Diss. ETH No. 25780

MULTIVARIATE METHODS FOR HETEROGENEOUS HIGH-DIMENSIONAL DATA IN GENOME BIOLOGY

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
BRITTA VELTEN
M.Sc., Heidelberg University
born on 16.06.1992
citizen of Germany

accepted on the recommendation of
Prof. Dr. Peter Bühlmann, examiner
Dr. Wolfgang Huber, co-examiner
Dr. Oliver Stegle, co-examiner

2019

Britta Velten: *Multivariate methods for heterogeneous high-dimensional data in genome biology,* 2019

DOI: 10.3929/ethz-b-000333437

To my parents

Abstract

Technological advances have transformed the scientific landscape by enabling comprehensive quantitative measurements, thereby increasingly facilitating data-driven research. This includes genome biology, where many data sets nowadays comprise a collection of heterogeneous high-dimensional data modalities, collected from different assays, tissues, organisms, time points or conditions. An important example are multi-omics data, i. e. data combining measurements from multiple biological layers. Jointly, such data promise to provide a better and more comprehensive understanding of biological processes and complex traits. A critical step to realize these promises is the development of statistical and computational methods that facilitate moving from the data to sound conclusions and biological insights. For this purpose, an integrative analysis that combines information from different data modalities is essential.

In this thesis, we propose novel methods that provide a multivariate approach to data integration, and we apply them in the context of multi-omics studies in precision medicine and single cell biology. Given a collection of different data modalities on a set of samples, we aim at addressing two main questions: First, how can we obtain an (unbiased) overview of the main structures that are present in the data, both within and across data modalities? And second, how can we use all data to predict a response of interest and identify relevant features, whilst taking the heterogeneity of the features into account?

The first question is important in all exploratory data analysis and leads us to unsupervised methods for data integration. Finding hidden structures in the data can give important insights into biological and technical sources of variation and yield an informative low-dimensional data representation. To this end, we introduce multi-table methods and latent factor models that can capture main axes of variation and co-variation in the data. Based on this, we present a novel factor method, multi-omics factor analysis (MOFA), to integrate information from different data modalities. By sparsity assumptions on the factor loadings, MOFA decomposes variation into axes present in all, some, or single modalities and promotes interpretable factors with a direct link to molecular drivers. MOFA combines a statistical model that accommodates different data types and missing data with a scalable inference algorithm, thereby ensuring a broad applicability. Once learnt, the factors enable a range of downstream analyses, including identification of sample subgroups, outlier detection and data imputation. We demonstrate its flexibility and potential to generate biological insight by applying MOFA to a multi-omics study on chronic lymphocytic leukaemia as well as a multi-omics single cell data set.

The second question leads us to supervised methods that enable building predictive models and selecting features relevant for a response of interest. Reliable methods for this purpose would have far-reaching consequences in many applications. For example, it would be extremely useful for decisions in clinical care if treatment outcome or disease progression could be predicted from available molecular or clinical data. Furthermore, the identification of important molecular markers could give insights into underlying biological mechanisms and eventually open up new treatment options. For this purpose, we turn to penalized re-

gression methods and, based on this, develop a method for penalized regression that takes into account additional information on the features to adapt the relative strength of penalization in a data-driven manner. Such additional information in form of external covariates is available in many applications and can for example encode structural knowledge on the data, e. g. different assay types, or provide information on a feature's variance, frequency or signal-to-noise ratio. We show that incorporating informative covariates can improve prediction performance in penalized regression, and we investigate the use of important covariates in genome biology such as the omics or tissue type.

Zusammenfassung

Dank technologischer Fortschritte haben Wissenschaftler inzwischen in vielen Bereichen Zugang zu umfangreichen quantitativen Messungen, wodurch daten-getriebene Ansätze in der Forschung immer wichtiger werden. Das trifft insbesondere auf die Genombiologie zu. Hier bestehen Datensätze heutzutage oft aus mehreren hoch-dimensionalen heterogenen Merkmalsgruppen, die mit verschiedenen Verfahren, in unterschiedlichen Geweben und Organismen, zu mehreren Zeitpunkten oder unter verschiedenen Bedingungen erhoben wurden. Ein wichtiges Beispiel sind multi-omische Daten, d.h. Daten, die Messungen verschiedener Arten von Biomolekülen kombinieren. Gemeinsam können solche Daten ein besseres und umfassenderes Verständnis von biologischen Prozessen und komplexen Merkmalen vermitteln. Dabei ist ein entscheidender Schritt jedoch die Entwicklung von statistischen und rechnerischen Methoden, die es uns ermöglichen, von den Daten zu fundierten Schlussfolgerungen und biologischen Einsichten zu gelangen. Hierbei sind integrative Ansätze wichtig, die Informationen aus verschiedenen Merkmalsgruppen kombinieren.

Diese Arbeit enthält neue Methoden für multivariate Ansätze zur integrativen Datenanalyse und wendet sie im Rahmen von multi-omischen Studien in der Präzisionsmedizin und Einzelzellbiologie an. Ausgehend von einer Sammlung verschiedener molekularer Daten für eine Reihe von Stichproben wollen wir insbesondere zwei Fragen aufwerfen: Erstens, wie können wir uns einen (unvoreingenommenen) Überblick über die wichtigsten Strukturen in den Daten verschaffen, die innerhalb einer Merkmalsgruppe oder gruppen-übergreifend vorliegen? Und zweitens, wie können wir auf Basis aller Daten eine Zielgröße von Interesse vorhersagen und relevante Merkmale identifizieren - unter Berücksichtigung der Heterogenität der Merkmale?

