraction of deformations from the measurements but prevents the derivation of a DEM. The condition can be broken on purpose. Resolutions for all three systems are 0.75 m in range direction and $4.4\,\frac{m}{km} * \mathrm{distance\ to\ radar}$ in cross-range direction. The range resolution is a consequence of operating the radar with a bandwidth of 200 MHz. Nontrivial angular resolution results from processing data gathered by an antenna moved either along a line or any other well known trajectory.

IBIS-ArcSAR: As explained in [110], spatial distributions of amplitudes and phases are extractable from measurements made by an omnidirectional antenna that rotates along a circular trajectory. Prototypes of an instrument of this type featuring continuous wave stepped frequency transmitters and a step motor driving a robotic arm have been developed by IDS and did lead to the commercially available IBIS-ArcSAR system, whose specifications are documented in [101]. The instrument performs similar to the IBIS-FM in practical applications w.r.t. resolution and accuracy but, as [155] records, the antenna's more complex motion pattern induces an elevation-dependent error term that results in a non-negligible defocusing not encountered in the rail-based systems.

GPRI: The Gamma Portable Radar Interferometer (GPRI) is a terrestrial radar interferometer with real aperture [191]. It can be deployed for deformation measurements as well as temporally dense monitoring of profiles and the generation of DEMs. For completion of the latter task, a nonzero baseline is required. For the GPRI this baseline is implemented in form of two separate receiver antennas mounted on a rotating tower that is fixed atop a tripod. The emitted radiation is concentrated in a certain direction orthogonal to the main axis of the antennas, which is why the instrument tower holding transmission and receiver antennas needs to be set into rotary motion to acquire an SLC. The acquisition time for one complete 180 degree acquisition is 30 seconds for instruments of the second generation.

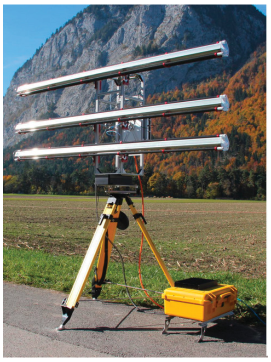


Figure 5.5: The GPRI system (Image: AP Swiss).

SSR: Manufactured by the Australian company Groundprobe and designed specifically for slope stability assessment, this system features a parabolic antenna that is rotatable horizontally and vertically. Its pencil beam is moved over the slope to sample it at regular intervals. With measured amplitude and phase as well as the time of flight and given the angles, a digital elevation model can be reconstructed from which regions of particular interest to the risk assessment might be extracted and automatically monitored [88]. The sampling method is in itself similar to the one used by laserscanners and especially in comparison to competing systems can be interpreted as less areal and more pointwise.

One advantage of the RAR approach over the SAR-systems is the suppression of effects induced by moving machinery necessary for day-to-day mining business. Due to the radar wave not being sharply focused in SAR, strongly reflective material can have an influence even on neighboring resolution cells [91, p.42] and these so-called sidelobes can lead to false alarms.

5.1.2 Typical processing chain

The minimum requirement for assessing surface deformations are unwrapped interferograms. These are available only after a sequence of post-processing steps that takes as input an ordered set of SLCs and converts them to interferograms via multilooking and complex conjugation. A global optimization routine then searches for that spatial distribution of phase differences which is simultaneously well explainable by the observed interferogram and has a low amount of large phase gradients between neighboring pixels [72]. Several algorithms exist, some of them easily interpretable in an RKHS framework and therefore especially suitable as a basis for explanations. Other interferometric byproducts exist that are not of immediate relevance in the above pipeline but provide either helpful estimations of a measurement's reliability or further insight into an observed phenomenon's dynamics. The information compiled in this section is well known generally but its presentation is scattered in the literature due to textbooks focusing mostly on spaceborne SAR. Nonetheless, almost everything apart from a few computational details may be found for example in [91]. [1]

SLCs and MLIs: Averaging the raw data to improve the signal-to-noise ratio (SNR) leads to SLCs. In each pixel of an SLC z, phases and amplitudes are encoded in the form of a complex number

$$z_{kl} = a + ib = |z_{kl}|e^{i\varphi_{kl}}.$$

where k and l are indices quantifying row and column. Calculating the absolute value of z_{kl} via $|z_{kl}| = \|z_{kl}\|_{\mathbb{C}} = \sqrt{\langle z_{kl}, z_{kl}\rangle_{\mathbb{C}}} = \sqrt{z_{kl}^* z_{kl}}$ and its angle φ_{kl} as $\varphi_{kl} = \operatorname{atan}\left(\mathfrak{Im}(z_{kl})/\mathfrak{Re}(z_{kl})\right)$ from the representation of z_{kl} in the complex plane amounts to extracting amplitude and phase, see figure 5.6. This is more than just a computational sleight of hand; the algebra of complex numbers mimics the behavior of the real backscattering process in some important aspects [91, p. 89]. The result of adding two complex numbers in neighboring pixels corresponds to that complex number, one would have recovered as a result of measuring jointly the surface patches associated to those two pixels. In this sense, addition of complex numbers imitates constructive and destructive interference of backscattered waves; the same cannot be said for example about the alternative candidate procedure of simply adding amplitudes and phases.

This relation can be used to form Multi-Look Images (MLI) which exhibit worse geometric resolution but a superior signal-to-noise ratio. To generate an MLI $z_M \in$

[1] Parts of the material in this subsection are translated excerpts from an unpublished report on results of a measurement campaign that was carried out on initiative of the Swiss office for the environment. The report was compiled together with professor Martin Funk, ETHZ. The excerpts presented here were written by the author of this monograph.

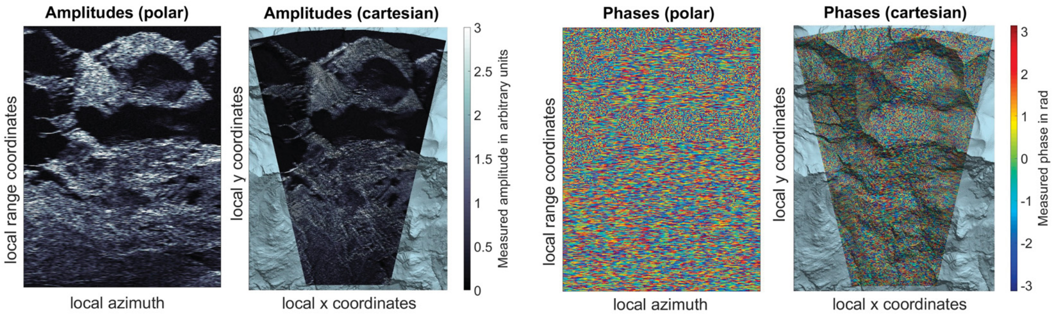


Figure 5.6: The amplitudes and phases are encoded in SLCs as the absolute values and angles of complex numbers. The axis of SLCs correspond to range and cross-range and are indicative of a polar geometry that first has to be transformed into the Cartesian plane or onto a DEM.

$\mathbb{C}^{m_{new}} \otimes \mathbb{C}^{n_{new}}$, one effectively applies a boxcar filter to an SLC z resulting in

$$(z_M)_{kl} = \sum_{o=0}^{n_r-1} \sum_{p=0}^{n_c-1} z_{(k-1)n_r+o+1,(l-1)n_c+p+1} \tag{5.2}$$

for individual entries in the MLI matrix z_M. Here n_r and n_c are the desired numbers of looks in range and cross-range direction and problems pertaining to out-of-bounds indices arising at the border of the SLC are obviously solvable by some convention or demanding $m_{new}n_r \leq m, n_{new}n_c \leq n$. For computational reasons one may reformulate equation 5.2 to make use of linear operators to derive the equivalent equation 5.3 posed in terms of matrix multiplications,

$$z_M = Mz = M_1 \otimes M_2 z = M_1 z M_2^* \tag{5.3}$$

$M_1,$ the m_{new} index matrix for range summation

$M_2,$ the n_{new} index matrix for cross-range summation,

where, as before, the asterisk denotes forming adjoints. $(M_1)_{kl}$ is 1 iff $l \in [(k-1)n_r+1, kn_r]$ and $(M_2)_{kl}$ is 1 iff $l \in [(k-1)n_c+1, kn_c]$. They are 0 otherwise. This is easily implementable and computationally efficient if Multi-looking is to be performed on a batch of SLCs as the matrices M_1,M_2 have to be formed only once upon which they are readily available for the subsequent matrix multiplications.

Interferograms: The interferometric phase φ, given here the symbol φ_{Int} for clarity, is the temporal difference between two spatially corresponding pixels in two different SLCs z_1 and z_2. It can be calculated via pointwise multiplication of the SLC z_1 with the elementwise complex conjugated $\overline{z_2}$ of z_2. The choice of order is inessential as it changes only the phase's sign and and leaves all other characteristics of the complex number untouched although the one presented here seems to be the most widespread. The individual entries of the complex interferogram z_{Int} are

$$\begin{aligned}(z_{\text{Int}})_{kl} = (A_{\text{Int}})_{kl} e_{kl}^{i(\varphi_{\text{Int}})} &= (A_1)_{kl}(A_2)_{kl} e^{i[(\varphi_1)_{kl} - (\varphi_2)_{kl}]} \\ &= (z_1)_{kl}(z_2)_{kl}^* \qquad\qquad =(z_1 \circ \overline{z_2})_{kl}.\end{aligned}$$

Calculating $z_1 \circ \overline{z_2} = z_{\text{Int}} \in \mathbb{C}^m \otimes \mathbb{C}^n$ ($\circ$ is the pointwise Hadamard product, not matrix multiplication) has the advantage that it produces complex numbers that can be coherently added in a second step. As alluded to during the explanation of MLIs, this improves the interferometric phase's signal-to-noise-ratio. Choosing natural numbers n_r and n_c determining the number of looks in range and cross-range direction, one arrives at a formula almost entirely analogous to the one for MLIs. For the interferogram $z_{\text{MInt}} \in \mathbb{C}^{m_{new}} \otimes \mathbb{C}^{n_{new}}$ generated via simultaneous multilooking one finds the components and the whole matrix to have the form

$$(z_{\text{MInt}})_{kl} = \sum_{o=0}^{n_r-1} \sum_{p=0}^{n_c-1} (z_1 \circ \overline{z_2})_{(k-1)n_r+1+o,(l-1)n_c+1+p}$$
$$z_{\text{MInt}} = M_1(z_1 \circ \overline{z_2})M_2^*$$

where M_1 and M_2 are index matrices as defined before. More information can be found in [91, p. 93].

There are still two problems. First, there may be regions in which the interferometric phase seems to be almost completely randomly distributed; those regions are said to exhibit low coherence. The other problem is that the values extractable for φ_{Int} are bound to lie in $[-\pi, \pi]$ even though the true phase change that occured between the acquisition times is $\varphi_{\text{Int}} + k2\pi$ for some $k \in \mathbb{Z}$. As even in the absence of noise one would be able to reconstruct the true interferometric phase only up to a multiple of 2π due to $\varphi_{\text{Int}} = \text{mod}(\varphi_{\text{Int}}^{\text{true}}, 2\pi)$, it is necessary to employ additional information to resolve this ambiguity; a task that is known as unwrapping. Both problems are illustrated in figure 5.7.

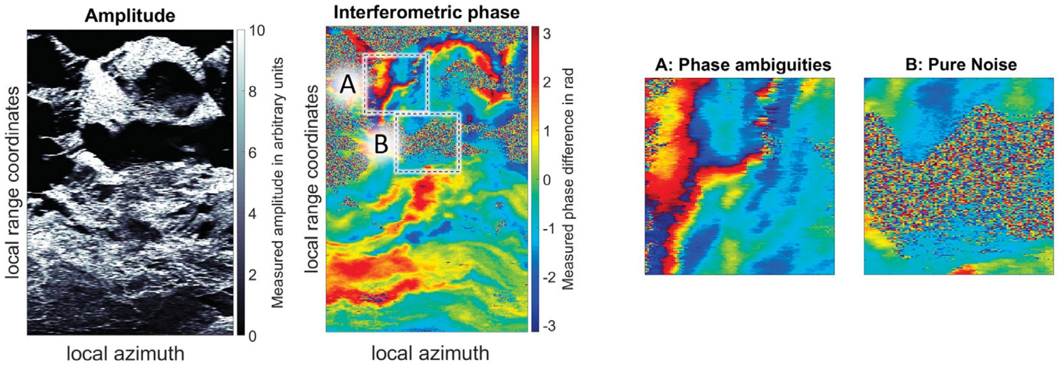


Figure 5.7: The amplitudes and phases of an interferogram. The two panels on the right side are magnifications of the areas outlined in the phase image. They show clear phase jumps (A) of 2π that do not reflect the true underlying phase and are due to the ambiguities in the measurement process. In areas of especially low backscattered intensity, the measured phase can be pure noise (B). Note that the interferometric phase is almost nowhere zero even though the covered scene is composed mostly of stable areas; this is due to the influence of atmospheric effects.

Unwrapped interferograms: It is often possible to resolve the phase ambiguities and the corresponding uncertainties regarding the sign of a pixels motion. This can be seen in figure 5.8 on the right hand side, which exhibits an unwrapped version of the interferogram previously shown in figure 5.7; note the different colorscale.

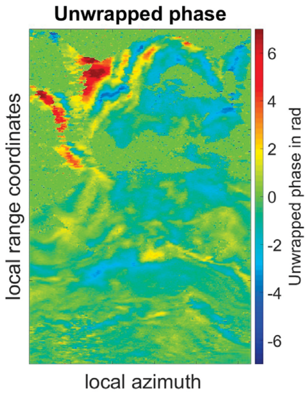


Figure 5.8: The range of the phase value after unwrapping spans multiple ranges of 2π.

Most of the errors in sign and discontinuities are gone in the regions exhibiting good SNR and a spatially coherent interferometric phase. The most common methods for unwrapping search paths between regions separated by phase discontinuities along which the phase gradients vary smoothly. Along those paths, the phase gradient can be integrated consistently meaning that the result of the integration is independent of the specific path taken. An example of this type of method is the branch-cut algorithm first proposed by Goldstein et al. [77]. It is also often coupled with a denoising step that smoothes out the spatial Fourier spectrum [76] and a phase gradient estimation based on a minimum cost flow problem that minimizes a global measure of error based on network programming [41]. The latter method was employed to generate figure 5.8. Even though approaches based on local integration of phase gradients are fast, the choice of integration path is a consistent source of uncertainty and global methods based on the ℓ^p norm minimization

$$\hat{\varphi}_{\text{Unw}} = \underset{\varphi_{\text{Unw}} \in \mathbb{R}^{m\times n}}{\operatorname{argmin}} \|\mathcal{W}\nabla\varphi_{\text{Int}} - \nabla\varphi_{\text{Unw}}\|_p \tag{5.4}$$

have been designed to sidestep these problems [72]. Here ∇ denotes the discrete gradient, φ_{Int} the wrapped phase, $\hat{\varphi}_{\text{Unw}}$ the best guess for the unwrapped phase and $\mathcal{W}$ is the wrapping operator $\mathcal{W}\varphi = \operatorname{mod}(\varphi, 2\pi)$. The p-norm $\|\cdot\|_p$ is understood to act on tuples of functions as $\|(f,g)\|_p^p = \|f\|_p^p + \|g\|_p^p$. For $p = 2$, essentially a least squares estimator results, that is both fast and global but suffers from oversmoothing and spreads residual errors over the whole interferogram instead of concentrating them in isolated spots as should be clear from the discussion surrounding subsection 4.4.1. One may as well formulate the problem of phase unwrapping as an inference problem in a reproducing kernel Hilbert space by posing the following problem.

$$\hat{\varphi}_{\text{Unw}} = \underset{\varphi_{\text{Unw}} \in \mathcal{H}_\varphi}{\operatorname{argmin}} \|\hat{\nabla}\varphi_{\text{Int}} - \nabla\varphi_{\text{Unw}}\|_{\mathcal{H}_N}^2 + \|\varphi_{\text{Unw}}\|_{\mathcal{H}_\varphi}^2 \tag{5.5}$$

Here $\hat{\nabla}\varphi_{\text{Int}}$ is a guess for the phase gradient, $\mathcal{H}_N$ the noise RKHS with inner product

$$\langle f, g\rangle_{\mathcal{H}_N} = \sum_{j=1}^{mn} \frac{f_j g_j}{\sigma_j^2}$$

and σ_j^2 the noise variance as derivable for example from coherence maps. $\mathcal{H}_\varphi$ is the Sobolev-type RKHS with inner product

$$\langle f, g\rangle_{\mathcal{H}_\varphi} = \langle \nabla f, \nabla g\rangle_{\mathcal{H}_K}$$

for some appropriately smooth kernel K. As usual, the first term enforces fidelity to the data whereas the second one regularizes the solution to have apriori likely properties amounting to simultaneous smoothing and denoising during unwrapping.

Interferograms after atmospheric corrections: Even after unwrapping and potentially denoising, the interferogram retains artifacts that do not relate to real deformation and are instead stemming from short- and long-term meteorological changes. Separating the so called atmospheric phase screen (APS) from deformation is still an open research question and the whole of section 5.2 is dedicated to its solution. As this is an ill-posed problem without exact solution and nontrivial ground truth is usually unavailable outside of image regions known to be stable, a multitude of approaches has been developed to deal with this problem.

The APS is temporally and spatially highly variable; timewise it consists of a turbulent part that changes within minutes and a long term periodic trend that operates in the frequency region of $1/\text{day}$. The most basic correction methods consist of simply averaging the interferometric phase in time to low-pass filter out the highly frequent part of the APS. This approach is termed stacking and the results can be seen in figure 5.9. Other approaches focus entirely on the spatial aspects and try to fit parametric or nonparametric models to predict and subtract the APS. Physically motivated simulations have been successfully tested for spaceborne SAR [79] but seem to be infeasible for TRI due to the small-scale nature of the local meteorological fluctuations. All these methods have a probabilistic interpretation that will be made explicit when a RKHS-based framework is introduced to solve the spatiotemporal signal-splitting problem in a stochastically rigorous setting in section 5.2

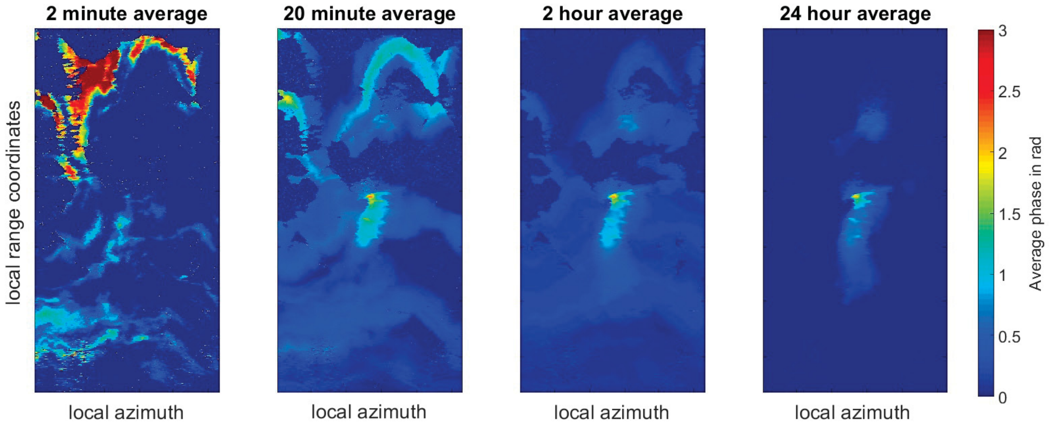


Figure 5.9: The longer one averages in time, the less prominent atmospheric effects induced by turbulent meteorological changes become. However, it is typically not possible to reduce them to zero and long term changes in temperature and humidity prevent convergence of the sequence of averages to the true underlying deformations.

Coherence images: Two SLCs describing the same real-world surface patch at different times are not necessarily sensibly comparable. Specifically this is the case when the surface under consideration has undergone significant transformations that have altered its reflective properties drastically. Changes in humidity and water content or the random movements of vegetation due to wind can therefore already deteriorate the comparability of two image regions [91, p. 98],[15]. The coherence $\gamma, |\gamma| \in [0, 1]$ is a measure for the systematicity of joint phase variations in two image regions and is used to exclude unsuitably incoherent and noise-dominated areas from further processing. It is a function of the expected value $E[\cdot]$ of the interferometric phase φ_{Int} which is estimated by considering all values in a moving window. If z_1 and z_2 denote two complex-valued random variables, whose phase difference forms the interferometric phase to be evaluated w.r.t. its coherence γ, then

$$\gamma = \frac{E[z_1 z_2^*]}{\sqrt{E[|z_1|^2]E[|z_2|^2]}} \qquad \gamma \in \mathbb{C}. \tag{5.6}$$

Several convenient properties carry over from the calculation of the real correlation coefficient to the complex coherence unperturbedly. Note for example, that for two complex random variables z_1 and z_2, if $z_2 = cz_1$, then $E[|z_2|^2] = |c|^2 E[|z_1|^2], E[z_1 z_2^*] = c^* E[|z_1|^2]$ and consequently

$$\gamma = \frac{E[z_1 z_2^*]}{\sqrt{E[|z_1|^2]E[|z_2|^2]}} = \frac{c^* E[|z_1|^2]}{cE[|z_1|^2]} = c^* c^{-1} \tag{5.7}$$

has modulus $|c^* c^{-1}| = |ae^{-i\varphi}[ae^{i\varphi}]^{-1}| = |e^{-2i\varphi}| = 1$ just as in the purely real case. Therefore a coherence with absolute value close to 1 indicates linear relationships between the random variables, although the coefficient relating both may be complex and encode not only scaling but phase shifts as well.

Coherence is high, if neighboring pixels behave similarly. This is typically indicative of the existence of reliable backscatterers. Often one does not distinguish between the complex coherence γ and its modulus $|\gamma|$. Usually no confusion arises between the two and we will follow this convention; to avoid ambiguities we will explicitly mention the coherence being complex when necessary. In figure 5.10 a coherence estimate $|\hat{\gamma}|$ on the basis of a two minute interferogram can be seen. Regions lying in the radar shadow are dominated by noise and therefore black, regions featuring vegetation are still relatively noisy and are assigned a grayish color (bottom parts). Areas containing objects that scatter incoming waves in a consistent manner are brightly colored and indicate high quality of the interferometric phase.

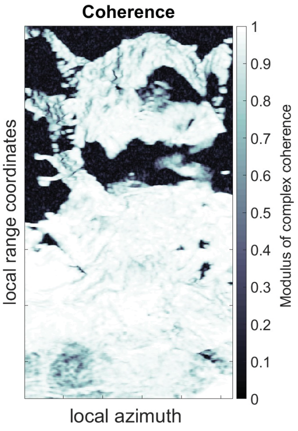


Figure 5.10: The coherence image provides information about phase quality.

Under certain simplifying assumptions, the coherence is the only quantity necessary to compute the variance of the interferometric phase and closed form expressions exist for $\sigma^2_{\varphi_{\text{Int}}}(\gamma)$ [14]. Even more, it is shown in [15] that under the assumption of finite and identical signal-to-noise ratios SNR in the two SLCs forming an interferogram and supposing an absence of changes of the imaging process itself,

$$|\gamma| = \frac{\text{SNR}}{\text{SNR}+1} \qquad \text{or equivalently} \qquad \text{SNR} = \frac{|\gamma|}{1-|\gamma|}.$$

This shows that the coherence is useful to examine the interferometric phase's reliability and by providing covariance matrices helps to define inner products in RKHS for which norm $\|\cdot\|_{\mathcal{H}_N}$ and likelihood of phase-noise are in one-to-one correspondence. Coherence is estimated via a moving window approach and an efficient implementation based on matrix multiplication is possible similarly to what was presented in the previous explanations, see page 256 for more details. Points exhibiting especially reliable [2] phase information are often called persistent scatterers (PS); see page 264 for more detailed procedures on how to determine them.

Derived products: The unwrapped and corrected interferograms featuring primarily deformation rates need to be interpreted in the light of the initial monitoring task. Aggregating information to aid decisions is not part of the generic radar interferometric set of methods and needs to be done in a way tailored to the specific task at hand. TRI data are particularly suitable for a subsequent time series analysis of individual pixels or whole regions. Due to the data being in the form of a sequence of matrices, tools from computer vision like optical flow or segmentation are as easily applied as extraction of statistical moments or special features like the coherence that are indicative of the surveyed surface's state.

5.1.3 Practical applications and limitations

Its ability to autonomously measure deformations over wide areas and with high temporal repeatability has lead to TRI being deployed primarily for purposes of monitoring medium- to large-scale natural phenomena like landslides, potential icefalls and volcanic activity as well as structural health of geotechnical constructions in the mining industry. Typically, the results of TRI measurements are then temporally dense, two-dimensional spatial maps quantifying deformation velocities and time series derived from these maps. Most instruments listed in subsection 5.1.1 can also be operated in a way that suppresses the formation of any angular resolution enabling them to perform measurements of a one-dimensional slice of a structure of interest directly in their field of view with very high frequency. Even though the practical performance of TRI is often satisfactory, certain theoretical problems recurrently appear in almost all applications. Most prominently, these are the effects induced by atmospheric changes between two acquisitions, unknown stochastic properties of the data, and a lack of interpretability stemming from an unclear relationship between measured changes in phase and underlying changes of the topography.

Since terrestrial radar interferometers are suitable for surveying and monitoring of various objects with differing characteristics and behavior, also the mathematical models representing those objects are different. It is common to perform complementary measurements with instruments other than TRI to derive the parametersets necessary to harmonize observations and models of either objects or measurement processes. We will call such a problem-specific strategy determining the mode of

[2] Reliability here is not meant in the classical engineering geodetic sense of Baarda [12] but in the more colloquial one implying high quality and low proportions of noise

gathering information a monitoring concept and explain its components for a few selected representative examples encountered in the current literature.

For one, there are **validation measurements** to verify the correctness of data provided by TRI. Apart from their use for detecting errors — particularly important in the early stages of instruments as young as TRIs — they provide additional value in the sense that the coordinates they deliver can serve as reference points for phase unwrapping. The latter is a problem of integrability and as such depends on the differentiability of the interferometric phase, which is why local decorrelations can negatively impact the success probability of extracting from the interferometric phase an absolute phase difference that is independent of the path of integration. Pixels in the interferogram with unambiguously identifiable deformations are particularly suitable reference points for unwrapping and as such can diminish the potential for contradictions between several separately unwrapped image regions. The ground truth derived in this way further helps to quantify deviations and artifacts encountered in TRI data and supports the development of stochastic models that enable statistical inference and uncertainty quantification also for derived products like deformation rates and failure probabilities.

Instead of investing time and financial resources into gathering data that are completely redundant in the worst case, it often seems more desirable to conduct **complementary measurements** to gather the information necessary to apply corrections and uniquely determine model parameters. Compiling meteorological data via thermometer, barometer and hygrometer to estimate and subsequently subtract the atmospheric phase screen [165] falls into this category, as does the deployment of ultrasound sensors and stream gauges for quantification of the phase induced by melting snow and other meteorological phenomena unrelated to kinematic changes of the ground [163]. Gathering physical and geometrical properties for a joint processing with TRI data seems to be fairly widespread and is implemented in different variants. In the framework of already existing projects, the following instruments and surveying principles have been used in conjunction with TRI:

- Terrestrial laser scanners (TLS) to generate three-dimensional models of steep surfaces with the intention of georeferencing the deformations extracted from TRI data [194].
- Spaceborne SAR interferometry to offer a different, independent line of sight and in combination with TRI-data the possibility to derive two dimensional vectors to quantify ground motions.
- GNSS to monitor the movement of particularly important points in a geodetic network that cannot reliably be surveyed by TRI.
- Geological field measurements, infrared tomography and thermochemical devices to infer information about composition of and processes in geological formations [134].
- Measurements of mechanical tension for structural health monitoring and inclinometers, hydraulic as well as borehole measurements and geoelectrics for

geotechnical monitorings [29].

- Documentation of human construction together with measurements of precipitation to find a causal connection between particularly rapid deformations and their triggers [29].

As far as possible and accessible, theoretical geotechnical models have increasingly been tied into the monitoring concepts; see for example [194]. In that source, a finite difference model has been designed to model the dynamics of a landslide that occurred in the province of Belluno, Italy. It has been changed regularly based on TRI data to minimize the contradictions between predicted and observed movements. Simple prediction procedures based solely on deformation measurements have thereby been replaced by a more algorithmic approach in which that physical model is derived that is most accurately explaining the observations. It is then used for further predictions. Apart from this inclusion of models, in practice one often encounters activities of a supporting nature, e.g. the installation of corner reflectors to increase partially the SNR in natural scenes that otherwise exhibit only a low proportion of stable backscatterers [152].

The possibilities for application of and practical experiences with TRI are the subject of disproportionately many articles owing probably to TRI still being considered as a technique in the experimental stage. While deformation monitoring of geometrical surface changes is still the primary area of application, the dependence of the interferometric phase on material properties of the reflecting surfaces has received attention sporadically as well and found its use in vegetation and land cover estimation. Since there are by now several companies originating from academic institutions dealing with monitoring of geohazards via TRI, it is to be expected that TRI will be deployed for a wider spectrum of tasks going beyond what it was originally designed for. Eight representative applications are listed below including their results and further references.

1. **Landslides:** Each of the systems previously presented is used for monitoring potential landslide areas. While for the SSR, these are primarily the technically created slopes of open pit mines [137] and for FastGBSAR these are dykes [167], GPRI, IBIS and LiSALab have been used for monitoring of a diverse set of regions with varying characteristics, see e.g. [194], [34], [47]. Publications are focused particularly on comparisons between data gathered with classical geodetic measurements and TRI.

2. **Monitoring of dykes:** Metasensings FastGBSAR was tested in August 2012 in an area specifically designed to simulate dyke failure and enable early warning. The result of these measurements were a sequence of time series that describe the kinematic progress of the artificially induced formation of crevasses [167].

3. **Early warning against volcanic geohazards:** IBIS-L has been deployed within the frame of the project ’Exupery’ under the auspices of the German program ’Geotechnologien’ as part of a monitoring system to enable early warning and rapid response to volcanic activity [166]. After creation of an artificial spatial baseline, the ground based SAR was able to generate a DEM and deliver a common reference system allowing inclusion of other observations and the geometric components of the monitoring scheme, that furthermore featured temperature measurements, seismic recordings and chemical analysis of gaseous emissions. For further information on volcanic monitoring with TRI and their integration with geomorphological surveys, consult [105].

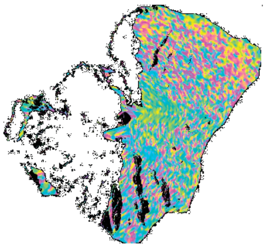


Figure 5.11: TRI was used for deformation monitoring of Stromboli. Image: [6]

4. **Glacial monitoring:** Several articles were published by or in cooperation with expert staff from Swiss Gamma Remote Sensing AG on the topic of glacial monitoring. These publications showcase functionality and practical applicability of their in-house development ’GPRI’ in the context of real-world measurement campaigns including outlet glaciers in Greenland [201] as well as several ones in Switzerland. The results indicate fast decorrelations induced by melting processes and atmospheric effects as the main hindrances preventing reliable deformation data, see [163] and [191] .

5. **Cliff stability monitoring:** Martino and Mazzanti [134] used IBIS-L to supplement TLS and infrared tomography thereby deriving statements about cliff stabilities on Mount Pucci, Italy. The GBSAR extracted local motions and the results were embedded into a framework that also employed geotechnical and continuum mechanical calculations. The authors identify as problems of this in principle successfully approach especially foreshortening effects and an almost invariably insufficient set of possible instrument positions suitable for cliff monitoring.

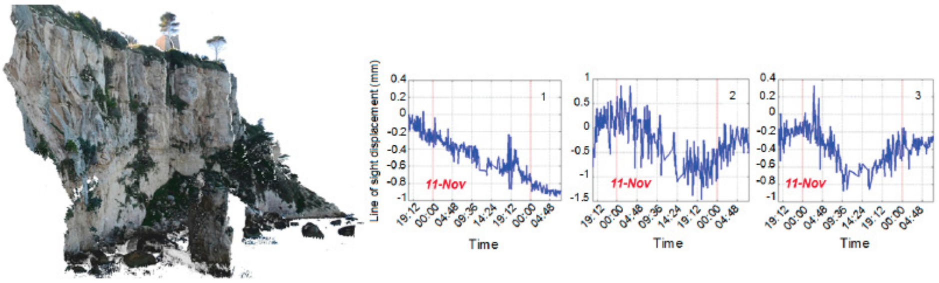


Figure 5.12: TLS and TRI can be combined to visualize three-dimensional deformations. The image is from [134].

6. **Monitoring of snow depth:** For the case of mainly dry snow, accumulation of snow induces changes of a region's topography that can be sensed almost analogously to surface deformations. Early warning of persons and localities potentially threatened by avalanches by means of TRI was tackled in [130], however the measurements to gauge snow-depth changes via TRI were delivering values that were at best moderately close to the ground truth as influences of material properties — especially those of wet snow — on penetration depth and phase retardation have not been fully understood yet.
7. **Monitoring of infrastructure and historic building fabric:** The leaflets of Nhazca S.r.l. and Metasensing list several projects related to man-made structures. Among them are analysis of long- and short-term dynamics of railroad bridges under different load configurations, of hydroelectric dams and of road beds and road superstructures. School buildings, high-rise buildings, towers, pillars, power plants, storage tanks and interim storages are mentioned as well as examples of objects potentially suitable for monitoring with GBSAR, especially in the RAR mode of operation. A specific application detailing concept, execution and results of monitoring a bridge is described in [122].
8. **Digital elevation models for visualization purposes:** Gamma Remote Sensing AG published a paper, in which the process of estimating a digital elevation model of a glacier from GPRI data was described in detail. In comparison to the official swisstopo DHM25 model provided by national authorities in 1992, height differences of approximately 40 m were recorded that may be traced back to melting processes and glacial recession [191]. The standard deviation of the difference between interferometrically generated height model and the one provided by governmental institutions was around 3.4 m at stable locations unaffected by glacial phenomena.

