Machine learning and geodesy: A survey

Originally published in:
Journal of Applied Geodesy 15(2), https://doi.org/10.1515/jag-2020-0043

Jemil Butt*, Andreas Wieser, Zan Gojcic, and Caifa Zhou

Machine learning and geodesy: A survey

https://doi.org/10.1515/jag-2020-0043
Received October 7, 2020; accepted January 29, 2021

Abstract: The goal of classical geodetic data analysis is often to estimate distributional parameters like expected values and variances based on measurements that are subject to uncertainty due to unpredictable environmental effects and instrument specific noise. Its traditional roots and focus on analytical solutions at times require strong prior assumptions regarding problem specification and underlying probability distributions that preclude successful application in practical cases for which the goal is not regression in presence of Gaussian noise.

Machine learning methods are more flexible with respect to assumed regularity of the input and the form of the desired outputs and allow for nonparametric stochastic models at the cost of substituting easily analyzable closed form solutions by numerical schemes. This article aims at examining common grounds of geodetic data analysis and machine learning and showcases applications of algorithms for supervised and unsupervised learning to tasks concerned with optimal estimation, signal separation, danger assessment and design of measurement strategies that occur frequently and naturally in geodesy.

Keywords: Geodesy, Adjustment theory, Machine learning, Hilbert spaces, Kernel methods

1 Introduction

One widely adopted definition of machine learning describes it as the study of algorithms whose performance on a specific task increases with experience [22]. Here experience is usually quantified by the amount of data and performance is measured by a function that includes intuition on when a result of the algorithm could be considered desirable.

In this generality, the definition encompasses actions as simple as taking the arithmetic mean $\bar{x} = n^{-1}\sum_{k=1}^{n} x_k$ of a dataset $\{x_k\}_{k=1}^{n}$ as an estimator for the expected value $\mu_X = E[X]$ of the random variable X with independent samples X_k and realizations x_k. This is due to the fact that $\bar{x}$ is a solution to the optimization problem

$$\bar{x} = \underset{\hat{x}\in\mathbb{R}}{\operatorname{argmin}} \sum_{k=1}^{n} (\hat{x} - x_k)^2 \tag{1}$$

and its performance as measured by the variance of the residual random variable $\bar{X} - X = n^{-1}\sum_{k=1}^{n} X_k - X$ for any new independent observation X

$$\begin{aligned}\sigma^2_{\bar{X}-X} &= E\left[\left(\frac{1}{n}\sum_{k=1}^{n} X_k - X\right)^2\right] - E\left[\frac{1}{n}\sum_{k=1}^{n} X_k - X\right]^2 \\ &= \frac{1}{n}E[X^2] + E[X^2]\end{aligned} \tag{2}$$

gets better (lower standard deviation) with increasing sample size n. Although typically one expects from machine learning algorithms a more complex interaction with the data, the above example is instructive in the sense that the task is to be expressed in the language of a probabilistically motivated optimization problem upon which it is solved employing numerical routines. The exact way in which this optimization is carried out will be of no concern in this article; instead priority is given to an intuitive explanation of the correspondence between learning task and equivalent optimization problem as the authors have found lacking clarity in this area to be the main impediment to understanding machine learning algorithms acting in high dimensional or even infinite dimensional settings.

Presumably due to missing or unclear links failing to explain the relation between classical adjustment theory and machine learning algorithms, the latter are rarely used in geodesy. Not considering photogrammetry and remote sensing, whose ties particularly to computer vision are undeniably strong and result in extensive usage of learning algorithms, references are rather limited.

In the field of physical geodesy, numerical reasons arising during manipulation of ill-conditioned normal matrices for example in gravity field estimation [20] have lead to widespread use of regularization which — as we will show in section 2 — is mathematically equivalent to statistical inference with a prior on some function space. Apart from this slightly hidden link, individual publications have addressed directly concrete applications ranging from system identification employing neural-networks [24] to the use of support vector machines for velocity

***Corresponding author: Jemil Butt,** ETH Zürich, Institute of Geodesy and Photogrammetry, Stefano-Franscini-Platz 5, CH-8093 Zürich, Switzerland, e-mail: jemil.butt@geod.baug.ethz.ch
Andreas Wieser, Zan Gojcic, Caifa Zhou, ETH Zürich, Institute of Geodesy and Photogrammetry, Stefano-Franscini-Platz 5, CH-8093 Zürich, Switzerland, e-mails: andreas.wieser@geod.baug.ethz.ch, zan.gojcic@geod.baug.ethz.ch, caifa.zhou@geod.baug.ethz.ch

field interpolation in the context of landslide monitoring [30] and thereby hinted at some of the potential of machine learning methods for typical geodetic core-tasks. [29] stake out the role artificial intelligence might have to play in geodesy and list several algorithms; however, they focus more on possible future developments whereas we want to make explicit mathematical equivalences and differences in perspective between the data analytical approaches taken in geodesy and machine learning.

One distinguishes machine learning tasks regarding the given inputs and the desired outputs. When a set of independent variables x_k and corresponding response variables y_k is given in the form of a sequence $\{(x_k, y_k)\}_{k=1}^n$ and the algorithm is supposed to closely emulate the mapping $f : x_k \mapsto y_k$ the task is said to be supervised [18, pp. 26–28]. When only a sequence $\{(x_k)\}_{k=1}^n$ is given and structure is to be found without further guidance the task is called unsupervised. Many intermediate shades exist between the two extremes; e. g. reinforcement learning in which an algorithm — designed to find optimal strategies in a stochastically changing environment — receives positive or negative feedback but no ground truth or optimal strategy is known that could serve to construct reference values y_k [34]. This scheme of clustering machine learning tasks can be contrasted with a more output oriented one, in which a task is called regression if the output is numerical or classification if it is categorical to name only the two most common formats [18, pp. 26–28].

The premises of geodetic data analysis as embodied by what is known as adjustment theory are typically narrower [5]: Measurements are sequences of real numbers $\{y_k\}_{k=1}^n$ and there exists a set of parameters $\{\lambda_k\}_{k=1}^m$ such that its transform $A(\{\lambda_k\}_{k=1}^m)$ by some function A resembles $\{y_k\}_{k=n}^n$ apart from a residual term that is assumed to be entirely stochastic in nature [26, p. 137]. Several extensions exist most notably among them collocation; see e. g. [23, 6]. The above problem is a supervised regression problem one could equally well tackle with different methods. In the next section classical least squares solutions for a very basic estimation task are rederived from different starting points. This will reveal differences in philosophies between geodesy and machine learning regarding how to pose a problem even though the calculations ultimately yield the same equation. The equivalence of adjustment to an algorithm that may be considered as belonging to machine learning (Gaussian process regression / Kriging) and one that surely does so (optimization in reproducing kernel Hilbert spaces / splines) is shown and augmented with a Bayesian interpretation. Section 3 is devoted to toy examples from geodesy that defy being solved by adjustment theory and require algorithms from machine learning that at first glance might seem obscure in this setting but will be demonstrated to work reasonably well and arise naturally when the viewpoint developed in section 2 is taken.

In those toy problems a dataset containing total station observations is subjected to a kernel based time series analysis to separate signal from noise, classified by a support vector machine (SVM) as stable or instable and split into maximally independent parts by kernel independent component analysis (K-ICA). We hasten to note that the examples presented in this paper are of an illustrative nature before closing with a discussion of the results and an outlook on potentially interesting and worthwhile future applications.

2 Adjustment and machine learning

We proceed by applying adjustment theory to a simple 1 dimensional regression / interpolation problem. By tackling the same task with geostatistical and functional analytic methods, the connections to statistical inference and deterministic function approximation are highlighted. This allows to coach adjustment theory in a learning framework. Both adjustment and machine learning procedures make use of the same words but their meanings often differ considerably. To alleviate the confusion we will always define the quantities appearing in this chapter strictly mathematically and we try to keep with the usual notational customs of the respective fields as far as no contradictions arise. Furthermore, we hope that Table 1 provides a guideline to translate terminology between machine learning and adjustment based approaches to estimation and urge the reader to briefly skim over it before entering the next section. However, it is by no means complete and the reader will have to fill in some of the missing pieces him or herself as he or she advances through the text.

The mode of presentation is geared towards paralleling that of earlier survey articles establishing links between processing schemes in geodesy and various other disciplines of science; we specifically recommend [16].

2.1 Regression / interpolation problem

Suppose n observations $\{y_k\}_{k=1}^n$ are given together with the locations $\{x_k\}_{k=1}^n \subset \mathcal{X}$ at which they were performed. The goal is to estimate the values $y(x)$ even for unobserved locations $x \in \mathcal{X}$, see Fig. 1 for an illustration.

A typical set of assumptions and procedures to derive a solution within an adjustment theoretic framework would consist in the items listed below.

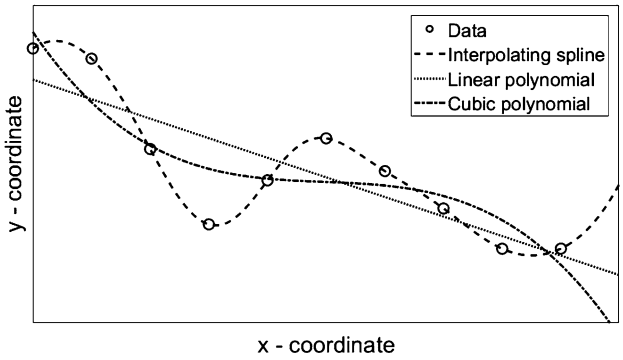

Figure 1: A one dimensional illustration of the regression / interpolation problem posed above. Without prior knowledge it is not clear which estimator — represented here by the various broken lines — is the most appropriate.

i) Assume there is an underlying deterministic function of x depending linearly on a set of m parameters λ, i. e. $(y_{\text{true}})_i = \sum_{j=1}^{m} \lambda_j g_j(x_i) = A\lambda$ with the $n \times m$ Matrix A having entries $(A)_{ij} = g_j(x_i)$.
ii) The deviations between y_{true} and y are due to measurement noise which is assumed to be multivariate Gaussian with expected value zero and covariance matrix Σ_v, preferably diagonal.
iii) Minimize the weighted sum of squares $v^T\Sigma_v^{-1}v$ of residuals $v(\lambda) = A\lambda - y$ by choosing the optimal set λ^* of parameters λ.