Die erste Frage ist ein wichtiger Schritt in jeder explorativen Datenanalyse und führt uns zu unüberwachten Methoden der Datenanalyse. Das Auffinden verborgener Strukturen in den Daten kann wichtige Erkenntnisse zu biologischen und technischen Ursachen liefern, die Variationen in den Daten zugrunde liegen. Zudem ermöglicht es oft eine informative niedrig-dimensionale Darstellung der Daten. Daher führen wir 'Multi-table' Methoden und latenten Variablenmodelle ein, die es ermöglichen, die wichtigsten Achsen der Variation und Kovariation in den Daten zu erfassen. Basierend darauf stellen wir eine neue Methode, multi-omische Faktoranalyse (MOFA), zur Integration von Informationen aus verschiedenen Merkmalsgruppen vor. Mittels Sparsamkeitsannahmen an die Faktorladungen zerlegt MOFA die Variation in den Daten in Faktoren, die in allen, mehreren oder einer einzelnen Merkmalsgruppe eine Rolle spielen, und begünstigt interpretierbare Faktoren, die direkt mit molekularen Markern in Verbindung gebracht werden können. Um MOFA vielseitig anwendbar zu machen, entwickeln wir ein statisches Modell, das mit verschiedenen Datentypen und fehlenden Werten umgehen kann, und kombinieren es mit einem skalierbarem Algorithmus zur Inferenz. Nachdem die Faktoren extrahiert wurden, können sie für eine Reihe nachfolgender Analysen genutzt werden, z.B. zur Identifizierung von Untergruppen oder Ausreißern in den Stichproben oder zur Imputation fehlender Werte. Wir zeigen in Anwendungen auf eine multi-omische Studie zur chronischen lymphatischen Leukämie sowie

auf einen multi-omischen Einzelzell-Datensatz, dass MOFA wichtige biologische Einsichten vermitteln kann und nützliche weiterführende Analysen eröffnet.

Die zweite Frage führt uns zu überwachten Methoden, die es ermöglichen, prädiktive Modelle für eine Zielgröße zu erstellen und relevante Merkmale auszuwählen. Verlässliche Verfahren zur Vorhersage und Variablenselektion hätten in vielen Bereichen weitgehende Konsequenzen. Zum Beispiel wäre es für Entscheidungen in der klinischen Praxis äußerst nützlich, wenn wir ein Behandlungsergebnis oder einen Krankheitsverlauf mithilfe verfügbarer molekularer oder klinischer Daten vorhersagen könnten. Zudem könnten Einsichten in relevante molekulare Marker wichtige biologische Zusammenhänge herstellen und letztlich neue Therapieansätze oder diagnostische Tests eröffnen. Daher wenden wir uns penalisierten Regressionsmethoden zu und entwickeln darauf aufbauend eine Methode, die zusätzliche Informationen über die Merkmale berücksichtigt, um die relative Stärke der Penalisierung daten-getrieben anzupassen. Solche zusätzlichen Informationen in Form von externen Kovariaten sind in vielen Anwendungen verfügbar: Sie können beispielsweise strukturelles Wissen über die Daten kodieren, wie z.B. welche Messpunkte mit welchem Verfahren gemessen wurden, oder Informationen über die Varianz, Frequenz oder das Signal-Rausch-Verhältnis eines Merkmals geben. Wir zeigen, dass die Einbeziehung informativer Kovariaten die Vorhersageleistung der penalisierten Regression verbessern kann. In Anwendungen untersuchen wir wichtige Kovariaten in der Genombiologie wie das zur Messung verwendete molekulare Verfahren oder den Gewebetyp.

Acknowledgements

First of all, I would like to express my deepest gratitude to my mentor Wolfgang Huber for providing me with all the support, advice and inspiration over the last years that strongly contributed to both my scientific and personal growth.

I am very grateful to all members of my thesis advisory committee - Lars Steinmetz, Oliver Stegle, Enno Mammen and Peter Bühlmann - for their advice throughout my PhD. In particular, many thanks to Oliver Stegle and his groups at EBI and EMBL, especially Damien Arnol, Ricard Argelaguet and Danila Bredikhin, for exciting and fruitful collaborations and many interesting discussions on factor models and single cell biology. Special thanks to Ricard for sharing all the ups and downs on our 'MOFA ride'. Moreover, additional thanks to Enno Mammen for giving me the opportunity to participate in interesting workshops, seminars and informal discussions as part of the Research Training Group 1953.

Furthermore, I would like to thank Susan Holmes for hosting me at the Statistics Department of Stanford university and teaching me a lot about multi-table methods and heterogeneous data. Thanks to Kris Sankaran, Claire Donnat, Lan Nguyen, Pratheepa Jeganathan and Christof Seiler for being very welcoming.

I am grateful to Thorsten Zenz, Sascha Dietrich and Marina Lukas for fruitful collaborations and interesting discussions on leukaemia that gave me a valuable clinical perspective on my research.

Moreover, thanks to Henrik Kaessmann and Margarida Moreira for giving me the opportunity to contribute to an exiting project on evolutionary changes in the developmental trajectories of gene expression.

Of course, a big 'thank you' to all past and current members of the Huber group for making the time at EMBL not only a scientifically exciting but also a very enjoyable time and for always being extremely supportive.

Last but not least, I would like to thank my parents for their incredible love, support and encouragement and my brother Lars for sharing all his life experiences as the older while giving the younger the strongest feeling of confidence and respect. Finally, many thanks to Lennart for always being at my side over the last years, all the way here and beyond.

List of publications

This thesis is based on the following articles:

Argelaguet*, R., **Velten*, B**, Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W. & Stegle, O. Multi-Omics Factor Analysis - a framework for unsupervised integration of multi-omic data sets. *Molecular Systems Biology* (2018)

Velten, B & Huber, W. Adaptive penalization in high-dimensional regression and classification with external covariates using variational Bayes. *submitted, available as arXiv preprint arXiv:1811.02962*

In addition, I contributed to several data analysis projects during my PhD that have been published or are currently in preparation or under review:

Dietrich*, S., Oleś*, M., Lu*, J., Sellner, L., Anders, S., **Velten, B**, Wu, B., Hüllein, J., da Silva Liberio, M., Walther, T., *et al.* Drug-perturbation-based stratification of blood cancer. *The Journal of Clinical Investigation* **128**, 427–445 (2018)

This study combined ex-vivo drug response screens with molecular profiling in order to explore determinants of drug response in blood cancers and obtain a better disease stratification. Using multivariate penalized regression I investigated genetic determinants of drug response and assessed the explanatory power of different molecular and clinical data types for the response to different drugs.