There exists an extensive supply of commercial and openly accessible software for processing and evaluation of data gathered via satellite radar interferometry. The main providers of open source software and freeware in that area are either public research institutions like ESA, NASA, and TU Delft or collaborations of motivated individuals. As an example of software originating from the latter constellation, one may notice RAT (RAdar Tools); developed under the supervision of and with contributions from Andreas Reigber, who codeveloped the software and augmented it with E-learning tutorials while working at DLR [115].

Commercial packages are offered by suppliers of programs for professional visualization of geodata or industry-oriented companies that sell hardware and monitoring services. The collections of programs typically offer the full functionality needed to process raw data all the way up to differential interferograms. Programs that go beyond this and support the user by assisting with examination and interpretation of the data seem to be implemented only seldomly. Another example is GMTSAR, which offers access to a preprocessor for every type of satellite currently in operation, an InSAR processor for coregistration, interferogram calculation and derivation of the topographic phase. A postprocessor enables coherence estimation, complex filtering, the plotting of displacement maps, georeferencing, and provides a set of scripts for two-pass interferometry and time series analysis [171].

For TRI, however, the author failed in identifying openly accessible software. This may have at least two reasons.

1. The necessity for a concerted effort to construct and operate TRIs is significantly smaller than the one for successful completion of a satellite mission. Consequently, in TRI there is no large publicly subsidized institution that considers itself responsible for developing and sharing high-quality program packages. Market shares are split between small private providers whose primary goal of generating revenue conflicts with making openly accessible the software that is often sold in conjunction with their hardware.
2. Apart from conversion of data in proprietary and company specific formats to SLCs and other accepted standard products, further processing can be done analogously to the spaceborne case. In principle, it is therefore possible to perform further processing steps with already existing software; the exactly known spatial baseline and a lack of any topographic phase even imply a significant simplification.

The analysis of TRI data has received a surprisingly small amount of attention up until now and is done essentially by visual inspection. Even though TRI has the potential to provide measurements with high spatial and temporal resolution and thereby acquire a large set of mutually correlated time series, papers rarely deal with TRI data from this perspective. In a 40 month long investigation of an area subject to frequently occurring landslides, Mazzanti et al. were able to identify typical patterns indicative of onsetting landslide activity. Exploiting this knowledge allowed to diagnose creep movements of soil and connect the time immediately preceding an event to a decrease in surface acceleration [29]. Presupposing a power law for velocity and acceleration during slope motion, they were able to successfully predict 8 out of 10 landslides with temporal deviations of less than 2 hours via back-analysis. Manually tuning a finite-impulse-response filter further increased the prediction accuracy. Failure cases dominated by strongly nonlinear creep processes still evaded any kind of reliable prognosis.

A different group of researchers considers the sensitivity of the interferometric phase to meteorological conditions as a disadvantage that might be mitigated by applying methodology from computer vision to the intensity images. Crosetto et al. utilize methods for sub-pixel exact image registration to examine the geometric content of a sequence of amplitude images. They claim that, unlike the normal interferometric procedure, their approach is not adversely affected by aliasing, atmospheric influences and is not restricted to inferring motion only in line-of-sight [44]. Obvious disadvantages are the need for corner reflectors in case of insufficient amounts of natural backscatterers of high quality and a cross range sensitivity that is linearly decreasing with distance.

The main obstruction to extracting true deformations from sequences of interferograms are the atmospheric effects that affect the interferometric phase φ_{Int} at every pixel and form an APS overlaying all of the image [157]. If no corrections are applied to the original measurements, the results of data processing are systematically wrong due to the effects of air pressure, air moisture and temperature [166]. Extensive and spatially low-frequent atmospheric effects can entirely mask the deformations [152]. Up to first order, they are proportional to the lengths of the propagation path and can lead to errors of magnitudes up to $1\ \mathrm{mm}/\mathrm{km}$ if the partial pressure of water vapor changes only slightly from $0\ \mathrm{mbar}$ to $0.2\ \mathrm{mbar}$ with temperature and total pressure as in the definition of the standard atmosphere and remaining constant. To the best of the author's knowledge, estimation and subtraction of the APS is the most serious non-profane limitation faced in practice and no reliable solution accounting for the highly complicated spatiotemporal structure of the APS has been proposed as of yet.

5.2 Mitigation of atmospheric effects

The preceding section concluded with the claim that atmospheric artifacts (in form of the APS) are the main obstruction in TRI. A brief investigation into the reason for their occurrence will reveal certain characteristics that imply the APS to be inaccessible to modelling via physical models. A closer examination of real-world data will validate these suspicions and provides a first opportunity to design and test stochastic models for the APS and noise components of interferograms. It is then possible to develop different optimization problems in RKHS, whose solutions coincide with estimators for deformation, APS and noise respectively. As soon as this is done, the machinery assembled in chapters 3 and 4 directly delivers efficient procedures to construct and evaluate these spatiotemporal splines. Due to the approach being fairly abstract, an analysis of the estimator's behavior in several special situations will aid the tangibility of its results; the section closes with a reinterpretation of commonly used APS correction methods in the RKHS framework allowing the derivation of the hidden stochastic assumptions underlying them.

5.2.1 The atmospheric phase screen

The theory of electromagnetic waves provides well-known formulas relating attenuation, phase delay, and trajectories of propagating waves to material properties of the medium they travel in. These physical models have been extensively studied for classical geodetic measurements; they explicitly depend on temporally and spatially highly variable meteorological quantities. This indicates that forward modelling of the APS based on an understanding of atmospheric dynamics supplemented by measurements of meteorological parameters is unfeasible and needs to be replaced by a data-driven approach. This conclusion is supported by the theoretical considerations developed by researchers primarily involved with the mechanics of turbulent fluids, who have proposed early on to employ statistical tools for handling such ill-posed problems [142]. Nonetheless, the physical measurement process has certain implications even for the statistical model that restricts the potential behavior and as such is useful as prior knowledge.

The propagation of radar waves is governed by Maxwell's equations, which are a set of four differential or integral equations linking the behavior of the electric field $E : \mathbb{R}^4 \to \mathbb{R}^3$ and the magnetic field $B : \mathbb{R}^4 \to \mathbb{R}^3$. In differential form and in absence of any matter or electric charges, they are [66, p. 3]

$$\langle \nabla, E \rangle = 0 \qquad \nabla \times E = -\mu_0 \partial_t B \tag{5.8}$$

$$\langle \nabla, B \rangle = 0 \qquad \nabla \times B = -\epsilon_0 \partial_t E \tag{5.9}$$

where μ_0, ϵ_0 are the permeability and permittivity of the vacuum. As is the usual notational custom in physics, the del operator $\nabla = [\partial_x, \partial_y, \partial_z]^T$ is treated as a normal vector subjectable to inner products $\langle \cdot, \cdot \rangle : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$ and cross products $\cdot \times \cdot : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ and whose components interact with any function by forming its derivative. Whereas the left hand side equations in 5.8 and 5.9 claim that electric and magnetic fields are divergenceless, the right hand side equations imply that a changing electric field induces a magnetic field with nontrivial curl and vice versa. Once initiated, the resulting chain of interlocking electric and magnetic fields propagates through the medium [175, p. 137]; this is then called an electromagnetic wave. By the usual identities for curl and divergence [32, p. 161] one finds

$$\begin{aligned} \nabla \times \nabla \times E &= -\mu_0 \nabla \times \partial_t B && = -\mu_0 \epsilon_0 \partial_t^2 E \\ \nabla \times \nabla \times B &= \epsilon_0 \nabla \times \partial_t E && = -\mu_0 \epsilon_0 \partial_t^2 B. \end{aligned}$$

Since $\nabla \times \nabla \times g = \nabla \langle \nabla, g \rangle - \Delta g$, it is clear from Maxwell's equation that both E and B satisfy the wave equation $\sum_{k=1}^{3} \partial_{x_k}^2 g - \mu_0 \epsilon_0 \partial_t^2 g = 0$. The same equation is rewritten slightly with the help of the negative semidefinite self-adjoint unbounded operator Δ, the componentwise Laplace operator, as

$$\partial_t^2 g = c^2 \Delta g \qquad c = (\mu_0 \epsilon_0)^{-1/2} \tag{5.10}$$

where c is the speed of light in vacuum [51, p. 250]. Intuitively the wave equation 5.10 can be solved via a modification of the functional calculus defined in subsection 2.2.2 by constructing the positive semidefinite square root $\sqrt{-\Delta}$ of the negative Laplacian. For this, note that $g(t, x) = \exp\left(t \frac{i}{\sqrt{\mu_0 \epsilon_0}} \sqrt{-\Delta}\right) g_0$ satisfies the wave equation, and real valued solutions to the problem

$$\begin{aligned} \partial_t^2 g &= \Delta g \\ g(0, x) &= g_0 \qquad \partial_t g_0 = g_0' \end{aligned}$$

can be found by taking the real part resulting in terms involving sines and cosines of the operator $\sqrt{-\Delta}$. More information on the explicit construction of wave operators and solutions to second order partial differential equations of hyperbolic type and their dependence on the initial conditions can be found in [53, pp. 377-397] and [99, pp. 267-272]. By renaming $h = [g, \partial_t g]^T, A = [0, I; \Delta, 0], h_0 = [g_0, g_0']^T$, the problem may also be written as the abstract Cauchy problem

$$\begin{aligned}\partial_t h &= Ah\\ h(0,x) &= h_0\end{aligned}$$

and subsequently solved with the help of the one-parameter semigroup generated by A leading to $h = \exp(tA)h_0$ [154, pp. 219-222]. The simplest solutions to the scalar versions of the wave equation are plane waves which one may represent as the real part of the complex exponential

$$g = a\exp\left(i\left(\langle k, x\rangle_{\mathbb{R}^3} - \omega t + \varphi_0\right)\right) = z\exp\left(i\left(\langle k, x\rangle_{\mathbb{R}^3} - \omega t\right)\right) \tag{5.11}$$

for angular frequency $\omega = 2\pi f$ and some complex $z \in \mathbb{C}$ that encodes both the phase shift φ_0 and the amplitude a. The vector $k \in \mathbb{R}^3$ is called the wave vector and determines the main direction of propagation; planes of constant phase are exactly those orthogonal to k as for any $x^\perp \in \{k\}^\perp$ one has $\langle k, x + x^\perp\rangle - \omega t = \langle k, x\rangle - \omega t$.

If during the propagation of an electromagnetic wave the medium's material properties change, then the wave's frequency $f = [2\pi]^{-1}\omega$ stays constant whereas the wavelength λ changes in a way that guarantees that $\lambda f = v$ where $v = (\epsilon\mu)^{-1/2}$ is the propagation speed in a medium with permittivity ϵ and permeability μ [66, p. 4]. The ratio $n = c/v$ of speed of propagation in vacuum to speed of propagation in the medium is called the index of refraction and for the wavelength the relation in equation 5.12 holds.

$$\lambda = \frac{v}{f} = \frac{1}{f\sqrt{\mu\epsilon}} = \frac{c}{nf} \tag{5.12}$$

Basic wave dynamics are illustrated in figure 5.13. Furthermore, at the boundaries (both fuzzy and hard) between regions of different indices of refraction, at least two effects can occur.

1. **Refraction:** According to Fermat's principle, light interpreted as a particle takes those trajectories which lead to extrema in the optical path length $l_{\text{optic}} = \int_{s_0}^{s_1} n(s)ds$. Snell's law follows and any change of refractive index implies the the possibility of a small deflection of the propagation direction [51, p. 321].

2. **Reflection:** From Huygens principle one may derive the reflection coefficients describing the energy ratio R of incident and reflected light which is

$$R = \left|\frac{a_{\text{reflected}}}{a_{\text{incident}}}\right|^2 = \left|\frac{n_1 - n_2}{n_1 + n_2}\right|^2$$

irrespective of polarization [66, pp. 43-45] for the case of a planar wave interacting with a planar boundary whose normal is parallel to the wavevector. At the interface, reflection occurs that is stronger if the gradient of n is bigger. The reflections can constructively interfere and form clear-air echos if the reflective structures have dominant spatial frequencies of order $\lambda/2$ [54, pp. 13-15].

Refraction changes the trajectory of the path taken by light and as such primarily affects the phase via bending of the ray and slight differences in path length and acquisition geometry. The influence is demonstrated to be negligible in [17]. Reflection leads to spurious signals especially in areas purportedly hidden in radar shadow. It is responsible for deteriorating SNR in regions already exhibiting unfavorable conditions. In what follows, we will largely ignore these two effects or consider them as subsumed under the general random field model proposed later on.

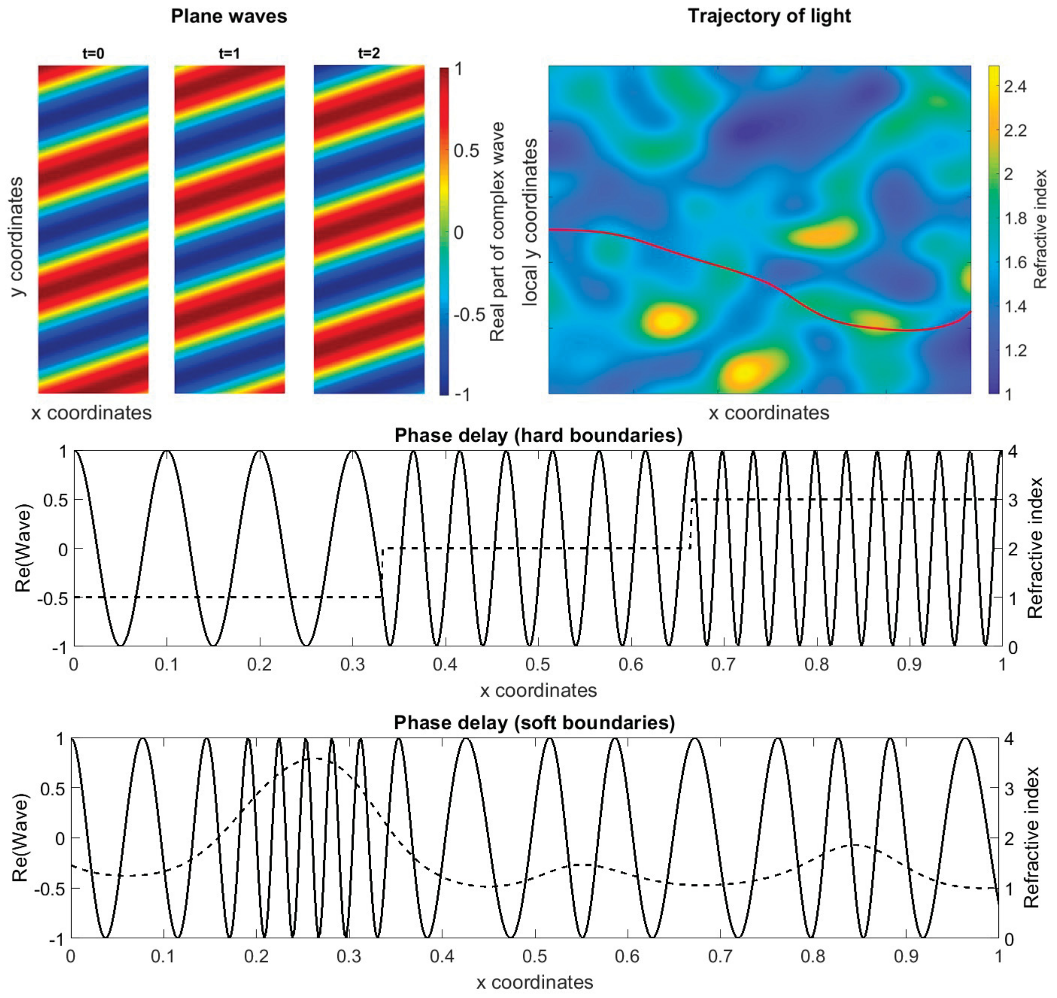


Figure 5.13: The top left panels show the evolution of a free 2D plane wave over time t. On the top right the influence of the index of refraction n on the propagation of a light ray is highlighted in red. The curved trajectory is calculated by a differential version of Snell's theorem that at each timestep estimates the curvature of a single ray on the basis of the refraction occurring at a plane oriented normal to the gradient of the continuous index of refraction. The lower two plots illustrate the wavelengths dependence on n, which is plotted as a dashed line. Note that in these plots the gradients in n are unrealistically high to yield deflections and phase delays visible to the naked eye.

Figure 5.13 indicates that a change in refractive properties of a medium induces a change in phase between transmitted and received signal. If the angular frequency

ω stays constant and the time of travel t increases as measured by the optical path length s_{optic}, one may write a time increment Δt as $(\Delta t/\Delta s)\Delta s = (\Delta s/\Delta t)^{-1}\Delta s$ which in the limit becomes $[v(s)]^{-1}ds$. Then

$$t = \int_{s_0}^{s_1} \frac{1}{v(s)} ds = \frac{1}{c}\int_{s_0}^{s_1} \frac{c}{v(s)} ds = \frac{1}{c}\int_{s_0}^{s_1} n(s) ds = \frac{s_{\text{optic}}}{c} \tag{5.13}$$

for s the spatial coordinate. Scaling up the path length locally by multiplying it with $n(s)$ is therefore equivalent to taking into consideration the reduced propagation velocity $v(s)$. A plane wave g with angular frequency ω, wave vector k, phase shift φ_0, and amplitude a starting at s_0 is consequently found in state

$$g(s,t) = \exp\left(i\varphi(s,t)\right) = \exp\left(i\left(\langle s_{\text{optic}}, k\rangle - \omega t + \varphi_0\right)\right) \tag{5.14}$$

at position s and time t as implied by equation 5.11. Comparing the phase terms $\varphi_{\text{received}}$ and $\varphi_{\text{transmitted}}$ of a wave, reflected at geometrical distance s and measured at the origin $s_0 = 0$, at a certain instant of time t yields

$$\phi_t = \varphi_{\text{received}} - \varphi_{\text{transmitted}} = \varphi(t, 2s) - \varphi(t, 0) = 2ks_{\text{optic}}$$

This only holds, however, if the optical path length stays constant in time. In TRI this is typically not the case and measurements taken at different times t_1, t_2 exhibit different phase values ϕ_{t_1}, ϕ_{t_2} due to changes in optical wavelength induced by atmospheric dynamics. As the values ϕ_{t_1} and ϕ_{t_2} are recorded in the SLCs as phases, the interferometric phase φ_{Int} depends on time via changes of refractive index even in complete absence of any deformation according to

$$\begin{aligned}\varphi_{\text{Int}} = \phi_{t_2} - \phi_{t_1} &= 2k\left[s_{\text{optic}}(s,t_2) - s_{\text{optic}}(s,t_1)\right] \\ &= 2k\int_{s_0}^{s} \left[n(\tilde{s},t_2) - n(\tilde{s},t_1)\right] \\ &= 2k\int_{s_0}^{s} \Delta n(\tilde{s}) d\tilde{s} \end{aligned} \tag{5.15}$$

where $\Delta n(\tilde{s})$ is the difference of indices of refraction at position $\tilde{s}$ between times t_2 and t_1. For general radio meteorological use covering a broad range of standard atmosphere conditions and radio frequencies of up to 30 GHz, Smith and Weintraub [183] propose the two term equation 5.16 linking scaled up refractivity $N = 10^6(n-1)$, temperature T, partial pressure e of water vapor, and total pressure p.

$$N = \frac{77.6}{T}\left(p + 4.81 * 10^3 \frac{e}{T}\right) \tag{5.16}$$

T : Absolute temperature in K

p : Total pressure $p_{\text{dry}} + e$ in mbar

e : Partial pressure of water vapor in mbar

The refractivities N predicted with this model align closely with ones derived in a later metastudy [169, eq 9] ($< 0.15\%$ deviation for $T \in [250, 310]$ K, $e \in [0, 100]$ mbar, predicted N ranging between 245 and 895) . Choosing $T = 288$ K, $p = 1013.25$ mbar and $e = 0$ mbar produces $N \approx 273$ for the standard atmosphere. A TRI measurement extending over a distance of $s = 1$ km through standard atmosphere leads to a phase

$$\phi = 2k \int_{0 \text{ km}}^{1 \text{ km}} n(\tilde{s}) d\tilde{s} = \frac{4\pi}{\lambda} \left(1 + 273.0146 * 10^{-6}\right) * 1000, \qquad \lambda \approx 17.6 \text{ mm}$$

whose wrapped value $\mod(\phi, 2\pi) \approx 2.44$ rad is associated to the corresponding pixel in the SLC. Changing the temperature by 1 K or the partial pressure of water vapor to 0.2 mbar along the whole propagation path already changes the observed phase by 0.7 rad $\cong$ 1 mm. For the phases ϕ_{t_1}, ϕ_{t_2} at different times, it is possible to rearrange equation 5.15 to

$$\begin{aligned} \phi_{t_2} - \phi_{t_1} = 2k \int_{s_0}^{s} \Delta n(\tilde{s}) d\tilde{s} &= 2k|s - s_0| \int_{s_0}^{s} \frac{\Delta n(\tilde{s})}{|s - s_0|} d\tilde{s} \\ &= 2k|s - s_0| \overline{\Delta n} \end{aligned}$$

with $\overline{\Delta n}$ the mean value of difference in refractive indices along the LOS at both epochs. Operating on the unwrapped interferometric phase $\varphi_{\text{Int}} = \phi_{t_2} - \phi_{t_1}$ only and ignoring the effects of the wrapping operator, variance propagation can be used to derive a first approximation to interferometric phase variability dependent on the variability of meteorological quantities. For the sake of argument, assume that the temperature and water vapor are the same everywhere in space and just vary temporally and that no variations occur in the total pressure. Presupposing standard atmosphere during the first epoch and denoting by c_1 and c_2 the constants appearing in equation 5.16, the relationship is

$$\begin{aligned} \overline{\Delta n} &= n_{\text{standard}} - \left[1 + 10^{-6} \frac{c_1}{T} \left(p + c_2 \frac{e}{T}\right)\right] \\ \sigma_{\overline{\Delta n}} &= 10^{-6} \sqrt{\left(\frac{\partial}{\partial T} \overline{\Delta n}\right)^2 \sigma_T^2 + \left(\frac{\partial}{\partial e} \overline{\Delta n}\right)^2 \sigma_e^2} \,|_{T,e=\text{standard atmosphere}} \\ &= 10^{-6} \sqrt{\left[\frac{c_1}{T^2} \left(p + 2c_2 \frac{e}{T}\right)\right]^2 \sigma_T^2 + \left[\frac{c_1 c_2}{T^2}\right]^2 \sigma_e^2} \,|_{T,e=\text{standard atmosphere}} \\ &= 10^{-6} \sqrt{\tilde{c_1} \sigma_T^2 + \tilde{c_2} \sigma_e^2} \\ \sigma_{\varphi_{\text{Int}}} &= 2k \frac{\|s - s_0\|}{10^6} \sqrt{\tilde{c_1} \sigma_T^2 + \tilde{c_2} \sigma_e^2} \end{aligned} \tag{5.17}$$

where $\tilde{c_1} = [c_1 T^{-2}(p + 2c_2 e T^{-2})]^2 \approx 0.9$ and $\tilde{c_2} = [c_1 c_2 T^{-2}]^2 \approx 20.25$. For $s_0 = 1 \text{ km}, \sigma_T = 1 \text{ K}, \sigma_e = 0.1 \text{ mbar}, \lambda = 17.6 \text{ mm}$, one derives an interferometric phase variance of $4\pi\lambda^{-1}10^{-3}\sqrt{0.94 + 0.21} \approx 0.76$ rad. This translates to a variance of estimated deformations of size 1.1 mm. It is quite clear that realistically the temperature and water vapor changes are not distributed homogeneously and instead have to be treated as random fields implying $\Delta n(\cdot)$ to be a random field as well. This is illustrated in figure 5.14 which also plots systematically the influence of temperature and water vapor on the interferometric phase per $10\ \mathrm{m}$.

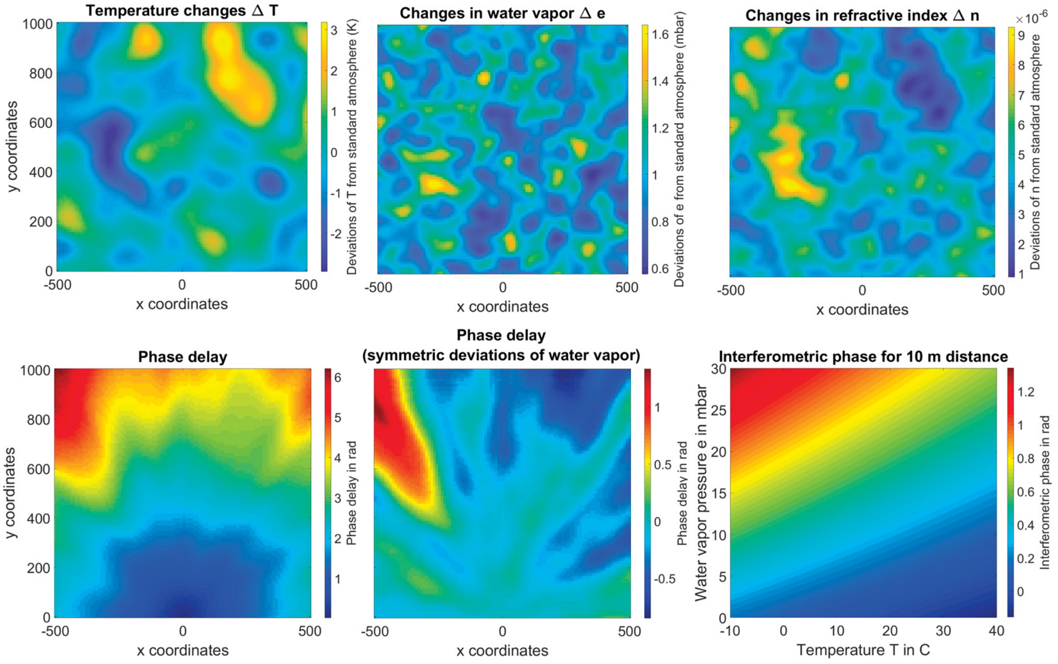


Figure 5.14: The top panels show two-dimensional simulations of temperature and water vapor changes (both assumed to be Gaussian random fields) and the changes they induce for the index of refraction. The resultant phase as calculated via the line integration in equation 5.15 is plotted in the bottom left image. The instrument position is assumed to be at $[0,0]^T$. Since the standard atmosphere only allows positive deviations of water vapor, phase delay is dominantly positive in that case. To prevent the false impression of a nonnegative phase delay being the standard case, the middle panel shows what happens when standard atmosphere is not assumed and deviations of water vapor are therefore allowed to be both positive and negative. The bottom right image shows the unwrapped interferometric phase between a measurement of $10\ \mathrm{m}$ length under standard atmosphere conditions and a measurement of the same length under the meteorological conditions indicated on the axes. Due to linearity, these values may be scaled up by distance d and wrapped as $\mathcal{W}\left(d\phi(T_2, e_2)\right) - \mathcal{W}\left(d\phi(T_1, e_1)\right)$ to derive a guess for the interferometric phase between measurements taken during conditions (T_2, e_2) and (T_1, e_1).

The highly dynamical nature of meteorological phenomena and the complexity already exhibited by the simple simulations in figure 5.14 suggest that an approach based on forward modelling of the atmospheric phase screen will fail. Neither can one expect measurements of temperature and water vapor distributed densely enough to form initial conditions that enable solving the associated differential equations nor would there be any confidence in a solution derived in this way since the Navier Stokes equations are known to be typically ill-posed with solutions exhibiting sensitive dependence on the input data [46, 27].

This seems to be a problem that is encountered routinely when dealing with turbulent media. To summarize Monin and Yagloms reasoning for tackling the problem from a statistical perspective [142, pp. 1-30], the behavior of turbulent media is so complicated and disorderly that the underlying velocity field is typically a nonsparse superposition of infinitely many base-vector fields and no analytical closed-form expression may be derived. The latter would also not likely to be helpful due to the sensitivity of the solutions to boundary conditions; instead the fields generated by virtually identical boundary conditions are best perceived as realizations of a random field whose persistent statistical features are to be emphasized. Modern day understanding of the 'problem of turbulence' is that it is one of determining a one-parameter family of probability distributions $P_t(\cdot)$ over the phase space of solenoidal vector fields indexed by time t. In principle, a nonlinear solution operator U_t determining a one parameter semigroup may be found demonstrating $P_t(\cdot)$ to be uniquely defined by $P_0(\cdot)$ and the dynamics [142, p. 8].

Instead of calculating $P_t(\cdot)$ directly, Monin suggests an approach akin to the method of moments [142, p. 8]. The PDE governing the system is used to derive evolution equations for the statistical moments; they correspond uniquely to a probability distribution. The advantage is that one may focus on first and second order statistical moments and still arrive at telling information about the system's global dynamics. From first principles based on cascades of self-similar eddies, power laws for the spectrum of covariance functions of quantities like velocities and kinetic energy can be deduced [142, p. 15]. These ideas are closely related to the investigations carried out in subsection 2.2.4 for the heat equation. Nonetheless, this is still not satisfactory in the context of our purposes. Even a self-similar theory sufficient for characterizing turbulent flows in regions far from any boundary interactions is bound to fail as a stochastic model for the APS in TRI, in which the topography and its interaction with air flow, temperature exchange and humidity transport takes on a central role.

Still, the general approach of deriving hypotheses for the shape of the statistical moments based on physical considerations is a promising one since the sheer amount of measurements generated by TRI allows to choose unknown parameters in a data-driven way and facilitates a semi-empirical model. The following hypotheses form the basis of later considerations.

1. TRI data contain the effects of deformation, atmosphere and noise. The atmospheric term (APS) is dominated by the influence of the phase delay integrated along the propagation path; other effects like ray bending and reflection are ignorable to the point, that no explicit stochastic model is needed for them. All involved quantities are spatiotemporal random fields.
2. The turbulent field $\Delta n(t, s)$ of changes in refractive index is ergodic [70] with an expected value of zero. Since integration is a linear operation commuting with expectation, also the first moment of the APS vanishes whereas the second moment is nontrivial and found by a double integration of the second moments of $\Delta n(t, s)$.

We remark that these hypotheses are debatable and counterarguments may be found. Their merit lies in suggesting a stochastic model for TRI data that is flexible enough to account for the complicated nature of the behavior witnessed in practice while still restricted enough to allow for inference.

5.2.2 Bisgletscher case study

In the preceeding sections, hypotheses regarding behavior of the APS have been motivated mostly by theoretical considerations. To alleviate this, an extensive TRI dataset is now introduced. It features challenging topography with height differences of several km, distances of up to 8 km, significant surface displacements and spans several months. It has originally been gathered to evaluate the suitability of TRI for glacial monitoring and early warning purposes. Duration and spatial dimension are helpful to form conjectures of the limiting average behavior of the APS and check the validity of the propositions put forward in the last subsection. This can be done by calculating statistical moments like expected values and variances and quantifying their dependence on distance to the instrument and topographical features.

In the mountainous regions of Switzerland, glacial icefalls and the avalanches potentially induced by them pose a constant risk towards rural communities in their vicinity. One such potentially dangerous glacier is the Bisgletscher, which is located in the southern part of Switzerland in canton Valais and may put the Mattervalleys critical transportation infrastructure in form of the local cantonal road and railway line at risk. Installation of classical geodetic measurement devices like GNSS receivers or prisms for total station measurements in this case is both expensive and dangerous. It is wasteful as well in the sense that the instruments often suffer fatal damage and glacial activity indicating a potential event is best quantified by a dense spatial distribution of glacial velocities [56] [55]. To sidestep these problems and complement the cameras already installed in the region, a GPRI terrestrial radar interferometer was installed in 2014 on the slope opposite to the Bisgletscher. It was operated almost continuously for 3 months during summertime to gain insight into the suitability of TRI for deformation monitoring in alpine regions. The setup is illustrated in figure 5.15. [3]

The radar was placed and in the vicinity of a mountain hut that provided power supply and data infrastructure. Apart from a few system failures due to power spikes and outages, the GPRI was continuously operated between mid July and mid September and acquired an SLC every 2 minutes during that time leading to a dataset containing approximately 65.000 SLCs. Each of the interferograms has a dimension of at least 1300×260 [range×cross-range] pixels covering an area, which spans $15\ \mathrm{km}^2$. The geometry is outlined in figure 5.16.

The different types of scattering mechanisms as well as the geometrical properties of the acquisition geometry clearly affect the distribution of the interferometric phase. Examining short- and long-term behavior reveals good agreement between the actual data and the hypotheses stating an expected value of APS + noise of zero,

[3] The project was carried out in collaboration with Prof. Martin Funk (ETHZ) who took the organizational lead, Prof. Martin Truffer (University of Alaska) who provided the instrument and his expertise, and Dr. Jan Beutel from the ETHZ Computer Engineering Group who provided the infrastructure necessary to remotely access the instrument. It was partially funded by the Swiss Office for the Environment. Georg Eisler from Swiss Alpine Club section Uto and Renato Schaller from the Domhütte mountain hut provided logistical support.