We arrive at the following-Gauss Markov model [26, p. 137]:

$$
\begin{aligned}
&A\lambda - y = v \qquad E[v] = 0 \qquad E[v_i v_j] = (\Sigma_v)_{ij}\\
&y \in \mathbb{R}^n \qquad y = [y_1, \dots, y_n]^T \qquad y_k = k\text{-th observation}\\
&\lambda \in \mathbb{R}^m \qquad \lambda = [\lambda_1, \dots, \lambda_m]^T \qquad \lambda_k = k\text{-th parameter}\\
&A \in \mathbb{R}^n \otimes \mathbb{R}^m \qquad A = [a_1, \dots, a_n]^T\\
&\quad \text{with } a_i = [a_{i1}, \dots, a_{im}]^T \text{ and } (A\lambda)_k = a_k^T\lambda.
\end{aligned}
$$

Bar some technicalities regarding invertibility of A^TA the solution can be written as the estimator $\hat{y}_{\text{true}}(x) = \sum_{j=1}^{m} \lambda_j^* g_j(x)$ where the parameters are optimal in the sense of being a minimizer for the discrepancy measure $\|\Sigma_v^{-1/2}(\hat{y}_{\text{true}} - y)\|^2$, i. e.

$$
\begin{aligned}
\lambda^* &= \operatorname*{argmin}_{\lambda \in \mathbb{R}^m} \|\Sigma_v^{-1/2}(A\lambda - y)\|_{\ell^2}^2\\
&= \operatorname*{argmin}_{\lambda \in \mathbb{R}^m} \|\widetilde{A}\lambda - \widetilde{y}\|_{\ell^2}^2\\
&= \widetilde{A}^+\widetilde{y} \qquad (3)
\end{aligned}
$$

where $\widetilde{A} = \Sigma_v^{-1/2}A$, $\widetilde{y} = \Sigma_v^{-1/2}y$ and $\widetilde{A}^+$ is the pseudoinverse of $\widetilde{A}$ [33, p. 218]. Therefore the well known formula for λ^* is $\lambda^* = (A^T\Sigma_v^{-1}A)^{-1}A^T\Sigma_v^{-1}y$. Then λ^* is an m-dimensional vector, $A\lambda^*$ is an n-dimensional vector representing estimations of the noiseless y_{true} at the observed locations and $\hat{y}_{\text{true}}(\cdot) = \sum_{j=1}^{m} \lambda_j^* g_j(\cdot)$ is a function of $x \in \mathcal{X}$. The family $\{g_j(\cdot)\}_{j=1}^m$ of functions used to approximate y only enters the problem formulation via the matrix A where each row of A is a row vector a_k^T of the possibly nonlinear functions $g_j(\cdot)$ acting on x; $(a_k^T)_j = g_j(x_k)$. The estimators $\hat{y}_{\text{true}}(\cdot)$ for different choices of function classes $\{g_j\}_{j=1}^m$ (linear, cubic) are shown in Fig. 1.

The probabilistic interpretation is quite straightforward. Under the assumption that the stochastic model of v being multivariate Gaussian with $E[v] = 0$ and $E[v \otimes v^*] = \Sigma_v$ is correct, one may write [27, p. 68]

$$
\begin{aligned}
L(\lambda, v) &= f_v(v|\lambda)\\
&= (2\pi)^{-n/2}\sqrt{\det \Sigma_v}\exp\left[-\frac{1}{2}v(\lambda)^T\Sigma_v^{-1}v(\lambda)\right]\\
\log f_v(v|\lambda) &= c_1 - c_2\left([\Sigma_v^{-1/2}v(\lambda)]^T[\Sigma_v^{-1/2}v(\lambda)]\right)\\
&= c_1 - c_2\|\Sigma_v^{-1/2}(A\lambda - y)\|_{\ell^2}^2 \qquad (4)
\end{aligned}
$$

where c_1 and c_2 are constants and $f_v(v|\lambda)$ is the conditional probability density function of the random variable v representing the residuals due to measurement error given parameters λ and the distributional information about their means and covariances. Since $\log(\cdot)$ is a monotonous function, the maximizer of $\log f_v$ is also the maximizer of f_v and the likelihood $L(\lambda, v)$ implying that the least squares solution is a maximum likelihood estimator. Note at this point that the likelihood $L(\lambda, v) = f_v(v|\lambda)$ is proportional to $f_\lambda(\lambda|v)$ via Bayes rule [27, p. 60]

$$
f_\lambda(\lambda|v) = f_v(v|\lambda)f_\lambda(\lambda)\left[\int_{-\infty}^{\infty} f_v(v|\lambda)f_\lambda(\lambda)d\lambda\right]^{-1}
$$

under an assumed uniform distribution for λ. Then the maximum likelihood estimate is actually the Bayesian maximum a posteriori estimate. Equation (4) consequently establishes a link between maximum likelihood estimation, norm minimization and, in special cases, Bayesian inference.

2.2 Adjustment as a learning task

The adjustment approach to interpolation can be identified as a supervised regression problem. The set of tuples $\{(x_k, y_k)\}_{k=1}^n$ are the training data, the goal is to approximate the input-output relation between the $\{x_k\}_{k=1}^n$ and $\{y_k\}_{k=1}^n$ where the decision variable is the target vector λ. Essentially nothing changes, if the pretext of an artificial

interpolation problem is dropped; the solution of a linear adjustment problem in Gauss-Markov form can always be written as [13, p. 93]

$$\hat{y}(\cdot) = \sum_{k=1}^{m} \lambda_k^* g_k(\cdot) \tag{5}$$

$$\lambda^* = \underset{\lambda \in \mathbb{R}^m}{\operatorname{argmin}} \|A(x)\lambda - y\|_{\mathcal{H}}^2 \tag{6}$$

where $\langle f, g\rangle_{\mathcal{H}} = \langle \Sigma^{-1} f, g\rangle_{\ell^2}$ is the inner product in some Hilbert space. Here we wrote $A(x)$ to explicitly document that the design matrix contains nonlinear features in x — a notion that is quite straightforward to interpret in the interpolation case as

$$A(x) = \begin{bmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \vdots & & \vdots \\ g_1(x_n) & \dots & g_m(x_n) \end{bmatrix}$$

in this case contains e. g. polynomials in x. However in arbitrary abstract adjustment problems, it might not always be easy to identify what the independent variable $\{x_k\}_{k=1}^n$ corresponds to, if just $A(x)$ as a matrix of features is provided. When for example the levelling problem (7)

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix}}_{A(x)} \underbrace{\begin{bmatrix} H_1 \\ H_2 \\ H_3 \end{bmatrix}}_{\lambda} \approx \underbrace{\begin{bmatrix} H_1^* \\ \Delta h_1 \\ \Delta h_2 \\ \Delta h_3 \end{bmatrix}}_{y} \tag{7}$$

$$\begin{aligned} H_1^* &: \text{ Approximately known height} \\ H_k &: \text{ Heights to be determined} \\ \Delta h_k &: \text{ Measured height difference} \end{aligned} \tag{8}$$

is given, it is quite hard to interpret the rows of $A(x)$ as nonlinear features of some scalar x. However, we might always resort to the mental trick of considering the rows of $A(x)$ as linear features of a vector valued independent variable $x \in \mathbb{R}^m$. Concretely this means having as training data $\{(x_k, y_k)\}_{k=1}^n = \{([1\ 0\ 0]^T, H_1^*), ([1\ -1\ 0]^T, \Delta h_1), \dots\}$ and approximating a function $f : \mathbb{R}^3 \to \mathbb{R}$ that maps the x_k to the y_k linearly, i. e. $f(x_k) = f([x_k^1, x_k^2, x_k^3]^T) = H_1 x_k^1 + H_2 x_k^2 + H_3 x_k^3$. But this equation just defines a hyperplane in $\mathbb{R}^4$ indicating that the adjustment problem has been reduced to a simple regression in a higher dimensional space.

We present for comparison a geostatistical and a functional analytic approach, that both enjoy some popularity in the machine learning community under the names of Gaussian process regression and splines in reproducing kernel Hilbert space. The equations will largely be identical but the spirit is noticeably different.

2.3 Adjustment, geostatistics and splines

In geostatistics, to solve the interpolation problem, one would assume the observations y_k to be realizations of a stochastic process $\{Y(x_k)\}_{k=1}^n$ with $Y(x) \in L^2(\Omega)$ a square integrable random variable for all $x \in \mathcal{X}$ and an estimator $\hat{Y}(x)$ for $Y(x)$ in general is sought. Assemble this estimator as a function of the given random variables $\{Y(x_k)\}_{k=1}^n$ in such a way as to minimize the expected square loss $E[\left(\hat{Y}(x) - Y(x)\right)^2]$ which is the error variance of the estimation.

It can be proven [25] that for a zero-mean Gaussian process the best predictor $\hat{Y}$ functionally dependent on some set $Y_k = Y(x_k)$, $k = 1, \dots, n$ is the conditional expectation, which is furthermore linear in its arguments.

$$\hat{Y}(x) = E[Y(x)|Y_1, \dots, Y_n] \tag{9}$$

$$\hat{Y}(x) = \sum_{k=1}^{n} \alpha_k Y_k \tag{10}$$

Presupposing knowledge of the mean-zero joint Gaussian distribution, denote by $\sigma(Y(x_1), Y(x_2))$ the covariance $E[Y(x_1)Y(x_2)]$ of the two random variables $Y(x_1), Y(x_2) \in L^2(\Omega); x_1, x_2 \in \mathcal{X}$. To find these α for which $\hat{Y}(x) = \sum_{k=1}^n \alpha_k Y_k$ is the conditional expectation, minimize

$$E[\left(\hat{Y}(x) - Y(x)\right)^2] = \sigma(\hat{Y} - Y, \hat{Y} - Y) =: \sigma_\alpha^2(v(x)).$$

Since the covariance $\sigma(\cdot, \cdot)$ is bilinear in its arguments, this amounts to solving $\partial/\partial\alpha_k \sigma_\alpha^2(v(x)) = 0$, $k = 1, \dots, n$ with

$$\begin{aligned} \sigma_\alpha^2(v(x)) &= \sigma\left(\sum_{i=1}^n \alpha_i Y_i - Y, \sum_{j=1}^n \alpha_j Y_j - Y\right) \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \sigma(Y_i, Y_j) - 2\sum_{i=1}^n \alpha_i \sigma(Y_i, Y) + \sigma(Y, Y) \end{aligned} \tag{11}$$

This immediately implies

$$\frac{\partial}{\partial \alpha_k} \sigma_\alpha^2(v(x)) = 2\left[\sum_{i=1}^n \alpha_i \sigma(Y_i, Y) - \sigma(Y_i, Y)\right] \overset{!}{=} 0$$

and α consequently satisfies

$$\underbrace{\begin{bmatrix} \sigma(Y_1, Y_1) & \dots & \sigma(Y_1, Y_n) \\ \vdots & \ddots & \vdots \\ \sigma(Y_n, Y_1) & \dots & \sigma(Y_n, Y_n) \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}}_{\alpha} = \underbrace{\begin{bmatrix} \sigma(Y_1, Y(x)) \\ \vdots \\ \sigma(Y_n, Y(x)) \end{bmatrix}}_{\Sigma_x} \tag{12}$$

The above formulae are known as the simple Kriging equations [10, p. 152]. Solving this system leads to the optimal

choice of coefficients α for assembling the simple Kriging predictor $\hat{Y}_{\mathrm{SK}} = \sum_{k=1}^{n} \alpha_k Y_k$ out of measurements Y_k, $k = 1, \ldots, n$ and finally

$$\hat{Y}_{\mathrm{SK}}(x) = \alpha^T \{Y_k\}_{k=1}^n = \Sigma_x^T \Sigma^{-1} \{Y_k\}_{k=1}^n. \tag{13}$$

In the case where also the mean function is unknown, needs to be estimated and has form $h(x) = \sum_{l=1}^{m} \beta_l g_l(x)$, the universal Kriging system [10, p. 168] arises instead:

$$\begin{bmatrix} \Sigma & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \mu \end{bmatrix} = \begin{bmatrix} \Sigma_x \\ A_x \end{bmatrix} \tag{14}$$

where α, Σ, Σ_x are defined as in equation (12), μ is some m-dimensional Lagrange multiplier, $(A)_{ij} = g_j(x_i)$ and $(A_x)_j = g_j(x)$ defines a column vector. For a fixed $x \in \mathcal{X}$, the optimal estimator is the universal Kriging predictor $\hat{Y}_{\mathrm{UK}} = \sum_{k=1}^{n} \alpha_k Y_k$ with the α chosen to satisfy the system of linear equations specified above.