Cardoso-Moreira, M., Halbert, J., Valloton, D., **Velten, B**, Chen, C., Shao, Y., Liechti, A., Ascenção, K., Rummel, C., Ovchinnikova, S., *et al.* Molecular innovation and conservation across mammalian organ development. *resubmitted*

This study provides a comprehensive comparison of gene expression across development in different species and tissues to investigate mammelian organ evolution. Using Gaussian processes I clustered time series of gene expression data over development from different species and tissues to facilitate comparison of developmental trajecotries and identify phylogenetic changes.

Lukas*, M., **Velten*, B**, Sellner, L., Tomska, K., Huellein, J., Walther, T., Wagner, L., Muley, C., Wu, B., Oleś, M., *et al.* Survey of ex vivo drug combination effects in chronic lymphocytic leukemia reveals synergistic drug effects and genetic dependencies. *in preparation*

This study provides a map of drug-drug interations in chronic lymphocytic leukaemia based on ex-vivo drug response screens. I performed the data analysis in this project starting from the raw data with conceptual input by Marina Lukas, Wolfgang Huber and Thorsten Zenz. In particular, the analysis includes the identification of molecular determinants of drug combination responses, cluster analyses and quantification of synergistic effects.

* denotes equal contribution as first author.

Contents

CHAPTER 1

Introduction: Premises, promises and challenges of heterogeneous data in genome biology

Homogeneous data are all alike; all heterogeneous data are heterogeneous in their own way.

The Anna Karenina principle

Over the last decades, technological advances have transformed the way how research is conducted in many disciplines. Scientists nowadays have access to unseen amounts of data, making data-driven approaches in research increasingly common. Prominent examples are physics or astronomy where petabytes of data are generated and stored each year, e.g. by particle detectors at CERN or telescopes around the globe. Likewise, many other disciplines have undergone a similar 'data revolution'. This includes genome biology, where the transition from scarce and qualitative data to vast amounts of quantitative data has been game-changing and set biology on an equal footing with particle physics or astronomy in terms of data resources [177].

While this development promises to lead to new insights and enhance our understanding of a given system or process, it also entails plenty of challenges and pitfalls. Apart from the infrastructural demands to store and process petabytes of data, their analysis opens up new statistical challenges and opportunities. Critical to unlock the potential of modern data collections is the development of statistical and computational methods that enable to move from the data to sound conclusions and scientific discoveries. In particular, such methods have to be reliable, powerful and adaptive. To arrive there, it is essential to take properties of the data into account, and interdisciplinary approaches become important that bring together statisticians, computer scientists and domain experts.

In this thesis, we present novel computational and statistical methods motivated by applications in genome biology. A particular goal is the integrative analysis of multiple high-dimensional data sets, e. g. comprising measurements from different assays. These two properties - high-dimensionality and heterogeneity - are common to many modern data collections, both in genome biology and elsewhere. Data is increasingly collected by combining different (high-throughput) technologies and joining collaborative efforts, where measurements are obtained from various labs or locations. In this chapter, we will introduce some motivating examples from genome biology with major focus on multi-omics studies and discuss the statistical challenges arising from integrative approaches to such data. Based on this, we provide an outline of the thesis at the end of the chapter.

1.1 The omics revolution in genome biology

Genome biology aims at understanding the genetic and molecular mechanisms underlying the functioning of an organism. For this purpose, a rich resource of high-dimensional heterogeneous data is nowadays generated by so-called (multi-)omics technologies. These technologies enable scientists to study biological processes and systems at unprecedented detail and can provide a comprehensive picture of multiple molecular layers. For the resulting data, it is of particular importance to take an integrative approach in the analysis, as the different molecular layers are inherently linked and play together in the regulation of biological processes. In this section, we will introduce some basic molecular layers and their interplay, starting from the 'central dogma of biology' as postulated by Crick in 1958 [41], and then turn to omics technologies and multi-omics approaches as important tools to study these layers. Using examples from precision medicine and single cell biology, we discuss some of the promises such data hold.

1.1.1 The central dogma of molecular biology

All organisms consist of cells as basic units, that each contain a 'construction plan' in their genetic information. This information is stored in the DNA (deoxyribonucleic acid), which consists of the four 'letters' A, G, C and T, encoded by the bases adenine (A), guanine (G), cytosine (C) and thymine (T). Importantly, this information can be inherited: In a process called replication the information in the DNA is copied, producing two identical DNA molecules that can be passed on to daughter cells. In order to give rise to a living organism with all its observable and measurable characteristics (also referred to as phenotypic traits), the DNA needs to be decoded. The central dogma describes the flow of the genetic information from nucleic acids to proteins (Figure 1.1). In most organisms, DNA is transcribed into RNA (ribonucleic acid), which is composed of the four letters A, U, C and G, using the base uracil (U) instead of thymine. RNA is then translated into proteins, composed of different amino acids. Proteins are the cells' building blocks and execute a range of functions, e. g. they are essential regulators, messengers, transporters and structural components of the cell.

Jointly, the genome (i. e., the complete set of DNA molecules), transcriptome (i. e., the complete set of RNA molecules) and proteome (i. e., the complete set of proteins) thus form the basic molecular layers in the cell. While still the central dogma remains the unifying principle of molecular biology, information flow between these layers is subject to many regulatory processes that have been discovered in the past 50 years. Therefore, the abundance of proteins, transcripts, and other molecular species cannot be expected to be well correlated, creating a need for multi-layered measurements. In particular, only a small percentage of the DNA directly encodes proteins and many elements have regulatory functions instead. Different types of RNA molecules exist that are not translated and, amongst others, act as catalytically active components of the translation machinery and regulators of gene expression. Furthermore, changes in how the DNA is packed or chemically modified determine the way in which genes are transcribed (or expressed). As this information can be inherited among multiple generations of cells or even organisms, it is also referred to as 'epigenetic' information or 'epigenome'.