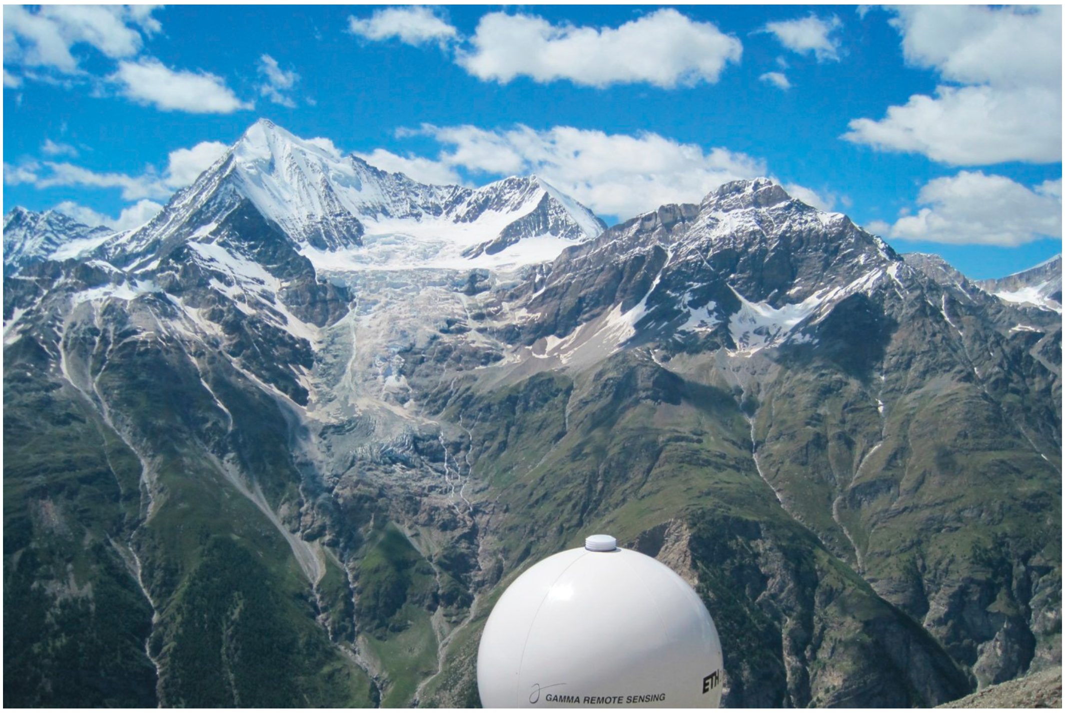


Figure 5.15: In the background, the Bisgletscher is visible with its two steep scarps and its suspension glaciers adhering to the prominent pyramidal form of the Weisshorn peak. The foreground shows the radome covering the GPRI for reasons of weather protection.

as for randomly chosen stable points s and times t

$$\mu_{\text{temporal}} = \sum_{j=1}^{n_{\text{time}}} (\varphi_{\text{Int}})_{s,j} = \epsilon_t \approx 0$$

However, the data does not conclusively support the hypothesis of mean-square ergodicity in the first moment as defined for example in [185, p. 170] because spatial and temporal averages, interpreted as ensemble averages of realizations of random fields, differ significantly. One has

$$\mu_{\text{spatial}} = \sum_{k=1}^{n_{\text{row}}} \sum_{l=1}^{n_{\text{col}}} (\varphi_{\text{Int}})_{kl,t} = \epsilon_s \not\approx 0$$

and no type of convergent behavior is observable when the area used for averaging is continually increased. This may be due to the systems boundaries not being set in a way that the atmospheric dynamics average out over space in the given geometry or because no conservation laws hold for the refraction index for specific moments in time. In the above, $\epsilon_t < 10^{-1}$ rad in nearly 100% of the cases but $\epsilon_s < 10^{-1}$ rad only in 30% of the cases. One day worth of interferograms has been used for these calculations and areas of actual motion have been masked out beforehand.
When directly plotting averages of interferograms over an increasing amount of time like in figure 5.17, it becomes apparent that over a relatively short period of

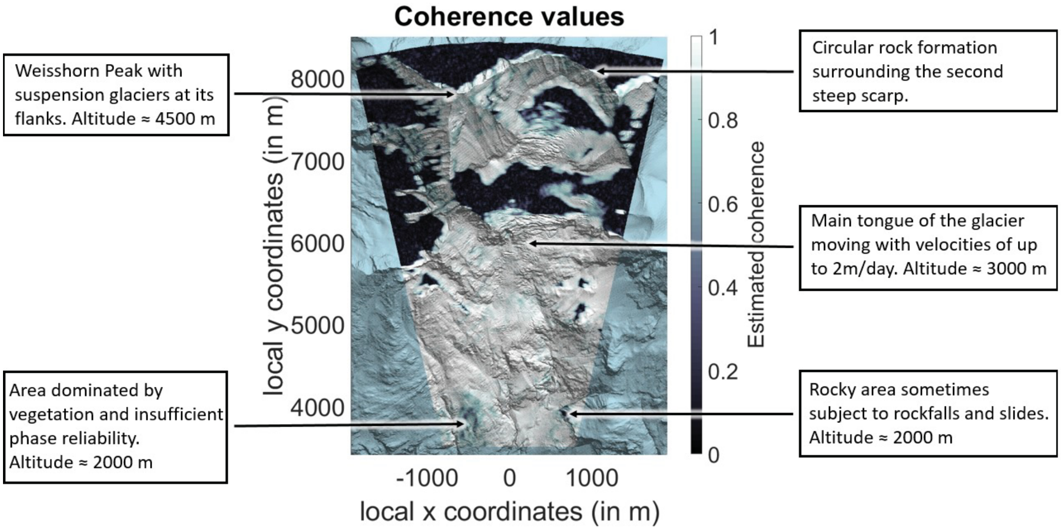

Coherence values
Weisshorn Peak with suspension glaciers at its flanks. Altitude ≈ 4500 m
Area dominated by vegetation and insufficient phase reliability. Altitude ≈ 2000 m
Circular rock formation surrounding the second steep scarp.
Main tongue of the glacier moving with velocities of up to 2m/day. Altitude ≈ 3000 m
Rocky area sometimes subject to rockfalls and slides. Altitude ≈ 2000 m
local y coordinates (in m)
8000
7000
6000
5000
4000
-1000
0
1000
local x coordinates (in m)
Estimated coherence
1
0.8
0.6
0.4
0.2
0

Figure 5.16: The area surveyed by the radar. Several distinct features of the landscape have direct consequences in the form of radar shadows, low coherence and increased noise levels. The radar is installed 3.5 km away from the lower edge of shown terrain at coordinates $[0, 0]^T$ (not visible).

time, the temporally highly frequent atmospheric effects of locally turbulent behavior cancel out and give way to the ones connected to slow and systematic changes in temperature, water vapor and total pressure. These are persistent, spatiotemporally autocorrelated and depend on topographical features.

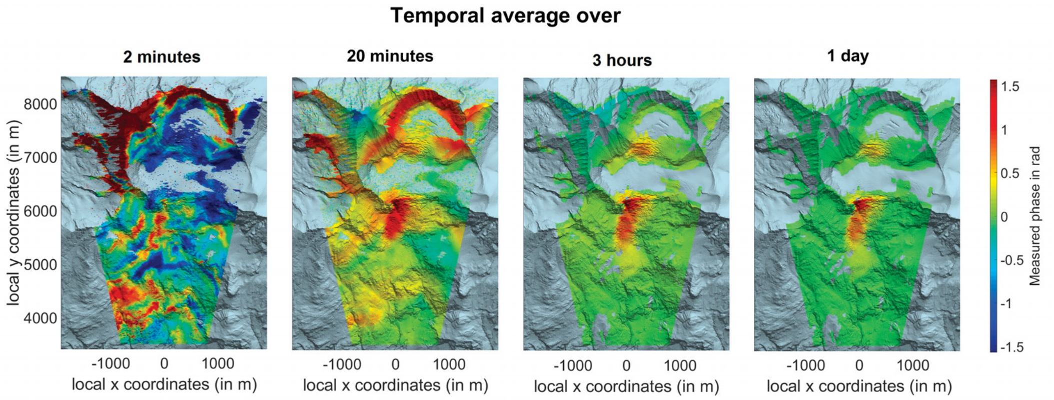


Figure 5.17: Temporal averages over 2 minutes, 20 minutes, 3 hours, and 1 day respectively. Notice the decrease in turbulent behavior and high spatial frequencies as well as the significant influence of the topography.

Based on these observations, one may concretize the initial hypotheses and state the decomposability of data into realizations of random fields corresponding to

- pure noise (negligible spatial and temporal autocorrelation)
- turbulent APS (significant spatial but negligible temporal autocorrelation)
- laminar APS (significant spatial and temporal autocorrelation)
- deformation (structure depending on observed object).

Analyzing the validity of the second hypothesis regarding the integral nature of the APS requires more of a spatial approach. Integration as a linear operator does commute with expectation meaning that the first moment stays unchanged by this hypothesis. Instead what will increase with integration length are the fluctuations of the values and by virtue of that also the variance of the interferometric phase. Figure 5.18 indicates that this is actually the case in practice by plotting phase variances as a function of distance to the instrument for different populations of pixels.

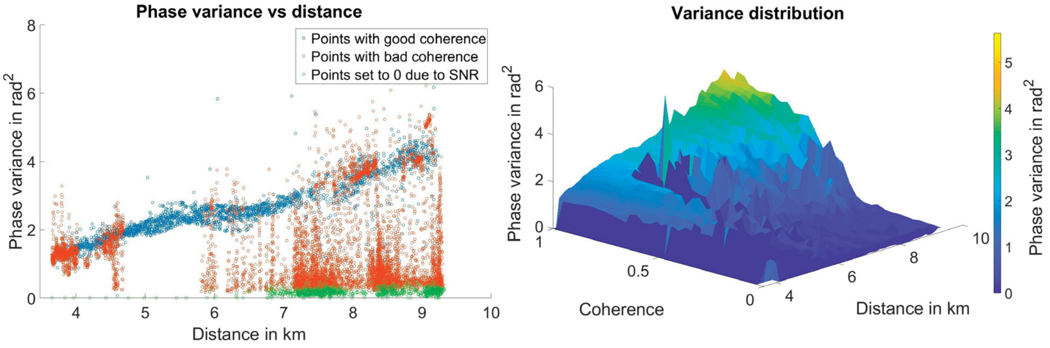


Figure 5.18: The scatterplot shows the phase variance of different populations of pixels and their respective distances to the instrument. Only a random subset of the points is plotted. The right plot demonstrates that the phase variance is not simply due to incoherent fluctuations associated to higher noise on longer propagation paths as phase variance increases though coherence stays good. Indeed, the random fluctuations of incoherent points are resulting in phase variances higher than those of coherent points only for small distances as APS induced phase-changes lessen in magnitude. 24 h worth of data were used to calculate the phase variances. Phases of points with an especially low SNR are set to zero before unwrapping. SNR is estimated for each epoch separately leading to these low-SNR points being unwrapped occasionally and having non-zero phase variances.

Similarly to the first order moments, the estimated average central third order moment approaches zero more reliably the more data is used and one finds for a specific day (01.08.2014) the values -0.92 rad^3, -0.09 rad^3 and -0.01 rad^3 when $10, 100$ and 720 interferograms are used for estimation. The third order moment can deviate consistently and significantly from zero for individual pixels, however, indicating that the underlying probability distribution is not really Gaussian and prediction accuracy could be gained by incorporating statistics of higher order. Since inference procedures that take third order moments into consideration are not covered in the RKHS approach to optimal estimation, attention will be restricted to second order moments.

Figure 5.18 implies that to predict the variance of the interferometric phase, one should incorporate the distance to the instrument and consequently a first ad hoc model could be of the form

$$\sigma^2_{\varphi_{\text{Int}}} = \alpha_0 + \alpha_1 h + \alpha_2 r + \alpha_3 az \tag{5.18}$$

$$h: \text{ Altitude} \qquad r: \text{ Range} \qquad az: \text{ Azimuth}$$

where dependence on azimuth and elevation are included to allow for more flexibility in cross-range direction and influence of topographical information respectively. The relationship is only approximately supposed to hold for interferometric phases primarily affected by APS and explicitly neglects measurements with bad coherence as they would necessitate the inclusion of an SNR-based variance component. A simple least squares fitting of model 5.18 to the empirical phase variances finds $\alpha_0 = -0.35\ \mathrm{rad}^2, \alpha_1 = 2.1 * 10^{-4}\ \mathrm{rad}^2/\mathrm{m}, \alpha_2 = 3.6 * 10^{-4}\ \mathrm{rad}^2/\mathrm{m}, \alpha_3 = 1.7 * 10^{-3}\ \mathrm{rad}$ and α_2 as the factor quantifying propagation distance to be the most important as measured via a sensitivity analysis of prediction performances. Figure 5.19 shows this model and the results in two exemplary situations.

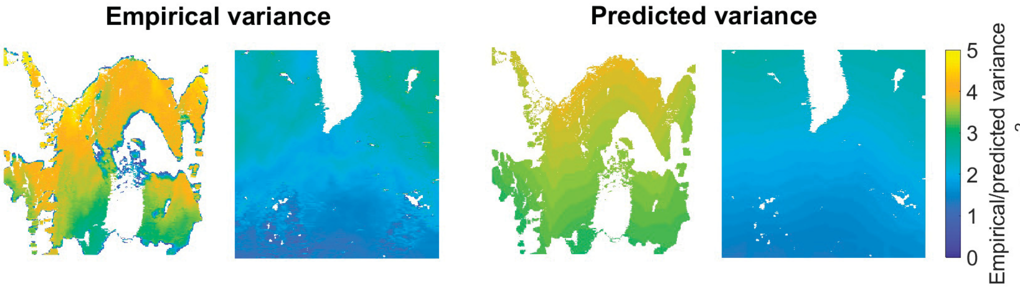


Figure 5.19: Empirical variances and the variances predicted by the simple ad hoc model described in equation 5.18 for two regions corresponding to the topmost part and the bottom of the area surveyed by the radar. Estimations can be invalid (negative) owing to the least squares nature of the estimator. Only pixels that neither move nor show bad SNR are plotted.

Note that as a means for inference, this stochastic model for $\sigma^2_{\varphi_{\text{Int}}}$ is insufficient. It demonstrates, however, that the general proposition of the APS being formed via line integration is supported by data. The linear relation between variance and distance hints at it behaving similar to a Wiener process, see subsection 2.1.1 for a discussion of the integral nature of the Wiener process. Based on these observations, we will elevate the hypotheses 1 and 2 from subsection 5.2.1 to the status of assumptions that will be tacitly presupposed in the further developments of a stochastic model suitable for large scale inference.

5.2.3 A stochastic model for TRI

Before actually formulating and solving the problem of estimating deformations from TRI data, the qualitative hypotheses have to be reframed in a quantitative setting. Based on the last subsection's findings, a stochastic model is presented that consists of a joint normal distribution of the deformations, APS, and noise whose superposition forms the measurements. The stochastic perspective based on random fields and random functions is complemented by an equivalent functional one making use of reproducing kernel Hilbert spaces; abstract splines are reinterpreted in this context. The impact of measurement principle and geometry on the probability distribution of the data is included into the stochastic model by introducing an auxiliary random field of refraction changes; the line integrals through this field form the APS.

Ignoring amplitude information, unwrapped interferograms can be interpreted as functions $m : U \to \mathbb{R}, U \subset \mathbb{R}^3$ that associate to a certain point in space the phase of the signal scattered back by the region surrounding that point. Alternatively, it is possible to map each of these points into a plane of constant height and interpret the result of this projection as functions $U \to \mathbb{R}, U \subset \mathbb{R}^2$. In both cases, the interferograms m are then functions that — although defined on an uncountable domain U — are only known through some of their values. This mirrors the construction in which a function m is associated with a finite set of values $Am \in \mathbb{R}^n$ via the measurement operator A during the formulation of abstract splines in subsection 3.1.2 . In accordance with the introductory remarks about turbulence stated in subsection 5.2.1, it seems sensible to consider m to be a realization of a random field $M : \Omega \times U \to \mathbb{R}$ with Ω some probability space and $m(\cdot) := M(\omega, \cdot) : U \to \mathbb{R}$ or equivalently but with slightly different terminology as a realization of a random function M.

Associating to the random function M a Gaussian probability measure on function space as done throughout chapter 3 and constructing the corresponding reproducing kernel Hilbert space $\mathcal{H}_M$, the following statements are consistent with the assumptions put forward in subsections 5.2.1 and 5.2.2.

1. The RKHS $\mathcal{H}_M$ of measurements $m(\cdot)$ is a direct sum of the RKHSs $\mathcal{H}_D, \mathcal{H}_P, \mathcal{H}_N$ containing the deformation functions $d(\cdot)$, the atmosphere functions $p(\cdot)$ and the noise $n(\cdot)$ respectively. One has
$$\mathcal{H}_M = \mathcal{H}_D \oplus \mathcal{H}_P \oplus \mathcal{H}_N$$
where D, P, N are random fields containing $d(\cdot), p(\cdot), n(\cdot) : U \to \mathbb{R}$.

2. Setting $X \in \{D, P, N\}$ as a placeholder and assuming $X_\cdot(\cdot) : \Omega \times U \ni (\cdot, u) \mapsto X_u \in L^2(\Omega)$ to be mean-zero and square integrable, $E[X_u] = 0, E[X_u^2] < \infty \, \forall u \in U$, the reproducing kernel for X can be written as
$$K_X(u, v) = \mathrm{Cov}(X_u, X_v) = E[X_u X_v]$$
i.e. the respective kernels are just the second order moment functions detailing the autocorrelations of the involved random fields.

3. Sequences of interferograms are spatiotemporal functions $m(\cdot) : U \to \mathbb{R}, U = S \times T$ where S is space and T is time. Using superscripts s and t to denote spatial or temporal parts of the random fields or their realizations, we propose a tensor product decomposition of $\mathcal{H}_X$. If assumed to hold, $\mathcal{H}_X$ will be written as $\mathcal{H}_X^\otimes$ with
$$\mathcal{H}_X^\otimes = \mathcal{H}_X^s \otimes \mathcal{H}_X^t$$
for $X \in \{M, D, P, N\}$ and the tensor product symbol indicating the full RKHS of spatiotemporal functions. This placeholder notation will be used throughout the rest of this section.

Remark The above statements will be interpreted as the qualitative versions of the statements found on page 234. We will tacitly assume their correctness or at least usefulness for the rest of the chapter as they form our description of a stochastic model for TRI. Statement number 3 can be omitted for most of the theoretical considerations and has primarily implications for the factorizability of equations and therefore the computational cost of the algorithm's practical implementation.

In the context of what was said about the stochastic implications of algebraic decomposability of an RKHS into tensor products and direct sums, statement 1 means that the measurements M are a superposition of deformations D, atmosphere P and noise N. All three of them are spatiotemporal random fields that may show complicated autocorrelated behavior but are uncorrelated with each other, i.e.

$$E[D_u P_v] = E[D_u N_v] = E[P_u N_v] = 0 \quad \forall u, v \in U.$$

Assuming Gaussianity of the probability measure on the space of functions is akin to asserting that one only deems relevant the first and second order moments as the Gaussian distribution is the maximum entropy distribution [42, p. 413] given fixed mean values and covariances. It is therefore the distribution including the least amount of additional assumptions given those two pieces of information. Using only second order statistics is justifiable primarily from a functional and algorithmic perspective as the objective of minimizing terms of type

$$\|Ad - y\|_{\mathcal{H}_1}^2 + \|Bd\|_{\mathcal{H}_2}^2 \qquad\qquad y \in \mathbb{R}^n \text{ observations}$$

w.r.t. deformation functions $d(\cdot) \in \mathcal{H}_D$ is a reasonable one for estimating $d(\cdot)$ and second order statistics are sufficient for that. It is also clear that including j-th order moments or cumulants would lead to exploding computational costs and memory requirements since it would be necessary to store and manipulate j-tensors with n^j elements.

Decomposability of the Hilbert space $\mathcal{H}_X^{\otimes}$ of spatiotemporal functions into a tensor product of Hilbert spaces $\mathcal{H}_X^s$ of spatial functions and Hilbert spaces $\mathcal{H}_X^t$ of temporal functions implies that every function $x(\cdot) \in \mathcal{H}_X^{\otimes}$ may be written as a superposition of base functions $\varphi_i^s \in \mathcal{H}_X^s$ and $\varphi_j^t \in \mathcal{H}_X^t$ as

$$x(u, v) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \alpha_{ij} \varphi_i^s(u) \varphi_j^t(v); \qquad\qquad \sum_{i,j=1}^{\infty} |\alpha_{ij}|^2 < \infty \tag{5.19}$$

where α forms an infinite second order coefficient tensor [7, pp. 358-361]. Equivalently, the reproducing kernel $K_X^{\otimes}$ of $\mathcal{H}_X^{\otimes}$ is just the product of the separate kernels

$$K_X^{\otimes}((s_1, t_1), (s_2, t_2)) = K_X^s(s_1, s_2) K_X^t(t_1, t_2) \tag{5.20}$$

which precludes certain types of complicated correlation patterns. The advantage of demanding a Hilbert space to be of this restricted class are that statements can be derived regarding representation and truncation of multivariable functions and the factorization of problems into subproblems.

The local and global structure of $\mathcal{H}_M = \mathcal{H}_D \oplus \mathcal{H}_P \oplus \mathcal{H}_N$ is determined by the structure of the constituent Hilbert spaces which are in turn determined by the Kernels K_D, K_P, K_N. It is to be expected that K_N is white noise with pointwise varying variance and no autocorrelations whatsoever. The comments on generalized stochastic processes from chapter 3 apply because K_N as defined on a continuous domain is not actually a function but a functional — however, when N is considered on the set of pixels, the distinction vanishes and no complications arise when $K_N^{\otimes}(u, v)$ is just considered as $\sigma_{uv}^2 \delta_{uv}$ with δ_{ij} the Kronecker delta.

The kernel $K_D^{\otimes}$ should reflect the spatiotemporal behavior expected of the deformation. Depending on the mechanism behind it, it may be written as $K_D^{\otimes} = D_D^s K_D^t$ with K_D^s and K_D^t both smooth for e.g. creep processes or as $K_D^{\otimes} = K_D^s K_D^t$ with K_D^s ragged and long correlation length on K_D^t to model spatially localized but temporally relatively persistent motion as encountered for example during monitoring of glacial iceflows. The choice of $K_D^{\otimes}$ is a design decision and the author sees few possibilities to infer it from data as deformations seldomly occur in TRI data without being accompanied by APS and noise. Whereas it might be sensible to first learn jointly $K_P + K_N$ with kernel inference by training on stable regions and then split any guess for K_M based on regions where D, P and N are nonzero into $K_P + K_N$ and a guess for K_D, the same will not work in early warning applications where the deformations are seldomly observed and too much adherence to observed data could cause a bias against recognizing deformations at their onset. Especially for rare events, the best choice of $K_D^{\otimes}$ needs to be discussed critically and in all likelihood on a case-by-case basis.

The stochastic model and its associated Hilbert space formulation are summarized below compactly for convenience and easier reference. The (phase) measurements M are the zero-mean random field

$$\begin{aligned} M_{\cdot}^{\cdot} : \Omega \times (S \times T) \ni (\omega, v) &\mapsto M_v^{\omega} \in \mathbb{R} \qquad (5.21) \\ \Omega :&\ \text{Some probability space} \\ S \times T :&\ \text{Subset of } \mathbb{R}^4 \text{ indexing the measurements} \\ v = (s, t) &\ \ \text{Element of } S \times T; s, t \text{ space, time index} \end{aligned}$$

If interpreted in the sense of functions $S \times T \to \mathbb{R}$ chosen at random via $X_{\cdot}^{\cdot} : \Omega \ni \omega \mapsto X_{\cdot}^{\omega} \in \mathcal{H}_X$, $\mathcal{H}_X$ being an RKHS with RK K_X, the notation $x(\cdot)$ or plainly x will be used instead of $X_{\cdot}^{\cdot}$. Then one may reformulate the equation $M_{\cdot}^{\cdot} = D_{\cdot}^{\cdot} + P_{\cdot}^{\cdot} + N_{\cdot}^{\cdot}$ in terms of Hilbert space valued random variables:

$$\begin{aligned} & m = d + p + n \qquad (5.22) \\ d \in \mathcal{H}_D &\quad \text{RKHS containing deformations} \\ p \in \mathcal{H}_P &\quad \text{RKHS containing phase screens} \\ n \in \mathcal{H}_N &\quad \text{RKHS containing noise functions} \\ m \in \mathcal{H}_M &\quad \mathcal{H}_M = \mathcal{H}_D \oplus \mathcal{H}_P \oplus \mathcal{H}_N \end{aligned}$$

where $\oplus$ denotes the orthogonal direct sum of Hilbert spaces and $\mathcal{H}_X$ has RK K_X

s.t. $K_X(\cdot,\cdot) : (S \times T)^2 \ni (v_1, v_2) \mapsto K_X(v_1, v_2) = E\left[X_{v_1}^{\cdot} X_{v_2}^{\cdot}\right] \in \mathbb{R}$ from which it is straightforward to show that $K_M = K_D + K_P + K_N$.

There are several possible approaches to generate structure identification procedures using the theorems and discussions contained in chapter 4. Apart from numerical issues that can arise when evaluating the solutions to optimal estimation problems in an RKHS framework, the main problem still left is one of kernel inference. The kernels K_D, K_P, K_N need to be estimated from data or, if no other way is found, prescribed on a reasonably sound physical basis. As indicated before, $K_P + K_N$ is indirectly observed via measurements on stable areas and can be separated due to the white noise assumption that distinguishes K_N from K_P. Nonetheless, the exact way in which inference of the reproducing kernels is carried out, i.e. the objective function to be minimized, has a significant impact on the form of the inferred kernels and consequently on the estimators that are derived using these kernels.

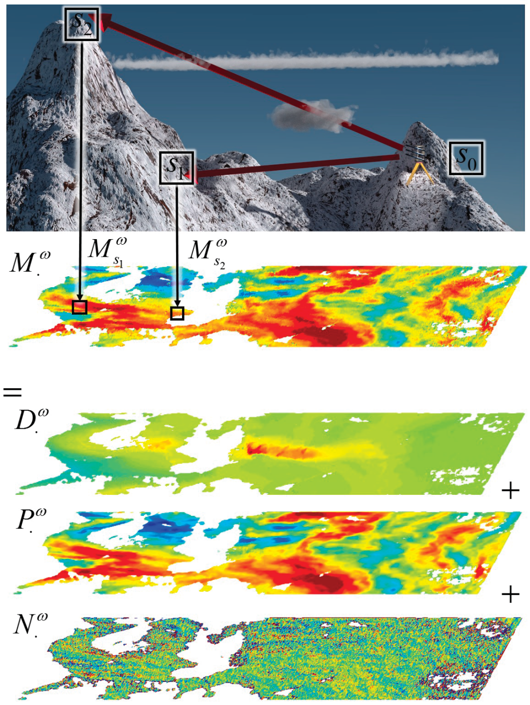


Figure 5.20: The actual measurements M are a superposition of deformations D, atmospheric phase screen P strongly autocorrelated due to measurement geometry, and uncorrelated noise N.

As implied by figure 5.20 $p(s,t)$, the phase screen associated to index (s,t), subsumes all atmospheric effects $\Delta p(\cdot)$ influencing the electromagnetic wave along its propagation path linking location of instrument s_0 with the location of backscattering $s \in S$. This will be reflected by including the functional relationship

$$p(s,t) = \int_{s_0}^{s} \Delta p(r,t) dr = \varphi_s \Delta p(\cdot, t) \tag{5.23}$$

$$\Delta p(\cdot, t) : \overline{\text{ch}}(S \cup s_0) \ni s \mapsto \Delta p(s,t) \in \mathbb{R} \tag{5.24}$$

where $p(\cdot) \in \mathcal{H}_P, \Delta p(\cdot) \in \mathcal{H}_{\Delta P}, \overline{\text{ch}}(S \cup s_0)$ is the closure of the convex hull of $S \cup s_0$ and $\mathcal{H}_{\Delta P}$ is the RKHS with RK $K_{\Delta P}$. $K_{\Delta P}$ and K_P are related via

$$\begin{aligned} K_P(v_1, v_2) &= E\left[\varphi_{s_1} \Delta p(\cdot, t_1) \varphi_{s_2} p(\cdot, t_2)\right] \\ &= \varphi_{s_1} \otimes \varphi_{s_2} K_{\Delta P}\left((\cdot, t_1), (\cdot, t_2)\right) \\ &= \int_{s_0}^{s_1} \int_{s_0}^{s_2} K_{\Delta P}\left((r_1, t_1), (r_2, t_2)\right) dr_1 dr_2 \end{aligned} \tag{5.25}$$

in which we understand $\varphi_{s_1} \otimes \varphi_{s_2}$ to act on the Mercer decomposition $K_{\Delta P}\left((\cdot, t_1), (\cdot, t_2)\right) = \sum_{i=1}^{\infty} \lambda_i e_i(\cdot, t_1) \otimes e_i(\cdot, t_2)$ of $K_{\Delta P}$.

The Moore-Aronszajn theorem immediately asserts that to K_P there corresponds a unique RKHS $\mathcal{H}_P$ with RK K_P and $\exists$ a random field $\{\overline{P}^{\cdot}_{(s,t)} : \Omega \to \mathbb{R}, (s,t) \in S \times T\}$ whose covariance function is K_P. However $\{\overline{P}^{\cdot}_{(s,t)}, (s,t) \in S \times T\}$ and $\{P^{\cdot}_{(s,t)}, (s,t) \in S \times T\}$ may differ from each other in moments of order higher than 2 [129, thm. 6.1].

5.2.4 Reproducing kernels arising in TRI

Whereas the correlation structure of the deformations is to a certain degree an unknowable design parameter to be prescribed by prior knowledge or assumptions, the covariance functions of noise and APS can reasonably be inferred from observations via kernel inference and compared to the theoretical model based on line integration. Several flexible approaches are possible; they are employed to guess the involved correlation functions which in turn are identified with the reproducing kernels of RKHS. The solutions to small-scale toy problems give intuitive insight into the suitability of the different stochastic models.

To illustrate the approaches listed in section 4.2, Bochner's theorem and kernel inference are contrasted with an approach making use of the line integration model. Suppose the problem were to infer the correlation structure of the APS in the regions outlined in red in figure 5.21. The region is assumed for now to be free of noise and deformations after some elementary preprocessing.

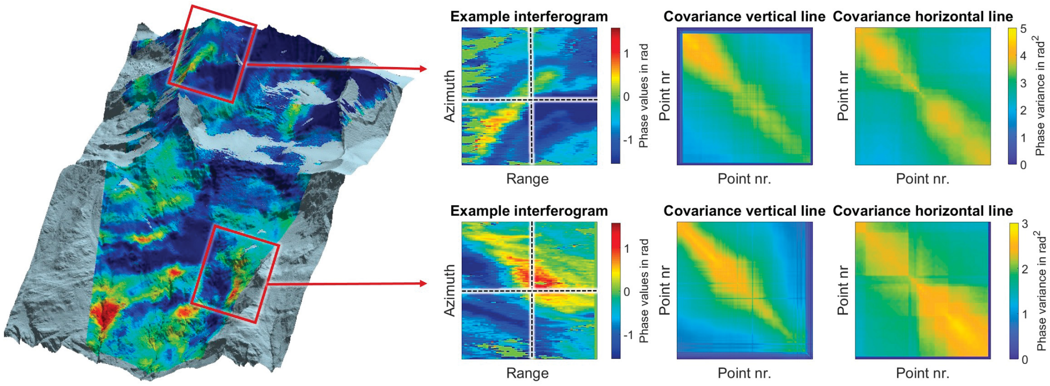


Figure 5.21: The TRI data in the subregions outlined in red are dominated by APS and noise. Two example interferograms detailing these regions are shown in the second column. The two lines indicate the two sets of points, for which the empirical covariance matrices are plotted in the four right images. For the calculation of the covariance matrices 24 h worth of data were used.

A first, purely ad hoc least squares based idea consists in presupposing the existence of a radial second order stationary kernel $K(u,v) = C(u-v) = C(\Delta r)$ that is linearly related to the observed empirical covariance matrix $S \in \mathbb{R}^n \otimes \mathbb{R}^n \cong \mathbb{R}^{n^2}$ via

$$S = AC. \tag{5.26}$$

Here A is a linear operator encoding double integration $(Af)(u,v) =$

$\int_{s_0}^{u}\int_{s_0}^{v} f(\tilde{u},\tilde{v})d\tilde{u}d\tilde{v}$ and if $C(\cdot)$ is represented as a discrete vector $\hat{C}$ in $\mathbb{R}^m$, A can be written as a matrix in $\mathbb{R}^{n^2} \otimes \mathbb{R}^m$ and $C = A^+S \in \mathbb{R}^m$ would be an obvious candidate providing sample values of the function $C(\cdot)$ that could be used for estimation of $C(\cdot)$ via for example the fitting of splines.