By solving system (14) via substitution, the coefficient vector $\alpha = [\alpha_1, \ldots, \alpha_n]^T$ is found explicitly and the estimator can be decomposed into three components.

$$\begin{aligned} \alpha &= \Sigma^{-1} \left[\Sigma_x - A(A^T \Sigma^{-1} A)^{-1} \left(A^T \Sigma^{-1} \Sigma_x - A_x \right) \right] \\ \hat{Y}_{\mathrm{UK}}(x) &= \underbrace{\Sigma_x^T \Sigma^{-1} \{Y_k\}_{k=1}^n}_{\hat{Y}_1(x)} \\ &\quad + \underbrace{A_x^T (A^T \Sigma^{-1} A)^{-1} A^T \Sigma^{-1} \{Y_k\}_{k=1}^n}_{\hat{Y}_2(x)} \\ &\quad - \underbrace{\Sigma_x^T \Sigma^{-1} A (A^T \Sigma^{-1} A)^{-1} A^T \Sigma^{-1} \{Y_k\}_{k=1}^n}_{\hat{Y}_3(x)} \end{aligned} \tag{15}$$

Comparing the above terms to equations (13) and (3), we find that

$$\hat{Y}_1(x) = \hat{Y}_{\mathrm{SK}}(x) \quad \hat{Y}_2(x) = \hat{Y}_{\mathrm{Adjustment}}(x)$$

and $\hat{Y}_3(x)$ is a cross term accounting for the fact that the estimated mean $\hat{Y}_{\mathrm{Adjustment}}(x)$ needs to be subtracted for normalization. An alternative way of writing (15) would therefore be

$$\hat{Y}_{UK}(x) = \hat{Y}_{\mathrm{Adjustment}}(x) + \hat{V}_{\mathrm{SK}}(x) \tag{16}$$

where $V(x) = Y(x) - \hat{Y}_{\mathrm{Adjustment}}(x)$, the residual after subtraction of the estimated mean function $h(x) = A_x^T (A^T \Sigma^{-1} A)^{-1} A^T \Sigma^{-1} \{Y_k\}_{k=1}^n$. We find the main difference to adjustment to be the existence of an estimation term for a stochastic component owing to the fact that what we want to estimate is only somewhat correlated to the measurements.

Summarizingly, from the geostatistical perspective, the inclusion of randomness results in a more flexible model for the predictions and residuals. This contrasts with the randomnesses role as a cover term to subsume unwanted and unmodelled effects in terms of deviations from the parametric model in classical adjustment.

In the approach described above, we minimized

$$E[(\hat{Y}(x) - Y(x))^2] = \|\hat{Y}(x) - Y(x)\|^2_{L^2(\Omega)}$$

pointwise for each $x \in \mathcal{X}$ separately to derive a predictor $\hat{Y}(x)$ because we took as fundamental the notion of a random variable and its variance. It is possible to abstract from this situation by introducing spaces $\mathcal{H}(\mathcal{X})$ of functions $f : \mathcal{X} \to \mathbb{R}$ with Gaussian measures on these spaces [21] which allow writing the probability of having a randomly drawn $f \in \mathcal{H}(\mathcal{X})$ in the subset $Q \subset \mathcal{H}(\mathcal{X})$ as

$$P(f \in Q) = \int_Q d\nu(f)$$

where the right hand side is an integral through function space against some measure ν. Under certain assumptions [14, p. 9] $\mathcal{H}(\mathcal{X})$ turns out to be completely determined by its covariance operator — an infinite dimensional analogue of the covariance matrix satisfying $C_f : \mathcal{H}(\mathcal{X}) \ni g \mapsto C_f g = E[\langle f, g \rangle_{\mathcal{H}} f] \in \mathcal{H}(\mathcal{X})$ which in turn is completely specified once the second moment function $K(x_1, x_2) = E[f(x_1) f(x_2)]$ is known [3, p. 29].

The space $\mathcal{H}(\mathcal{X})$ can be shown to be the reproducing kernel Hilbert space $\mathcal{H}_K$ with reproducing kernel $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In this function space the norm $\|f\|_{\mathcal{H}_K}$ is inversely related to its probability of occurrence [14, p. 19]. This leads one to formulate the estimation problem globally for all $x \in \mathcal{X}$ simultaneously as

$$\sigma_f = \underset{f \in \mathcal{H}_K : Lf = \{y_k\}_{k=1}^n}{\operatorname{argmin}} \|f\|^2_{\mathcal{H}_K} \tag{17}$$

where $\sigma_f(\cdot)$ is then called an interpolating spline. The operator $L : \mathcal{H}_K \to \mathbb{R}^n$ goes by the name of measurement operator and relates the function $f(\cdot)$ to the observed values y_k, $k = 1, \ldots, n$. It is simple evaluation in this case, i. e. $(Lf)_j = y_j = L_j f$. Minimization of $\|f\|_{\mathcal{H}_K}$ is reasonable as the whole problem (17) then translates to finding that function f which is y_k at the positions x_k and is most likely as described by some Gaussian measure on the Hilbert space $\mathcal{H}_K$.

If we decide to drop the interpolating conditions and replace them with the constraint that Lf be "close" to the observations $\{y_k\}_{k=1}^n$ as measured for example in the ℓ^2-norm, the smoothing spline equation (18) ensues.

$$\sigma_f = \underset{f\in\mathcal{H}_K}{\operatorname{argmin}} \|Lf - \{y_k\}_{k=1}^n\|_{\ell^2}^2 + \|f\|_{\mathcal{H}_K}^2 \tag{18}$$

It balances fidelity to the data and likelihood of the chosen function. The explicit solution is given by [4, p. 161]

$$\sigma_f(\cdot) = \sum_{j=1}^n \lambda_j L_j K(\cdot,\cdot) \tag{19}$$

$$\lambda = (\Sigma + I)^{-1}\{y_k\}_{k=1}^n \tag{20}$$

where I is the $n \times n$ unit matrix and $(\Sigma)_{ij} = K(x_i, x_j)$. For a specific $\sigma_f(x)$ one gets

$$\sigma_f(x) = \sum_{j=1}^n \lambda_j K(x_j, x) = \Sigma_x^T \Sigma^{-1}\{y_k\}_{k=1}^n$$

under interpolating conditions — this is just the simple Kriging estimator if $K(x_1, x_2) = E[Y(x_1)Y(x_2)]$. Extensions to account for unknown means are standard and ultimately yield the same predictions $\hat{Y}_{\text{Spline}}(x) = \sigma_f(x)$ as universal Kriging [3, pp. 88–91].

Finally notice that at the $\{x_k\}_{k=1}^n$ for which observations are available

$$\hat{Y}_{\text{Spline}}(x_k) = \hat{Y}_{\text{UK}}(x_k) = \hat{Y}_{\text{Adjustment}}(x_k) + \hat{V}_k = y_k$$

where $\hat{V}$ is the residual $A\lambda^* - \{y_k\}_{k=1}^n$ from the adjustment procedure. This allows the conclusion that splines and Kriging as representers of machine learning approaches on the one hand and adjustment as a representer of classical geodetic techniques on the other hand are basically equivalent bar the philosophical difference of what is considered uninteresting noise to be discarded and what is not.

Another difference is that in the adjustment formulation, the decision variable is a parameter vector λ which determines a function f whereas in the machine learning formulation the function f is itself the decision variable to be determined via optimization.

2.4 Connection to other norm-based algorithms

Defining estimators as solutions to norm minimization problems is a common method of formalization in both geodesy and machine learning. In geodetic estimation tasks the quantity to be optimized is often the likelihood of residuals whose assumed Gaussian distribution yields the classical least squares formulations. Estimation tasks arising in machine learning seem to less often make a strict distinction between deterministic signal and random noise and at times avoid making use of distributional assumptions altogether. Instead, they communicate an estimator's desirability via an objective function that is not in all cases stochastically motivated. This leads to a wider variety of estimators whose properties are less well-known but interesting nonetheless. As shown in the previous equations (14) and (18), norm minimization tasks of the type

$$\sigma_f = \underset{f\in\mathcal{H}_K}{\operatorname{argmin}} \|Af - y\|_{\ell^2}^2 + \|f\|_{\mathcal{H}_K}^2 \tag{21}$$

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 correspondence extends uniquely to an adjustment problem $A\lambda - y = v$ with white noise v on the measurements and a prior that favors small lengths of the coefficient vector λ. 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 opens up interpretations of further machine learning methods that are similar in flavour to the abstract spline problem (21). Consider for example

$$\text{Ridge regression: } \sigma_f = \underset{f\in\mathbb{R}^m}{\operatorname{argmin}} \|Af - y\|_{\ell^2}^2 + \alpha\|Bf\|_{\ell^2}^2$$

$$\text{LASSO: } \sigma_f = \underset{f\in\mathbb{R}^m}{\operatorname{argmin}} \|Af - y\|_{\ell^2}^2 + \alpha\|f\|_{\ell^1}$$

$$\text{Elastic net: } \sigma_f = \underset{f\in\mathbb{R}^m}{\operatorname{argmin}} \|Af - y\|_{\ell^2}^2 + \alpha_1\|f\|_{\ell^2}^2 + \alpha_2\|f\|_{\ell^1}$$

[15, 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 the ℓ^p norms are nonnegative functions of f; consequently minimizing them is equivalent to maximizing a likelihood. This holds since for any 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. 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. 2 below for some sketches of the respective norms and densities in the instructive 1 dimensional case.