Together, these mechanisms provide the necessary flexibility to give rise to various different cell types and tissues (starting from the same genetic code) and to react to environmental influences. At the same time, defects in any of these layers and regulatory processes can give rise to disorders and diseases. Investigating the interplay of the different layers

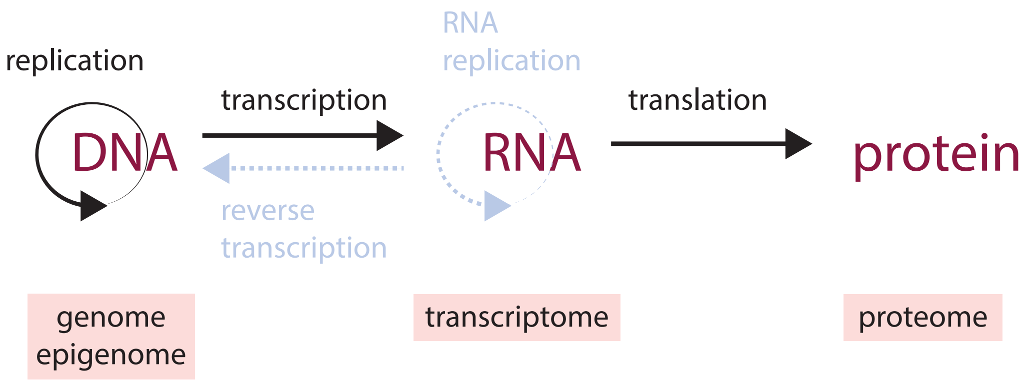


Figure 1.1: The central dogma of molecular biology. The genetic information that is stored in the DNA is transcribed into RNA, which then in turn is translated into proteins. In a process called replication the information in the DNA is copied, yielding two identical DNA molecules. In some viruses, different processes have been discovered, i.e. the reverse transcription from RNA to DNA and the direct replication of the RNA. The central dogma states that there is no information transfer from proteins to nucleic acids (i. e. no arrow back from the right-hand side).

is thus essential to understand the biological mechanisms underlying the development and functioning of an organism in health and disease.

1.1.2 What is (multi-)omics?

The advent of next-generation sequencing enabled researchers in genome biology to study the different biological layers at unprecedented detail. Nowadays, it is possible to measure the abundances and activities of thousands of biomolecules at a reasonable cost and effort. The term 'omics' is typically used for the comprehensive quantitative description of a class of molecules in a given sample that is obtained by these technologies. Important examples are genomics, epigenomics, transcriptomics and proteomics, which investigate the basic components introduced in context of the central dogma. In addition, metabolomics, microbiomics and approaches based on imaging, perturbation studies or other types of phenotypic profiling (phenomics) can contribute to the variety of omics (Table 1.1).

After some preprocessing, omics data can often be represented by a matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$ containing p molecular features measured on n samples, e. g. cell lines or tissue samples. Frequently, p is of the order of several ten or hundred thousand (e. g. expression levels of 20,000 genes, methylation marks at 500,000 CpG sites in the genome or the mutational status at a million genomic sites). The number of samples is often comparatively small and commonly ranges in the hundreds or less, with some exceptions (especially in the field of single-cell biology (see Section 1.1.4), where the number of samples can reach up to a million cells).

By now, large scale research initiatives, such as the Human Genome Project [98], the Cancer Genome Atlas [201] or the Human Microbiome Project [140], as well as individual groups have generated rich data resources of various omics types across different organisms and conditions. Several platforms provide the infrastructure to deposit and access such data including for example Ensemble [212], ArrayExpress [115], Gene Expression Omnibus [53], UniProt [193], Reactome [106] or EMPIAR [100].

Table 1.1: Short overview of common omics types

Genomics	study of DNA sequences and genetic variants, e. g. using whole genome or targeted sequencing
Epigenomics	study of structural and chemical modifications of the DNA such as methylation or histone modifications, e. g. using Chip-Seq, bisulfite sequencing, ATAC-seq or related methods
Transcriptomics	study of gene expression by measuring the relative abundances of RNA molecules, e. g. using micro-arrays or RNA sequencing
Proteomics	study of protein levels, modifications and structures, e. g. using mass spectrometry
Metabolomics	study of metabolite levels, e. g. using mass spectrometry
Microbiomics	study of presences and abundances of microorganisms, e. g. using 16S or shotgun sequencing
Phenomics	study of observable characteristics and traits, e.g. using high-throughput imaging or perturbation experiments

While many studies focus on a particular molecular layer, or omics type, technological advances enable nowadays to *simultaneously* profile different layers in high-throughput on the same samples, resulting in so-called multi-omics data. These can often be represented as a collection of matrices $\mathbf{Y}^{(m)} \in \mathbb{R}^{n \times p_m}$, one for each omics type m, with (usually distinct) molecular features in columns and (common) samples in rows. Multi-omics approaches are important, as biological processes and complex traits typically arise from interactions of many molecular layers. Hence, it is essential to take multiple biological layers into account in order to gain a comprehensive understanding of the underlying biology [85, 107, 162]. Therefore, with decreasing costs and increasing automation, multi-omics studies become more and more common across domains, including medicine [29, 47, 77, 99, 139], microbiology [110] and, most recently, single-cell biology [40].

In addition to the combination of various omics technologies, data is often collected from different time points, species, tissues, locations, batches, conditions or perturbations, adding to the complexity of the data. For example, gene expression data in different human tissues were collected by the GTEx consortium [126] and recent projects generated single cell expression data from various tissues of mice [83, 157].

1.1.3 Multi-omics approaches for precision medicine

A major challenge in clinical care is the heterogeneity of treatment outcome and disease progression across patients. Often it is unclear, why some patients respond well to a certain treatment while others do not; or why for some patients a disease is much more aggressive than for others. Precision medicine aims to find better patient stratifications and eventually arrive at molecularly informed personalized treatment decisions for individual patients (Figure 1.2). To achieve this goal, the key is a better understanding of the molecular sources and characteristic markers that underlie the observed variability between patients. Here, multi-omics data provide a rich resource: By combination of different technologies, researchers hope to tackle the complexity of human diseases [4, 35, 82, 108]. This has motivated comprehensive omics profiling on large patient cohorts to study disease heterogeneity. For example, we and others aimed at finding better disease stratification and molecular markers of drug sensitivity in cancer by combining molecular profiling with drug response screens on cell lines [12, 74, 99] or primary cancer cells from individual patients [6, 47, 68].