Apart from the problem of providing only values, through which a positive definite function is to be fitted carefully, figure 5.22 clearly indicates this procedure to be unreliable. Furthermore the shape of the guess $\hat{C} \in \mathbb{R}^m$ of $C(\cdot)$ looks highly unexpected and with its strong fluctuations is both implausible and unlikely to allow fitting of a positive definite function as the Toeplitz matrix generated by $\hat{C}$ is not positive semidefinite due to off diagonal terms being larger than their diagonal counterparts. This approach is therefore unsuitable for inference of the involved covariance functions. It is thus interesting to investigate the more sophisticated models explained in section 4.2 and what they tell about the correlation structure of TRI data.

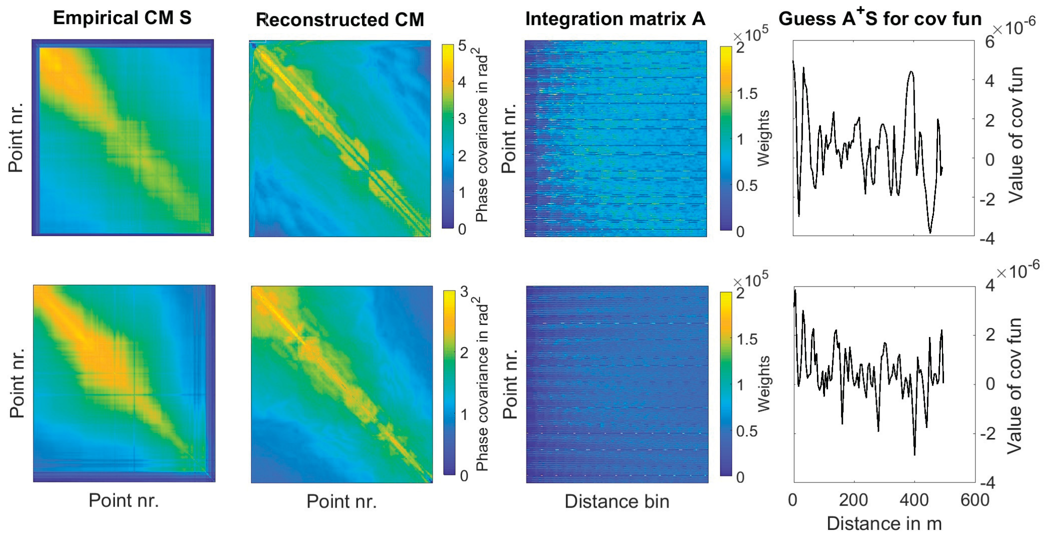


Figure 5.22: Guesses for underlying radial stationary covariance functions using equation 5.26. The empirical covariance matrices are extracted from the regions outlined on the previous figure by evaluating 24 h of TRI data. The guesses for the covariance functions in the last column look highly implausible and the reconstructed covariance matrices in the second column are very irregular compared to the original empirical covariance matrices. The correlation structures encoded in $\hat{C}_1 = A_1^+S_1$ and $\hat{C}_2 = A_2^+S_2$ do not coincide with what one would expect from atmospherically driven processes and are very dissimilar to each other.

§ Estimation by Bochners theorem

Given an empirical covariance matrix $S \in \mathbb{R}^{n_\sharp} \otimes \mathbb{R}^{n_\sharp} \cong \mathbb{R}^{n_\sharp^2}$ deviating from the true one $(K_{\text{true}}(v_i,v_j))_{i,j=1}^{n_\sharp}$, $n_\sharp$ number of persistent scatterers, in a nonsystematic manner and $K_{\text{true}} = L_\otimes C_{\text{true}} = L \otimes LC_{\text{true}}, C_{\text{true}} \in \mathcal{P}(\text{Space})$ one may propose for K_{true} the estimator $\tilde{K}$ as defined below.

$$\tilde{K} = \underset{K=L_\otimes C, C\in\mathcal{P}(\text{Space})}{\text{argmin}} \| \operatorname{Re}(S - A_\otimes K) \|_{L^2}^2 \tag{5.27}$$

Again, $A = \bigoplus_{i=1}^{n_\sharp} A_i$ is evaluation at points $\{v_i\}_{i=1}^{n_\sharp} \subset$ Space, $A_\otimes = A \otimes A$, and L is the bounded linear operator relating a s.o.s. random field with covariance $C(\cdot - \cdot)$ [4] to the instationary one with covariance function $K(\cdot, \cdot)$. Inclusion of this assumption allows C to be written as a weighted integral of complex exponentials $\chi_v = \exp(2\pi i \langle v, \cdot \rangle)$ against a non-normalized probability measure. Approximation via

$$\tilde{C} = \sum_{l=1}^{n_{\text{exp}}} \tilde{\gamma}_l \chi_l \qquad \tilde{\gamma} = (\tilde{\gamma}_l)_{l=1}^{n_{\text{exp}}} \in \mathbb{R}_+^{n_{\text{exp}}} \tag{5.28}$$

for $\chi_l = \chi_{\omega_l}$ for some choice of wavevectors $\{\omega_l\}_{l=1}^{n_{\text{exp}}}$ is then a reasonable option as by positivity of coefficients γ_l, $\tilde{C}$ is positive definite (see equation 4.1.11). Consequently $\tilde{K} = L_\otimes \tilde{C}$ is also positive definite and a valid reproducing kernel.

The right hand side of equation 5.27 can be simplified sufficiently to allow expressing the problem in a more amenable form. Considering S and $A_\otimes K$ as column vectors in the tensor spaces $\mathbb{R}^{n_\sharp} \otimes \mathbb{R}^{n_\sharp}$ or $\mathbb{C}^{n_\sharp} \otimes \mathbb{C}^{n_\sharp}$ respectively and introducing the notation $L_\otimes \chi_l = \psi_l$ as well as $A_\otimes \psi_l = \Psi_l \in \mathbb{C}^{n_\sharp} \otimes \mathbb{C}^{n_\sharp}$ we find

$$\begin{aligned} \| \operatorname{Re}(S - A_\otimes K) \|_{L^2}^2 &= \|S - \operatorname{Re}\left(\sum_{l=1}^{n_{\text{exp}}} A_\otimes \psi_l \gamma_l \right) \|_{L^2}^2 \\ &= \|S - \operatorname{Re}\left[\Psi_1 \cdots \Psi_{n_{\text{exp}}} \right] \gamma \|_{L^2}^2 \\ &= \|S - \Psi \gamma\|_{L^2}^2 \end{aligned} \tag{5.29}$$

where Ψ is as defined by equation 5.31. The optimal choice $\tilde{\gamma} \in \mathbb{R}_+^{n_{\text{exp}}}$ to minimize expression 5.29 and by extension solve the nonnegative least squares problem

$$\tilde{\gamma} = \underset{\gamma \in \mathbb{R}_+^{n_{\text{exp}}}}{\operatorname{argmin}} \|S - \Psi\gamma\|_{L^2}^2 = \operatorname{nnls}(S, \Psi) \tag{5.30}$$

$$\Psi = \begin{bmatrix} \operatorname{Re}(\Psi_1)_{11} & \cdots & \operatorname{Re}\left(\Psi_{n_{\text{exp}}}\right)_{11} \\ \vdots & \ddots & \vdots \\ \operatorname{Re}(\Psi_1)_{n_\sharp n_\sharp} & \cdots & \operatorname{Re}\left(\Psi_{n_{\text{exp}}}\right)_{n_\sharp n_\sharp} \end{bmatrix} \tag{5.31}$$

S : Column vector of empirical covariances

is well known to be approximable using iterative standard methods from convex optimization [37, p. 72].

To reduce the computational burden arising later during evaluation of $\tilde{K}$, coefficients $\tilde{\gamma}_l$ so small as to be virtually negligible w.r.t. their contribution to $\tilde{K} = \sum_{l=1}^{n_{\text{exp}}} \tilde{\gamma}_l \psi_l$ will be discarded. Truncation of $\tilde{\gamma}$ leads to the reduced coefficient vector $\tilde{\gamma}_{\text{red}}$ and the reduced positive definite function $\tilde{C}_{\text{red}} = \sum_{l=1}^{n_{\text{exp}}} (\tilde{\gamma}_{\text{red}})_l \chi_l$. Using the $L_2 - L_2$ operator norm $\|A\|_{\text{op}} = \sup_{\|x\|_2 = 1} \|Ax\|_2$ it is possible to assemble the

[4] This notation indicates that C is a function acting on tuples of arguments via $C : (s, t) \mapsto C(s - t)$

following inequalities.

$$\begin{aligned} \|\tilde{K} - \tilde{K}_{\text{red}}\|_{L^2} =& \|L_\otimes \left(\tilde{C} - \tilde{C}_{\text{red}}\right) \|_{L^2} \\ \le& \|L_\otimes\|_{\text{op}} \|\tilde{C} - \tilde{C}_{\text{red}}\|_{L^2} \qquad (5.32) \\ \|\tilde{K}\|_{L^2} \le& \|L_\otimes\|_{\text{op}} \|\tilde{C}\|_{L^2} \qquad (5.33) \end{aligned}$$

To ensure negligibility, it is then demanded that

$$\begin{aligned} \|L_\otimes\|_{\text{op}} \|\tilde{C} - \tilde{C}_{\text{red}}\|_{L^2} \le& 10^{-2} \|L_\otimes\|_{\text{op}} \|\tilde{C}\|_{L^2} \\ \Leftrightarrow \qquad 10^{-2} \ge& \frac{\|\tilde{C} - \tilde{C}_{\text{red}}\|_{L^2}}{\|\tilde{C}\|_{L^2}} \\ \ge& \sqrt{\frac{\sum_{l=1}^{n_{\text{exp}}} \left(\tilde{\gamma}_l - (\tilde{\gamma}_{\text{red}})_l\right)^2}{\sum_{l=1}^{n_{\text{exp}}} \left(\tilde{\gamma}_l\right)^2}} \qquad (5.34) \end{aligned}$$

Effectively this limits the upper bound for the energy of $\tilde{K} - \tilde{K}_{\text{red}}$ to 1% of that of $\tilde{K}$. Calculating $\tilde{K}^s_M = \tilde{K}^s_P + \tilde{K}^s_N$ requires the evaluation of the more complicated terms $\psi^s_j(\cdot,\cdot) = \varphi_. \otimes \varphi_. \chi^s_j(\cdot - \cdot)$, where $\varphi_.$ is the line integration introduced in equation 5.23.

$$K^s_M(s_n, s_m) = \underbrace{\sum_{l=1}^{n_{\text{exp}}} \psi^s_l(t_n, t_m)\gamma^s_{P,l}}_{K^s_P(s_n,s_m)} + \underbrace{\sum_{l=1}^{n_s} \delta^s_l(s_n, s_m)\gamma^s_{N,l}}_{K^s_N(s_n,s_m)} \qquad (5.35)$$

In the above, $\delta^s_l(s_n, s_m)$ is a Kronecker-delta type function that is 1 if $s_l = s_n = s_m$ and 0 otherwise. With the same notation as before and using the fact that the dual group $\widehat{\mathbb{R}^n}$ of $\mathbb{R}^n$ is $\widehat{\mathbb{R}}^n$ implying its characters χ^s_j to satisfy

$$\chi^s_j(s_l) = \prod_{k=1}^{n} \exp(2\pi i \omega^k_j s^k_l) = \exp(2\pi i \langle \omega_j, s_l \rangle) \qquad (5.36)$$

$s_l = (s^k_l)^n_{k=1}, \omega_j = (\omega^k_j)^n_{k=1}$ a direct computation is possible.

$$\begin{aligned} A_k \otimes A_l \psi^s_j &= \varphi_{s_k} \otimes \varphi_{s_l} \exp(2\pi i \langle \omega_j, \cdot - \cdot \rangle) \\ &= \int_{s_0}^{s_k} \int_{s_0}^{s_l} \exp(2\pi i \langle \omega_j, u - v \rangle) du dv \\ &= g_j(s_0, s_k) \overline{g_j(s_0, s_l)} \qquad (5.37) \\ g_j(s_0, s_k) &= \frac{\|s_k - s_0\|_{\ell^2}}{2\pi \langle \omega_j, s_k - s_0 \rangle} \left[e^{2\pi i \langle \omega_j, s_k \rangle} - e^{2\pi i \langle \omega_j, s_0 \rangle} \right] \qquad (5.38) \end{aligned}$$

If $\langle \omega_j, s_k - s_0 \rangle = 0$ then $g_j(s_0, s_k)$ is trivially $\|s_k - s_0\|_{\ell^2}$. It immediately follows by means of equations 5.30 and 5.31 that

$$\gamma_M^s = (\tilde{\gamma}_P^s, \tilde{\gamma}_N^s) = \operatorname{nnls}(S, \Psi^s) \tag{5.39}$$

$$\Psi^s = \begin{bmatrix} f_1(s_1,s_1) & \cdots & f_{n_{\exp}}(s_1,s_1) & \delta_1(s_1,s_1) & \cdots & \delta_{n_s}(s_1,s_1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ f_1(s_{n_s},s_{n_s}) & \cdots & f_{n_{\exp}}(s_{n_s},s_{n_s}) & \delta_1(s_{n_s},s_{n_s}) & \cdots & \delta_{n_s}(s_{n_s},s_{n_s}) \end{bmatrix} \tag{5.40}$$

$$f_j(s_k, s_l) = \operatorname{Re}(g_j(s_0,s_k)\overline{g_j(s_0,s_l)}) \tag{5.41}$$

The coefficients $\{\tilde{\gamma}_{P,l}^s\}_{l=1}^{n_{\exp}}$ and $\{\tilde{\gamma}_{N,l}^s\}_{l=1}^{n_s}$ are then inserted into formula 5.35 completing step 1 of the estimation process followed by truncation of the coefficient vectors. Some results of this procedure can be seen in figure 5.23.

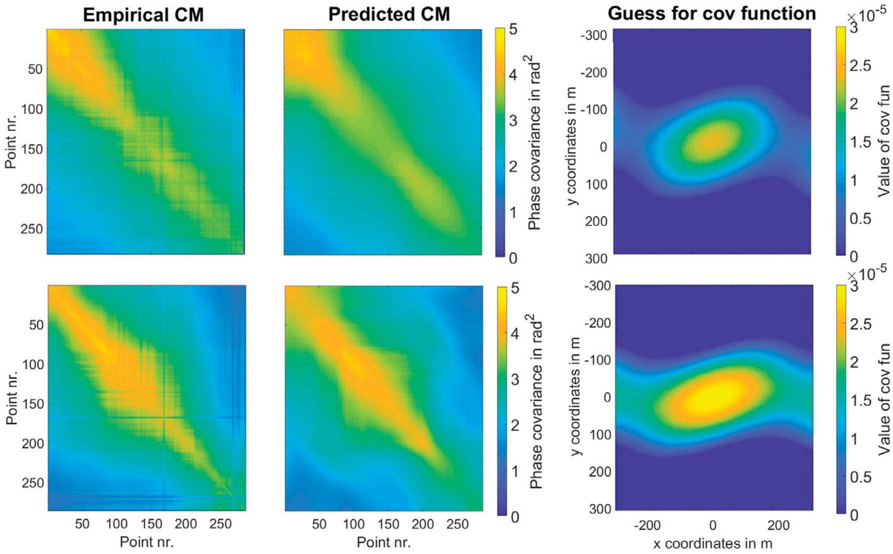


Figure 5.23: For the same empirical covariance matrices S_1 and S_2 as in figure 5.22, the Bochner type approach provides an estimate of an underlying stationary covariance function $C(\cdot)$ and the model predictions $A_\otimes L_\otimes C(\cdot)$ are positive semidefinite. The third column shows the inferred underlying stationary covariance functions at altitude $z = 0$, i.e. it shows an assortment of values $C(x, y, 0)$ detailing the covariance of any two points having respective coordinate separation of x ,y, and 0. The parameter vector γ has approximately 340 entries and highest frequencies of 1 cycle per 500 m.

The derivation of a stationary covariance function C with $L_\otimes C = K$ and $A_\otimes K$ close to S via Bochner's theorem is pleasing from a theoretical point of view and as figure 5.23 indicates, the results can be reasonable. However, some of the the practical disadvantages are prohibitive. For example, it is unclear, how to choose the right set of wavevectors ω for the complex exponentials $\chi_\omega(\cdot)$ and the results of inference can change drastically depending on that choice. The expansion $C(\cdot) = \int_{\omega \in \widehat{\mathbb{R}^n}} \chi_\omega(\cdot) d\nu(\omega)$ is only guaranteed to hold if the domain of integration is the whole dual group $\widehat{\mathbb{R}^n}$ and not only a discrete subset $\{\omega_l\}_{l=1}^{n_{\exp}}$. Furthermore the closing comments from subsection 4.2.1 hold: Even in the case where approximation to the underlying ground truth is good in the ℓ^2-sense and

$$\widetilde{C}(\cdot) = \sum_{l=1}^{n_{\text{exp}}} \tilde{\gamma}_l \chi_l(\cdot) \quad \text{satisfies} \quad \|A_\otimes L_\otimes \widetilde{C}(\cdot) - S\|_F^2 \quad \text{is small}$$

it is by no means guaranteed that the covariance matrix $A_\otimes L_\otimes \widetilde{C}(\cdot)$ derived from $\widetilde{C}$ is likely in the statistical sense given the observations summarized in the empirical covariance matrix S. The truncation induces persistent long range oscillations in the correlations between far away regions and harms this model's credibility further.

§ Estimation using parametrized families of kernels

The most obvious way to avoid regularity problems is to choose $C(\cdot)$ from a set of well-behaved kernels. Again, the main assumption will be that the relation

$$K_P(s_i, s_j) = (L_\otimes C(\cdot - \cdot))_{(s_i, s_j)} = \int_{s_0}^{s_i} \int_{s_0}^{s_j} C(u - v) du dv$$

holds where $K_P(\cdot,\cdot)$ is the kernel for $\mathcal{H}_P^s$, L is the operator of line integration and $C(\cdot)$ is to be inferred from a class of parametric kernels. If the dependence is to be made explicit, $C(\cdot)$ is also written as $C_\theta(\cdot)$ where $\theta \in \Theta$ are the parameters. The goal is to determine θ in such a way as to minimize the discrepancy between empirical and predicted covariance matrices as, for the sake of simplicity, measured by the Frobenius norm.

$$\theta_{\text{opt}} = \underset{\theta \in \Theta}{\text{argmin}} \ \|A_\otimes L_\otimes C_\theta(\cdot) - S\|_F^2 \tag{5.42}$$

When θ_{opt} is found, the kernel $K_{\theta_{\text{opt}}}(\cdot,\cdot) = L_\otimes C_{\theta_{\text{opt}}}(\cdot,\cdot)$ is considered the best guess and $\mathcal{H}_P^s$ is set to $\mathcal{H}_{K_{\theta_{\text{opt}}}}$. To keep the optimization 5.42 feasible, the covariance functions $C_\theta(\cdot - \cdot)$ are allowed to be a part of only 3 restrictive families with two parameters each. The main properties of these covariance models are outlined below and described more comprehensively in [39, pp. 84-101].

1. Name: Squared exponential covariance function

   Equation: $C_\theta(s_i - s_j) = \theta_1 \exp\left(-\frac{\|s_i - s_j\|^2}{\theta_2^2}\right), \quad (\theta_1, \theta_2) \in \mathbb{R}_+ \times \mathbb{R}_+$

   Corresponds to: Δn is an infinitely differentiable field of changes in refraction index.

2. Name: Exponential covariance function

   Equation: $C_\theta(s_i - s_j) = \theta_1 \exp\left(-\frac{\|s_i - s_j\|}{\theta_2}\right), \quad (\theta_1, \theta_2) \in \mathbb{R}_+ \times \mathbb{R}_+$

   Corresponds to: Δn is an everywhere continuous but nowhere differentiable field of changes in refraction index.

3. Name: Spherical covariance function

Equation: $C_\theta(s_i - s_j) = \mathbb{1}_{[\|s_i-s_j\| \leq \theta_2]}\theta_1 \left(1 - \frac{3}{2}\frac{\|s_i-s_j\|}{\theta_2} + \frac{1}{2}\frac{\|s_i-s_j\|^3}{\theta_2^3}\right)$, $(\theta_1, \theta_2) \in \mathbb{R}_+ \times \mathbb{R}_+$, $\mathbb{1}_x = 1$ if statement x is true and 0 otherwise

Corresponds to: Δn exhibits a behavior in between being smooth and discontinuous and is an irregular and, near the origin, structurally scale invariant field of changes in refraction index similar to fractal fields governed by a power law variogram $E[(\Delta n(s_i) - \Delta n(s_j))^2] = \|s_i - s_j\|^\alpha$.

Even though only two parameters are involved, the range parameter θ_2 enters the equations for $C_\theta(\cdot)$ nonlinearly. Since evaluation of the covariance functions can be made relatively cheaply with the help of numerical methods, a brute force approach is adopted to estimate θ_2. It consists in evaluating the objective function in equation 5.42 for a range of positive values θ_2 and following up with a refinement step. Once a specific value θ_2 is chosen, the optimal θ_1 can be calculated in closed form as the term

$$\theta_1^{\text{opt}} = \underset{\theta_1 \in \mathbb{R}_+}{\operatorname{argmin}} \ \|\theta_1 K_{\theta_2} - S\|_F^2$$

where $K_{\theta_2} = A_\otimes L_\otimes C_{\theta=(1,\theta_2)}(\cdot - \cdot)$. Then

$$\begin{aligned}\|\theta_1 K_{\theta_2} - S\|_F^2 &= \langle \theta_1 K_{\theta_2} - S, \theta_1 K_{\theta_2} - S\rangle_F \\ &= \theta_1^2 \langle K_{\theta_2}, K_{\theta_2}\rangle_F + \langle S, S\rangle_F - 2\theta_1 \langle K_{\theta_2}, S\rangle_F \\ &= \theta_1 \|K_{\theta_2}\|_F^2 - 2\theta_1\langle K_{\theta_2}, S\rangle_F + \|S\|_F^2\end{aligned}$$

Taking the derivative $\partial_{\theta_1}\|\theta_1 K_{\theta_2} - S\|_F^2$ and equating it to zero to determine the minimum of this convex problem yields

$$\partial_{\theta_1}\|\theta_1 K_{\theta_2} - S\|_F^2 = 2\theta_1\|K_{\theta_2}\|_F^2 - 2\langle K_{\theta_2}, S\rangle_F \overset{!}{=} 0$$

which is immediately resolved and establishes

$$\theta_1^{\text{opt}} = \frac{\langle K_{\theta_2}, S\rangle_F}{\langle K_{\theta_2}, K_{\theta_2}\rangle_F}.$$

No need exists therefore to brute force search for θ_1^{opt} as it is uniquely determined once θ_2 has been chosen. The brute force algorithm then takes the form

1. Choose a sequence of ranges $\{\theta_{2j}\}_{j=1}^{n_{bf}} \subset \mathbb{R}_+$ and calculate the sequence $\{K_{\theta_{2j}}\}_{j=1}^{n_{bf}}$. Calculate $\{\theta_{1j}\}_{j=1}^{n_{bf}} = \{\langle K_{\theta_2}, S\rangle_F \|K_{\theta_{2j}}\|_F^{-2}\}_{j=1}^{n_{bf}}$.
2. Evaluate the sequence $\{\|\theta_{1j}K_{\theta_{2j}} - S\|_F^2\}_{j=1}^{n_{bf}}$ and determine the parameters $\hat{\theta_1}^{\text{opt}}, \hat{\theta_2}^{\text{opt}}$ leading to its least value. They are approximators for the optimal tuple $\theta^{\text{opt}} = (\theta_1^{\text{opt}}, \theta_2^{\text{opt}})$.
3. Repeat step 1 and 2 with a new sequence of ranges $\{\theta_{2j}\}_{j=1}^{n_{bf}}$ that is centered around $\hat{\theta_2}^{\text{opt}}$.

Although not very involved, this procedure did lead reliably to parameter estimates arbitrarily close to the optimal ones for the data we tested it on. We observed the existence of a well defined global optimum of $\|\theta_1 K_{\theta_2} - S\|_F$ w.r.t θ_2 in all cases where S had more than 1 element.

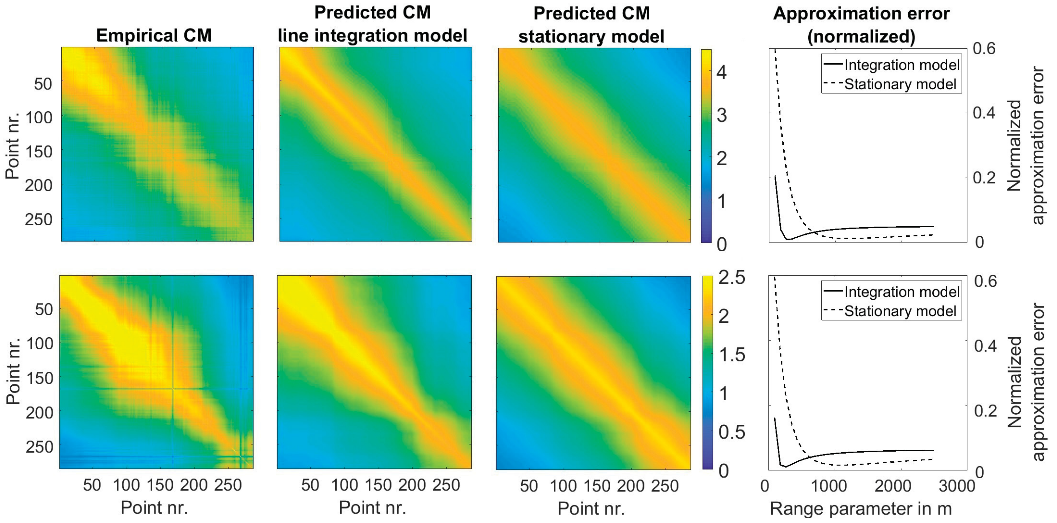


Figure 5.24: The best choices of the range parameters are relatively easy to determine via brute force; the simplicity of the empirical relationships between θ_2 and $\|\theta_1 K_{\theta_2} - S\|_F$ exhibited above are representative for the data we investigated . Notice that the predicted covariance matrices exhibit features that are deviating more from the empirical covariance matrices compared to the Bochner approach but have stronger regularity properties. The third column features a best fit of a stationary covariance model and allows the comparison of the two covariance models $K(\cdot,\cdot) = C(\cdot - \cdot)$ and $K(\cdot,\cdot) = L_{\otimes}C(\cdot - \cdot)$. The results directly imposing a stationary covariance on the APS are overly smooth and (apart from stationarity) show too strong long-distance correlations compared with the observed behavior. The integration model performs better as quantified by the normalized error measure $\epsilon = \|K - S\|_F^2 \|S\|_F^{-2}$. The minimal error $\epsilon_{\min}$ is (0.0092,0.0081) for the two different regions respectively using the line integration model and (0.0162,0.0123) using the stationary model. The underlying covariance function is the exponential one in both cases.

If one abandoned the model $K_P = L_{\otimes}C$ and would instead directly fit one of the aforementioned stationary covariance models $C(\cdot - \cdot) : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$ to the data, a lot of computation will be avoided. The approximation quality, however, is visibly worse. See figure 5.24 for a comparison between the covariance models featuring path integration and those that do not. In reality, the distribution of points from which the empirical covariance matrix is to be generated is less geometrically regular and more scattered. Then the stationary models are even less able to capture the correlation structure of the point set; presumably since the full 3-dimensional variation patterns are not isotropic and the line integration model has a natural built-in anisotropy distinguishing range and azimuth. See figure 5.25 for a concrete example featuring an irregular distribution of points.

The best fitting of the three covariance function models is the squared exponential one under the no-line-integral assumption and the exponential covariance in the line-integral case. This reflects the regularity properties: The APS is rather smooth and this can be either modelled directly by the covariance of a smooth process or by having its derivative at least continuous and having the smoothness emerge during the act of integration.

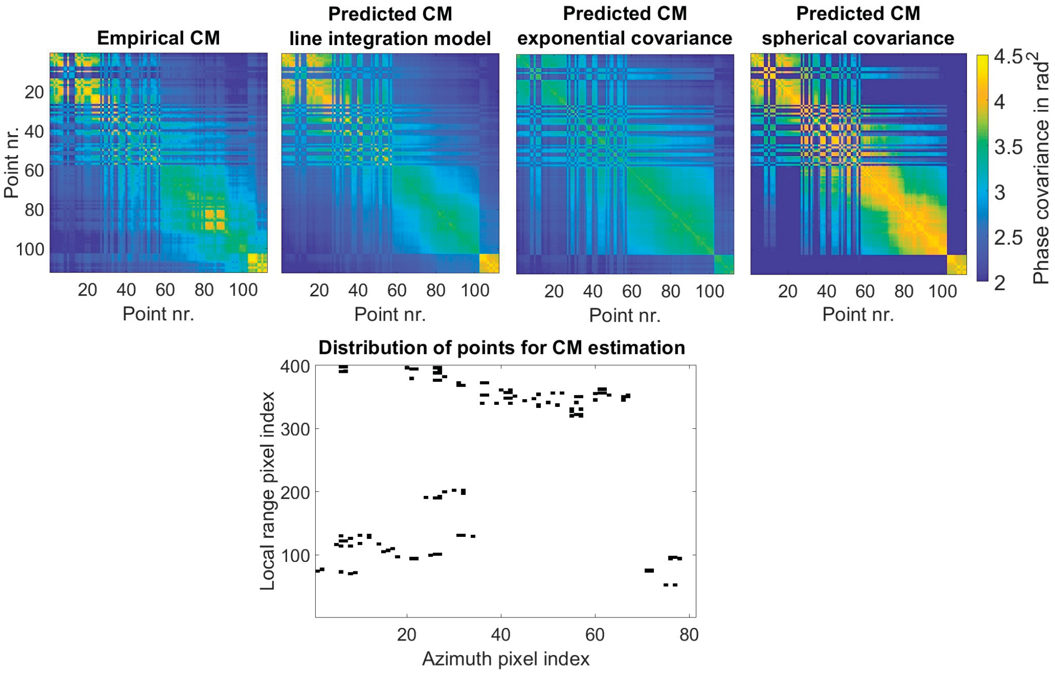


Figure 5.25: The best fits of three different covariance models to an empirical covariance matrix, where the points don't lie on a line. The normalized approximation errors ϵ are (from left to right) 0.007, 0.011 and 0.065 respectively. Line integration was performed against an exponential covariance function to generate the image in the second panel. The bottom row illustrates the underlying geometric distribution of the points whose correlation structure is to be reconstructed. The point enumeration scheme is derived from matrix vectorization, i.e. the ordering is column-by-column and within a column row-by-row.

The most expensive step in the procedure described above is the calculation of the covariance matrices with entries $\varphi_{s_k} \otimes \varphi_{s_l} C(\cdot - \cdot)$. Analytical expressions for these terms are not available and numerical approximation of the integral

$$K_P(s_k, s_l) = \int_{s_0}^{s_k} \int_{s_0}^{s_l} C(r_1 - r_2) dr_1 dr_2 \tag{5.43}$$

is costly in terms of the number of function evaluations, if not a scheme more sophisticated than discretization on a regular grid and subsequent averaging is employed. In one dimension, n-point Gauss quadrature rules of type

$$\int_{-1}^{1} f(u) du \approx \sum_{i=1}^{n} f(u_i) w_i \tag{5.44}$$

are optimal in the sense of being exact for polynomials of degree $2n - 1$, if the u_i are the n roots of the n-th (orthogonal) Legendre polynomial on $[-1, 1]$ and the w_i are the integrals over the associated Lagrange interpolation polynomials [73, § 5.3].