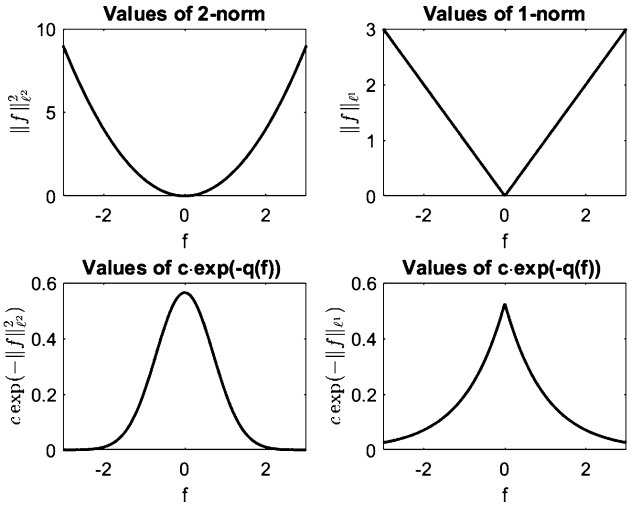

Figure 2: The $\|\cdot\|_{\ell^1}$ and $\|\cdot\|_{\ell^2}$ norms and their corresponding probability densities associated with the Gaussian and Laplacian distribution. Note the Laplacian's heavier tails.

The Gaussian pdfs derivative at its mean is zero; the pdfs 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. 2:

I When minimizing the ℓ^2-norm or equivalently maximizing the likelihood under a Gaussian pdf, small residuals are considered almost irrelevant since the gradient of $\|\cdot\|^2_{\ell^2}$ around 0 is zero. Large deviations are punished disproportionately strong: During minimization decreasing a big residual is considered more favourable 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.

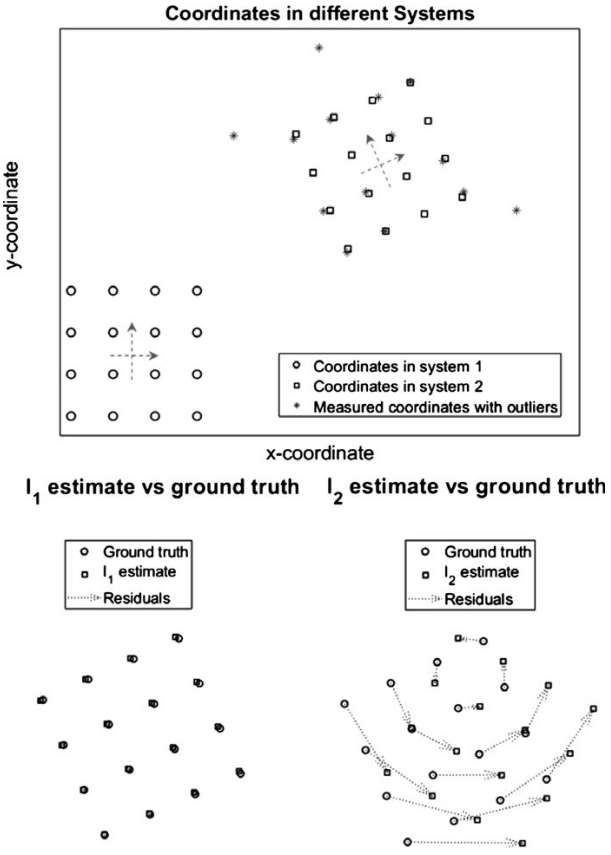

Figure 3: The ℓ^1-norm based estimation is much more robust than the ℓ^2-norm based estimation. The lower panels show the residuals between coordinates in system 2 and the coordinates transformed into system 2 from system 1 via transformations derived from Eq. (22). The scale in the lower panels is identical.

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.

We demonstrate performance difference of ℓ^2 and ℓ^1-norm based estimation in the presence of outliers for the typical geodetic task of inferring a Helmert transformation with fixed scale from coordinate measurements in Fig. 3. We briefly sketch the algorithm used to find the optimal transformation $A(\lambda^*)$ with

$$\lambda^* = \underset{\lambda=[x_A,y_A,\varphi_A]\in\mathbb{R}^3}{\operatorname{argmin}} \|A(\lambda)x - y\|_{\ell^p} \quad p = 1, 2 \tag{22}$$

that maps the coordinates x in system 1 onto the coordinates y in system 2:

Table 1: Correspondence of terminology in machine learning and adjustment.

Terminology or quantity	Role in ML	Role in adjustment
Target vector y in $\|Aq - y\| \to \min$.	y is a vector of observations y_k, $k = 1, \dots, n$.	y is a vector of observations y_k, $k = 1, \dots, n$.
Decision variable q in $\|Aq - y\| \to \min$.	$q = f$ is the vector of function values to be estimated; no parametric form is assumed.	$q = \lambda$ is a vector of parameters used to construct a function $f(\cdot) = \sum_{j=1}^{m} \lambda_j g_j(\cdot)$.
Operator A in $\|Aq - y\| \to \min$.	A is an operator that maps f onto measurements of f and emulates the way that observations y are generated. Af then typically is f evaluated at points x_k, $k = 1, \dots, n$. A is called the measurement operator.	A is a matrix whose rows are vectors of (nonlinear) transformations of the points x_k, $k = 1, \dots, n$. $A\lambda$ then typically is $f(\cdot) = \sum_{j=1}^{m} \lambda_j g_j(\cdot)$ evaluated at the points x_k. A is called the design matrix.
Term $\|Bq\|_r^r$ in $\|Aq - y\|_p^p + \|Bq\|_r^r \to \min$.	With $q = f$, $\|Bf\|_r^r$ is a regularization term that includes a prior on the function f into the estimation of f. The exact nature of the prior depends on norm r and energy operator B.	Since $q = \lambda$ is a vector of parameters, there seems hardly any justification for penalizing terms $\|B\lambda\|_r^r$. With $B = \alpha I$, $\alpha > 0$ and $r = 2$ they may be introduced for numerical reasons under the name of Tikhonov regularization.
Randomness and residuals	The quantity f to be estimated is assumed to come from a stochastic process. Unmodelled effects can be pushed onto f during estimation but f is very flexible. The residuals $v = Af - y$ are random too; f and v are distinguishable only via their correlation structure.	The quantities λ to be estimated are assumed to have fixed deterministic values. Randomness is a property of the residuals $v = A\lambda - y$ that act as a flexible catch all term subsuming all effects unaccounted for by the parametric model.
Features	A feature is a potentially infinite dimensional vector that contains (nonlinear) transformations of the input variable x, i. e. $g(x) = \{g_j(x)\}_{j=1}^{\infty}$ for some set of functions g_j.	During the construction of the design matrix A, the concept is used implicitly. Each of its rows can be interpreted as a feature in some input variable x.
Representations	A representation of a dataset $\{y_k\}_{k=1}^{n}$ is a choice of basis functions $\{g_j\}_{j=1}^{\infty}$ such that each y_k is representable as a combination of g_j's. A representation can be determined automatically by solving an optimization problem.	The choice of a good representation is left to the practitioner, whose responsibility it is to either determine a set of function $\{g_j\}_{j=1}^{m}$ such that $\sum_{j=1}^{m} \lambda_j g_j(x_k)$ approximates y_k or derive them from the geometrical of physical configuration of the task.

Note that many special procedures exist, which is why our explanations are geared to a proper description of only a simple subset of tasks that might be formulated as the minimization of discrepancy and irregularity measures. We consider these to be a good first order approximation to many commonly encountered problems in both fields.

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) [7, p. 294].

We close this section by stating in the following Table 1 an approximate correspondence between terminology and some quantities roles in machine learning and geodetic estimation.

3 Learning algorithms and toy applications

After having related adjustment theory to learning, we proceed to explain three algorithms which do not have an exact analogue within the bounds of the adjustment framework. Since therefore necessarily the arguments and calculations deviate from classical material, intuition is provided, as to why the methods work and how they are to be applied in practice. To underline the latter, brief and simple — but from a least squares perspective nontrivial — toy examples from geodesy are tackled in a fashion emphasizing approximate interrelations between concrete task and methodological approach rather than rigour.

The preceding section made use of a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to represent an estimator $\hat{f}$ for a function f in terms

of basis functions $K_x(\cdot) := K(x,\cdot)$ evaluated at the sample points $\{x_k\}_{k=1}^n$. The stochastically optimal choice for this so called kernel function turned out to be given by the covariance function $K(x_1,x_2) = E[F_{x_1}F_{x_2}]$ where $\forall c \in \mathcal{X}$, $F_x : \Omega \ni \omega \mapsto F_x^\omega \in \mathbb{R}$ was a square integrable random variable indexed by the space variable $x \in \mathcal{X}$.

This view immediately suggests to generalize the finite dimensional covariance matrix Σ_F of a random vector F taking values in $\mathbb{R}^n$, $n < \infty$ and satisfying

$$\begin{aligned}\langle \Sigma_F g, h\rangle_{\mathbb{R}^n} &= \langle E[F\otimes F^*]g,h\rangle_{\mathbb{R}^n}\\ &= E[\langle F,g\rangle_{\mathbb{R}^n}\langle F,h\rangle_{\mathbb{R}^n}]\ \ \forall g,h\in\mathbb{R}^n,\end{aligned} \tag{23}$$

towards the typically infinite dimensional covariance operator C_F exhibiting an exactly analogue relationship [3, p. 29]. This can be done by defining it as the selfadjoint positive definite kernel operator $C_F : \mathcal{H}_K \ni g \mapsto (C_F g)_{(\cdot)} := \int_{\mathcal{X}} K(x,\cdot)g(x)dx \in \mathcal{H}_K$.

As such $K(\cdot,\cdot)$ takes the role of a function determining the entries in an infinite dimensional covariance matrix that will intuitively be recognized by the practical geodesist as a natural extension of the already known frameworks to function space valued estimation problems. For the remainder of the paper we term this way of thinking about a kernel the covariance-interpretation. There is, however, a second radically different perspective onto kernels that is used concurrently in machine learning [32, p. 39] and emphasizes the meaning of $K(x,\cdot) = \phi_x(\cdot) \in \mathcal{H}_K$ as a Hilbert space valued nonlinear feature of $x \in \mathcal{X}$.

An instructive way to illustrate this consists in two separate steps that are roughly sketched for the special case of a Gaussian kernel $K(x_1,x_2) = \exp(-\|x_1 - x_2\|^2)$ for $x_1, x_2 \in \mathcal{X} \subset \mathbb{R}$.