1.1.4 Multi-omics approaches at single cell resolution

Up to now, most omics measurements originate from bulk samples, i. e. they start from a mix of heterogeneous cells taken from a tissue or cell culture of interest. This was necessary as a substantial amount of starting material is required for most omics technologies to be applicable. However, recent technological advances have made it possible to explore many common omics types at the level of individual cells [169], including the transcriptome [114, 132, 153], genome [60, 147], epigenome [23, 172] or protein abundances [11]. For a multi-omics approach several of these techniques have to be applied to the exact same cell. While challenging, such protocols have been developed over the last four years, e. g. based on the isolation and physical separation of different biomolecules in a cell (e. g. [7, 38]) or by advanced molecular biology strategies (e. g. [46, 178]). This now enables researchers to jointly apply up to three different omics technologies at single-cell resolution [40], facilitating e. g. joint profiling of transcriptome and genome [46, 130], epigenome and transcriptome [7, 36, 38, 81] or transcriptome and surface proteins [178].

Gaining single-cell resolution provides valuable additional information compared to bulk measurements. While latter can only reveal the average of a molecular feature in an often heterogeneous cell population, single cell resolution enables to study the feature's distribution across cells and deconvolve the contribution of different cell types. For example, differences in gene expression between tumour samples can often be largely explained by different degrees of immune cell infiltration, which may mask tumour-cell specific effects [166]. Multi-omics approaches at single cell level thus enable to explore the regulation and interplay of different biological layers in cell fate decisions, e. g. in disease formation or development, or cell-cell interactions.

1.2 Statistical challenges

Despite the increasing availability of large data collections, such as multi-omics studies, their full potential has not yet been realized. A bottleneck is the integrative analysis of all data. Different data types, such as binary, count or continuous data, as well as different data qualities come along with each technology. Integrative approaches need to combine the different sources of information and uncover their relationships whilst taking this heterogeneity into account. Currently, many analyses are based on marginal associations between individual features such as in genome-wide association studies (GWAS) [13] or quantitative trait loci (QTL) studies [34, 79]. While this can lead to important insights, it ignores much of the available information. A multivariate approach could empower the analyses and provide a global picture beyond pairwise, and often spurious, associations. However, it also encounters additional statistical and computational challenges, and we will discuss important examples in this section.

By now, different strategies have been developed for integrating all data into a joint statistical model [18, 85, 162]. Depending on the aim of the analysis, we can broadly distinguish unsupervised and supervised integration. Unsupervised integration aims at identifying the major structures in the data in an unbiased way without guidance by labels or response variables. Supervised integration aims at finding the relation of the measured features to

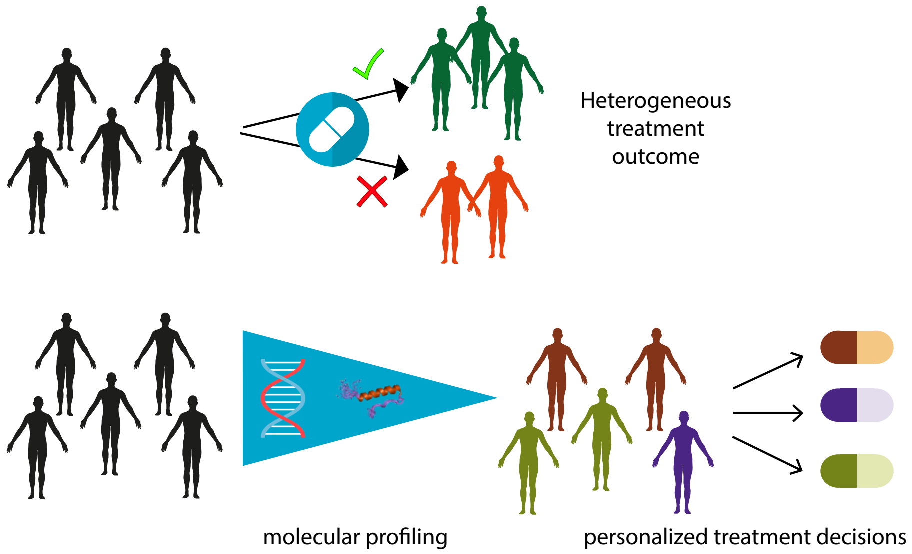


Figure 1.2: Illustration of multi-omics approaches to personalized medicine. Treatment outcome varies from one patient to another. Molecular profiling could enable clinicians to arrive at a better disease stratification and take molecularly informed treatment decisions for the individual patient.

a specific response variable. Both approaches need to deal with the high-dimensionality of the data set, its heterogeneity as well as other common properties such as missing values or complex correlation structures.

1.2.1 High-dimensionality

In a statistical sense, high-dimensional data refers to a setting where the number of samples n is much smaller than the number of features p ($n \ll p$). It might seem favourable to have a comprehensive set of features providing information. However, as not all features are necessarily informative, separating signal from noise can be challenging. Importantly, the large feature space brings along the curse of dimensionality, implying that the observations become very sparse. This leads to non-identifiability of traditional estimates, results in an increased multiple-testing burden in marginal approaches and renders multivariate methods prone to over-fitting, i. e. yielding estimates with large variance and bad prediction performance on new data. Furthermore, due to the sparsity of observations reliable estimates of covariances are difficult to obtain.

To tackle these problems it is important to employ some sort of regularization on the model complexity in order to obtain identifiable, generalizable and interpretable models. Common tools to restrict the model complexity include regularization via penalized likelihood approaches or Bayesian priors as well as dimensionality reduction and feature selection. Often, regularization is based on the assumption that the data can be explained by a sparse model, where only a relatively small number of features play a role. Alternatively, the data can often be reasonably assumed to lie on a much lower dimensional manifold embedded in the high-dimensional space that encodes the most fundamental structures in the data, yielding a model that is sparse after a suitable transformation of the feature space. In addition, regularization can also be useful to incorporate known properties of the data, such as smoothness along time or space or functional relationships. Apart from the need for regularization, the increasing size of data also makes the computational scalability of the algorithms an important pre-requisite for the applicability of a method.

Despite these challenges, the increasing dimensions of the data also open up new opportunities for statistical modelling by adaptive and data-driven approaches. An example is Empirical Bayes [54], where suitable Bayesian priors are learnt from the data itself, sharing information across different features or data sets.