According to a theorem by [78], the w_i can be found efficiently from the eigenvalues and eigenvectors of the matrix J encoding the three term recurrence relation $(k + 1)P_{k+1}(x) = (2k + 1)xP_k(x) - kP_{k-1}(x)$ satisfied by Legendre polynomials P_k in the following way:

$$J = \begin{bmatrix} 0 & \alpha_1 & 0 & \dots & 0 \\ \alpha_1 & 0 & \alpha_2 & 0 & \dots \\ 0 & \alpha_2 & 0 & \alpha_3 & \dots \\ \vdots & 0 & \alpha_3 & \ddots & \dots \\ 0 & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{5.45}$$

$$\alpha_k = \sqrt{\frac{k^2}{(2k+1)(2(k-1)+1)}} = \frac{k}{\sqrt{4k^2-1}} \tag{5.46}$$

$$u_k = k\text{-th eigenvalue of } J \tag{5.47}$$

$$w_k = \int_{-1}^{1} \left(e_k^{(1)}/\|e_k\|\right)^2 du = 2(e_k^{(1)})^2 \tag{5.48}$$

where $e_k^{(1)}$ is the first component of the k-th normed eigenvector e_k of J. For two-dimensional integrals, the tensor product of the n-point Gauss quadrature rule given by specifying evaluation points $u^{\oplus}$ and weights $w^{\otimes}$, while denoting by $\mathbf{1} \in \mathbb{R}^n$ a vector of ones, as

$$\int_{-1}^{1}\int_{-1}^{1} f(u_1,u_2)du_1du_2 \approx \sum_{i=1}^{n}\sum_{j=1}^{n} f(u_{ij}^{\oplus})w_{ij}^{\otimes} \tag{5.49}$$

$$u^{\oplus} = (\mathbf{1} \otimes \{u_i\}_{i=1}^{n}) \oplus (\{u_k\}_{j=1}^{n} \otimes \mathbf{1}) \in \mathbb{R}^{2n^2} \tag{5.50}$$

$$w^{\otimes} = \{w_i\}_{i=1}^{n} \otimes \{w_j\}_{j=1}^{n} \in \mathbb{R}^{n^2} \tag{5.51}$$

is not strictly optimal any more. However, it turns out to approximate double integrals to sufficient numerical accuracy for our purpose with only a modest amount of function evaluations. The appropriate coordinate diffeomorphism[5] $\phi : U = [-1,1]^2 \to [0,1]^2 = V$ inducing a pullback ϕ^* mapping integral 5.43 to integral 5.49 can be given by

$$\phi(u_1,u_2) = \begin{bmatrix} v_1(u_1,u_2) \\ v_2(u_1,u_2) \end{bmatrix} = \begin{bmatrix} \frac{1}{2}u_1 + \frac{1}{2} \\ \frac{1}{2}u_2 + \frac{1}{2} \end{bmatrix}. \tag{5.52}$$

With this definition and setting $u = (u_1,u_2), v = (v_1,v_2)$ the following sequence of integral substitutions holds:

$$\begin{aligned} K_P(s_k,s_l) &= \int_{s_0}^{s_k}\int_{s_0}^{s_l} C(r_1 - r_2)dr_1dr_2 \qquad (5.53) \\ &= c \iint\limits_{[0,1]^2} C(r_1(v_1,v_2) - r_2(v_1,v_2))dv_1dv_2 \end{aligned}$$

[5] A diffeomorphism is a differentiable one-to-one map with differentiable inverse.

$$
\begin{aligned}
&= c \iint\limits_{\phi^{-1}[0,1]^2} \phi^* C(r_1(v) - r_2(v)) \begin{vmatrix} \frac{\partial v_1}{\partial u_1} & \frac{\partial v_1}{\partial u_2} \\ \frac{\partial v_2}{\partial u_1} & \frac{\partial v_2}{\partial u_2} \end{vmatrix} du_1 du_2 \\
&= \frac{c}{4} \iint\limits_{[-1,1]^2} C(r_1(v(u)) - r_2(v(u))) du_1 du_2 \qquad (5.54)
\end{aligned}
$$

where $\phi^* : (V \to \mathbb{R}) \to (U \to \mathbb{R})$ is the usual pullback on functions [62, pp. 22 -25] and $c = \|s_k - s_0\|_{\ell^2} \|s_l - s_0\|_{\ell^2}$. Ultimately, the cubature rule yields the expression:

$$
K_P(s_k, s_l) = \frac{c}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} C\left(r_1(v(u_{ij}^{\oplus})) - r_2(v(u_{ij}^{\oplus}))\right) w_{ij}^{\otimes} \qquad (5.55)
$$

Here $C(\cdot - \cdot)$ might be any positive definite function; we have chosen it to be an exponential kernel. The error in the one dimensional case is bounded by the $2n$-th derivative $\frac{\partial^{2n}}{\partial u^{2n}} C(r_1(v(u)) - \cdot)$ [90, p. 325] implying it to be non-negligible for rapidly decaying covariance functions C. In our experiments, this did not seem to be of much concern, as the optimally inferred $K^s_{\Delta P}$ all exhibited smooth behavior, correlation lengths between 200 m and 1500 m and their integrals were well approximated using the tensor product of a 5-point Gauss rule. In comparing the 5-point Gauss rule to the trivial integration procedure consisting of taking the average of 10.000 function evaluations on a regular grid, one finds both method's approximation qualities to be quite similar. The trade-off for the potential (but in our experiments not significant) loss in numerical accuracy is an increase in speed of up to a factor of 400 for an interferogram with 1000 persistent scatterers and 300.000 pixels for which estimation is to be performed. This reduced the time for the calculation of the covariance matrices to 10 minutes on an office computer with 32.0 GB RAM and a 3.50 GHz processor. Once the covariance matrices have been calculated, APS estimation for a single interferogram can be executed in less than 1 second. Nonetheless, we suggest to recalculate the covariance matrices in regular intervals of several days to avoid long term changes in meteorological correlations to negatively impact the estimations although further investigations would be needed to derive a reliable rule of thumb.

§ Kernel Inference

Inferring a suitably flexible kernel from data is exactly the task, for which the algorithm described as 'nonparametric kernel inference with affine constraints' explained on page 193 was designed. In the case of the APS in TRI, the kernel

$$
K(\cdot, \cdot) : S \times S \ni (s_i, s_j) \mapsto K(s_i, s_j) \in \mathbb{R}, \quad s_i, s_j \in \mathbb{R}^3
$$

of the APS is to be inferred on the index set of spatial coordinates S, whose exact format is irrelevant for but will occasionally be assumed to consist of local cartesian coordinates centered around the instrument's position for the sake of simplicity. The observations are given as a sequence of interferograms evaluated at stable points $\{s_i\}_{i=1}^n \in \mathbb{R}^3$ with good signal-to-noise ratio to guarantee that the empirical covariance matrix (also denoted as S) is formed solely by values of realizations $P_\omega(\cdot), \omega \in \Omega$ where $P_.(\cdot)$ is the random function corresponding to the APS.

To keep arrangements simple and runtimes short, we will for now neither include any affine constraints nor give too much weight to the prior guess for the covariance function. To get an initial orthonormal basis $\{\varphi_i\}_{i=1}^{n_{\text{exp}}}, \varphi_i : \mathbb{R}^3 \mapsto \mathbb{R}$, set

$$K_{\text{prior}}(s_i, s_j) = K_{sq}^x(x(s_i), x(s_j))K_{sq}^y(y(s_i), y(s_j))K_{sq}^z(z(s_i), z(s_j))$$

where K_{sq}^x is the squared exponential kernel on the x-coordinates of the three-dimensional coordinates s_i, s_j and analogously for K_{sq}^y, K_{sq}^z. Employing the individual Mercer decompositions

$$K_{sq}^u(\cdot, \cdot) = \sum_{i=1}^{\infty} \lambda_i^u \varphi_i^u(\cdot) \otimes \varphi_i^u(\cdot), \quad u = x, y, z$$

allows to write

$$K_{\text{prior}}(\cdot, \cdot) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(\cdot) \otimes \varphi_i(\cdot)$$

where λ_i is the i-th element of the sequence $\{\lambda_j^x \lambda_k^y \lambda_l^z\}_{j,k,l=1}^{\infty}$ sorted in descending order and $\varphi_i(\cdot) = \varphi_j^x(x(\cdot)) \otimes \varphi_k^y(y(\cdot)) \otimes \varphi_l^z(z(\cdot))$ is the tensor product of the eigenfunctions corresponding to the eigenvalues $\lambda_j^x, \lambda_k^y, \lambda_l^z$ that form λ_i. Apart from providing the orthonormal basis, the prior will be made rather unimportant by setting to zero the parameter r in the objective function [6]

$$L(\gamma) = \log |C_\gamma| + \text{tr}\left(SC_\gamma^+\right) + r\left[-\log|\gamma| + \text{tr}\left(\Lambda^+\gamma\right)\right]$$
$$C_\gamma = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij} \psi_i \otimes \psi_j^*, \quad \psi_i \in \mathbb{R}^n, (\psi_i)_k = \varphi_i(s_k).$$

This sidesteps the need for an iteration and in accordance to the remarks in subsection 4.3.2, page 184, the explicit formula 5.56 is found.

$$\gamma = S_\psi = (\Psi)^+ S(\Psi^*)^+, \quad \Psi = [\psi_1, ..., \psi_{n_{\text{exp}}}] \in \mathbb{R}^n \otimes \mathbb{R}^{n_{\text{exp}}} \tag{5.56}$$

The solution is then $K(\cdot, \cdot) = \sum_{i,j=1}^{n_{\text{exp}}} \gamma_{ij} \varphi_i(\cdot) \varphi_j(\cdot)$ and the predicted covariance matrix is

$$C_\gamma = \sum_{i,j=1}^{n_{\text{exp}}} \psi_i \otimes \psi_j^* = \Psi \gamma \Psi^* = (\Psi\Psi^+) S (\Psi\Psi^+)^*$$

which is $P_\psi S P_\psi^*$ with P_ψ the orthogonal projection onto the range of Ψ. The results of this procedure are plotted in figure 5.26 as well as the results obtained when

[6] Note that γ is a positive semidefinite coefficient matrix in the equations surrounding 5.56 and not, as in the next paragraph, the coherence.

different amounts of regularization are used.

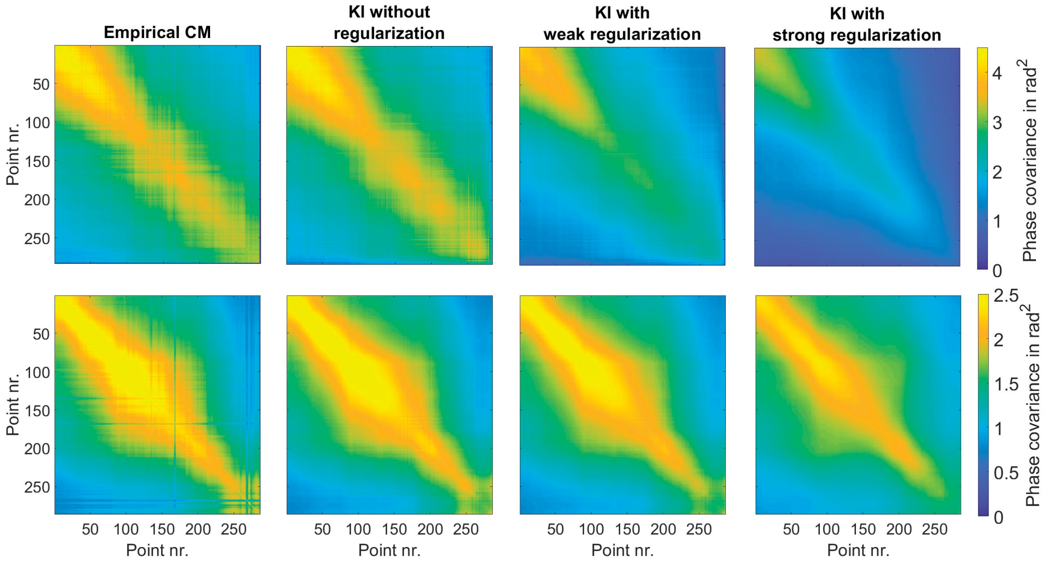


Figure 5.26: The results of kernel inference for the same empirical covariance matrices S_1 and S_2 as in figure 5.22 with and without regularization. When prior knowledge is ignored, the reconstructive performance of kernel inference exceeds that of both the parametric and the Bochner-type approach. Imposing stronger requirements on a kernel's regularity discourages sharp changes in the associated covariance matrices; care must be taken to ensure that this effect is avoided if improper (first row). Due to numerical problems, the regularization was introduced by means of moving along the geodesic path connecting S_ψ and Λ as parametrized in [96].

Remark Up until now only the negative log likelihood was used to assess the quality of the approximation. This is not wrong per se but incomplete as any kernel being identical to the empirical covariance matrix S at the measurement points has the best score under this metric. But these kernels obviously do not have the best generalization performance, so the negative log likelihood should be seen more as a measure of a models flexibility to emulate the structure encountered in the matrix S rather than a definite performance indicator. As discussed before, a kernel's suitability for purposes of inference and prediction depend also on that kernel's regularity. The only procedures guaranteeing this are kernel inference and the parametric inference scheme. For reasons of simplicity, the latter approach will be adopted for the practical implementations presented later during this chapter.

§ Estimation of noise variance

Typically one links coherence, phase noise variance and signal-to-noise ratio [91, p. 98]. The classical estimator for the coherence employs spatial averages and is, as already mentioned in subsection 5.1.2,

$$\hat{\gamma}_{\text{spatial}} = \frac{\sum_{j \in Nbd} z_1^j ({z_2}^j)^*}{\sqrt{\sum_{j \in Nbd} |z_1^j|^2 \sum_{j \in Nbd} |z_2^j|^2}} \tag{5.57}$$

for a certain spatial neighborhood Nbd of a fixed point. In this formula z_i^j denotes the complex number encoding amplitude and phase of the the j-th pixel in the i-th interferogram z_i and ergodicity is assumed to hold [91, p. 96]. In the privileged situation where TRI measurements are dense in time, one can estimate the complex correlation coefficient γ at a certain fixed pixel location j (dropped from the notation for economy of presentation) as the normalized off-diagonal entries of the complex covariance matrix

$$\begin{aligned} \Sigma_{[z_1,z_2]^T} &= E\left[\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \otimes \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}^*\right] \in \mathbb{C}^2 \times \mathbb{C}^2 \\ \hat{\Sigma}_{[z_1,z_2]^T} &= \frac{1}{n-1} \sum_{i=1}^{n-1} \left[\begin{bmatrix} z_i \\ z_{i+1} \end{bmatrix} \otimes \begin{bmatrix} z_i^* & z_{i+1}^* \end{bmatrix}\right] \end{aligned}$$

for $z_i := z_i^j; i = 1, ..., n$ denoting the complex number of the pixel at location j in the i-th epoch. This immediately leads to the alternative expression

$$\hat{\gamma}_{\text{temporal}} = \frac{n}{n-1} \frac{\sum_{i=1}^{n-1} z_i z_{i+1}^*}{\sum_{i=1}^{n} |z_i|^2}. \tag{5.58}$$

The interpretation of equation 5.58 is different from the interpretation of the usual estimator 5.57. Both estimation formulas for γ take on high values if an image patch in image 2 is a scaled and phase-shifted version of the same patch in image 1. But whereas 5.57 takes on low values if the interferometric phase varies randomly in a spatial image patch, expression 5.58 takes on low values only if the interferometric phase varies randomly in time. In the presence of other temporally high frequent sources of strong phase variation like atmospheric influences, the estimator 5.58 does not deliver a suitable indicator for phase reliability of individual pixels whereas 5.57 still does.

From a purely practical point of view, the assumptions regarding the pure noise component (see page 237) permit to estimate the phase noise as that part of the measurements which exhibits low temporal and low spatial correlation to its immediate vicinity. The author proposes the following estimators for noise and noise variance at location s_k.

$$N(s_k) = M(s_k) - \frac{1}{|Nbd(k)|} \sum_{j \in Nbd(k)} M(s_j) \tag{5.59}$$

$$\sigma^2_{N(s_k)} = \frac{1}{n_t} \sum_{l=1}^{n_t} \left[M(s_k, t_l) - \frac{1}{|Nbd(k)|} \sum_{j \in Nbd(k)} M(s_j, t_l) \right]^2 \tag{5.60}$$

where $Nbd(k)$ is a set containing the indices j of the locations s_j that form the spatial neighborhood of the location s_k.

5.2.5 Spatiotemporal deformation estimation

Given a Hilbert space model $\mathcal{H}_M$ for the measurements that decomposes into direct sums of Hilbert spaces $\mathcal{H}_D, \mathcal{H}_A, \mathcal{H}_N$, estimating deformations d when only a superposition of deformation, atmosphere and noise is measured can be formulated as the task of finding a certain spatiotemporal smoothing spline σ_d. The corresponding norm-minimization problem is at the same time a maximum likelihood problem when the kernels K_D, K_A, K_N are chosen to reflect the different phenomena's covariance functions. Several different formulations of this estimation problem are possible and correspond e.g. to interpolating the APS given some of its values, solving spatial and temporal filtering separately, or doing full spatiotemporal signal separation. Briefly surveying these formulations in order of ascending generality and focusing on computability leads to the tensor-spline construction explained in subsection 3.1.3. Assuming second order stationarity of all quantities involved yields immediate results; implementing the more justified line integral covariance model requires methods of numerical integration. The end result is an algorithm the author terms TRI-MAPS, where the latter part is an acronym for 'mitigation of atmospheric phase screen'. It produces three sequences of two-dimensional maps depicting deformation, atmosphere, and noise.

Suppose the Hilbert space model $\mathcal{H}_M = \mathcal{H}_D \oplus \mathcal{H}_P \oplus \mathcal{H}_N$ is given and the kernels K_D, K_P and K_N determining the reproducing kernel Hilbert spaces of deformations, atmospheric phase screen and noise are known. Without specifying a priori if the domain U of the functions in these four RKHS is space, time or spacetime, the following general interpolating spline contains all subcases relevant for deformation estimation.

$$\sigma_m = \underset{m \in A^{-1}a}{\operatorname{argmin}} \ \|m\|^2_{\mathcal{H}_M} \tag{5.61}$$

where $A : \mathcal{H}_M \to \mathbb{R}^n$ is the measurement operator of evaluation at $\{u_j\}_{j=1}^n \subset U$ and decomposes as $A = A_D \oplus A_P \oplus A_N, A_X : \mathcal{H}_X \to \mathbb{R}^n, X \in \{D, P, N\}$. The data is provided as an n-dimensional vector $a \in \mathbb{R}^n$ whose exact nature depends on that of U. By the remarks in subsection 3.1.2, the smoothing splines $\sigma_d, \sigma_p, \sigma_n$ corresponding to optimal estimators of deformation, APS, and noise can be derived via orthogonal projection of σ_m onto the respective subspaces $\mathcal{H}_X \overline{\boxslash} \mathcal{H}_M, X \in \{D, P, N\}$. This means

$$\sigma_d = \underset{d \in \mathcal{H}_D}{\operatorname{argmin}} \ \|A_D d - a\|^2_{A(\mathcal{H}_M/\mathcal{H}_D)} + \|d\|^2_{\mathcal{H}_D} \qquad = \Pi_{\mathcal{H}_D} \sigma_m \tag{5.62}$$

$$\sigma_p = \underset{d \in \mathcal{H}_D}{\operatorname{argmin}} \ \|A_P p - a\|^2_{A(\mathcal{H}_M/\mathcal{H}_P)} + \|p\|^2_{\mathcal{H}_P} \qquad = \Pi_{\mathcal{H}_P} \sigma_m \tag{5.63}$$

$$\sigma_n = \underset{d \in \mathcal{H}_D}{\operatorname{argmin}} \ \|A_N n - a\|^2_{A(\mathcal{H}_M/\mathcal{H}_N)} + \|n\|^2_{\mathcal{H}_N} \qquad = \Pi_{\mathcal{H}_N} \sigma_m \tag{5.64}$$

where the Hilbert spaces $A(\mathcal{H}_M/\mathcal{H}_X)$ have kernel $(A \otimes A)[K_{Y_1}(\cdot,\cdot) + K_{Y_2}(\cdot,\cdot)] \in \mathbb{R}^n \otimes \mathbb{R}^n$ with Y_1, Y_2 the elements of the set $\{D, P, N\}/\{X\}$.

The probabilistic interpretation is simply that the objective functions in equations 5.62 to 5.64 are the negative log likelihoods of Gaussian processes whose minimizers correspond to those elements σ_x in an RKHS $\mathcal{H}_X$ of functions that maximize jointly both the likelihood of σ_x and of the residuals $A\sigma_x - a$. It was mentioned in chapter 3 that given a certain set of assumptions, splines σ_x minimize the expected square loss and as such are analogues to conditional expectations $E[X(\cdot)|Am = a] =: E[X|a]$. Before any processing happens, a decision has to be made about which data to include for the estimation step and which data to discard. Ignoring data corresponds to the strong assumption that asserts this data's lack of informativity for the specific estimation purpose. This can be made formal using the concept of conditional independence of random variables.

Definition 5.2.1 Two random variables M_1 and M_2 are conditionally independent given a third random variable M_3 if

$$f_{M_1,M_2|M_3}(m_1, m_2|m_3) = f_{M_1|M_3}(m_1|m_3)f_{M_2|M_3}(m_2|m_3)$$

where f_Y is the probability density function of expression Y. This is written as $M_1 \coprod M_2|M_3$. M_1, M_2, M_3 can also be sets of random variables [116, p. 24].

Theorem 5.2.2 *Conditional independence has the following three properties.*

Weak union: $M_1 \coprod M_2, M_3|M_4 \Rightarrow M_1 \coprod M_2|M_3, M_4$

Decomposition: $M_1 \coprod M_2, M_3|M_4 \Rightarrow M_1 \coprod M_2|M_4$

Intersection: $(M_1 \coprod M_2|M_3, M_4) \wedge (M_1 \coprod M_4|M_2, M_3) \Rightarrow M_1 \coprod M_2, M_4|M_3$

The original statements including details on their interpretation and their proofs can be found in the literature concerned with graphical models and causality, e.g. in [116, p. 25]. Theorem 5.2.2 can be used to prove that from $X \coprod Y_1, ..., Y_n|Z_1, ..., Z_m$ it follows that

$$f_{X|Y_1,...,Y_n,Z_1,...,Z_m}(x|y_1, ..., y_n, z_1, ..., z_m) = f_{X|Z_1,...,Z_m}(x|z_1, ..., z_m)$$

whenever the conditional distribution is well defined. Renaming $Y_1, ..., Y_n$ to Y, $Z_1, ..., Z_m$ to Z and compressing the subscript indicators for the probability density functions and the notation for any realizations in the same way, this can be deduced from

$$f_{X|Y,Z}(x|y, z) = \frac{f_{X,Y|Z}(x, y|z)}{f_{Y|Z}(y|z)} \overset{X \coprod Y|Z}{=} \frac{f_{X|Z}(x, z)f_{Y|Z}(y|z)}{f_{Y|Z}(y|z)} = f_{X|Z}(x|z)$$

when the denominator is nontrivial. This subsequently implies that the conditional expectation $E[X|Y = y, Z = z] =: E[X|y, z]$ can be written as

$$\begin{aligned} E[X|y_1, ..., y_n, z_1, ..., z_m] &= \int_\Omega X(\omega)f_{X|Y,Z}(x|y_1, ..., y_n, z_1, ..., z_m)d\omega \\ &= \int_\Omega X(\omega)f_{X|Z}(x|z_1, ..., z_m)d\omega \\ &= E[X|z_1, ..., z_m] \end{aligned} \tag{5.65}$$

Recalling that $E[X|z_1, ..., z_m]$ and $E[X|y_1, ..., y_m, z_1, ..., z_m]$ as conditional expectations are in the Gaussian case the typical Kriging estimators, this allows one to conclude that data $y_1, ..., y_n$ can be ignored for optimal estimation under the specific assumption of $X \coprod Y_1, ..., Y_n | Z_1, ..., Z_m$.

As concrete examples, consider two special cases. Let $M_{\bar{s},\bar{t}} := \{M(s_i, t_j)\}_{i=1,j=1}^{n_s n_t}$ be the set of all random variables associated to measurements $M : \Omega \times S \times T \to \mathbb{R}$. For a specific spatial position $s \in S$, denote by $M_{s,\bar{t}} := \{M(s, t_j)\}_{j=1}^{n_t}$ the 'pillar' of all measurements at different times $t_1, ..., t_n$ at location $s \in S$ and denote by $M_{\bar{s},t} := \{M(s_i, t)\}_{i=1}^{n_s}$ the 'slice' of all measurements at different locations taken at time $t \in T$. Suppose that for all $s \in S$ it holds that

$$M_{s,t} \coprod \frac{M_{\bar{s},\bar{t}}}{M_{\bar{s},t}} \mid \frac{M_{\bar{s},t}}{M_{s,t}} \qquad (5.66)$$

where the fraction notation is to be interpreted as set subtraction, i.e. $\frac{X}{Y} = X \setminus Y$.

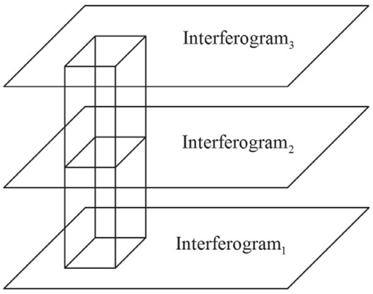


Figure 5.27: Different ways to cluster sequences of interferograms

Expression 5.66 then means that a specific random variable $M_{s,t}$ is independent of all other random variables $M_{\bar{s},\bar{t}} \setminus M_{\bar{s},t}$ that lie in different time 'slices' (interferograms) given knowledge of all other measurements $M_{\bar{s},t} \setminus M_{s,t}$ in the time slice that contains $M(s, t)$. By making recursive use of the intersection property and the symmetry of conditional independence, one then finds

$$\left(\frac{M_{\bar{s},\bar{t}}}{M_{\bar{s},t}} \coprod M_{s_1,t} \mid \frac{M_{\bar{s},t}}{M_{s_1,t}} \right) \wedge ... \wedge \left(\frac{M_{\bar{s},\bar{t}}}{M_{\bar{s},t}} \coprod M_{s_n,t} \mid \frac{M_{\bar{s},t}}{M_{s_n,t}} \right)$$
$$\Rightarrow M_{s_j,t} \coprod \frac{M_{\bar{s},\bar{t}}}{M_{\bar{s},t}} \quad \text{for } j = 1, ..., n \ .$$

This implies that time slices are independent of each other and optimal estimation of $M_{s,t}$ via the conditional expectation $\hat{M}_{s,t} = E[M_{s,t} | M_{\bar{s},\bar{t}} \setminus M_{s,t}]$ can be reduced to $\hat{M}_{s,t} = E[M_{s,t} | M_{\bar{s},t} \setminus M_{s,t}]$ meaning that the interferogram $M_{\bar{s},t}$ at time t contains all relevant information about $M_{s,t}$. Analogously it is possible to derive

$$M_{s,t_j} \coprod \frac{M_{\bar{s},\bar{t}}}{M_{s,\bar{t}}} \quad \text{for } j = 1, ..., n$$

based on the assumption $M_{s,t} \coprod \frac{M_{\bar{s},\bar{t}}}{M_{s,\bar{t}}} \mid \frac{M_{s,\bar{t}}}{M_{s,t}}$. This corresponds to the assertion that the time evolution of a single pixel contains all information necessary to characterize that pixel thereby reducing a spatiotemporal estimation to a purely temporal one:

$$\hat{M}_{s,t} = E[M_{s,t} | M_{\bar{s},\bar{t}} \setminus M_{s,t}] = E[M_{s,t} | M_{s,\bar{t}} \setminus M_{s,t}]$$

The assumptions are in general too strong to be accepted; nonetheless they are implicit when instead of the full spatiotemporal situation only spatial or temporal aspects are respected. In the same spirit, it will be shown that for example the stacking of interferograms is stochastically optimal only under the hypothesis that the APS behaves as spatiotemporal white noise and the deformation is constant.

Given a sequence of unwrapped interferograms interpreted as a function from $S \times T \to \mathbb{R}$ and $n_\sharp = n_s n_t$ measurements where n_s is the number of pixels per interferogram and n_t is the number of interferograms, define

A^s, A^t the spatial and temporal evaluation operators for locations $\{s_i\}_{i=1}^{n_s}$ or times $\{t_j\}_{j=1}^{n_t}$ respectively.

$A = A^s \otimes A^t$ the spatiotemporal evaluation operator $Af = \{f(s_i, t_j)\}_{i=1,j=1}^{n_s,n_t}$.

$B\mathcal{H}_K$ the RKHS with reproducing kernel $B \otimes BK$ for any bounded linear operator B.

$\mathcal{H}_{K_M}/\mathcal{H}_{K_X}$ the quotient RKHS with reproducing kernel equal to $K_M - K_X$ where $\mathcal{H}_{K_M} = \mathcal{H}_{K_Y} \oplus \mathcal{H}_{K_X}$.

$\mathcal{H}_X^\otimes = \mathcal{H}_X^s \otimes \mathcal{H}_X^t$ the tensor product RKHS with reproducing kernel $K_X^s K_X^t$.

In general all involved quantities receive the additional superscript s or t if their explicitly spatial or temporal nature is to be highlighted. The special case where only one interferogram is used for estimation amounts to fixing a certain time t_0 in the spatiotemporal problem

$$\sigma_d = \underset{d \in \mathcal{H}_D^\otimes}{\operatorname{argmin}} \ \|Ad - a\|^2_{A\mathcal{H}_M^\otimes/\mathcal{H}_D^\otimes} + \|d\|^2_{\mathcal{H}_D^\otimes}, \tag{5.67}$$

restricting the involved operators accordingly and solving the minimization problem only for $t = t_0$. The estimator for the deformation is then

$$\sigma_d^s = \underset{d^s \in \mathcal{H}_D^s}{\operatorname{argmin}} \|A^s d^s - a^s\|^2_{A^s\mathcal{H}_P^s \oplus A^s\mathcal{H}_N^s} + \|d^s\|^2_{\mathcal{H}_D^s} \tag{5.68}$$

$$= \underbrace{[(K_D^s(s_i, \cdot))_{i=1}^{n_s}]^T \left[((K_M^s)(s_i, s_j))_{i,j=1}^{n_s}\right]^{-1}}_{\Xi^s : \mathbb{R}^{n_s} \to \mathcal{H}_D^s} a^s \tag{5.69}$$

Equivalently for a fixed location $s = s_0$ the solution to a temporal filtering problem can be written analogously with obvious change of notations.

$$\sigma_d^t = \underset{d^t \in \mathcal{H}_D^t}{\operatorname{argmin}} \|A^t d^t - a^t\|^2_{A^t\mathcal{H}_P^t \oplus A^t\mathcal{H}_N^t} + \|d^t\|^2_{\mathcal{H}_D^t} \tag{5.70}$$

$$= \underbrace{\left[\left(K_D^t(t_i, \cdot)\right)_{i=1}^{n_t}\right]^T \left[\left((K_M^t)(t_i, t_j)\right)_{i,j=1}^{n_t}\right]^{-1}}_{\Xi^t : \mathbb{R}^{n_t} \to \mathcal{H}_D^t} a^t \tag{5.71}$$

By passing from the assumption $d(\cdot) \in \mathcal{H}_D$ with RK K_D to the less general assumption $d(\cdot) \in \mathcal{H}_D^s \otimes \mathcal{H}_D^t =: \mathcal{H}_D^\otimes$, i.e. $d(s_0, t_0) = \sum_{i=1}^\infty d_i^s(s_0) \otimes d_i^t(t_0), \|d\|_{\mathcal{H}_D^\otimes} < \infty, d^s \in \mathcal{H}_D^s, d^t \in \mathcal{H}_D^t$, equations 5.68 and 5.70 can be joined to form the spatiotemporal tensor spline equation 5.72

$$\begin{aligned}\sigma_d &= \Xi^s \otimes \Xi^t a \\ &= \underset{d\in\mathcal{H}_D^{\otimes}}{\operatorname{argmin}} \|A^s \otimes A^t d - a\|^2_{\bigotimes_{i\in\{s,t\}} A^i\mathcal{H}_P^i \oplus A^i\mathcal{H}_N^i} + \|d\|^2_{\mathcal{H}_D^{\otimes}} \\ &+ \|A^s \otimes \mathrm{id}_{\mathcal{H}_D^t}\, d\|^2_{A^s\mathcal{H}_D^s \otimes \mathcal{H}_D^t} + \|\,\mathrm{id}_{\mathcal{H}_D^s} \otimes A^t d\|^2_{\mathcal{H}_D^s \otimes A^t\mathcal{H}_D^t}.\end{aligned} \tag{5.72}$$

Theorem 81.3 in [21] together with the abstract spline representation of the solution in terms of superpositions of kernel functions shows that, given trivial nullspace $\ker(B) = \{0\}$, equation 5.72 is equivalent to the following expression.

$$\begin{aligned}\sigma_d &= \Xi^s \otimes \Xi^t a \\ &= \underset{d\in\mathcal{H}_D^{\otimes}}{\operatorname{argmin}} \|A^s \otimes A^t d - a\|^2_{A_{M/D}\mathcal{H}_M^{\otimes}/\mathcal{H}_D^{\otimes}} + \|d\|^2_{\mathcal{H}_D^{\otimes}} \\ \mathcal{H}_M^{\otimes} &= (\mathcal{H}_D^s \oplus \mathcal{H}_P^s \oplus \mathcal{H}_N^s) \otimes \left(\mathcal{H}_D^t \oplus \mathcal{H}_P^t \oplus \mathcal{H}_N^t\right)\end{aligned} \tag{5.73}$$

The interpretation is as for normal abstract splines. The first term measures fidelity to the data by quantifying the likelihood of the gap between predicted values $A^s \otimes A^t d$ and an actually observed a. The second term assesses the likelihood of d itself by smoothness criteria derived from assumptions on the function space $\mathcal{H}_D^{\otimes}$.

§ Special cases

Three special cases are particularly useful in applications. Although the estimators derived in this subsection are not strictly optimal in the sense defined above, they share the property of being practically feasible and can circumvent problems arising due to apriori unknown kernels.