1. Rewrite $-\|x_1 - x_2\|^2$ as $-\|x_1\|^2 + 2\langle x_1,x_2\rangle - \|x_2\|^2$ and substitute this term in the exponential expression for $K(x_1,x_2)$ to derive

$$\begin{aligned}K(x_1,x_2) &= c(x_2)e^{2x_1x_2-x_1^2}\\ &= c(x_2)\sum_{n=0}^{\infty}\frac{H_n(x_2)}{n!}x_1^n = \sum_{n=0}^{\infty}\alpha_n(x_2)x_1^n\end{aligned} \tag{24}$$

   where $H_n(\cdot)$ is the n-th Hermite polynomial [31, p. 456] and $\alpha_n(x_2) := c(x_2)H_n(x_2)/n!$ is a function solely depending on x_2.
2. Notice that $K(x_1,x_2)$ is effectively a linear superposition of monomials $x_1^n, n \in \mathbb{N}_0$ where the coefficient vector $\{\alpha_n\}_{n\in\mathbb{N}_0}$ depends on the exact value of x_2. As x_2 is varied to x_2' the coefficient vector changes as well resulting in $K(x_1,x_2')$ being a different linear combination of powers of x_1. When $x_2 \in \mathcal{X}$ is not fixed at all, then $K(x_1,\cdot) = \phi_{x_1}$ is a function from $\mathcal{X}$ to $\mathbb{R}$ and we have at the same time $\phi_{x_1}(\cdot)$ as an element of a (reproducing kernel) Hilbert space $\mathcal{H}_K$ [32, p. 39] and as an infinite set of $\mathcal{X}$-parametrized powers of x_1. As $\phi_{x_1}(\cdot)$ contains nonlinear information about x_1 it is called a (nonlinear) feature of x_1.

It should now be clear that $\phi_x(\cdot) \in \mathcal{H}_K$ is an infinite dimensional representation of $x \in \mathcal{X}$ that for specific choices of $K(\cdot,\cdot)$ can even encode all the information possibly to be known about $x \in \mathcal{X}$ [12]. We will call this the feature-interpretation of a kernel in what follows.

The reader is advised to not mix up both interpretations as the implied objects of investigation are different. The covariance-interpretation assumes the measured objects to be (nonlinear) functions $f : \mathcal{X} \to \mathbb{R}$ with $f \in \mathcal{H}_K$ subjectable to linear operations only. In contrast to this, the feature-interpretation assumes the measured objects to be $x \in \mathcal{X}$ and embeds them nonlinearly in $\mathcal{H}_K$ for some reproducing kernel Hilbert space $\mathcal{H}_K$. Both perspectives rely on the algebraic and geometric properties of the involved reproducing kernel Hilbert spaces (RKHS), whose internal structure as determined by the kernel $K(\cdot,\cdot)$ impacts the form of estimators and feature embeddings alike.

We proceed to apply both high dimensional embedding philosophies to supervised and unsupervised problems and start with the more familiar one of interpreting kernels as generators of covariance matrices.

3.1 Application 1: signal separation for total station data when the covariances are known

Suppose a total station was set up as depicted in Fig. 4 to monitor the movement of a prism. To not unnecessarily complicate this example, it will be assumed to hold that no tilt occurs and movement of the prism is constrained to purely lie in the x-direction implying a one dimensional formulation to be sufficient. The measurements are supplied in the form of a time series of x-coordinates.

Given: A sequence of times $t_j \in T$ and corresponding measurements $m_j \in \mathbb{R}$ of the x-coordinate in the format $\{(t_j,m_j)\}_{j=1}^n$ where n is the number of measurements.

Goal: Split the signal into separate parts that are in a stochastically reasonable way optimally identifiable with noise, atmospheric influences and true x-coordinate.

Assumption: The measurements m_j are realizations of square integrable random variables $M_{t_j} : \Omega \ni \omega \mapsto M_{t_j}^\omega \in \mathbb{R}$

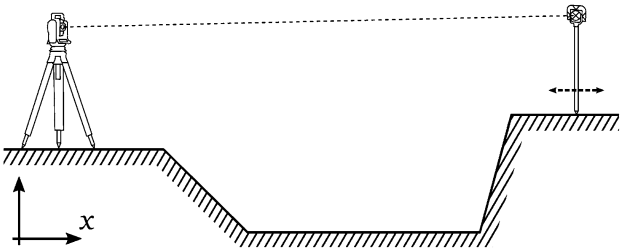

Figure 4: The setup for the signal separation problem in section 3.1. The measurements of the prisms changing x-coordinates contain atmospheric influences and noise.

for all $t_j \in T$; i. e. $\{M_t : t \in T\}$ is a stochastic process and separable in the following way:

$$M_t = N_t + A_t + X_t \tag{25}$$

where the stochastic processes N_t, A_t, X_t correspond to noise, atmosphere and x-coordinate respectively and are independent from each other. Their covariance functions (=kernels K_N, K_A, K_X) are assumed to be either known approximately or inferrable from the different time scales of N_t, A_t, X_t that find their expression in the decay characteristics of the kernels.

Main idea: The measurements $M_\cdot : \Omega \ni \omega \mapsto M_\cdot^\omega \in \mathbb{R}^T$ are assumed to lie in some infinite dimensional Hilbert space $\mathcal{H}_M$ with kernel $K_M = K_N + K_A + K_X$ which implies that $\mathcal{H}_M$ is the direct sum of the Hilbert spaces containing pure noise, atmospheric influences and x-coordinates in the sense that $\mathcal{H}_M = \mathcal{H}_N \oplus \mathcal{H}_A \oplus \mathcal{H}_X$. For a more rigorous account, see [21]. An optimal interpolating spline σ_m is found such that $\sigma_m(t_j)$ perfectly coincides with the measurements at times t_j.

$$\sigma_m(\cdot) = \underset{m \in \mathcal{H}_M : m(t_j) = m_j}{\operatorname{argmin}} \|m\|^2_{\mathcal{H}_M} \tag{26}$$

$$\sigma_m(\cdot) = \sum_{j=1}^{n} \lambda_j K_M(t_j, \cdot) \text{ with } \lambda = (K_M^{ij})^{-1} m \tag{27}$$

whereby K_M^{ij} is the matrix with entries $K_M(t_i, t_j) \in \mathbb{R}^n \otimes \mathbb{R}^n$ and $m \in \mathbb{R}^n$ is the n-dimensional vector containing the measurements. Notice that $\sigma_m(\cdot)$ is not a number but a function of $t \in T$. Subsequently $\sigma_m(\cdot) \in \mathcal{H}_M$ will be orthogonally projected onto the subspaces $\mathcal{H}_N, \mathcal{H}_A$ and $\mathcal{H}_X$ to yield the optimal estimators $\sigma_n = \Pi_N \sigma_m$, $\sigma_A = \Pi_A \sigma_m$, $\sigma_x = \Pi_X \sigma_m$.

Results: When reliable covariance information is available, the results are stochastically optimal under a Gaussian process assumption [28, p. 27]. Also on a purely visual level the outcome of applying the estimation procedure above to an exemplary dataset seems reasonable — for an example see Fig. 5. The algorithm is quite robust to misspecification of the kernels as long as the correlations structures of the individual mixture components to be separated are appropriately encoded in the kernels.

The extension to vector valued and tensor valued splines is straightforward and much effort has been put forward to guarantee practical computability and stability of the numerical schemes that nowadays incorporate many ideas from finite element analysis and spectral theory, see [4].

3.2 Application 2: signal separation for total station data when the covariances are unknown

In what follows, the requirements on prior knowledge are relaxed and the covariance structure is no longer assumed to be known. In only presupposing the measurements as being made up of statistically independent parts, we pass from a supervised to an unsupervised learning problem that has no analytical solution anymore. Before describing and applying K-ICA in this setting, it is instructive to explain the most common measures for characterizing independence of random variables.

Two random variables $X, Y : \Omega \to \mathbb{R}$ are called independent if — bar some technicalities concerning measurability and continuity — their joint probability density function $f_{XY}(x, y)$ factors into the product of its marginals; i. e. $f_{XY}(x, y) = f_X(x) f_Y(y)$ [1, p. 91]. One writes $X \perp\!\!\!\perp Y$ in this case. It is well known that two jointly multivariate Gaussian distributed random variables X and Y are uncorrelated if and only if they are independent [27, p. 71]. Generally, however, two random variables X, Y may have zero correlation without necessarily being independent. In the non-Gaussian case the covariance function

$$\operatorname{cov}(X, Y) = \int_{\mathbb{R}^2} (x - E[X])(y - E[Y]) f_{XY}(x, y) dx dy$$

is therefore a necessary but insufficient indicator of independence and needs to be replaced by the entropy-based mutual information $I(X, Y)$.

The mutual information between two random variables is defined as [19]

$$I(X, Y) = \int_{\mathbb{R}^2} f_{XY}(x, y) \log \left[\frac{f_{XY}(x, y)}{f_X(x) f_Y(y)} \right] dx dy \tag{28}$$

where the expression is evaluated as an appropriate limit in any pathological cases. It can be shown that $I(X, Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$ and otherwise $I(X, Y) > 0$.

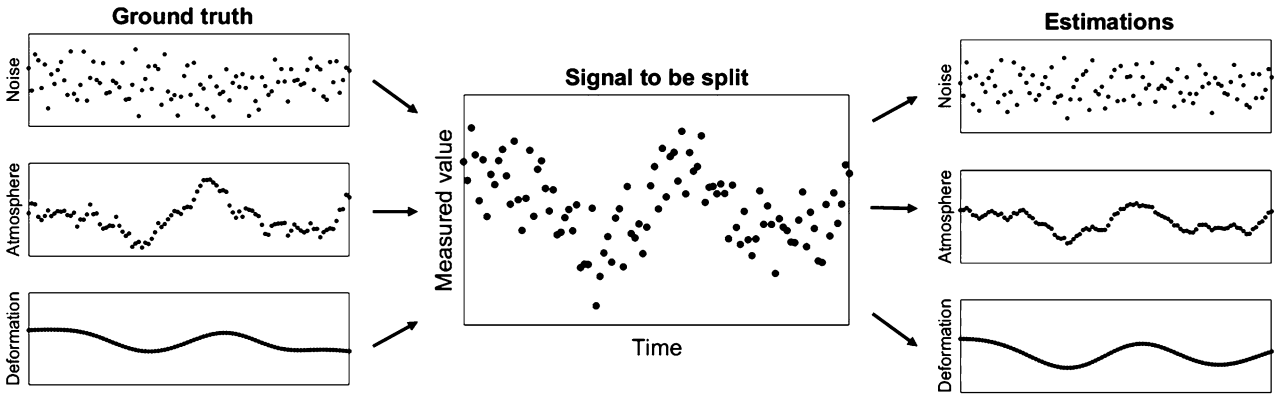

Figure 5: An example of signal separation. The left panels show the true underlying ground truth (synthetic data), that is superimposed to generate the signal plotted in the center. This time series is the input for the RKHS based estimation framework outlined in section 3.1 whose output are the estimations visible on the right side. The scale is the same for the six outer plots.