1.2.2 Heterogeneity

With the term heterogeneity we describe a data collection, whose features differ in their statistical properties. For example, they might be of different data types (e. g. categorical or continuous) or display distinct correlation and noise structures. As exemplified by multi-omics data, modern data sets often consist of such heterogeneous modalities due the availability of diverse technologies and increasingly collaborative efforts. They can provide both complementary and redundant information on the samples, and their integration can be the key to reconstruct a more complete picture of the underlying system by re-assembling the parts accessible by a single technology or sensor. In addition, their integration can help to mitigate the noise of each single source. In order to benefit from the data as a whole, it is therefore strongly advisable to adopt a proper integration strategy instead of analysing each data modality in isolation from the rest. For this, however, methods need to be able to jointly model measurements that may follow very different statistical distributions and uncover relations of heterogeneous feature sets collected from different sources. This motivates the development of new methods in several fields such as multi-table analysis in statistics, multi-view learning in the machine learning community or multi-way analysis in the case of matching features across all modalities.

1.2.3 Correlation structures and missing values

Besides high-dimensionality and heterogeneity, other properties of the data often need to be taken into account in the analysis such as incompleteness of the data or complex correlation structures.

Missing values are present in most real data sets and can occur both missing at random or with some (usually unknown or vaguely known) pattern. Their presence can reduce the power of an analysis and bias the results. To cope with missing values, it is therefore important to understand potential causes for the incompleteness of the data at hand and adopt a suitable imputation approach or an explicit model of the missing data points.

Strong correlations are very common in real data, both between samples and between features. This can be problematic, as many commonly applied statistical methods assume independence or only allow for limited dependency structures. As a result, analyses of such data can lead to invalid conclusions, if based on methods whose assumptions on the (in)dependence are not met. To alleviate this problem, it is often helpful to analyse the major factors underlying the correlation structures, which can then be accounted for in the statistical analysis. Furthermore, correlations between features can result in unstable feature selection and reduced power in multivariate approaches. Here, approaches based on clusters or hierarchies of the features can become necessary.

1.3 Outline of the thesis

The thesis encompasses two novel methods, that we developed for the integrative analysis of high-dimensional heterogeneous data in genome biology in an unsupervised (Chapter 4) as well as in a supervised context (Chapter 5). We demonstrate these methods in applications to omics studies with main focus on personalized medicine and single cell biology. To lay the ground for this work the following two introductory chapters will outline existing strategies to integrate high-dimensional data both in an unsupervised manner using factor models (Chapter 2) and in a supervised manner in the framework of penalized regression (Chapter 3).

CHAPTER 2

Factor models for unsupervised data integration

Finding latent structures in data is an essential step in order to make sense of one's data, obtain meaningful visualizations and uncover major sources of variation, which can originate both from technical factors such as batch effects as well as from biological factors such as different disease states or cell types. While originally samples are profiled in very high-dimensional feature spaces, the core underlying biology can often be represented by much lower dimensional manifolds, which we aim to reconstruct. Already with a single data modality, an unsupervised analysis can provide important insights into these underlying structures. When multiple data modalities are available on the same samples, each can provide its own 'view' onto this manifold and their integration can help to identify the most important latent structures in a more accurate and complete manner. Thereby, we can eventually gain a better understanding of the samples' distribution and the connections of different features or data types.

In this chapter, we will introduce some major techniques with the aim of recovering latent structures. Our main focus lies on factor models that uncover latent structures in form of continuous axes or factors. Starting with methods for a single data modality we will then investigate their extensions to multiple data modalities. This introduction will lay the grounds for Chapter 4, where we propose a novel method for the unsupervised integration of multi-omics data. Following the terminology in the field of multi-view learning or multi-table analysis we will also refer to the different data modalities as views or tables.

2.1 Identifying latent structures from a single data modality

Due to omni-presence of high-dimensional datasets nowadays, a large collection of methods has been developed that aim to find a meaningful low-dimensional data representation starting from a single data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$. A first class of methods are linear methods based on matrix factorization such as principal component analysis (PCA) [93], factor analysis or non-negative matrix factorization [149]. In addition, there are non-linear methods including principal curves [86], Gaussian process latent variable models [117], auto-encoders [89] as well as graph-based approaches. Prominent examples of the latter are Isomap [180], diffusion maps [39], t-distributed stochastic neighbour embedding (t-SNE) [129], locally linear embedding [164], Hessian eigenmaps [49], or uniform manifold approximation and projection (UMAP) [135], which all try to preserve either global or local properties of the data encoded in a neighbourhood graph.

Here, we will focus on linear factor models, which have proven powerful tools in genomics. There are of course examples where non-linear techniques are required to capture structures present in the data. However, in many applications to real data, non-linear counterparts do not outperform linear methods such as PCA [194], and linear methods can already provide important insights and a meaningful data representation. Furthermore, they provide a direct link from the factors to the features via the weight matrices, which can enhance interpretability of the model.

In essence, we will look into models that share the basic form

$$\mathbf{Y} = \mathbf{Z}\mathbf{W}^T, \tag{2.1}$$

where the data $\mathbf{Y} \in \mathbb{R}^{n \times p}$ is decomposed into a set of continuous factors $\mathbf{Z} \in \mathbb{R}^{n \times k}$ and weights $\mathbf{W} \in \mathbb{R}^{p \times k}$. Here, k is typically much smaller than p resulting in dimensionality reduction. Depending on the absence or presence as well as the nature of a probabilistic model in the decomposition (2.1) this includes as special cases factor analysis, PCA and probabilistic or Bayesian PCA which we will introduce below.