Case 0: Deformations constant, APS white noise

This situation is purely hypothetical and serves more as an example than a practically relevant set of conditions and solutions. If $\mathcal{H}_D^{\otimes} = \mathcal{H}_D^s \otimes \mathcal{H}_D^t$ with $\mathcal{H}_D^s$ spatially uncorrelated white noise and $\mathcal{H}_D^t$ the semi-Hilbert space of constants with semi-reproducing kernel $K_D^t(t_i, t_j) = 1$, $\mathcal{H}_P^{\otimes} = \mathcal{H}_P^s \otimes \mathcal{H}_P^t$ is spatiotemporally white noise and $\mathcal{H}_N = \emptyset$, then

$$\begin{aligned}\sigma_d &= \Xi^s \otimes \Xi^t a \\ &= \underset{d\in\mathcal{H}_D^{\otimes}}{\operatorname{argmin}}\ \|A^s \otimes A^t d - a\|^2_{A\mathcal{H}_P} + \underbrace{\|d\|^2_{\mathcal{H}_D^{\otimes}}}_{0} \\ &= \underset{d\in\mathcal{H}_D^{\otimes}}{\operatorname{argmin}}\ \sum_{i=1}^{n_s}\sum_{j=1}^{n_t} (d(s_i,t_j) - a_{ij})^2 \qquad\qquad d \in \mathcal{H}_D^{\otimes} \Rightarrow d(s_i,t_j) = d(s_i) \\ &= \underset{d=\{d(s_i)\}_{i=1}^{n_s}}{\operatorname{argmin}}\ \sum_{i=1}^{n_s}\left(\sum_{j=1}^{n_t}(d(s_i) - a_{ij})^2\right)\end{aligned} \tag{5.74}$$

From the last line it follows that $d(s_i,t) = (n_t)^{-1}\sum_{j=1}^{n_t} a_{ij}$ which is the temporal average for each pixel s_i individually. This confirms the initial considerations that established a purely temporal approach as sufficient when no spatial dependence can be found between any of the random variables.

Case 1: Deformations confined in space

When no valid information about the expected smoothness of potential deformations is available at all and K_D therefore unknown, estimation of $d(\cdot)$ is still possible when the deformations are known to be zero outside a fixed spatial domain $V \subset S$. Given this information, σ_{d+p} reduces to σ_p in $(S \cap V^C) \times T$. It is then natural to calculate σ_p based on data available at spacetimes $(S_M \cap V^C) \times T_M$ and estimate $d(\cdot)$ by $\hat{d}(\cdot) = \sigma_m - \sigma_p - \sigma_n$ on $(S_M \cap V) \times T_M$. A subscript M indicates the restricted sets of positions and times, for which measurements are available. Since $n(\cdot)$ is assumed to be uncorrelated noise with mean zero and $m(\cdot)$ was measured directly this simplifies to

$$\hat{d}(\cdot) = m(\cdot) - \sigma_p(\cdot) \tag{5.75}$$

σ_p can be calculated with aid of equation 5.73 by setting σ_d to zero.

$$\begin{aligned} \sigma_p(\cdot) &= \Xi_P^s \otimes \Xi_P^t a \qquad (5.76)\\ \Xi_P^s &= \left[(K_P^s(s_i,\cdot))_{i=1}^{n_s}\right]^T \left[((K_P^s + K_N^s)(s_i,s_j))_{i,j=1}^{n_s}\right]^{-1} \\ \Xi_P^t &= \left[\left(K_P^t(t_i,\cdot)\right)_{i=1}^{n_t}\right]^T \left[\left((K_P^t + K_N^t)(t_i,t_j)\right)_{i,j=1}^{n_t}\right]^{-1} \end{aligned}$$

For practical evaluation it will be necessary to use a tensor representation of the data a, achievable e.g. via singular value decomposition, to write explicitly

$$\sigma_p(\cdot) = \sum_{j=1}^{\min(n_s,n_t)} \Xi_P^s a_j^s \otimes \Xi_P^t a_j^t. \tag{5.77}$$

K_P and K_N need to be estimated from the data, which is to be restricted to $(S \cap V^C) \times T$ first. The complete implementation consists of the three steps

1. Choice of stable persistent scatterers $\{s_i\}_{i=1}^{n_s} \subset S \cap V^C$.
2. Inferring kernels K_P and K_N from data.
3. Performing APS interpolation and denoising in $S \cap V$

As all points in $S \cap V^C$ are stable by assumption, step 1 reduces to finding a suitable metric to evaluate the noise level associated to a pixel. The amplitude dispersion index (ADI) [61]

$$ADI(s) = \frac{\mu(Amp(s))}{\sigma(Amp(s))} \tag{5.78}$$

detailing the ratio of expected amplitudes $\mu(Amp)$ to amplitude standard deviations $\sigma(Amp)$. is a useful measure in this regard since high, stable amplitudes indicate the presence of a single object in the resolution cell that dominates the backscattered signal. Additions of smaller complex numbers corresponding to random processes affecting the reflected wave in a noise-like way then leave the phase relatively unperturbed, see [91, p. 51] for more details. From a sequence of interferograms, an ADI map is easily generated. Thresholding and thinning of points with good ADI, typically < 0.25 [61], leads to a set of stable pixels with reliable phases — these will be used as persistent scatterers and their coordinates form the sequence $\{s_i\}_{i=1}^{n_s} \subset S \cap V^C$.

Inferring K_N is done as described in the previous subsection, i.e. K_N is estimated from phase deviations of a pixel from that pixel's neighborhood's average as

$$K_N^s(s_i, s_j)K_N^t(t_m, t_n) = \frac{\delta_{ij}\delta_{mn}}{n_t} \underbrace{\sum_{l=1}^{n_t} \left(a_{il} - \frac{1}{|Nbd(i)|} \sum_{p \in Nbd(i)} a_{pl} \right)^2}_{n_t \sigma_i^2}$$

where a_{pl} are the measurements (interferometric phases) at pixels with index p at time l and $Nbd(i)$ is a neighborhood of pixel i. Typically, $K_N \approx 0$ on the persistent scatterers which permits ignoring the noise when trying to extract the correlation structure of the APS. For K_P^s, the parametric approach is chosen and it is estimated as

$$K_P^s(s_i, s_j) = \int_{s_0}^{s_i} \int_{s_0}^{s_j} K_{\Delta P}(r_1, r_2) dr_1 dr_2 \tag{5.79}$$

$$K_{\Delta P}(r_i, r_j) = c_1^{\text{opt}} \exp\left(-\frac{\|r_i - r_j\|_{\mathbb{R}^3}}{c_2^{\text{opt}}} \right) \tag{5.80}$$

$$(c_1^{\text{opt}}, c_2^{\text{opt}}) = \underset{(c_1, c_2) \in \mathbb{R}_+^2}{\text{argmin}} \ \|\{K_P^s(s_i, s_j)\}_{i,j=1}^{n_s} - S\|_F^2 \tag{5.81}$$

where S is the empirical covariance matrix for the persistent scatterers at $\{s_i\}_{i=1}^{n_s}$. Fitting K_P^t to the data can be done by adopting the parameter c_2 in an exponential covariance model and setting $c_1 = 1$ to guarantee a correlation model close to the observed one. One then has everything to directly implement equation 5.77. In figure5.28 the spatial APS interpolation using the integral covariance model was performed for single interferograms averaged in time. The term $\hat{p} = \sigma_p(\cdot)$ was first calculated and then subtracted from the data yielding $\hat{d} + \hat{n}$. Subsequently $\hat{n}$ was estimated by comparing $\hat{d}(s) + \hat{n}(s)$ to the average value of its neighbors. This is summarized in equation 5.82.

$$\sigma_p(\cdot) = \sum_{i=1}^{n_s} \lambda_i K_P(s_i, \cdot) \tag{5.82}$$

$$\begin{bmatrix}\lambda_1\\ \vdots \\ \lambda_{n_s}\end{bmatrix} = \begin{bmatrix}\int_{s_0}^{s_1}\int_{s_0}^{s_1} K_{\Delta P}(r_r, r_2)dr_1dr_2 & \dots & \int_{s_0}^{s_1}\int_{s_0}^{s_{n_s}} K_{\Delta P}(r_r, r_2)dr_1dr_2\\ \vdots & \ddots & \vdots \\ \int_{s_0}^{s_{n_s}}\int_{s_0}^{s_1} K_{\Delta P}(r_r, r_2)dr_1dr_2 & \dots & \int_{s_0}^{s_{n_s}}\int_{s_0}^{s_{n_s}} K_{\Delta P}(r_r, r_2)dr_1dr_2\end{bmatrix}^{-1}\begin{bmatrix}a_1\\ \vdots \\ a_{n_s}\end{bmatrix}$$

Notice that the APS+noise estimation using the integral covariance is better than the one provided for example by simple Kriging. There are clear effects of the topography and anisotropies stemming from the measurement geometry that a standard Kriging approach operating only on pixels or their three-dimensional coordinates is not able to model. Performance differences are made more formal in the results section. Figure 5.29 aims to highlight visually the quality of the APS interpolation.

Figure 5.28 also confirms that especially the turbulent parts of the APS are hard to predict. This is why, as more interferograms are averaged together and the correlation structure of the averaged APS gets smoother, both the APS estimations and the deformation estimations get more reliable.

Case 2: Deformations confined in time

This is in a sense the most general situation for which estimation is still possible in the framework presented in this monograph. Here no apriori information about the behavior of noise and atmospheric effects is supposed to be accessible and the spatial extent of the deformations is unknown. Assuming the deformations to occur only outside a certain time interval $V \subset T$, all measurements $m(v_i), v_i \in S_M \times (T_M \cap V)$ consist only of noise and APS enabling estimation of their covariance functions. The set of empirical covariance functions $K_P^s, K_P^t, K_N^s, K_N^t$ needs to be completed by adding K_D^s, K_D^t based on knowledge of the underlying physics and implied spatiotemporal correlation structure of the motion phenomena to be observed. Equation 5.73 can then be used directly.

It is worth noting, however, that the quality of the solution to this ill posed signal extraction problem depends on the validity of assumptions regarding the motion patterns encoded in the choice of K_D^s and K_D^t. Two problems may arise:

i) $\mathcal{H}_D^{\otimes} := \mathcal{H}_D^s \otimes \mathcal{H}_D^t$ with RK $K_D^s K_D^t$ does not contain the deformation functions $d(\cdot)$. If the prespecified structure of the deformation is 'too erroneous', the estimated deformations $\sigma_d(\cdot) \in \mathcal{H}_D^{\otimes}$ differ from the real ones systematically.

ii) $K_D^s K_D^t$ is 'too similar' to $K_P^s K_P^t$ or $K_N^s K_N^t$. If the spatiotemporal structure of the deformation is virtually indistinguishable from that of the APS or noise, the amount of confidence awarded to the results of signal separation needs to be evaluated critically.

The author suggest to choose the squared exponential kernel $K_{\text{sq}}(t_1, t_2) = c_0 \exp\left(-(\|t_2 - t_1\|/r_0)^2\right), c_0, r_0$ constants, to model smooth behavior as $f \in \mathcal{H}_{K_{\text{sq}}}$ implies f to be infinitely differentiable. Jagged and irregular behavior will be associated to the exponential kernel $K_{\exp}(s_1, s_2) = c_0 \exp\left(-\|s_2 - s_1\|/r_0\right),\ c_0, r_0$ constants, as $f \in \mathcal{H}_{K_{\exp}}$ implies f to be continuous but not differentiable [39, pp. 88-89]. For deformations like locally incohesive glacial motion that are nonrapid

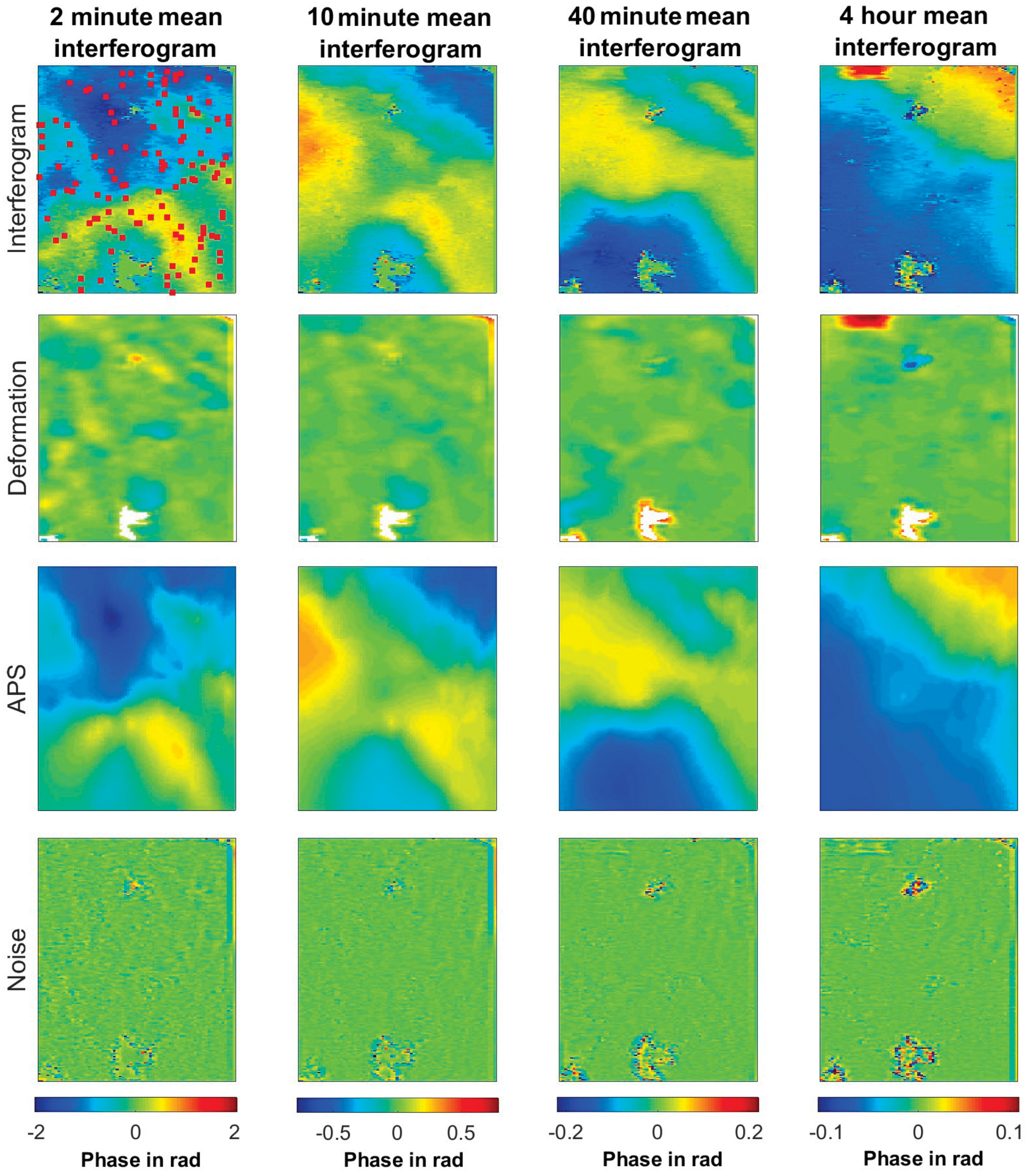


Figure 5.28: Example of APS and deformation estimation performance using the RKHS approach. The region featured in the interferogram is mostly stable. Any nonzero guesses for the deformation are erroneous and optimally the predicted APS plus noise values should be close to the original interferogram. The red squares in the first image showcase the basepoints (PSs), over which the APS is interpolated. A clearer image of the estimator's performance and topography dependence can be found in figure 5.29. The zone of persistent phase change identifiable in the upper part of the stacked interferogram covering 4 hours is of unknown origin but may actually be due to local motion as the area under question is an inclined ice-covered slope topographically close to the glacial scarp. It is unlikely to be induced by locally confined melting or snow accumulation processes as the detected phase changes persistently reappear during several consecutive days.

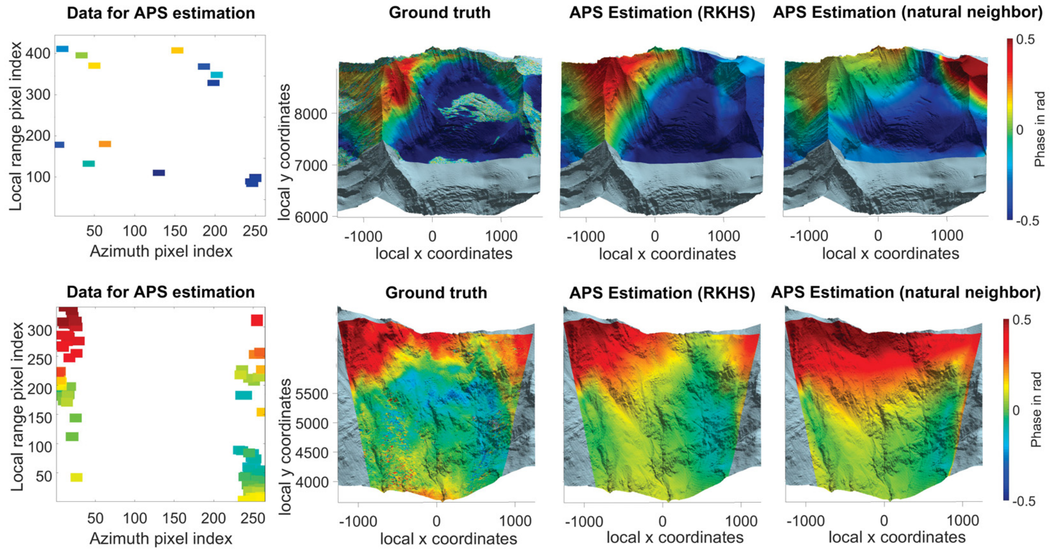


Figure 5.29: The data points in the left panels provide values of the APS at certain locations. These values are used to predict the APS at all other locations using the RKHS approach and the standard method of natural neighbor interpolation. The ground truth is plotted in the second column; as the regions are mostly stable in the time interval considered, the sum of APS and noise is equal to the whole interferogram. The differences in interpolation performance are especially noticeable in regions with severe topographical changes.

w.r.t the sampling frequency it seems reasonable to assume strong autocorrelation in time but not in space, i.e. $d(\cdot) \in \mathcal{H}_D^{\otimes}$ with RK $K_D = K_{\text{exp}}^s K_{\text{sq}}^t$. Together with the often turbulent nature of the APS this guarantees good separability. Other deformation phenomena like slow creep processes affecting widespread regions of homogeneous soil are represented more faithfully by elements of an RKHS with RK $K_D = K_{\text{sq}}^s K_{\text{sq}}^t$. in this case signal separation will likely suffer from problem ii) in the above list since the assumed spatial structures of APS and deformation are similar.

Since it was not assumed that the points used for estimation were stable or of high phase quality, it is not necessary to restrict attention to the persistent scatterers. Instead one may directly write down the solution to

$$\sigma_d(\cdot) = \underset{d \in \mathcal{H}_D^{\otimes}}{\operatorname{argmin}} \; \|A^s \otimes A^t d - a\|^2_{A(\mathcal{H}_M^{\otimes}/\mathcal{H}_D^{\otimes})} + \|d\|^2_{\mathcal{H}_D^{\otimes}} \tag{5.83}$$

after inference of kernels has been performed as in case 1. Then

$$\begin{aligned} \sigma_d(\cdot) &= \Xi^s \otimes \Xi^t a \\ \Xi^s &= \left[(K_D^s(\cdot, s_i))_{i=1}^{n_s}\right]^T \underbrace{\left[(K_M^s(s_i, s_j))_{i,j=1}^{n_s}\right]}_{\Sigma_M^s} \\ \Xi^s &= \left[\left(K_D^t(\cdot, t_i)\right)_{i=1}^{n_t}\right]^T \underbrace{\left[\left(K_M^t(t_i, t_j)\right)_{i,j=1}^{n_t}\right]}_{\Sigma_M^t} \end{aligned} \tag{5.84}$$

$$\begin{aligned} K^s_M(s_i,s_j) &= K^s_D(s_i,s_j) + \int_{s_0}^{s_i}\int_{s_0}^{s_j} K^s_{\Delta P}(r_1,r_2)dr_1dr_2 + \delta_{ij}\sigma_i^2 \\ K^t_M(t_i,t_j) &= K^t_D(t_i,t_j) + K^t_P(t_i,t_j) + \delta_{ij} \end{aligned}$$

Employing theorem 3.31 from subsection 3.1.3, it is possible to evaluate $\sigma_d(\cdot)$ at all locations $\{s_i^{\text{eval}}\}_{i=1}^{n^s_{\text{eval}}}$ and times $\{t_j^{\text{eval}}\}_{j=1}^{n^t_{\text{eval}}}$ simultaneously in a computationally efficient way. When denoting by σ_d^{eval} the $\mathbb{R}^{n_s^{\text{eval}}} \otimes \mathbb{R}^{n^t_{\text{eval}}}$ matrix with entries $\left(\sigma_d^{\text{eval}}\right)_{ij} = \sigma_d(s_i^{\text{eval}}, t_j^{\text{eval}})$, then

$$\sigma_d^{\text{eval}} = (Q^s)^* \left(\Sigma^s_M\right)^+ a \left(\Sigma^t_M\right)^+ \left(Q^t\right) \tag{5.85}$$

$$\begin{aligned} a \in \mathbb{R}^{n_s} \otimes \mathbb{R}^{n_t} &: \text{ The phase observations in form of a } n_s \times n_t \text{ matrix} \\ Q^s \in \mathbb{R}^{n_s} \otimes \mathbb{R}^{n^s_{\text{eval}}} &: \text{ Matrix with entries } (Q^s)_{ij} = K^s_D(s_i, s_j^{\text{eval}}) \\ Q^t \in \mathbb{R}^{n_t} \otimes \mathbb{R}^{n^t_{\text{eval}}} &: \text{ Matrix with entries } (Q^t)_{ij} = K^t_D(t_i, t_j^{\text{eval}}) \end{aligned}$$

Equation 5.85 speeds up the evaluation significantly and reduces the memory requirements to a level that can be handled by ordinary office computers. As a matter of fact, the author was unable to implement and evaluate computationally unoptimized estimators[7] at all on ordinary personal computers for the spatiotemporal case. The results of using σ_d^{eval} as a deformation estimate are collected in figure 5.30.

It is again necessary to use numerical integration to calculate the double integrals in an efficient way. Please note that generally the performance of the deformation estimate will be worse compared to the one exhibited in case 1, equation 5.82. This is due to the fact that the assumptions are weaker: the method just developed presupposes virtually no prior structure and is usable quite generally whereas the spline estimator for case 1 relied heavily on the deformations only occurring in a confined region in space.

§ TRI MAPS

To conclude, we propose to integrate cases 1 and 2 as alternative subprocesses into an algorithm called TRI-MAPS, that tackles the mitigation of atmospheric phase screens in terrestrial radar interferometry.

TRI MAPS

1. For a series of interferograms, calculate the ADI maps and associate to each pixel its three-dimensional coordinates.
2. Perform parametric kernel inference to determine the spatiotemporal kernel $K_P^{\otimes} = K_P^s K_P^t$ and the noise kernel $K_N^{\otimes} = K_N^s K_N^t$ from measurements on the identified PS..

[7] Here, an estimator is considered to be 'unoptimized' if it is of the same form as in equation 3.15 and the involved matrices are not expressed as tensor products of simpler matrices.

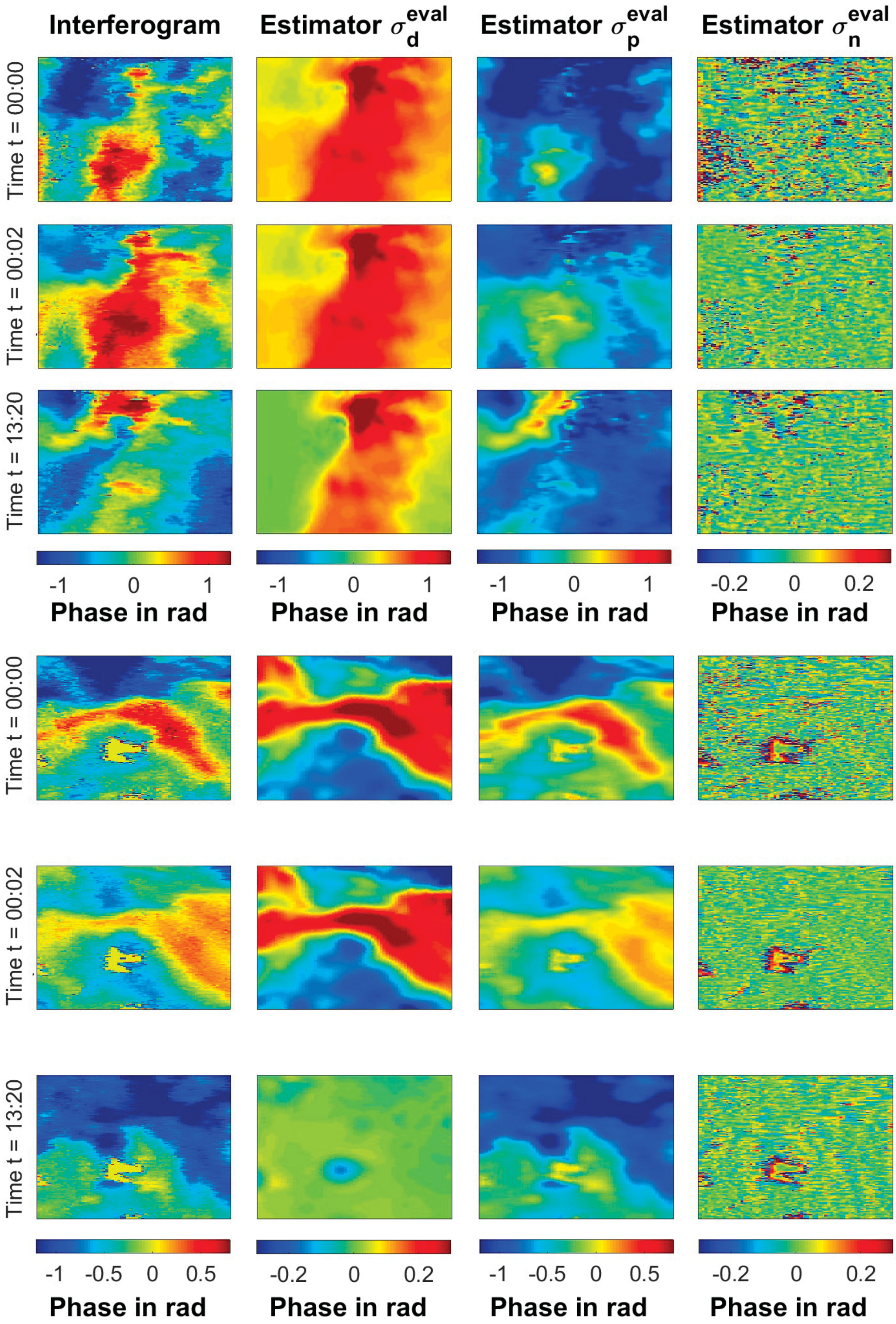


Figure 5.30: Unwrapped but not otherwise preprocessed interferograms at three different times and the tensor splines for deformation, APS and noise respectively. No stability assumptions have been made for the scene; the spatiotemporal signal separation was executed as described in the sequence of steps named as 'case 2' starting on page 265. The first three rows show the glacial tongue for which deformation actually occurs whereas the second set of rows show the tensor spline estimations for a stable area. Comparing the estimations recorded in them to the average of 24 h of interferograms, one can notice a tendency to overestimate the size of small deformations and to underestimate the size of big, locally confined deformations. The results are nonetheless significantly better than simple averaging in time as documented on later occasion in figure 5.33. Note the different color scales.

3. Depending on stable areas being known or unknown, either solve

$$\sigma_d = m - \sigma_p - \sigma_n \qquad \sigma_p(\cdot) = \underset{p \in A^{-1}a \cap \mathcal{H}_P^{\otimes}}{\operatorname{argmin}} \|p\|^2_{\mathcal{H}_P^{\otimes}}$$

for averaged interferograms or

$$\sigma_d(\cdot) = \underset{d \in \mathcal{H}_D^{\otimes}}{\operatorname{argmin}} \|A^s \otimes A^t d - a\|^2_{A(\mathcal{H}_M^{\otimes}/\mathcal{H}_D^{\otimes})} + \|d\|^2_{\mathcal{H}_D^{\otimes}}$$

for the whole sequence where the kernel $K_D^{\otimes} = K_D^s K_D^t$ represents prior knowledge about the deformation process.

4. If the interferograms are too big to solve the problems in step 3, apply step 2 and 3 locally to spatiotemporal subsets $\{sub_i\}_{i=1}^{n_{\text{part}}} \subset S \times T$ of interferograms. Then patch the local estimations together using an overlapping partition of unity $\{\varphi_i\}_{i=1}^{n_{\text{part}}}$ on the indexset $S \times T$ such that $\sum_{i=1}^{n_{\text{part}}} \varphi_i(\cdot) = \mathbb{1}_{S \times T}(\cdot)$ and each φ_i has support exactly on subset sub_i.

This concludes the description of algorithms for atmospheric correction. At this point we want to remark in the sense of an outlook that it is entirely possible to swap what one subjectively considers noise and what signal thereby shifting focus towards the APS as an entity of interest in itself. The amount of information that TRI data can provide about microlocal changes in meteorology and tropospheric dynamics is significant and to just eliminate it from further processing seems wasteful. In fact, by changing slightly the formulation of the abstract spline interpolation problem

$$\sigma_p(\cdot) = \underset{p \in A^{-1}a \cap \mathcal{H}_P}{\operatorname{argmin}} \|p\|^2_{\mathcal{H}_P} \tag{5.86}$$

$$\begin{aligned} A &: \text{ Measurement operator, evaluation at stable points } \{s_i\}_{i=1}^{n_s} \\ a &: \text{ Data consisting of } n_s \text{ atmospheric phase delays at locations } \{s_i\}_{i=1}^{n_s} \\ \mathcal{H}_P &: \text{ RKHS pf phase delays, has reproducing kernel } K_P \end{aligned}$$

to include instead of the APS $P : \Omega \times \mathbb{R}^3 \to \mathbb{R}$ the differential field $\Delta P : \Omega \times \mathbb{R}^3 \to \mathbb{R}$ and designating it as the target of estimation, one recovers problem 5.87.

$$\sigma_{\Delta P}(\cdot) = \underset{\Delta p \in (AL)^{-1}a \cap \mathcal{H}_{\Delta_P}}{\operatorname{argmin}} \|\Delta p\|^2_{\mathcal{H}_{\Delta P}} \tag{5.87}$$

$$\begin{aligned} L &: \text{ Operator of line integration from } s_0 \text{ to a variable } s \in \mathbb{R}^3 \\ \mathcal{H}_{\Delta P} &: \text{ RKHS of differential phase delays, has reproducing kernel } K_{\Delta P} \end{aligned} \tag{5.88}$$

where A, a are as before and $(AL\Delta P) = \{\int_{s_0}^{s_i} \Delta p(r)dr\}_{i=1}^{n_s}$. Relating the field ΔP of differential atmospheric phase delays to the field of refraction index changes $\Delta n : \Omega \times \mathbb{R}^3 \to \mathbb{R}$ via

$$\Delta P_{(\cdot)}(s) = \frac{4\pi}{\lambda} \Delta n_{(\cdot)}(s) \tag{5.89}$$

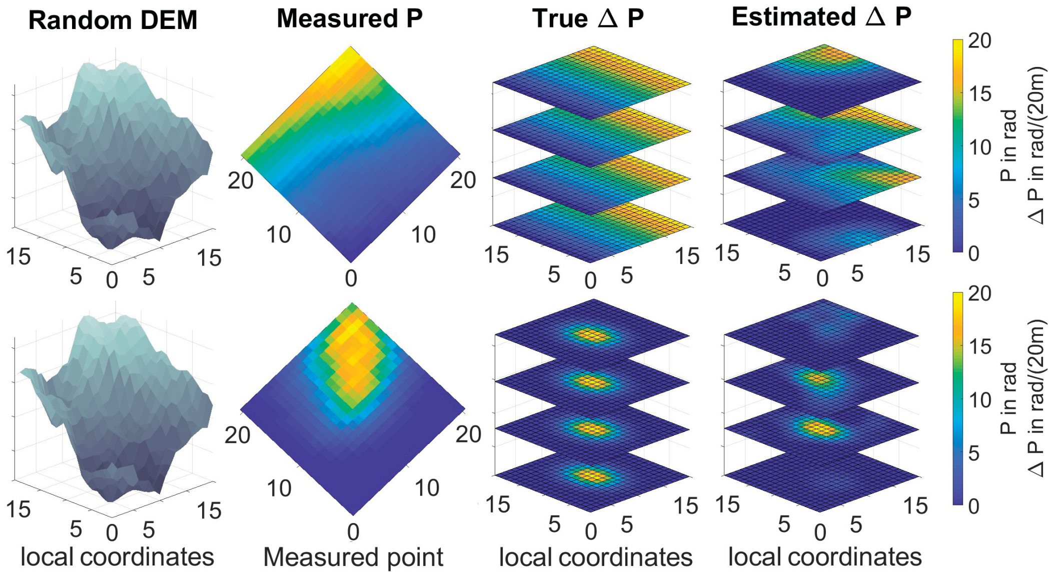


Figure 5.31: The left panel shows a randomly generated landscape. The instrument's position is $s_0 = [0, 0]^T$ and the measurements in the second column are generated via line integration from s_0 to the surface's coordinates s through the field ΔP illustrated in the third column. The estimations of ΔP derived as solutions to problem 5.87 are plotted in column 4. Note that the estimation errors grow rapidly when leaving the volume through which P was measured. The results are, however, promising.

for λ the wavelength of about 17.6 mm and $s \in S$ a location enables some type of TRI meteorology by solving the tomography problem 5.87 for the field of refraction changes $\Delta n_{(\cdot)}(\cdot)$. The solution to abstract spline problem 5.87 can be written down explicitly and in closed form. It is

$$\sigma_{\Delta p}(\cdot) = \sum_{i=1}^{n_s} \lambda_i \int_{s_0}^{s_i} K_{\Delta P}(\cdot, r)dr \tag{5.90}$$

$$\begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_{n_s} \end{bmatrix} = \begin{bmatrix} \int_{s_0}^{s_1}\int_{s_0}^{s_1} K_{\Delta P}(r_r, r_2)dr_1dr_2 & \dots & \int_{s_0}^{s_1}\int_{s_0}^{s_{n_s}} K_{\Delta P}(r_r, r_2)dr_1dr_2 \\ \vdots & \ddots & \vdots \\ \int_{s_0}^{s_{n_s}}\int_{s_0}^{s_1} K_{\Delta P}(r_r, r_2)dr_1dr_2 & \dots & \int_{s_0}^{s_{n_s}}\int_{s_0}^{s_{n_s}} K_{\Delta P}(r_r, r_2)dr_1dr_2 \end{bmatrix}^{-1} \begin{bmatrix} a_1 \\ \vdots \\ a_{n_s} \end{bmatrix}$$

or in more compact notation denoting $h(\cdot) = \int_{s_0}^{\cdot} K_{\Delta P}(\cdot, r)dr$ as

$$\sigma_{\Delta p}(\cdot) = [h_1(\cdot), ..., h_{n_s}(\cdot)] \begin{bmatrix} \langle h_1, h_1\rangle_{\mathcal{H}_{\Delta P}} & \dots & \langle h_1, h_{n_s}\rangle_{\mathcal{H}_{\Delta P}} \\ \vdots & \ddots & \vdots \\ \langle h_{n_s}, h_1\rangle_{\mathcal{H}_{\Delta P}} & \dots & \langle h_{n_s}, h_{n_s}\rangle_{\mathcal{H}_{\Delta P}} \end{bmatrix}^{-1} \begin{bmatrix} a_1 \\ \vdots \\ a_{n_s} \end{bmatrix}. \tag{5.91}$$

Figure 5.31 records some solutions for synthetic examples. Potentially, one might try to extract temperature and water vapor changes by inverting equation 5.16 but this constitutes a severely ill-posed problem that is left for future work.