Therefore from a theoretical perspective if one wanted to split a timeseries $\{m_j\}_{j=1}^n$ linearly into independent components $\{a_j\}_{j=1}^n$ and $\{x_j\}_{j=1}^n$ one could try to minimize the mutual information between the $\{a_j\}_{j=1}^n$ and the $\{x_j\}_{j=1}^n$ which are assumed to be realizations of random variables A and X with significantly different probability distributions. Typically estimating the mutual information empirically is hard, however, and it is more common to instead maximize a contrast function $\rho(A,X)$ that convincingly measures how different the distribution of A is from that of X. Such contrast functions ρ are typically derived from a Taylor expansion of $-I(\cdot,\cdot)$ in terms of features of probability distributions (e. g. third moment or higher order cumulants) that partially emulate the property of $-I(A,X)$ being biggest for f_A being very different from f_X [2].

It is entirely possible to apply the aforementioned infinite dimensional embedding of probability distributions into an RKHS $\mathcal{H}_K$ in this setting. An efficiently computable measure of dependence to be minimized is then given by the $\mathcal{H}_K$-correlation

$$\rho_{\mathcal{H}_K}(A,X) = \sup_{\|g_j\|_{\mathcal{H}_K}=1, j=1,2} \frac{E[g_1(A)g_2(X)]}{\sqrt{E[g_1(A)^2]E[g_2(X)^2]}} \tag{29}$$

for functions g_1, g_2 that are already centered in feature space $\mathcal{H}_K$ [2]. $\rho_{\mathcal{H}_K}$ is to be interpreted as the maximally achievable correlation between nonlinear transformations of the random variables A and X where optimization is carried out over the class of nonlinear transformations. Making use of the reproducing property

$$\langle g, \phi_P \rangle_{\mathcal{H}_K} = E_P[g(X)]$$

for $\phi_P = E_P[K(X,\cdot)]$ expectation w.r.t to the probability measure P the kernel trick allows computationally efficient finite dimensional implementation of this infinite dimensional problem.

Suppose, a total station S was set up as depicted in Fig. 6 to monitor the movement of two prisms P_1, P_2 mounted on a planar structure subject to a translational rigid, but time dependent change of coordinates. To keep the example simple, only the x-coordinates will be investigated to arrive again at a one dimensional formulation that parallels the one presented in section 3.1 but with increased difficulty due to the absence of any knowledge of the correlation structure of the signals to be separated.

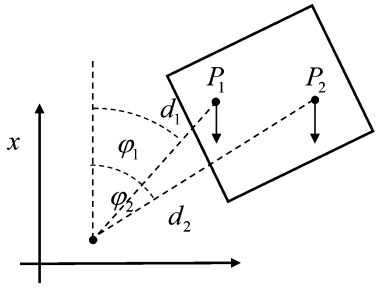

Figure 6: The x coordinate of the prisms P_1, P_2 are measured employing a total station positioned at S. The distances d_1, d_2 and angles φ_1, φ_2 are not assumed to be known.

Given: A sequence of times $t_j \in T$ and corresponding sequences of measurements $m_j^k \in \mathbb{R}$ of the x-coordinates of the prism P_k, $k = 1, 2$. The totality of measurements is summarized in the sequence $\{(t_j, m_j^1, m_j^2)\}_{j=1}^n$ where n is the number of measurements.

Goal: Split the signal $\{(m_j^1, m_j^2)\}_{j=1}^n$ into two separate parts that are in a stochastically reasonable way identifiable

with atmospheric influences and the true x-coordinates of P_1, P_2.

Assumption: The measurements m_j^k are realizations of square integrable random variables $M_{t_j}^k$ for all $t_j \in T$, i. e. $\{M_t^k : t \in T\}$ are two stochastic processes which we assume to be linear mixtures of deformations and atmospheric influences:

$$M_t^1 = q_{11}X_t + q_{12}A_t$$
$$M_t^2 = q_{21}X_t + q_{22}A_t$$

where the stochastic processes X_t and A_t correspond to x-coordinates and atmospheric influences respectively. In short vector notation and with obvious identifications, one may write

$$M_t = QY_t \tag{30}$$

instead. For this model to be reasonable, it is necessary that the whole planar structures motion is sufficiently well described by a translation to guarantee that the behaviour of the two prisms x-coordinates is identical. Furthermore the atmospheric conditions need to be constant over the whole spatial domain to ensure that their influence on the measurement series $\{m_j^k\}_{j=1}^n$ is representable as terms $q_{12}A_t$ and $q_{22}A_t$ linearly related to some underlying scalar A_t.

Further explanation: The atmospheric conditions are allowed to vary in time. We may calculate the entries of Q based on knowledge of the geometrical configurations and usual formulas for distance reduction of electrooptic measurements by noting that the atmospheric correction Δx_k satisfies

$$\begin{aligned}\Delta x_k &= \Delta d_k \cos\varphi_k \\ &= \underbrace{\alpha(\text{Temperature},\text{Pressure})}_{\hat{A}_t}\underbrace{d_k \cos\varphi_2}_{\hat{q}_{k2}}\end{aligned}$$

where α is a meteorological correction factor, see for example [35, p. 310]. This would allow us to solve the problem immediately by inverting Q and applying $Q^{-1} =: W$ to the sequences of measurements — however we do not want to do this but demand that the algorithm finds the most probable decomposition based not on physically or geometrically motivated knowledge but solely on the probabilistic assumption that the x-coordinates and atmosphere are stochastically independent of each other. Neither X_t nor A_t are allowed to be Gaussian since approximate stochastical independence will be achieved by maximizing some measure of non-Gaussianity [8].

Main idea: Since the measurements $M_t^k : \Omega \ni \omega \mapsto M_t^k(\omega) \in \mathbb{R}$ are supposedly both linear mixtures of X_t and A_t with $X_t \amalg A_t$, a 2×2 matrix W with $WM_t = \hat{Y}_t$ maximally independent in the sense of mutual information would solve the problem apart from the usual ambiguities encountered during ICA [17].

To approximately achieve this, minimize the $\mathcal{H}_k$-correlation

$$\rho_{\mathcal{H}_k}(\hat{X}_t, \hat{A}_t) = \sup_{f,g\in\mathcal{H}_k} \frac{E[f(\hat{X}_t)g(\hat{A}_t)]}{\sqrt{E[f^2(\hat{X}_t)]E[g^2(\hat{A}_t)]}}$$

for some kernel $k(s,t)$ determining the flexibility of permissible transformations $f, g \in \mathcal{H}_k$. As explained in Eq. (29), this is a measure of mutual dependency. To minimize it, employ the code provided by [2] consisting of iteratively executing the following three steps starting from an initial guess of W:

I Whiten the data and construct the kernel matrices K_l, $l = 1, 2$ with elements $(K_1)_{ij} = k(\hat{X}_{t_i}, \hat{X}_{t_j})$ and $(K_2)_{ij} = k(\hat{A}_{t_i}, \hat{A}_{t_j})$ for $i, j = 1, \ldots, n$ where $\hat{X}_t$ and $\hat{A}_t$ are derived from $\hat{Y}_t = WM_t$. Then center the kernel matrices.

II Solve the regularized kernel canonical correlation generalized eigenvalue problem

$$\begin{bmatrix} 0 & K_1K_2 \\ K_2K_1 & 0 \end{bmatrix}\begin{bmatrix}\lambda \\ \mu\end{bmatrix} = \rho^W \begin{bmatrix} (K_1+\alpha I)^2 & 0 \\ 0 & (K_2+\alpha I)^2 \end{bmatrix}\begin{bmatrix}\lambda \\ \mu\end{bmatrix}$$

for ρ^W to determine the $\mathcal{H}_k$ correlation dependent on the matrix W.

III Minimize $-1/2 \log \lambda^W$ where λ^W is the smallest of the generalized eigenvalues ρ^W by gradient descent on the manifold of orthogonal matrices.

Results: When the two underlying stochastic processes X_t and A_t generate non-Gaussian data and are governed by probability distributions which are reasonably well distinguishable via linear combinations of higher order statistical moments, the splitting achieved via kernel ICA is convincing. For an exemplary application to a simulated dataset, see Fig. 7.

The situation exhibited in Fig. 6 is not entirely realistic and would need to be modified for any actual application in practice — however the idea of splitting several sequences of measurements into maximally independent components is a promising one.

The framework is applicable whenever measurements generate for each point in time a whole vector of values and there are reasons to suspect that each entry in that vector is a linear mixture of quantities of actual interest. This

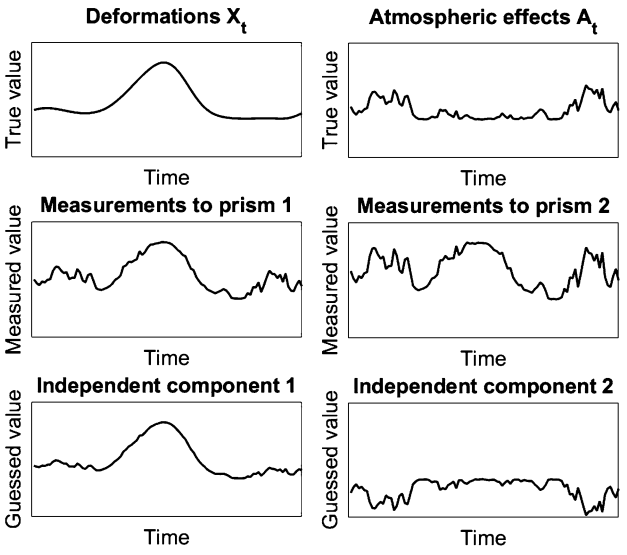

Figure 7: Simulated atmospheric effects and deformations in forms of time series (row 1) are mixed together with the matrix [1 0.5; 1 0.8] (row 2). The unimixing was done with the K-ICA procedure outlined in this section (row 3). After normalization, the estimated mixing matrix is [1 − 0.4; 1 − 0.7]. The scale is arbitrary but identical for all subplots.