2.1.1 Principal component analysis

One of the most commonly used methods for dimensionality reduction is principal component analysis (PCA), which was introduced by Hotelling in 1933 [93]. PCA is a linear method that performs an orthogonal transformation of the measured variables into a set of linearly uncorrelated variables. Each of these principal component axes is defined such that the projection onto the axis maximizes the variance under the constraint of orthogonality with respect to the previous components. Suppose we are given a data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$ with samples in rows, features in columns and column-wise zero means. Then the principal axes can iteratively be determined by the following maximization problems

$$u_1 = \arg \max_{||u||=1} u^T \mathbf{Y}^T \mathbf{Y} u \tag{2.2}$$

and for $r = 2, \dots, \min(n-1, p)$

$$u_r = \arg \max_{||u||=1} u^T \mathbf{Y}_r^T \mathbf{Y}_r u \tag{2.3}$$

with $\mathbf{Y}_r$ denoting the deflated data matrix $\mathbf{Y}_r = \mathbf{Y} - \sum_{s=1}^{r-1} \mathbf{Y} u_s u_s^T$. The complete set of principal components can be obtained as

$$\mathbf{T} = \mathbf{Y}\mathbf{U}, \tag{2.4}$$

where the principal axes $\mathbf{U}$ are given by the eigenvectors of the sample covariance matrix $\mathbf{S} = \frac{1}{n}\mathbf{Y}^T\mathbf{Y}$ and $\mathbf{T}$ contains the component scores. Hence, the solution can be found by an eigenvalue problem on the sample covariance matrix $\mathbf{S}$ or a singular-value decomposition (SVD) of the data matrix $\mathbf{Y}$. The first k principal components are equivalently characterized as the data representation in reduced dimensions that minimizes the total squared reconstruction error to the original data. I. e., the projection $\mathbf{Y}_k = \mathbf{Y}\mathbf{U}_k\mathbf{U}_k^T$ with $\mathbf{U}_k = [u_1, \dots, u_k]$ minimizes the Frobenius norm $||\mathbf{Y} - \hat{\mathbf{Y}}||_F$ among all matrices $\hat{\mathbf{Y}} \in \{\mathbf{M} \in \mathbb{R}^{n \times p} | \text{rank}(\mathbf{M}) \leq k\}$ [52]. Importantly, the solution of principal component analysis depends on the scaling of the individual variables: Variables with a higher variance will have stronger contributions to the direction of maximal variance. Extensions of PCA have been proposed which replace the covariance matrix with an arbitrary kernel matrix resulting in non-linear versions of PCA [167]. Depending on the choice of the kernel these have close relationships to some of the non-linear graph-based techniques mentioned above [194].

2.1.2 Sparse PCA

Motivated by improving interpretability of the principal axes, sparse variants of PCA have been developed. Here, only a small set of features has non-zero weights in the principal axes $\mathbf{U}$. This can make it easier to identify the relevant features contributing most to the variation in the data. For this purpose, the optimization problems solved by PCA, such as the maximal variance problem or the minimal reconstruction error, are modified with an L_1-constraint on the principal axes [104, 205, 217]. Apart from interpretability, constraints towards sparsity can also improve the properties of PCA in high-dimensions, where ordinary PCA yields an inconsistent estimator of the subspace with maximal variance [103].

2.1.3 Probabilistic interpretation of PCA

PCA itself is not based on an explicit probability model. However, a probabilistic formulation could give access to likelihood measures and a generative model or open up mixture model extensions and applications of Bayesian inference. Motivated by this, Tipping & Bishop suggested in 1999 a probabilistic interpretation of PCA, probabilistic principal component analysis (pPCA) [187]. In particular, they studied the factor model given by

$$y_i = \mathbf{W} z_i + \mu + \epsilon_i, \tag{2.5}$$

where $y_i \in \mathbb{R}^p$ denotes the i^{th} sample (corresponding to the i^{th} row of $\mathbf{Y}$), $\mu \in \mathbb{R}^p$ the intercept, $\mathbf{W} \in \mathbb{R}^{p \times k}$ the weight matrix, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 \mathbb{1}_p)$ an isotropic Gaussian noise term and $z_i \overset{iid}{\sim} \mathcal{N}(0, \mathbb{1}_k)$ the latent factors, or components, for $i = 1, \dots n$. From the resulting marginal model,

$$y_i \overset{iid}{\sim} \mathcal{N}(\mu, \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbb{1}_p), \tag{2.6}$$

the maximum likelihood estimator (MLE) of the $p \times k$ matrix $\mathbf{W}$ can be obtained as

$$\hat{\mathbf{W}}_{\text{MLE}} = \mathbf{U}(\boldsymbol{\Lambda} - \sigma^2 \mathbb{1}_k)^{\frac{1}{2}} \mathbf{R}. \tag{2.7}$$

Here, $\mathbf{U} \in \mathbb{R}^{p \times k}$ are the k principal eigenvectors of the sample covariance matrix $\mathbf{S}$, $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_k)$ contains the corresponding eigenvalues of $\mathbf{S}$ on the diagonal and $\mathbf{R} \in \mathbb{R}^{k \times k}$ denotes an arbitrary orthogonal rotation matrix. In particular, the maximum likelihood solution spans the k-dimensional principal subspace of the data. The individual weight vectors in $\hat{\mathbf{W}}_{\text{MLE}}$ do not directly correspond to the principal axes but are scaled by σ^2 and $\boldsymbol{\Lambda}$ as well as arbitrarily rotated. In contrast to the PCA solution, they are thus not necessarily orthogonal and are non-identifiable with respect to rotations (rotational ambiguity). In practice, an expectation-maximization (EM) algorithm [45] was suggested to derive the parameters in Equation (2.5) and avoid working with the full covariance matrix in high dimensions [187]. Here, the posterior mean for the latent factors is given by

$$\mathbb{E}[z_i | y_i] = (\mathbf{W}^T \mathbf{W} + \sigma^2 \mathbb{1}_k)^{-1} \mathbf{W}^T (y_i - \mu), \tag{2.8}$$

which can be interpreted as a ridge estimate (see Chapter 3).

Based on this probabilistic model, several Bayesian versions of PCA have been formulated. They impose regularizing priors such as the automatic relevance determination (ARD) prior [131] or sparsity promoting priors on the components in order to automatically determine the dimensionality of the latent space [19] or to obtain sparse solutions [80].