5.2.6 Results and comparisons

It is important but difficult to check the validity of any methodology for filtering of TRI-data as ground truth is available only for the regions known to be stable. Alternatively one may simulate deformation data, which acts as synthetic nonzero ground truth when mixed into measurements known to consist only of APS and noise. In a comparison between TRI-MAPS and more common methods based on the fitting of parametric models, TRI-MAPS exhibits superior performance as measured by the RMSE both in stable and unstable areas. It is possible to choose kernels and probabilistic assumptions in a way such that the RKHS based approach reproduces the easier models like stacking, Kriging, or topography induced phase variation models. Therefore one is able to write down the specific set of assumptions under which they are stochastically optimal.

A completely hypothesis free check of estimation performances is not possible given the state of technology. TRI is as of now the only ground based geodetic technique capable of surveying spatially densely areas of up to 100 km^2 in less than a minute and with high repetition rates. The spatial density of GNSS measurements and total station measurements (per minute) is not high enough and terrestrial laser-scanning exhibits a whole other set of unresolved problems and comparability issues when displacements vectors are to be extracted from point clouds — apart from the fact that these systems are influenced by atmospheric effects as well. Therefore, ground truth covering the whole area essentially needs to come from apriori knowledge about the deformation behavior that is occurring and not from other measurement systems.

When stable areas are surveyed, one may assume that any phase changes documented in interferograms are due to atmosphere and noise. The ground truth to compare the deformation estimation against should then be zero. The downside of this assumption is that the ground truth is always zero and the root mean square error is the $RMSE_1 = \|\sigma_d\|_{L^2}^2$ which implies that no algorithm, no matter how sophisticated or stochastically justified, would ever outperform the trivial but practically useless estimator that guesses a deformation of zero uniformly and independently of observed data. Evaluating performance on stable areas only would introduce a bias towards assessing those estimators as lacking in quality that actually at some point assert the presence of deformation in the TRI signal. Optimizing w.r.t. this performance measure would lead to estimators that consistently underestimate deformation and in a sense minimize the α-error of rejecting stability when it is actually the underlying ground truth.

To prevent this, the estimator also has to be tested on regions including nontrivial deformations $d_{\text{true}}(\cdot)$ to quantify $RMSE_2 = \|\sigma_d - d_{\text{true}}\|_{L^2}^2$. Since such regions do not exist in the data available for this dissertation and can hardly be provided at all, we will simply simulate d_{true} as a spatiotemporal random field d_{simu}, superimpose it on data $M = P + N$ coming from stable regions and assess the performance by comparing the estimations to this synthetic ground truth. The choice of probability distribution for the simulated field of deformations d_{true} impacts the estimation performance since the spline σ_d is explicitly constructed from what one considers apriori to be the correlation functions that determine the spatiotemporal structure of $d(\cdot)$. The RKHS approach shares this problem with the other estimators commonly used in the literature that all propose parametric or nonparametric models of equal likelihood on either the APS or the deformations. It should be noted that by tuning the probability distribution of d_{true}, one can make the spline estimator look almost arbitrarily bad or even stochastically optimal depending on how aligned the prior guess for $K_D^{\otimes}$ and the true underlying deformation structure are. We will therefore make explicit what type of deformation model we use and in how far the estimators are based on misspecified assumptions. This second performance measure is then related to the β-error of wrongly accepting stability and the relative importance of $RMSE_1$ vs $RMSE_2$ depends on the concrete application and should be subject to a risk-oriented discussion.

Figure 5.32 shows an exemplary simulated deformation field and its evolution over time. The spatiotemporal behavior is indicative of the correlation model that is used to generate samples of d_{true} for performance evaluation purposes.

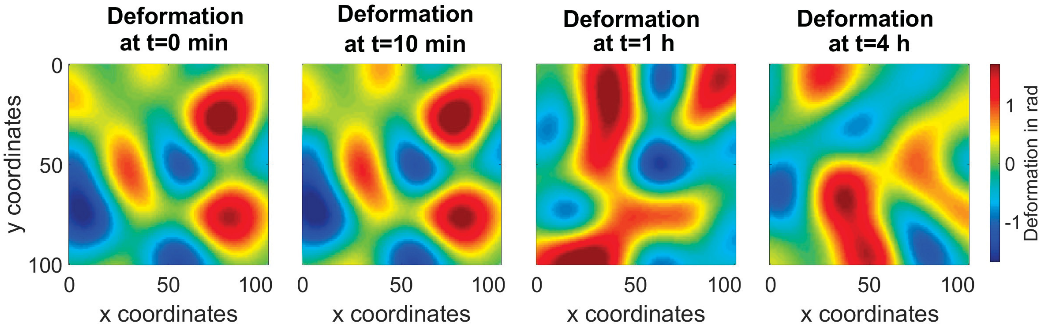


Figure 5.32: An exemplary deformation field detailing coordinate changes at different times. Each image reflects the sum of deformations that occurred during the 2-minute interval preceding the timestamp and is thereby of the same incremental form as the deformations underlying interferograms. The correlation model is squared exponential in space and time. It will be used generically to generate synthetic ground truth if nothing to the contrary is mentioned.

It was explained before that the APS consists of a turbulent part that approximately averages out over time and a persistent, smooth part that evolves over time scales of several hours and reflects changes in average meteorological parameters. The latter one is easier to predict and estimation performance gets better if the algorithms for APS estimation act on more data that they can potentially average. The graphs in figure 5.33 compare simple stacking, the RKHS approaches, second order polynomial fitting, and a multiple regression model based on elevation information (see [104]) in this respect.

Several methods other than the RKHS approach have been developed in the literature under different assumptions. They span the whole range from parametric to nonparametric estimators and from estimators that use only temporal information to ones that employ primarily spatial information. Approaches based on Kriging of the APS over stable points have become more widespread recently although in contrast to the RKHS method presented here, they incorporate only piecewise constant deterministic models for the deformation and second order stationary covariances [212, 13]. Most of the methods are data-driven and it is possible to find an RKHS embedding of the estimation problem in such a way that these correction methods emerge as splines in that framework enabling a clear stochastic interpretation.

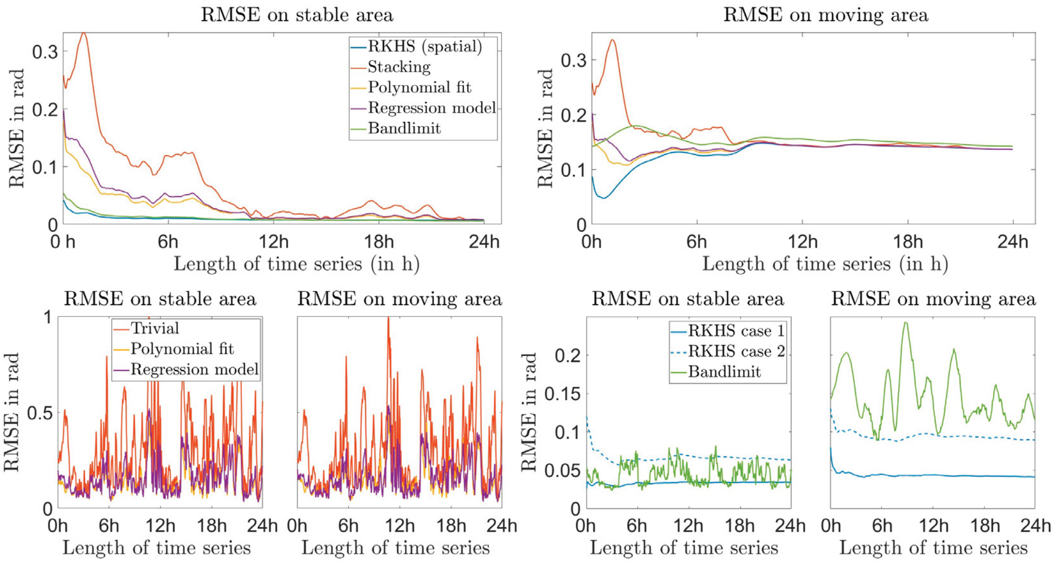


Figure 5.33: The performance of RKHS-based methods (spatiotemporal splines as in cases 1 and 2 in subsection 5.2.5), second order polynomial fitting, simple stacking, and other methods for extraction of deformations. The two plots on the top were generated by evaluating 24 h worth of data by passing averaged interferograms of increasing integration time to the algorithms. The region of interest is the same one as the bottom region plotted in figure 5.21. Nonzero ground truth has been simulated as described before and illustrated in figure 5.32. Note that the RMSE in the moving areas does not converge to zero but instead to the standard deviation of the movement; as from an averaged interferogram only a constant estimator for the deformations is extracted, faithful reconstruction of the time-varying ground truth is not possible. The second row presents the results of several algorithms for full spatiotemporal estimation of the sequence of two-dimensional deformations. The trivial model consists in taking the individual interferograms phase values directly as estimators for the deformation without further processing. Bandlimiting is described in subsequent explanations. Note the different length scales of the y-axis.

Stacking

Explanation: Interferograms are simply averaged in time; potential spatial correlations are ignored.

Equation: $\sigma_d(t) = \frac{1}{n_t}\sum_{j=1}^{n_t} m(t_i)$

RKHS embedding:

$$\begin{aligned}
\sigma_d &= \underset{d\in\mathcal{H}_D^t}{\operatorname{argmin}} \; \|A^t d - a^t\|^2_{A^t\mathcal{H}_P^t} \\
\mathcal{H}_M^t &= \mathcal{H}_D^t \oplus \mathcal{H}_P^t \\
\mathcal{H}_D^t &= \text{ Hilbert space of constants with RK } K_D^t(t_1,t_2) = 1[170, p.\ 71] \\
A^t\mathcal{H}_P^t &= \mathbb{R}^{n_t}, \text{ Hilbert space of white noise residuals}
\end{aligned}$$

Stochastic interpretation: This approach is optimal under the assumption that the atmosphere behaves as white noise and the deformations are constant in time. If the topography is trivial and the APS purely turbulent, the method can work. Otherwise it is not recommended.

Proof: We know that $d \in \mathcal{H}_D^t$ is a constant and at the same time $\sigma_d = \underset{d\in\mathcal{H}_D^t}{\operatorname{argmin}} \sum_{i=1}^{n_t}[d(t_i) - a_i^t]^2 = \underset{c_0\in\mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n_t}[c_0 - a_i^t]^2 = \frac{1}{n_t}\sum_{i=1}^{n_t} a_i$.

Polynomial fitting

Explanation: A second degree polynomial in range is fitted to explain the APS observed on stable persistent scatterers and predict it at other locations, see [152].

Equation: $\sigma_d(r, az) = m(r, az) - \sigma_p(r)$

$$\sigma_p(r) = [1 \;\; r \;\; r^2] \left(F^T F\right)^{-1} F^T a^s \quad , (F)_{ij} = (r_i)^{j-1}$$

RKHS embedding:

$$\begin{aligned} \sigma_p &= \underset{p \in \mathcal{H}_P^s}{\operatorname{argmin}} \; \|A^s p - a^s\|^2_{A^s \mathcal{H}_N^s} \\ \mathcal{H}_M^s &= \mathcal{H}_D^s \oplus \mathcal{H}_P^s \oplus \mathcal{H}_N^s \\ \mathcal{H}_D^s &= \text{ unspecified Hilbert space of deformations} \\ \mathcal{H}_P^s &= \text{ Hilbert space of second order polynomials in range.} \\ & \quad \text{ Has kernel } K_P^s(r_i, r_j) = c_0^2 + c_1^2 r_i r_j + c_2^2 r_i^2 r_j^2. \\ A^s \mathcal{H}_N^s &= \mathbb{R}^{n_s}, \text{ Hilbert space of white noise residuals} \end{aligned}$$

Stochastic interpretation: This approach is optimal under the assumption that the local phase delay varies linearly in range and therefore integrates to a second order polynomial that forms the APS. The stable points are known and residuals between predicted APS and observed phase values are uniformly white noise on these parts. If the topography is trivial, the APS highly regular and the area is small, the method can work.

Proof: We know that $p \in \mathcal{H}_P^s$ is of the form $p(r) = c_0 + c_1 r + c_2 r^2$ and at the same time $\sigma_p(r) = \underset{p \in \mathcal{H}_P^s}{\operatorname{argmin}} \sum_{i=1}^{n_s} \left[p(s_i) - a_i^s\right]^2 = \underset{c_0, c_1, c_2 \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n_s} \left[c_0 + c_1 r_i + c_2 r_i^2 - a_i^s\right]^2$ for which the solution is simply the usual expression $(F^T F)^{-1} F^T a^s$ known from adjustment theory.

High pass filtering

Explanation: The signal is split into a part with high spatial frequencies and a part with low spatial frequencies [34].

Equation: $\sigma_d = \left(I - \mathcal{F}^* \mathbb{1}_{[-b,b]} \mathcal{F}\right) m \quad \mathcal{F}$ Fourier transform

RKHS embedding:

$$\begin{aligned} \sigma_d &= \underset{d \in \mathcal{H}_D^s}{\operatorname{argmin}} \; \|d - m\|^2_{\mathcal{H}_P^s} \\ \mathcal{H}_M^s &= \mathcal{H}_D^s \oplus \mathcal{H}_D^t \\ \mathcal{H}_P^s &= \text{ Paley Wiener space of bandlimited functions with kernel} \\ & \qquad K_P^s(s_i, s_j) = \frac{\pi}{b} \operatorname{sinc}(b(s_i - s_j))[20, p.\ 304] \\ \mathcal{H}_D^s &= L^2(S)/\mathcal{H}_P^s, \text{ Hilbert space of deformations that have no} \end{aligned}$$

frequencies less than $|b|$

Stochastic interpretation: This approach is optimal if there is no noise, the deformations occur only locally and the APS is smooth. If the deformation and the APS have non-disjoint support in frequency space or the deformations behavior is unknown, it is not usable.

Proof: Notice that $P_{\mathcal{H}_P^s} = \mathcal{F}^* \mathbb{1}_{[-b,b]} \mathcal{F} : L^2(S) \to \mathcal{H}_P^s$ is an orthogonal projection as $P_{\mathcal{H}_P^s}^* = P_{\mathcal{H}_P^s}^2 = P_{\mathcal{H}_P^s}$ as one can easily check using unitarity of $\mathcal{F}$. The set of functions $f : P_{\mathcal{H}_P^s} f = f$ is precisely those f with frequencies contained in the interval $[-b, b]$ and therefore $P_{\mathcal{H}_P^s}$ is the orthogonal projection onto $\mathcal{H}_P^s$ [93, p. 103]. Consequently $P_{\mathcal{H}_P^s} m$ satisfies

$$P_{\mathcal{H}_P^s} m = \underset{p \in \mathcal{H}_P^s}{\operatorname{argmin}} \ \|p - m\|_{L^2(S)}^2$$

and $I - P_{\mathcal{H}_P^s}$ is the projection onto $L^2(S)/\mathcal{H}_P^s \cong \mathcal{H}_D^s$. The best reconstruction of m by functions in $\mathcal{H}_D^s$ is given by $(I - \mathcal{F}^* \mathbb{1}_{[-b,b]} \mathcal{F})m$. As $\| \cdot \|_{L^2(S)}$ and $\| \cdot \|_{\mathcal{H}_P^s}^2$ coincide for elements of $\mathcal{H}_P^s$ [20, p. 304], the claim follows.

These and other methods are illustrated in figure 5.34 which shows estimated atmospheric phase screens and deformations for two exemplary interferograms with known , but in the case of the second interferogram, synthetic, ground-truth. As is clearly visible and quantified more objectively in figure 5.33, the deformation estimations provided by stacking and parametric modelling of the APS based on observations of the phase on stable points are not very accurate. The method described as RKHS case 2 does not use any potentially inaccessible prior information about the stability of the observation points and the corresponding lack of assumptions on the one hand improves the generality under which it is applicable but also degrades estimation results. The RKHS approach that also integrates stability information performs best and exhibits realistic estimations of APS and deformations although bad signal-to-noise ratio and unwrapping errors can deteriorate the quality.

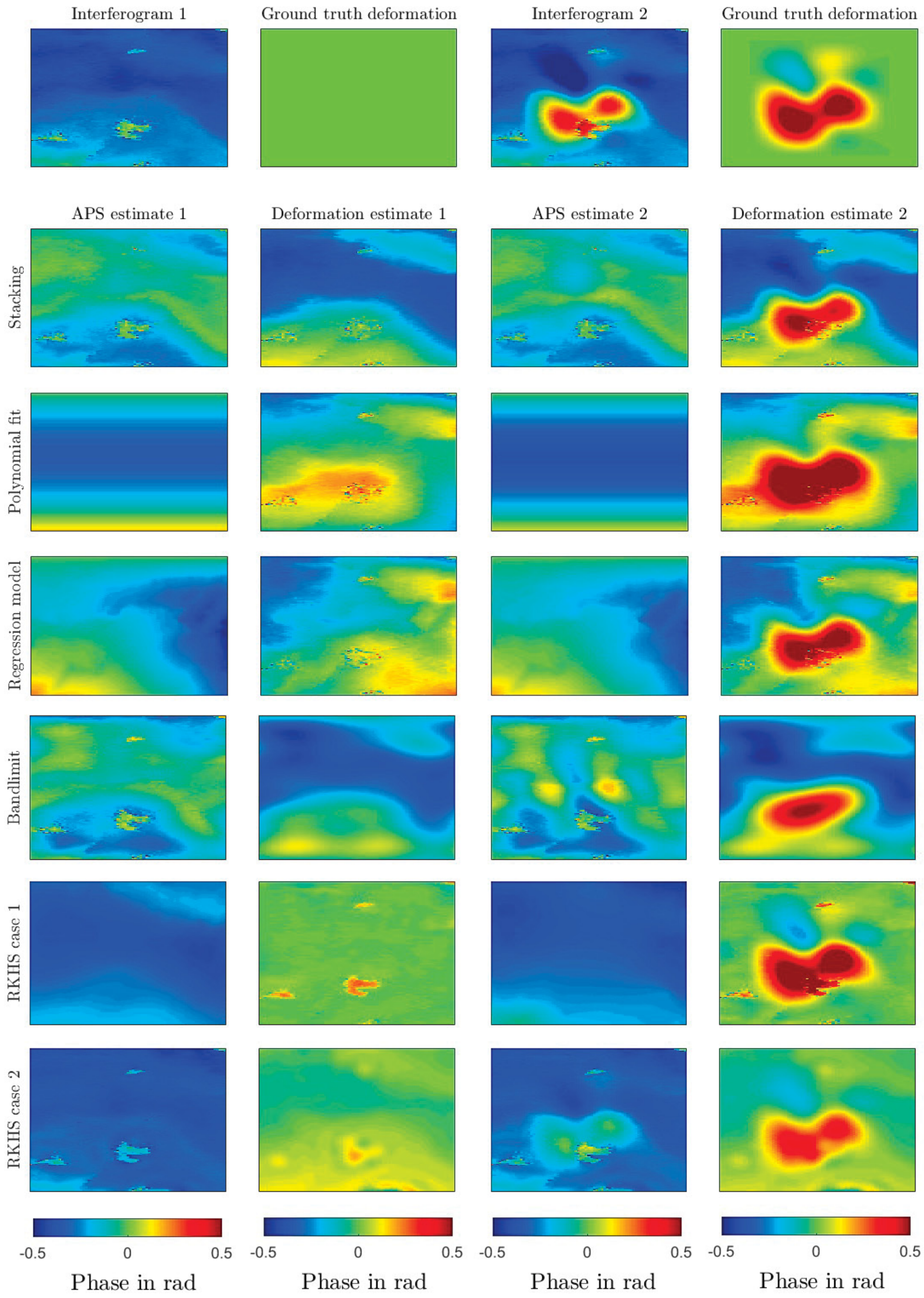


Figure 5.34: The estimators for APS and deformation previously described in the text. They act on 1 h of data whose average is plotted in the first row. The interferogram in column 3 is a superposition of the non-zero ground truth plotted in column 4 and the APS shown in column 1. The reference for the regression model featuring explicit dependence of the APS on the elevation is [104].

Chapter 6

Conclusions and outlook

This monograph investigated, explained, and advanced a Hilbert space-based approach to signal processing, in which spaces of functions are furnished with probability distributions to solve challenging and often ill-posed estimation problems by reformulating them as optimization tasks in infinite-dimensional functional spaces. Our main contributions are the introduction of a probabilistically motivated nonparametric algorithm for a data-driven choice of the Hilbert space most suitable for estimation and the application of the theory to the problem of separating signal and noise in terrestrial radar interferometry.

An introductory chapter aimed at familiarizing the reader with functional analytic notions and concepts within the context of low dimensional spaces for reasons of accessibility. After introducing some standard notation and basic theorems expounding Hilbert spaces and linear operators on them, light was shed on their relationship to signal processing by showing that optimal approximation tasks can equivalently be formulated as norm-minimization problems. In this sense, the choice of models for representation of signals, the choice of basis elements in a Hilbert space, optimality in the least-squares-sense and orthogonal projections are all related to problems of interpolation, smoothing, and efficient representation of observational data. With the help of spectral theory, we derived and presented new results that relate certain transforms of a covariance operator's spectrum to equivalently transformed stochastic processes. Employing the exact same functional calculus allowed to solve physically motivated deterministic differential equations as well; estimating a system's state subject to physical constraints and uncertainty in the observations was shown to be feasible and simple for exemplary diffusion-type problems. The duality between stochastic and deterministic perspectives is explorable especially well in the framework of reproducing kernel Hilbert spaces. They allow the coupling of differential equations, correlation structures and feature representations.

Several equivalent formulations of optimization problems were investigated and their relationship to geostatistical standard methods was made explicit. Extending these optimization problems such that their solutions could be expressed as abstract splines allowed the target of estimation to be related to a prior and to the given measurements nontrivially via linear operators. Direct sums, quotients, and tensor products of RKHS lead again to RKHS and widen the applicability of the abstract splines framework to include estimation of function tuples, equivalence classes and tensors. We presented new memory efficient and cheaply implementable equations for tensor splines that facilitate spatiotemporal estimation infeasible otherwise. In a similar spirit, sequential approximation schemes for inverses of nonfactorizable kernel operators were discussed primarily from the perspective of computational load and practicability before matters of kernel construction and design were touched upon. We proved the relevance of the RKHS approach by solving nonobvious geodetically motivated problems that involved vector fields, tomographic imaging, and nonlinear hypothesis testing.

We proposed a novel solution to the open question of how to actually infer the kernel uniquely determining an RKHS in the absence of prior knowledge of its parametric form. After reviewing abstract harmonic and operator-theoretic notions of positivity, we introduced a Wishart-type probability distribution over the convex cone of all kernels by expressing kernels as superpositions of simple tensors composed from the eigenfunctions of a prior guess's associated kernel operator. This representation led to a matrix-valued coefficient structure that has to be determined in the kernel's stead. Ensuring positive definiteness, a necessary and sufficient key requirement for a two-variable function to be a reproducing kernel and determine an RKHS, amounts to introducing positivity constraints on the coefficient matrix's spectrum. Under these spectral constraints, solving for the optimal coefficients is a numerically challenging task for which we suggested an iterative Fisher-scoring-type algorithm that, bar occasional convergence problems, leads to nonparametric, highly complex kernel estimations that are consistent with the observed data and likely under some freely selectable prior assumptions. We showed that they outperform popular parametric families of kernels in both reconstructive flexibility and estimation quality in practical applications.

Investigating our algorithm further revealed ties to semidefinite programming and enabled us to extend the kernel inference procedure to respect linear equality constraints and lessen the requirements on the basis functions into which the kernel is supposedly decomposable. We showed that this approach is suitable to solve instationary spatial estimation problems intractable with other methods and emphasized its generality as well as its practical applicability by expressing the problem of variance components estimation in terms of our kernel inference framework.

Assembling the tools devised previously in the monograph, we tackled and solved a difficult signal separation problem arising regularly during TRI-based deformation monitoring of geohazards in mountainous terrain. This problem featured turbulence induced artifacts and instationary long-term drifts highly autocorrelated in time and space with a strong but not deterministically expressible dependence on the topography. Both of these noise components differed from the deformation signal to be extracted only in terms of their probability distribution. By embedding spatiotemporal sequences of interferograms into infinite-dimensional RKHS exhibiting a tensorproduct structure, we prove that it is possible to explicitly pose and solve stochastically motivated optimization problems whose solutions are optimal estimators for the spatiotemporal evolution of deformations noisily observed by terrestrial radar interferometry.

In the most general formulation, our approach does neither assume the accessibility of prior knowledge about stable areas nor simple, parametrically describable correlation structures of the artifacts to be filtered out. Instead, it splits the full sequence of measurements into guesses for noise and signal based on a detailed analysis of the correlation structure of TRI-data resulting in a data-driven learning scheme that improves with the amount of data observed. Our method is neither expressible as the fitting of a parametric function nor as a simple geostatistical procedure of Kriging-type and, in fact, properly generalizes and contains both of these approaches that were previously proposed in the literature for solving the problem of signal separation in TRI-processing. We reformulated the most common previously proposed solutions in an RKHS framework enabling an explicit stochastic interpretation that allowed uncovering their underlying probabilistic assumptions and straightforward comparison. We found that in comparison to the widely employed temporal averaging, our method's root mean square error converges to zero more reliably and faster by almost a factor of 10. Although less drastically, the method also outperformed a host of other approaches ranging from topography-based regression models to spectrum-based filters as tested with the help of a real-world dataset gathered during a monitoring campaign in the Swiss Alps that features distances of up to $8\ \mathrm{km}$ and elevation differences of up to $3\ \mathrm{km}$.

The author hopes that the monograph suitably demonstrated the unifying power and potential of reproducing kernel Hilbert spaces for spatiotemporal statistics and its many uses for geodetic data. There is still much to improve and even more to discover. Questions like the following are both interesting from a theoretical perspective and practically relevant:

- Can one implement a holistic joint inference on complex radar interferometric data and avoid splitting amplitude and phase information?
- Is it possible to implement rigorous stochastically driven linear dynamics of probability distributions and use them to do inference for physical systems?
- Can one efficiently optimize over arguments in kernels to find optimal experiment designs?
- Are there better priors on covariances than the Wishart distribution?
- What exactly is the relationship between kernels, dynamical systems, physical laws and learning?

In the end, the effort put into them might very well contribute to a general framework for optimal monitoring and uncertainty quantification. At the very least, immediate improvements in specific instances of typical engineering geodetic tasks are to be expected. These prospects are well worth pursuing.

List of Figures

Bibliography

[1] J. Ádám, *A detailed study of the duality relation for the least squares adjustment in euclidean spaces*, Bulletin géodésique, 56 (1982), pp. 180–195.

[2] J. Adamek, H. Herrlich, and G. Strecker, *Abstract and Concrete Categories - The Joy of Cats*, Dover Publications, New York, 2009.

[3] N. I. Akhiezer and I. M. Glazman, *Theory of Linear Operators in Hilbert Space Vol I*, Courier Corporation, New York, 2013.

[4] ——, *Theory of Linear Operators in Hilbert Space Vol II*, Courier Corporation, New York, 2013.

[5] F. Alizadeh, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM Journal on Optimization, 5 (1995), pp. 13–51.

[6] G. Antonello, J. Fortuny, D. Tarchi, N. Casagli, C. Del Ventisette, L. Guerri, G. Luzi, F. Mugnai, D. Leva, and E. Srl, *Microwave interferometric sensors as a tool for space and time analysis of active volcano deformations: The stromboli case*, 2008.

[7] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American mathematical Society, 68 (1950), pp. 337–404.

[8] N. Aronszajn and K. T. Smith, *Characterization of positive reproducing kernels. applications to green's functions*, American Journal of Mathematics, 79 (1957), pp. 611–622.

[9] R. B. Ash, *Basic Probability Theory* -, Courier Corporation, New York, 2008.

[10] ——, *Basic Abstract Algebra - For Graduate Students and Advanced Undergraduates*, Courier Corporation, New York, 2013.

[11] M. Atteia, *Hilbertian Kernels and Spline Functions* -, Elsevier, Amsterdam, 2014.

[12] W. Baarda, *A testing procedure for use in geodetic networks.*, Delft, Kanaalweg 4, Rijkscommissie voor Geodesie, 1968., 1968.

[13] S. Baffelli, O. Frey, and I. Hajnsek, *Geostatistical analysis and mitigation of atmospheric phase screens in ku-band terrestrial radar interferometry*, in IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, July 2018, pp. 6504–6507.

[14] R. Bamler and P. Hartl, *Synthetic aperture radar interferometry*, Inverse Problems, 14 (1998), p. R1.

[15] R. Bamler and D. Just, *Phase statistics and decorrelation in sar interferograms*, in Proceedings of IGARSS '93 - IEEE International Geoscience and Remote Sensing Symposium, Aug 1993, pp. 980–984 vol.3.

[16] C. Batty, *Unbounded operators: Functional calculus, generation, perturbations*, Extracta Mathematicae, 24 (2009), pp. 99–233.

[17] B. R. Bean, E. J. Dutton, and C. R. P. L. (U.S.), *Radio meteorology -*, Dover Publications, New York, 1966.

[18] A. Ben-Israel and T. N. Greville, *Generalized Inverses - Theory and Applications*, Springer Science & Business Media, Berlin Heidelberg, 2nd ed. 2003 ed., 2003.

[19] C. v. d. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups - Theory of Positive Definite and Related Functions*, Springer Science & Business Media, Berlin Heidelberg, 2012.

[20] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics -*, Springer Science & Business Media, Berlin Heidelberg, 2011.

[21] A. Y. Bezhaev and V. A. Vasilenko, *Variational Theory of Splines -*, Springer Science & Business Media, Berlin Heidelberg, softcover reprint of hardcover 1st ed. 2001 ed., 2013.

[22] A. Bobrowski, *Functional Analysis for Probability and Stochastic Processes - An Introduction*, Cambridge University Press, Cambridge, 2005.

[23] G. Boffi and A. Wieser, *Dynamics-based system noise adaptation of an extended kalman filter for gnss-only kinematic processing*, in Proceedings of the 29th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2016), 29th International Technical Meeting of the Satellite Division of The Institute of Navigation, ION GNSS + 2016, 2016, pp. 554–563.

[24] K. Borre, *Mathematical Foundation of Geodesy - Selected Papers of Torben Krarup*, Springer Science & Business Media, Berlin Heidelberg, 2006.

[25] J. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization - Theory and Examples*, Springer Science & Business Media, Berlin Heidelberg, 2010.

[26] D. Bosq, *Linear Processes in Function Spaces - Theory and Applications*,

Springer Science & Business Media, Berlin Heidelberg, 2012.

[27] J. Bourgain and N. Pavlovic, *Ill-posedness of the navier-stokes equations in a critical space in 3d*, Journal of Functional Analysis, 255 (2008), pp. 2233 – 2247. Special issue dedicated to Paul Malliavin.

[28] S. P. Boyd and L. Vandenberghe, *Convex Optimization* -, Cambridge University Press, Cambridge, 2004.

[29] F. Bozzano, I. Cipriani, P. Mazzanti, and A. Prestininzi, *A field experiment for calibrating landslide time-of-failure prediction functions*, International Journal of Rock Mechanics and Mining Sciences, 67 (2014), pp. 69–77.

[30] I. N. Bronstejn, H. Mühlig, K. A. Semendjajew, and G. Musiol, *Taschenbuch der Mathematik (Bronstein)* -, Europa Lehrmittel Verlag, Haan-Gruiten, 2016.