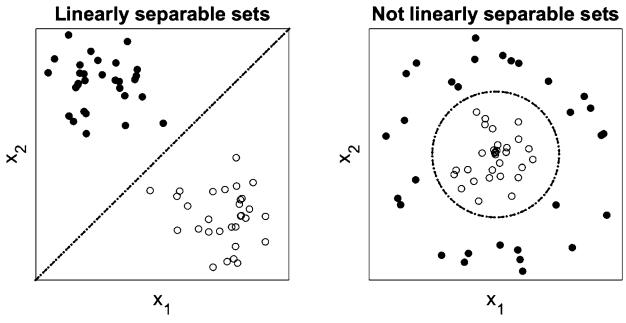

Figure 8: The left panel shows two sets χ_+ and χ_- which are linearly separable in $\mathbb{R}^2$ whereas the same cannot be said for the situation on the right side. Elements of χ_+ are marked as disks, elements of χ_- as circles.

may open up not only new purely statistical signal separation procedures but also suggest different mensuration strategies that are explicitly meant to measure only indirectly the quantities of interest in the form of easily accessible linear mixtures and infer them later on via optimization in a separate post-processing phase. This promotes a rather opportunistic viewpoint similar to the one taken in compressive sensing that contrasts starkly with the classical geodetic perspective, in which the quantities of interest are supposed to be the direct outcomes of measurements.

3.3 Application 3: classification of total station data

Suppose now that for a time series of noisily gathered coordinate measurements a decision is sought as to judge if it is indicative of harmful deformations or not. Assume further that only exemplary time series are provided to which a label is assigned that classifies them as belonging to a harmless (–1) or dangerous (+1) situation. One commonly used way to solve supervised classification tasks like this is to use support vector machines, whose basic principle we briefly outline in what follows.

Suppose the set of pairs $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^m$, $y_i \in \{-1, +1\}$ are the training examples and let $\chi_+ := \{x_i \in \chi : y_i = +1\}$, $\chi_- := \{x_i \in \chi : y_i = -1\}$ where χ is the set of all x-values in the training set. If it is possible to find a hyperplane $P(\mathbb{R}^m)$ in $\mathbb{R}^m$ which separates χ_+ from χ_- then $P(\mathbb{R}^m)$ is called a separating hyperplane and χ_+ and χ_- are termed linearly separable in $\mathbb{R}^m$. An illustration for the case $m = 2$ is found in Fig. 8.

The figure also exhibits an example, in which χ_+ and χ_- are not linearly separable in $\mathbb{R}^2$. However, after centering and a nonlinear transformation of type $\phi : (x_1, x_2) \mapsto x_1^2 + x_2^2$ associating to each x its distance to the origin, χ_+ and χ_- are linearly separable in the feature space $\mathbb{R}^1$ via a separating hyperplane — simply thresholding in this low dimensional case.

This suggests again a kernelization approach: Instead of trying to find a separating hyperplane $P(\mathbb{R}^m)$ in the input space $\mathbb{R}^m$ containing the x_i, map x_i into some infinite dimensional Hilbert space $\mathcal{H}_k$ of features by setting $\phi : \mathbb{R}^m \ni x \mapsto \phi(x) = k(x, \cdot) \in \mathcal{H}_k$ for some reproducing kernel $k(\cdot, \cdot)$ and search for a separating hyperplane in the RKHS $\mathcal{H}_k$ instead.

A hyperplane $P(\mathcal{H}_k)$ in $\mathcal{H}_k$ is completely specified by a normal vector $f \in \mathcal{H}_k$ and a (positive or negatively weighted) distance d to the origin.

$$P(\mathcal{H}_k) := \{g \in \mathcal{H}_k : \langle f, g \rangle_{\mathcal{H}_k} + d = 0\}$$

Depending on what side of $P(\mathcal{H}_k)$ a point $h = \phi(x) \in \mathcal{H}_k$ comes to lie, it is predicted to either belong to class χ_+ or χ_- via

$$\hat{y} = \operatorname{sign}\left(\langle f, h \rangle_{\mathcal{H}_k} + d\right) = \operatorname{sign}\left(\langle f, k(x, \cdot) \rangle_{\mathcal{H}_k} + d\right)$$

which also directly gives the decision rule for classifying previously unencountered inputs $x \in \mathbb{R}^m$ [32, p. 190]. Finding an approximately separating hyperplane $P(\mathcal{H}_k)$ in the

RKHS $\mathcal{H}_k$ that balances the number of misclassifications and the regularity of $P(\mathcal{H}_k)$'s backprojection into $\mathbb{R}^m$ is approximable by the optimization problem [15, p. 428]

$$(\hat{f}, \hat{d}) = \underset{f \in \mathcal{H}_k, d \in \mathbb{R}}{\operatorname{argmin}} \sum_{j=1}^{n} \left[1 - y_j \hat{y}(x_j)\right]_+ + \alpha \|f\|^2_{\mathcal{H}_k} \tag{31}$$

where $[\,\cdot\,]_+$ denotes the positive part and $\hat{y}(x) = \operatorname{sign}(f(x) + d)$. Consequently $L_j(f) = \left[1 - y_j \hat{y}_j(x_j)\right]_+$ is a positive functional of f for each j quantifying the classification error. The norm $\|f\|_{\mathcal{H}_k}$ has the same interpretation as in section 2.4. The positive parameter α balances fidelity to the data and regularity [15, p. 424]. Notice that it is entirely possible to swap $L_j(f)$ in equation (31) for a quadratic error term and recover the smoothing spline equation (18).

Equivalently one may solve the quadratic program

$$\begin{aligned} &\underset{\lambda \in \mathbb{R}^n}{\text{maximize}} \quad \sum_{j=1}^{n} \lambda_j - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j K(x_i, x_j) \\ &\text{subject to} \quad 0 \le \lambda_j \le C \ , \ \sum_{j=1}^{n} \lambda_j y_j = 0 \end{aligned} \tag{32}$$

for some positive C dependent on the parameter α from equation (31) [32, p. 205] [15, p. 420]. Efficient algorithms are available to solve this problem for the parameters $\{\lambda_j\}_{j=1}^n$ which then are used to assemble the class estimator

$$\hat{y}(x) = \operatorname{sign}\left(\sum_{j=1}^{n} \lambda_j y_j K(x, x_j) + d \right) \tag{33}$$

with $d = 1/y_i - f(x_i)$ for any $i = 1, \ldots, n$. This classifier is termed support vector machine and we will apply it immediately to the problem outlined before.

Let the time series in Fig. 9 be the input x for our classification problem; the sets χ_+ and χ_- providing exemplary time series associated to harmful and harmless situations are sampled there as well by listing some representatives.

Given: A sequence $\{x_j\}_{j=1}^n$ of deformation measurements $x_j = \{x_j^i\}_{i=1}^m \in \mathbb{R}^m$ at times $t_i \in T$ in the format $\{(t_i, x_j^i)\}_{i=1}^n$ where the sequence $x_j \in \mathbb{R}^m$ is the interesting part and the time information will regularly be discarded. There is furthermore a training set of examples $\{(x_j, y_j)\}_{j=1}^n$ where again each x_j is a time series and y_j is the corresponding label.

Goal: Emulate the input-output behaviour mapping time series onto danger assessments via the class prediction function $\hat{y}(x)$ defined in equation (33). It makes use of the RKHS $\mathcal{H}_k$ of functions on $\mathbb{R}^m$ with reproducing kernel $k(\cdot, \cdot)$ that maps pairs of time series onto a real number quantifying their similarity.

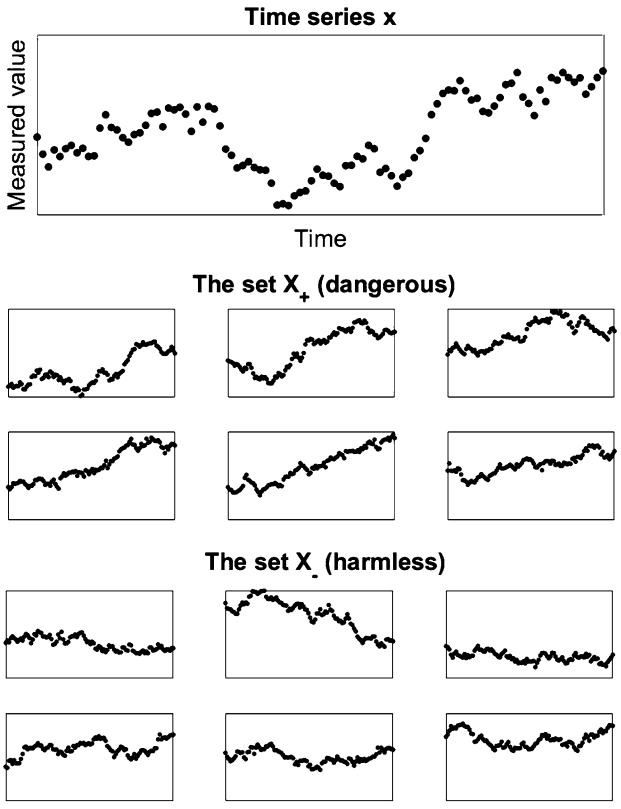

Figure 9: The uppermost panel exhibits the sequence of deformation measurements which are to be evaluated as either dangerous or not. Some of the training data is presented in the lower two panels. The scale is the same for all plots of training data.

Assumption: The set of time series χ_+ associated to dangerous behaviour is approximately linearly separable from the set χ_- after embedding it into the infinite dimensional Hilbert space $\mathcal{H}_k$ of features via the map

$$\phi : \mathbb{R}^m \ni c \mapsto \phi(x) = k(x, \cdot) \in \mathcal{H}_k.$$

Furthermore assume that the euclidean distance is a meaningful measure of closeness between time series. Usage of the linear kernel $k(x_1, x_2) = \langle x_1, x_2 \rangle_{\mathbb{R}^m}$ derived from the inner product in $\mathbb{R}^n$ is then justified.

Main idea: Solve the optimization problems specified in equations (31) or (32) to find a parameter vector λ and a constant d such that the classifier $\hat{y}$ assembled from λ and d according to equation (33) has both acceptable regularity and misclassification rate on the training set. Afterwards, apply $\hat{y} : \mathbb{R}^m \to \{-1, +1\}$ to unseen time series to classify them.

Results: Support vector machines usually perform reasonably well although more sophisticated methods exist for function approximation problems [9]. Table 2 summarizes

the SVM's behaviour in terms of errors of the first and second kind. For the estimation of empirical error probabilities the cycle of simulating ground truth, fitting an svm and classifying 100 randomly chosen time series was rerun 100 times while the amount of training examples was subjected to systematic change. Classification was done using the Matlab built-in "fitclinear".

Table 2: Performance of SVM's for the specific task outlined above.

error	samples			
	10	10^2	10^3	10^4
type I in %	6.6	1.8	0.6	0.2
type II in %	6.7	2.0	0.6	0.2

Empirically estimated probabilities of type I error (incorrect rejection of null hypothesis H_0) and type II error (failure to reject incorrect null hypothesis H_0). H_0 is the hypothesis that $y(x) = -1$.

We want to close this section with a few clarifying remarks regarding simulation methodology and a link to classical hypothesis testing.

This example is again purely synthetic. We randomly sampled from a stochastic process that corresponds to Brownian motion, each realization was considered to be a time series $x_j \in \mathbb{R}^m$ of deformation measurements. If the best fitting line through $\{(t_i, x_j^i)\}_{i=1}^m$ had positive slope, the situation was classified as dangerous and harmless otherwise. This generation rule for our synthetic ground truth was not communicated to the SVM however, which only received the labeled training examples and had to infer the rule by itself. Notice that even for the trivial finite dimensional kernel $k(x_i, x_j) = \langle x_i, x_j \rangle_{\mathbb{R}^m}$ the limit performance should be almost perfect separation since the underlying true classification rule is

$$Ax \geq 0 \Rightarrow y(x) = +1$$
$$Ax < 0 \Rightarrow y(x) = -1$$

where $A : \mathbb{R}^m \to \mathbb{R}$ is a linear operator consisting of a concatenation of line fitting and calculation of the derivative of that line — both operations being linear in the data. Therefore $A\chi_+$ is linearly separable from $A\chi_-$ in $\mathbb{R}^1$, and the underlying decision rule can be written as

$$\langle \tilde{f}, Ax_+ \rangle_{\mathbb{R}} \geq 0$$
$$\langle \tilde{f}, Ax_- \rangle_{\mathbb{R}} < 0$$

$\forall x_+ \in \chi_+$ and $x_- \in \chi_-$ where $\tilde{f}$ is any nonzero number. This implies for $f = A^T\tilde{f} \in \mathbb{R}^m$ the equivalent decision rule

$$\langle f, x_+ \rangle_{\mathbb{R}^m} \geq 0$$
$$\langle f, x_- \rangle_{\mathbb{R}^m} < 0$$

$\forall x_+ \in \chi_+$ and $x_- \in \chi_-$ because $\langle \tilde{f}, Ax \rangle_{\mathbb{R}} = \langle A^T\tilde{f}, x \rangle_{\mathbb{R}^m}$ for any $A : \mathbb{R}^m \to \mathbb{R}$. For a simple example like this, embeddings into infinite dimensional $\mathcal{H}_k$ are unnecessary. When the underlying classification rule (=failure mechanism in our example) is complicated or unknown and danger assessment is demanded based only on a sequence of measurements somewhat correlated with the reasons for critical behaviour, they may however prove helpful. [32] provide some examplary applications that go into this direction and demonstrate the usefulness of including kernel-based nonlinearities into estimation.

It is possible to establish that the inner-product-based decision rule for linear SVM's is the same as the Bayes rule

$$\log(f_{Y|X}(y = +1|x) f_{Y|X}^{-1}(y = -1|x)) \gtrless 0 \Rightarrow \hat{y}(x) = \pm 1$$

for some semiparametric probability density function f_Y whose parameters haven been inferred via Maximum Likelihood estimation [11]. This is obviously a form of likelihood ratio test as employed for comparing two statistical models in classical hypothesis testing.

4 Conclusion and outlook

In this paper, we investigated the interface between geodetic data analysis and machine learning algorithms. It turned out that adjustment as used in the geodetic community can be interpreted as a learning algorithm via proper relabeling of the terms occuring in the optimization task arising during maximum likelihood estimation under assumption of Gaussianity. This was exemplified in a simple application, in which adjustment, geostatistics and splines were employed for regression and interpolation purposes. They were shown to essentially agree when applicable. A table was provided that served as a guideline to translate between adjustment theoretic and machine learning motivated treatments of estimation problems.

Apart from the different role of stochasticity in both fields, one of the main differences is the focus on high dimensional embeddings of data. It was outlined, how infinite dimensional problems can be efficiently solved using kernels and some intuition was gathered by tackling a sequence of instructive albeit simple geodetic toy problems — not all of which were known to be easily solvable. The algorithms are shown to be demonstrably easy to implement with further examples freely available on GitHub.[1]

1 https://github.com/jemil-butt/ML_tutorials_geodesy

We speculate that an influx of ideas and procedures developed in the machine learning community into the set of methods finding widespread usage in geodesy is bound to be beneficial particularly in the following subfields:

Mensuration design: The existence of numerical algorithms for approximate optimization in connection to difficult nonlinear tasks with many decision variables implies the possibility to adapt measurement strategies dynamically as data comes in, as for example might be the case in monitoring scenarios. Furthermore solutions to previously untackled problems in estimation and inference might relax constraints usually imposed on instrument and campaign setups.

Data analysis: Regression and classification, supervised, unsupervised and reinforcement learning are tools of which only the first one is commonly exploited in the geodetic community. Whereas the impact of better classification methods as generalizations of rigorous hypothesis testing is to a certain degree predictable, unsupervised and reinforcement learning as frameworks for optimal decision making and pattern recognition in situations involving uncertainty provide exciting opportunities to solve new and seemingly ill posed estimation problems.

However, the at times less rigorously stochastic approach of machine learning algorithms implies weaknesses in diagnosing distributional characteristics and derived error bounds for the outputs. This is a limitation in need of rectification before widespread use becomes feasible in a discipline as dependent on reliability as geodesy. To a lesser degree, it is also expected that increasingly instruments may arise whose measurements only yield the target quantities after a costly optimization — a trade off between post processing and instrument complexity. As the performance of estimation and inference grows, physically motivated forward models for instrument errors or the behaviour of observed objects in general might to a certain degree gradually be replaced by data driven stochastic approximations. Visualization may be aided by classification and clustering algorithms which also regularly prove useful for data exploration and knowledge discovery. We see less potential in the less processing dominated domains of geodesy — those dealing particularly with the development of theoretical models or infrastructural and legislative aspects.

Acknowledgment: The authors acknowledge the work of the two anonymous reviewers who contributed to this paper by providing suggestions on content and formatting of the paper thereby improving its readability and correctness.

References

[1] R. B. Ash, *Basic Probability Theory*, Courier Corporation, New York, 2008.

[2] F. R. Bach and M. I. Jordan, *Kernel independent component analysis*, J. Mach. Learn. Res., 3 (2002), pp. 1–48.

[3] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Springer, Berlin Heidelberg, 2011.

[4] A. Y. Bezhaev and V. A. Vasilenko, *Variational Theory of Splines*, 1st ed., Springer, Berlin Heidelberg, 2013, softcover reprint of hardcover.

[5] K. Borre, ed., *The Adjustment Procedure in Tensor Form*, Springer, Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 17–21.

[6] K. Borre, ed., *Some Remarks About Collocation*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 261–272.

[7] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[8] J. Cardoso, *Blind signal separation: statistical principles*, Proceedings of the IEEE, 86 (1998), pp. 2009–2025.

[9] R. Caruana and A. Niculescu-Mizil, *An empirical comparison of supervised learning algorithms*, in Proc. 23 rd Intl. Conf. Machine learning (ICML'06, 2006, pp. 161–168.

[10] J.-P. Chiles and P. Delfiner, *Geostatistics – Modeling Spatial Uncertainty*, John Wiley & Sons, New York, 2012.

[11] V. Franc, A. Zien and B. Schölkopf, *Support vector machines as probabilistic models*, in Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, Omnipress, USA, 2011, pp. 665–672.

[12] K. Fukumizu, A. Gretton, B. Schölkopf and B. K. Sriperumbudur, *Characteristic kernels on groups and semigroups*, in in Advances in Neural Information Processing Systems 21, D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds., Curran Associates, Inc., 2009, pp. 473–480.

[13] E. W. Grafarend and B. Schaffrin, *Ausgleichungsrechnung in linearen Modellen*, BI-Wissenschaftsverlag, Mannheim, Wien, Zürich, 1993.

[14] M. Hairer, An Introduction to Stochastic PDEs, ArXiv e-prints, 2009.

[15] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*, Springer, Berlin Heidelberg, 2013.

[16] U. C. Herzfeld, *Least-squares collocation, geophysical inverse theory and geostatistics: a bird's eye view*, Geophysical Journal International, 111 (1992), pp. 237–249.

[17] A. Hyvärinen and E. Oja, *Independent component analysis: algorithms and applications*, Neural Networks, 13 (2000), pp. 411–430.

[18] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning – with Applications in R*, Springer, Berlin Heidelberg, 2013.

[19] A. Kraskov, H. Stögbauer and P. Grassberger, *Estimating mutual information*, Phys. Rev. E, 69 (2004) 066138, arXiv:cond-mat/0305641.

[20] J. Kusche and R. Klees, *Regularization of gravity field estimation from satellite gravity gradients*, Journal of Geodesy, 76 (2002), pp. 359–368.

[21] F. Larkin, *Gaussian measure in Hilbert space and applications in numerical analysis*, Rocky Mountain J. Math., 2 (1972), pp. 379–422.
[22] T. M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.
[23] H. Moritz, *Advanced least-squares methods*, Reports of the Department of Geodetic Science, 175, (1972).
[24] H. Neuner, Model selection for system identification by means of artificial neural networks, Journal of Applied Geodesy, 6 (2012), pp. 117–124.
[25] J. Neveu, *Processus aléatoires gaussiens*, 1968.
[26] W. Niemeier, *Ausgleichungsrechnung – Statistische Auswertemethoden*, Walter de Gruyter, Berlin, 2008.
[27] S. J. Press, *Applied Multivariate Analysis – Using Bayesian and Frequentist Methods of Inference*, *2nd Edition*, Courier Corporation, New York, 2012.
[28] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, 2006.
[29] A. Reiterer, U. Egly, T. Vicovac, E. Mai, S. Moafipoor, D. Grejner-Brzezinska and C. Toth, *Application of artificial intelligence in geodesy – a review of theoretical foundations and practical examples*, Journal of Applied Geodesy, 4 (2010), pp. 201–217.
[30] B. Riedel and M. Heinert, *An adapted support vector machine for velocity field interpolation at the Baota landslide*, in Proc. Application of Artificial Intelligence in Engineering Geodesy – 1st International Workshop, Vienna, 2008, pp. 101–116.
[31] S. Roman, *Advanced Linear Algebra*, Springer, Berlin Heidelberg, 2007.
[32] B. Schoelkopf and A. J. Smola, *Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, 2002.
[33] G. Strang, *Introduction to Linear Algebra*, Wellesley-Cambridge Press, Wellesley, 2016.
[34] M. Wiering and M. v. Otterlo, *Reinforcement Learning – State-of-the-Art*, Springer, Berlin Heidelberg, 2012.
[35] B. Witte and P. Sparla, *Vermessungskunde und Grundlagen der Statistik für das Bauwesen*, Vde Verlag GmbH, Berlin, Offenbach, 2015.