[31] J. R. Bunch, C. P. Nielsen, and D. C. Sorensen, *Rank-one modification of the symmetric eigenproblem*, Numerische Mathematik, 31 (1978), pp. 31–48.

[32] K. Burg, H. Haf, F. Wille, and A. Meister, *Vektoranalysis - Höhere Mathematik für Ingenieure, Naturwissenschaftler und Mathematiker*, Springer-Verlag, Berlin Heidelberg New York, 2. aufl. ed., 2012.

[33] G. Buttazzo, M. Giaquinta, and S. Hildebrandt, *One-dimensional Variational Problems - An Introduction*, Clarendon Press, Oxford, 1998.

[34] R. Caduff, A. Kos, F. Schlunegger, B. McArdell, and A. Wiesmann, *Terrestrial radar interferometric measurement of hillslope deformation and atmospheric disturbances in the illgraben debris-flow catchment, switzerland*, IEEE Geoscience and Remote Sensing Letters, 11 (2014), pp. 434–438.

[35] R. Caduff, F. Schlunegger, A. Kos, and A. Wiesmann, *A review of terrestrial radar interferometry for measuring surface change in the geosciences*, Earth Surface Processes and Landforms, 40 (2015), pp. 208–228.

[36] M. Capinski and P. E. Kopp, *Measure, Integral and Probability* -, Springer Science & Business Media, Berlin Heidelberg, 2013.

[37] D. Chen and R. Plemmons, *Nonnegativity constraints in numerical analysis*, World Scientific, 2009.

[38] P. Chigansky and M. Kleptsyna, *Exact asymptotics in eigenproblems for fractional Brownian covariance operators*, arXiv e-prints, (2016), p. arXiv:1601.05715.

[39] J.-P. Chiles and P. Delfiner, *Geostatistics - Modeling Spatial Uncertainty*, John Wiley & Sons, New York, 2012.

[40] O. Christensen, *An Introduction to Frames and Riesz Bases* -, Birkhauser,

Basel, 2016.

[41] M. Costantini, *A novel phase unwrapping method based on network programming*, IEEE Transactions on Geoscience and Remote Sensing, 36 (1998), pp. 813–821.

[42] T. Cover and J. Thomas, *Elements of Information Theory* -, John Wiley & Sons, New York, 2 ed., 2012.

[43] N. Cressie, *The origins of kriging*, Mathematical Geology, 22 (1990), pp. 239–252.

[44] M. Crosetto, O. Monserrat, G. Luzi, M. Cuevas-Gonzalez, and N. Devanthery, *A noninterferometric procedure for deformation measurement using gb-sar imagery*, Geoscience and Remote Sensing Letters, IEEE, 11 (2014), pp. 34–38.

[45] R. W. R. Darling, *Differential Forms and Connections* -, Cambridge University Press, Cambridge, 1994.

[46] R. G. Deissler, *Is navier-stokes turbulence chaotic?*, The Physics of Fluids, 29 (1986), pp. 1453–1457.

[47] C. Del Ventisette, E. Intrieri, G. Luzi, N. Casagli, R. Fanti, and D. Leva, *Using ground based radar interferometry during emergency: the case of the a3 motorway (calabria region, italy) threatened by a landslide*, Natural Hazards and Earth System Science, 11 (2011), pp. 2483–2495.

[48] J. Dixmier, *C*-algebras* -, North-Holland, Amsterdam, 1982.

[49] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

[50] R. Edwards, *Functional Analysis - Theory and Applications*, Courier Corporation, New York, 2012.

[51] W. C. Elmore and M. A. Heald, *Physics of Waves* -, Courier Corporation, New York, 2012.

[52] K.-J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations* -, Springer Science & Business Media, Berlin Heidelberg, 2006.

[53] L. C. Evans, *Partial Differential Equations* -, American Mathematical Soc., Heidelberg, 2010.

[54] F. Fabry, *Radar Meteorology - Principles and Practice*, Cambridge University Press, Cambridge, 2015.

[55] J. Faillettaz, M. Funk, and D. Sornette, *Instabilities on alpine temperate glaciers: New insights arising from the numerical modelling of allalingletscher (valais, switzerland)*, Natural Hazards and Earth System Science, 12 (2012), pp. 2977–2991. cited By 9.

[56] J. FAILLETTAZ, D. SORNETTE, AND M. FUNK, *Numerical modeling of a gravity-driven instability of a cold hanging glacier: Reanalysis of the 1895 break-off of altelsgletscher, switzerland*, Journal of Glaciology, 57 (2011), pp. 817–831. cited By 15.

[57] S. J. FARLOW, *Partial Differential Equations for Scientists and Engineers* -, Courier Corporation, New York, 2012.

[58] G. E. FASSHAUER, *Green's functions: Taking another look at kernel approximation, radialbasis functions, and splines*, in Approximation Theory XIII: San Antonio 2010, M. Neamtu and L. Schumaker, eds., New York, NY, 2012, Springer New York, pp. 37–63.

[59] F. FAZAYELI AND A. BANERJEE, *The matrix generalized inverse gaussian distribution: Properties and applications*, in European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851, ECML PKDD 2016, Berlin, Heidelberg, 2016, Springer-Verlag, pp. 648–664.

[60] J. C. FERREIRA AND V. A. MENEGATTO, *Eigenvalues of integral operators defined by smooth positive definite kernels*, Integral Equations and Operator Theory, 64 (2009), pp. 61–81.

[61] A. FERRETTI, C. PRATI, AND F. ROCCA, *Permanent scatterers in sar interferometry*, IEEE Transactions on Geoscience and Remote Sensing, 39 (2001), pp. 8–20.

[62] H. FLANDERS, *Differential Forms with Applications to the Physical Sciences* -, Courier Corporation, New York, revised ed. ed., 1963.

[63] G. B. FOLLAND, *A Course in Abstract Harmonic Analysis, Second Edition* -, CRC Press, Boca Raton, Fla, 2016.

[64] W. FÖRSTNER, *Ein verfahren zur schätzung von varianz und kovarianzkomponenten*, Allgemeine Vermessungs-Nachrichten, 86 (1979), pp. 446–453.

[65] S. FOUCART AND H. RAUHUT, *A Mathematical Introduction to Compressive Sensing* -, Springer Science & Business Media, Berlin Heidelberg, 2013.

[66] G. R. FOWLES, *Introduction to Modern Optics* -, Courier Corporation, New York, 1989.

[67] V. FRANC, V. HLAVÁČ, AND M. NAVARA, *Sequential coordinate-wise algorithm for the non-negative least squares problem*, in Computer Analysis of Images and Patterns, A. Gagalowicz and W. Philips, eds., Berlin, Heidelberg, 2005, Springer Berlin Heidelberg, pp. 407–414.

[68] K. FUKUMIZU, L. SONG, AND A. GRETTON, *Kernel Bayes' rule*, ArXiv e-prints, (2010).

[69] K. FUKUMIZU, B. SRIPERUMBUDUR, A. GRETTON, AND B. SCHÖLKOPF, *Characteristic kernels on groups and semigroups*, in Proceedings of the

21st International Conference on Neural Information Processing Systems, NIPS'08, USA, 2008, Curran Associates Inc., pp. 473–480.

[70] B. GALANTI AND A. TSINOBER, *Is turbulence ergodic?*, Physics Letters A, 330 (2004), pp. 173 – 180.

[71] T. W. GAMELIN AND R. E. GREENE, *Introduction to Topology - Second Edition*, Courier Corporation, New York, 2013.

[72] D. C. GHIGLIA AND L. A. ROMERO, *Minimum lp-norm two-dimensional phase unwrapping*, J. Opt. Soc. Am. A, 13 (1996), pp. 1999–2013.

[73] A. GIL, J. SEGURA, AND N. M. TEMME, *Numerical Methods for Special Functions -*, SIAM, Philadelphia, 1. aufl. ed., 2007.

[74] R. GILMORE, *Lie Groups, Lie Algebras, and Some of Their Applications -*, Courier Corporation, New York, 2012.

[75] F. GIROSO, M. JONES, AND T.POGGIO, *Priors,stabilizers and basis functions:from regularization to radial tensor and additive splines*, A.I. Memo MIT, (1993).

[76] R. M. GOLDSTEIN AND C. L. WERNER, *Radar interferogram filtering for geophysical applications*, Geophysical Research Letters, 25 (1998), pp. 4035–4038.

[77] R. M. GOLDSTEIN, H. A. ZEBKER, AND C. L. WERNER, *Satellite radar interferometry: Two-dimensional phase unwrapping*, Radio Science, 23 (1988), pp. 713–720.

[78] G. H. GOLUB AND J. H. WELSCH, *Calculation of gauss quadrature rules*, Mathematics of Computation, 23 (1969), pp. 221–s10.

[79] W. GONG, F. MEYER, P. W. WEBLEY, D. MORTON, AND S. LIU, *Performance analysis of atmospheric correction in insar data based on the weather research and forecasting model (wrf)*, in 2010 IEEE International Geoscience and Remote Sensing Symposium, July 2010, pp. 2900–2903.

[80] E. GRAFAREND, *Harmonic maps*, Journal of Geodesy, 78 (2005), pp. 594–615.

[81] M. D. GREENBERG, *Applications of Green's Functions in Science and Engineering -*, Courier Dover Publications, Mineola, New York, 2015.

[82] A. GRETTON, K. M. BORGWARDT, M. J. RASCH, B. SCHÖLKOPF, AND A. SMOLA, *A kernel two-sample test*, J. Mach. Learn. Res., 13 (2012), pp. 723–773.

[83] A. GRETTON, O. BOUSQUET, A. SMOLA, AND B. SCHÖLKOPF, *Measuring statistical dependence with hilbert-schmidt norms*, in Algorithmic Learning Theory, S. Jain, H. U. Simon, and E. Tomita, eds., Berlin, Heidelberg, 2005, Springer Berlin Heidelberg, pp. 63–77.

[84] A. GRETTON, A. SMOLA, O. BOUSQUET, R. HERBRICH, A. BELITSKI, M. AUGATH, Y. MURAYAMA, J. PAULS, B. SCHOLKOPF, AND N. LOGOTHETIS, *Kernel constrained covariance for dependence measurement*, in Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, January 2005, pp. 1–8.

[85] D. GRIFFEL, *Applied Functional Analysis* -, Courier Corporation, New York, 2012.

[86] L. GROSS, *Measurable functions on hilbert space*, Transactions of the American Mathematical Society, 105 (1962), pp. 372–390.

[87] L. GROSS, *Abstract wiener spaces*, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Contributions to Probability Theory, Part 1, Berkeley, Calif., 1967, University of California Press, pp. 31–42.

[88] GROUNDPROBE, *Groundprobe ssr*. https://www.groundprobe.com/product/ssr-xt/. Accessed: 2019-01-09.

[89] M. HAASE, *The Functional Calculus for Sectorial Operators* -, Springer Science & Business Media, Berlin Heidelberg, 2006.

[90] R. HAMMING, *Numerical Methods for Scientists and Engineers* -, Courier Corporation, New York, 2nd revised ed. ed., 1973.

[91] R. F. HANSSEN, *Radar Interferometry. Data Interpretation and Error Analysis*, Springer Science & Business Media, Berlin Heidelberg, 2006.

[92] T. HASTIE, R. TIBSHIRANI, AND J. FRIEDMAN, *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, Springer Science & Business Media, Berlin Heidelberg, 2013.

[93] HELMBERG, *Introduction to Spectral Theory in Hilbert Space- Dover Books on Mathematics*, Dover Publications, Inc., Mineola, New York, 1997.

[94] E. HEWITT, *Abstract Harmonic Analysis* -, Springer Science & Business Media, Berlin Heidelberg, 1st ed. 1970. 2nd printing 1994 ed., 1994.

[95] E. HEWITT AND K. A. ROSS, *Abstract Harmonic Analysis - Volume I: Structure of Topological Groups Integration Theory Group Representations*, Springer, New York, 2013.

[96] F. HIAI AND D. PETZ, *Riemannian metrics on positive definite matrices related to means*, ArXiv e-prints, (2008).

[97] T. HIDA, *White Noise - An Infinite Dimensional Calculus*, Kluwer Academic Publishers, Dordrecht, 1993.

[98] J. C. HOLLADAY, *A smoothest curve approximation*, Mathematical Tables and Other Aids to Computation, 11 (1957), pp. 233–243.

[99] S. S. HOLLAND, *Applied Analysis by the Hilbert Space Method - An In-*

troduction with Applications to the Wave, Heat, and Schrödinger Equations, Courier Corporation, New York, 2012.

[100] R. A. Horn, R. A. Horn, and C. R. Johnson, *Matrix Analysis* -, Cambridge University Press, Cambridge, 1990.

[101] G. IDS, *Ibis arcsar*. https://idsgeoradar.com/products/interferometric-radar/ibis-arcsar. Accessed: 2019-01-09.

[102] ——, *Ibis fm*. https://idsgeoradar.com/products/interferometric-radar/ibis-fm. Accessed: 2019-01-09.

[103] ——, *Ibis fs*. https://idsgeoradar.com/products/interferometric-radar/ibis-fs. Accessed: 2019-01-09.

[104] R. Iglesias, X. Fabregas, A. Aguasca, J. J. Mallorqui, C. Lopez-Martinez, J. A. Gili, and J. Corominas, *Atmospheric phase screen compensation in ground-based sar with a multiple-regression model over mountainous regions*, IEEE Transactions on Geoscience and Remote Sensing, 52 (2014), pp. 2436–2449.

[105] E. Intrieri, F. Di Traglia, C. Del Ventisette, G. Gigli, F. Mugnai, G. Luzi, and N. Casagli, *Flank instability of stromboli volcano (aeolian islands, southern italy): Integration of gb-insar and geomorphological observations*, Geomorphology, 201 (2013), pp. 60–69.

[106] E. Isaacson and H. B. Keller, *Analysis of Numerical Methods* -, Courier Corporation, New York, 2012.

[107] T. Kailath, *Rkhs approach to detection and estimation problems–i: Deterministic signals in gaussian noise*, IEEE Transactions on Information Theory, 17 (1971), pp. 530–549.

[108] R. H. Kasriel, *Undergraduate Topology* -, Dover Publications, New York, 2009.

[109] D. Kissinger, *Millimeter-Wave Receiver Concepts for 77 GHz Automotive Radar in Silicon-Germanium Technology* -, Springer Science & Business Media, Berlin Heidelberg, 2012.

[110] H. Klausing, *Feasibility of a synthetic aperture radar with rotating antennas (rosar)*, in 1989 19th European Microwave Conference, Sep. 1989, pp. 287–299.

[111] H. Klausing and W. Holpp, *Radar mit realer und synthetischer Apertur. Konzeption und Realisierung*, Oldenbourg Wissenschaftsverlag, Berlin Heidelberg, 1999.

[112] E. d. Klerk, *Aspects of Semidefinite Programming - Interior Point Algorithms and Selected Applications*, Springer Science & Business Media, Berlin Heidelberg, 2006.

[113] A. W. Knapp, *Advanced Real Analysis* -, Springer Science & Business Me-

dia, Berlin Heidelberg, 2008.

[114] K.-R. KOCH, *Parameter Estimation and Hypothesis Testing in Linear Models* -, Springer Science & Business Media, Berlin Heidelberg, 2013.

[115] G. KOENIG, A. REIGBER, AND T. WESER, *An e-learning tutorial for radar remote sensing with rat*, in ISPRS Archives: Tools and Techniques for E-Learning, Commission VI, WG VI/1 - VI/2, vol. 36(6)/W30, Potsdam, Germany, 06 2005, ISPRS.

[116] D. KOLLER, N. FRIEDMAN, AND F. BACH, *Probabilistic Graphical Models - Principles and Techniques*, MIT Press, Cambridge, 2009.

[117] H. KÖNIG, *Eigenvalue Distribution of Compact Operators* -, Birkhäuser, Basel, 2013.

[118] B. KOOPMAN, *Hamiltonian systems and transformation in hilbert space*, Proceedings of the National Academy od Sciences of the United States of America, 17 (1931), pp. 315–318.

[119] M. KREIN AND D. MILMAN, *On extreme points of regular convex sets*, Studia Mathematica, 9 (1940), pp. 133–138.

[120] J. KUELBS, F. LARKIN, AND J. A. WILLIAMSON, *Weak probability distributions on reproducing kernel hilbert spaces*, Rocky Mountain J. Math., 2 (1972), pp. 369–378.

[121] S. KUMAR, *Eigenvalue statistics for the sum of two complex wishart matrices*, EPL (Europhysics Letters), 107 (2014), p. 60002.

[122] B. L., A. BRUNETTI, C. C., AND M. P., *Structural health characterization of an old riveted iron bridgeby remote sensing techniques*, in Proceedings of the 7th International Conference on Structural Health Monitoring of Intelligent Infrastructure, Turin, Italy, 2015, Curran Associates, Inc., pp. 1695–1705.

[123] C. LANCZOS, *The Variational Principles of Mechanics* -, Courier Corporation, New York, 2012.

[124] F. LARKIN, *Gaussian measure in hilbert space and applications in numerical analysis*, Rocky Mountain J. Math., 2 (1972), pp. 379–422.

[125] F. W. LAWVERE AND S. H. SCHANUEL, *Conceptual Mathematics - A First Introduction to Categories*, Cambridge University Press, Cambridge, 2009.

[126] T. LI, E. K. WAH CHU, W.-W. LIN, AND P. C.-Y. WENG, *Solving large-scale continuous-time algebraic riccati equations by doubling*, Journal of Computational and Applied Mathematics, 237 (2013), pp. 373 – 383.

[127] W. A. LIGHT AND E. W. CHENEY, *Approximation Theory in Tensor Product Spaces* -, Springer, Berlin, Heidelberg, 2006.

[128] L. H. LOOMIS, *Introduction to Abstract Harmonic Analysis* -, Courier Cor-

poration, New York, 2013.

[129] M. LUKIC AND J. BEDER, *Stochastic processes with sample paths in reproducing kernel hilbert spaces*, Transactions of the American Mathematical Society, 353 (2001), pp. 3945–3969.

[130] G. LUZI, L. NOFERINI, D. MECATTI, AND G. MACULOSO, *Using a ground-based sar interferometer and a terrestrial laser scanner to monitor a snow-covered slope: Results from an experimental data collection in tyrol (austria)*, in 2009 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), vol. 47(2), Munich, Germany, 02 2011, IEEE, pp. 382–393.

[131] O. L. MAITRE AND O. M. KNIO, *Spectral Methods for Uncertainty Quantification - With Applications to Computational Fluid Dynamics*, Springer Science & Business Media, Berlin Heidelberg, 2010.

[132] O. L. MANGASARIAN ET AL., *Generalized support vector machines*, Advances in Neural Information Processing Systems, (1999), pp. 135–146.

[133] K. V. MARDIA AND R. J. MARSHALL, *Maximum likelihood estimation of models for residual covariance in spatial regression*, Biometrika, 71 (1984), pp. 135–146.

[134] S. MARTINO AND P. MAZZANTI, *Integrating geomechanical surveys and remote sensing for sea cliff slope stability analysis: the mt. pucci case study (italy*, Natural Hazards and Earth System Sciences, 14 (2014), pp. 831–848.

[135] D. MAURO, *Topics in Koopman-von Neumann Theory*, eprint arXiv:quant-ph/0301172, (2003).

[136] M. MAZEIKA, *The singular value decomposition and low rank approximation*, 2016.

[137] E. L. MCHUGH, J. DWYER, D. LONG, AND C. SABINE, *Mining publication: Applications of ground-based radar to mine slope monitoring*, tech. rep., Office of Mine Safety and Health Research, 2006.

[138] S. MEHROTRA, *On the implementation of a primal-dual interior point method*, SIAM Journal on Optimization, 2 (1992), pp. 575–601.

[139] R. MEIDAN, *Reproducing-kernel hilbert spaces of distributions and generalized stochastic processes*, SIAM Journal on Mathematical Analysis, 10 (1979), pp. 62–70.

[140] J. L. MELSA AND A. P. SAGE, *An Introduction to Probability and Stochastic Processes -*, Courier Corporation, New York, 2013.

[141] B. MENDELSON, *Introduction to Topology - Third Edition*, Courier Corporation, New York, 2012.

[142] A. S. MONIN AND A. M. YAGLOM, *Statistical Fluid Mechanics - Mechanics of Turbulence*, Courier Corporation, New York, 2007.

[143] O. MONSERRAT, M. CROSETTO, AND G. LUZI, *A review of ground-based {SAR} interferometry for deformation measurement*, {ISPRS} Journal of Photogrammetry and Remote Sensing, 93 (2014), pp. 40 – 48.

[144] G. MOREAUX, J. P. BARRIOT, AND L. AMODEI, *A harmonic spline model for local estimation of planetary gravity fields from line-of-sight acceleration data*, Journal of Geodesy, 73 (1999), pp. 130–137.

[145] H. MORITZ, *The variational method of physical geodesy*, Bulletin géodésique, 54 (1980), pp. I–II.

[146] K. MUANDET, K. FUKUMIZU, B. SRIPERUMBUDUR, AND B. SCHÖLKOPF, *Kernel mean embedding of distributions: A review and beyond*, Foundations and Trends in Machine Learning, 10 (2017), pp. 1–141.

[147] P. N. SWARZTRAUBER, *The approximation of vector functions and their derivatives on the sphere*, Siam Journal on Numerical Analysis - SIAM J NUMER ANAL, 18 (1981), pp. 191–210.

[148] M. NASHED AND G. WAHBA, *Generalized inverses in reproducing kernel spaces: An approach to regularization of linear operator equations*, SIAM J. Math. Anal, 5 (1974), pp. 974–987.

[149] Y. NESTEROV, *Introductory Lectures on Convex Optimization - A Basic Course*, Springer Science & Business Media, Berlin Heidelberg, 2013.

[150] J. NEVEU, *Processus aléatoires gaussiens.*, Séminaire de mathématiques supérieures University of Montreal, Montreal, 1968.

[151] W. NIEMEIER, *Ausgleichungsrechnung - Statistische Auswertemethoden*, Walter de Gruyter, Berlin, 2008.

[152] L. NOFERINI, M. PIERACCINI, D. MECATTI, G. LUZI, C. ATZENI, A. TAMBURINI, AND M. BROCCOLATO, *Permanent scatterers analysis for atmospheric correction in ground-based sar interferometry*, Geoscience and Remote Sensing, IEEE Transactions on, 43 (2005), pp. 1459–1471.

[153] F. W. J. OLVER, *NIST Handbook of Mathematical Functions Hardback and CD-ROM -*, Cambridge University Press, Cambridge, 2010.

[154] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations -*, Springer Science & Business Media, Berlin Heidelberg, 2012.

[155] M. PIERACCINI AND L. MICCINESI, *Arcsar: Theory, simulations, and experimental verification*, IEEE Transactions on Microwave Theory and Techniques, 65 (2017), pp. 293–301.

[156] C. C. PINTER, *A Book of Abstract Algebra - Second Edition*, Dover Publications, New York, 2012.

[157] L. PIPIA, X. FABREGAS, A. AGUASCA, C. LOPEZ-MARTINEZ,

S. DUQUE, J. MALLORQUI, AND J. MARTURI, *Polarimetric differential sar interferometry: First results with ground-based measurements*, IEEE Geoscience and Remote Sensing Letters, 6 (2009), pp. 167–171. cited By (since 1996)15.

[158] G. F. PIVARO, S. KUMAR, G. FRAIDENRAICH, AND C. F. DIAS, *On the exact and approximate eigenvalue distribution for sum of wishart matrices*, IEEE Transactions on Vehicular Technology, 66 (2017), pp. 10537–10541.

[159] B. PORAT, *Digital Processing of Random Signals - Theory and Methods*, Courier Dover Publications, Mineola, New York, 2008.

[160] S. J. PRESS, *Applied Multivariate Analysis - Using Bayesian and Frequentist Methods of Inference, Second Edition*, Courier Corporation, New York, 2005.

[161] J. RAMSAY AND B. W. SILVERMAN, *Functional Data Analysis -*, Springer Science & Business Media, Berlin Heidelberg, 2013.

[162] C. E. RASMUSSEN AND C. K. I. WILLIAMS, *Gaussian Processes for Machine Learning -*, MIT Press, Cambridge, 2006.

[163] P. RIESEN, T. STROZZI, A. BAUDER, A. WIESMANN, AND M. FUNK, *Short-term surface ice motion variations measured with a ground-based portable real aperture radar interferometer*, Journal of Glaciology, 57 (2011), pp. 53–60.

[164] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis -*, Courier Corporation, New York, reprint ed., 1990.

[165] S. ROEDELSPERGER, M. BECKER, C. GERSTENECKER, G. LAEUFER, K. SCHILLING, AND D. STEINECK, *Generation of a Digital Elevation Model with the Ground Based SAR IBIS-L*, in EGU General Assembly Conference Abstracts, D. N. Arabelos and C. C. Tscherning, eds., vol. 11 of EGU General Assembly Conference Abstracts, Apr. 2009, p. 4656.

[166] S. ROEDELSPERGER, M. BECKER, C. GERSTENECKER, G. LAEUFER, K. SCHILLING, AND D. STEINECK, *Digital elevation model with the ground-based {SAR} ibis-l as basis for volcanic deformation monitoring*, Journal of Geodynamics, 49 (2010), pp. 241 – 246. {WEGENER} 2008 - Proceedings of the 14th General Assembly of Wegener.

[167] S. ROEDELSPERGER, A. COCCIA, D. VICENTE, C. TRAMPUZ, AND A. META, *The novel fastgbsar sensor: Deformation monitoring for dike failure prediction*, 2013, pp. 420–423. cited By (since 1996)0.

[168] W. RUDIN, *Real and Complex Analysis -*, Tata McGraw-Hill, New York, 1987.

[169] J. RÜEGER, *Refractive index formulae for radio waves*, Proc. FIG XXII International Congress, Washington, D. C., (2002).

[170] S. SAITOH AND Y. SAWANO, *Theory of Reproducing Kernels and Applica-*

tions -, Springer, Berlin, Heidelberg, 2016.

[171] D. Sandwell, R. Mellors, X. Tong, M. Wei, and P. Wessel, *Open radar interferometry software for mapping surface deformation*, Eos, Transactions American Geophysical Union, 92 (2011), pp. 234–234.

[172] H. Schneider and G. P. Barker, *Matrices and Linear Algebra* -, Courier Corporation, New York, 2012.

[173] B. Schölkopf, R. Herbrich, and A. J. Smola, *A generalized representer theorem*, in Computational Learning Theory, D. Helmbold and B. Williamson, eds., Berlin, Heidelberg, 2001, Springer Berlin Heidelberg, pp. 416–426.

[174] B. Schölkopf and A. J. Smola, *Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, 2002.

[175] A. J. Schwab, *Begriffswelt der Feldtheorie - Praxisnahe, anschauliche Einführung*, Springer-Verlag, Berlin Heidelberg New York, 2013.

[176] F. Schwabl, *Statistical Mechanics* -, Springer Science & Business Media, Berlin Heidelberg, 2006.

[177] E. Serantoni, M. Muster, and A. Wieser, *Numerical structural identification of a cross-laminated timber slab using 3d laserscanning*, The e-journal of nondestructive testing & ultrasonics, (2018), p. 0065.

[178] R. Shankar, *Principles of Quantum Mechanics* -, Springer Science & Business Media, Berlin Heidelberg, 2012.

[179] S. Sharma and J. W. Cutler, *Robust orbit determination and classification: A learning theoretic approach*, Interplanetary Network Progress Report, 42 (2015), pp. 1–20.

[180] G. E. Shilov, *Linear Algebra* -, Courier Corporation, New York, 2012.

[181] ——, *Elementary Functional Analysis* -, Courier Corporation, New York, 2013.

[182] G. E. Shilov, G. E. Silov, and R. A. Silverman, *Elementary Real and Complex Analysis* -, Courier Corporation, New York, 1996.

[183] E. K. Smith and S. Weintraub, *The constants in the equation for atmospheric refractive index at radio frequencies*, Proceedings of the IRE, 41 (1953), pp. 1035–1037.

[184] A. Smola, A. Gretton, L. Song, and B. Schölkopf, *A hilbert space embedding for distributions*, in Algorithmic Learning Theory, M. Hutter, R. A. Servedio, and E. Takimoto, eds., Berlin, Heidelberg, 2007, Springer Berlin Heidelberg, pp. 13–31.

[185] K. Sobczyk and D. J. Kirkner, *Stochastic Modeling of Microstructures*

-, Springer Science & Business Media, Berlin Heidelberg, 2012.

[186] L. Song, J. Huang, A. Smola, and K. Fukumizu, *Hilbert space embeddings of conditional distributions with applications to dynamical systems*, in Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, New York, NY, USA, 2009, ACM, pp. 961–968.

[187] E. S.r.l., *Lisalab*. http://www.lisalab.com/engl/default.asp?seze=6. Accessed: 2019-01-09.

[188] F. Steinke and B. Schölkopf, *Kernels, regularization and differential equations*, Pattern Recognition, 41 (2008), pp. 3271–3286.

[189] A. G. Stove, *Linear fmcw radar techniques*, IEE Proceedings F - Radar and Signal Processing, 139 (1992), pp. 343–350.

[190] G. Strang, *Introduction to Linear Algebra* -, Wellesley-Cambridge Press, Wellesley, 2016.

[191] T. Strozzi, C. Werner, A. Wiesmann, and U. Wegmuller, *Topography mapping with a portable real-aperture radar interferometer*, Geoscience and Remote Sensing Letters, IEEE, 9 (2012), pp. 277–281.

[192] J. F. Sturm, *Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones*, Optimization Methods and Software, 11 (1999), pp. 625–653.

[193] D. Tarchi, H. Rudolf, G. Luzi, L. Chiarantini, P. Coppo, and A. J. Sieber, *Sar interferometry for structural changes detection: a demonstration test on a dam*, in IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No.99CH36293), vol. 3, June 1999, pp. 1522–1524 vol.3.

[194] G. Teza, M. Balzani, L. Noferini, A. Galgaro, G. Luzi, Ucelli.F, M. Pieraccini, S. Silvano, N. Zaltron, R. Genevois, G. Galvani, D. Mecatti, G. Macaluso, M. Pieraccini, and C. Atzeni, *Ground-based monitoring of high-risk landslides through joint use of laser scanner and interferometric radar*, International Journal of Remote Sensing, 29 (2008), pp. 4735–4756.

[195] K. C. Toh, M. J. Todd, and R. H. Tütüncü, *Sdpt3 —a matlab software package for semidefinite programming, version 1.3*, Optimization Methods and Software, 11 (1999), pp. 545–581.

[196] M. Tucsnak and G. Weiss, *Observation and Control for Operator Semigroups* -, Springer Science & Business Media, Berlin Heidelberg, 2009.

[197] A. N. Uddin, *Generalized Functionals Of Brownian Motion And Their Applications: Nonlinear Functionals Of Fundamental Stochastic Processes* -, World Scientific, Singapur, 2011.

[198] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Re-

view, 38 (1996), pp. 49–95.

[199] L. Vandenberghe, S. Boyd, and S.-P. Wu, *Determinant maximization with linear matrix inequality constraints*, SIAM Journal on Matrix Analysis and Applications, 19 (1998), pp. 499–533.

[200] C. Vinzant, *What is...a spectrahedron?*, Notices of the American Mathematical Society, 61(5) (2014), pp. 492–493.

[201] D. Voytenko, T. Dixon, C. Werner, N. Gourmelen, I. Howat, P. Tinder, and A. Hooper, *Monitoring a glacier in southeastern iceland with the portable terrestrial radar interferometer*, in Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International, July 2012, pp. 3230–3232.

[202] G. Wahba, *Spline interpolation and smoothing on the sphere. siam j sci stat comput 2:5-16*, Siam Journal on Scientific and Statistical Computing, 2 (1981).

[203] ——, *Spline Models for Observational Data* -, SIAM, Philadelphia, new. ed., 1990.

[204] L. Wang, Y. Li, J. Jia, J. Sun, D. Wipf, and J. Rehg, *Learning sparse covariance patterns for natural scenes*, in Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 05 2012, pp. 2767–2774.

[205] R. Weinstock, *Calculus of Variations - With Applications to Physics and Engineering*, Courier Corporation, New York, 1974.

[206] D. Werner, *Funktionalanalysis* -, Springer-Verlag, Berlin Heidelberg New York, 2011.

[207] A. Wilansky, *Topology for Analysis* -, Courier Corporation, New York, 2008.

[208] S. Willard, *General Topology* -, Courier Corporation, New York, 1970.

[209] B. Witte and P. Sparla, *Vermessungskunde und Grundlagen der Statistik fuer das Bauwesen* -, Vde Verlag GmbH, Berlin, Offenbach, 2015.

[210] H. Wolkowicz, R. Saigal, and L. Vandenberghe, *Handbook of Semidefinite Programming - Theory, Algorithms, and Applications*, Springer Science & Business Media, Berlin Heidelberg, 2012.

[211] Y. Xia, *The maxdet problem - algorithm and applications in machine learning*.

[212] J. Yue, Z. Qiu, X. Wang, and S. Yue, *Atmospheric phase correction using permanent scatterers in ground-based radar interferometry*, Journal of Applied Remote Sensing, 10 (2016).

[213] C. Zhou and A. Wieser, *Jaccard Analysis and LASSO-Based Feature Selection for Location Fingerprinting with Limited Computational Complexity*, Springer International Publishing, Cham, 2018, pp. 71–87.