2.1.4 Factor analysis

Strongly related to PCA is factor analysis, which dates back to 1904 [174]. Factor analysis is based on a probabilistic model that allows, in contrast to the pPCA model, for non-isotropic covariance matrices of the noise term. Introductions to factor analysis can be found in most textbooks on multivariate analysis, e. g. [161]. The basic model of factor analysis is given by

$$y_i = \mathbf{W} z_i + \mu + \epsilon_i, \tag{2.9}$$

where as above $y_i \in \mathbb{R}^p$ denotes the observation vector from the i^{th} sample, $\mathbf{W} \in \mathbb{R}^{p \times k}$ the weights and $z_i \in \mathbb{R}^k$ the factors' scores for $i = 1, \dots, n$. One assumes that the error terms fulfil $\mathbb{E}\,\epsilon_i = 0$, $\mathrm{Var}\ \epsilon_i = \boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \dots, \psi_p)$, that the factors obey $\mathbb{E}\, z_i = 0$, $\mathrm{Var}\ z_i = \mathbb{1}_k$ and that the factors and errors are uncorrelated. The variance $\mathrm{Var}(y_{ij})$ is thus decomposed as $\sum_{l=1}^k w_{jl}^2 + \psi_j$, where the first term is referred to as communality and the second term as the uniqueness or specific variance [84]. Due to the non-isotropic covariance structure of the error term, factor analysis focusses on capturing variation that is shared between features, trying to best reproduce correlations between features. This is in contrast to PCA, where also variation unique to a single feature is modelled via the principal components.

There are several methods for estimating the parameters in the factor analysis model. We will again focus on the maximum likelihood solution, similar to pPCA. Here, it is common to assume $\epsilon_i \sim \mathcal{N}(0, \boldsymbol{\Psi})$ with $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \dots, \psi_p)$ and $z_i \sim \mathcal{N}(0, \mathbb{1}_k)$ yielding the marginal model

$$y_i \overset{iid}{\sim} \mathcal{N}(\mu, \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}). \tag{2.10}$$

In particular, due to the non-isotropic covariance structure, the maximum likelihood solution is invariant with respect to scaling of the individual features. As no closed form solution is available, the estimation requires an iterative algorithm. For example, an EM-algorithm as in the pPCA model can be used [20], but here with a weighted ridge estimate in the E-step, i. e.

$$\mathbb{E}[z_i|y_i] = (\mathbb{1}_k + \mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W})^{-1} \mathbf{W}^T \boldsymbol{\Psi}^{-1} (y_i - \mu). \tag{2.11}$$

Per se the factor analysis model is non-identifiable: It is invariant under orthogonal transformations and different choices of rotations have been suggested to obtain more interpretable factors. An example is the varimax rotation, that maximizes the variance of squared loadings for each factor, thereby resulting in a clearer distinction between large and small loadings on a factor.

Another important choice is the number of factors to be included into the model. While in PCA higher-order principal component models will always include the lower-order principal components due to the isotropic noise model, in factor analysis the model with two factors can be very different from a model with a single factor [187].

As for pPCA, a Bayesian treatment of factor analysis with priors on the loading matrices can provide tools for determining the number of factors and to impose sparsity or other constraints, thereby in parts alleviating the problem of rotational ambiguity and model selection. In addition, this opens up the use of Bayesian inference methods based on Markov chain Monte Carlo (MCMC) methods such as Gibbs sampling [76] or deterministic approximate inference methods such as the Laplace approximation, variational Bayes [21, 105] or expectation propagation [142].

2.1.5 The role of factor models in genome biology

In genome biology, principal components and factor models are widely used, both for exploratory data analysis and down-stream analyses. PCA is the standard tool to obtain a low-dimensional visualization of the data. This initial exploration can already reveal possible batch effects or biological drivers of variation. In addition, the top principal components are heavily used for various downstream analyses, including clustering (e. g. [132]), regression or classification tasks (e. g. [77]) or further non-linear dimensionality reduction (e. g. [129]). This enables working in a much lower dimensional space, that is hoped to retain most of the biological signal but less of the technical noise, and thus can reduce computational costs or over-fitting. The number of principal components to keep in such downstream analyses can be picked based on resampling procedures and scree-plots of the variance explained by each component.

While PCA provides an important generic tool, the flexibility of factor models has also led to many adaptations and extensions for specific applications in genome biology. For example, models to incorporate known covariates [24, 72, 120, 176], non-Gaussian data types [24, 154], known feature annotations [24, 61] or sample relationships [71] have been suggested. Their applications range from the detection and modelling of confounding factors such as batch effects [72, 120] or population structure in genetic data [57, 152, 156] to decompositions of large data sets in order to define biological reference signatures, e. g. using somatic mutation data to define signatures of mutational processes [3].

Apart from their flexibility, the popularity of linear factor models is also due to the fact that they can provide a direct link from the factors to the molecular features via the weight matrices, thereby enabling to uncover the molecular underpinnings of the major sources of variation. To further improve interpretability, sparsity constraints have been imposed on the weights in many applications resulting in a small set of active features [24, 57].

2.2 Identifying latent structures from multiple data modalities

Suppose we are given multiple data tables or views on the same set of samples $\mathbf{Y}^{(m)} \in \mathbb{R}^{n\times p_m}$ for $m = 1, \dots, M$. In this setting, we can distinguish between latent structures that are present in several or all of the modalities and those unique to a single data modality. The methods discussed in the previous section therefore need to be adapted to explore not only variation within a modality and relationships of individual features but also covariation across modalities and relationships between whole data tables.

A first starting point to analyse such a data collection could be concatenating all data to a single data matrix $\mathbf{Y} = (\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(M)}) \in \mathbb{R}^{n\times \sum_m p_m}$ and apply single-view methods such as PCA. By this, we can arrive at a low-dimensional representation of the samples and study relationships between individual features. However, this approach does not directly provide us with information of relationships between the data modalities. In addition, data modalities might come along with different data types, e. g. continuous or categorical data, and can be of different quality or dimension. To account for this, methods tailored towards multiple data modalities have been proposed. While some methods, such as canonical correlation analysis [92], focus on the shared patterns of variation, other methods consider both shared and unique structures such as inter-battery factor models or group factor analysis [113, 197]. Here, we will introduce such methods, again with main focus on linear method based on matrix factorizations. We note, however, that non-linear dimensionality reduction techniques can also been extended to multiple views [43] and might be advantageous in the presence of strongly non-linear relationships in the data. We will start with introducing two important examples of multi-table analyses, i. e. canonical correlation analysis and co-inertia analysis, and then turn again to probabilistic interpretations as latent variable models.

In the following, we suppose we are given $\mathbf{Y}^{(m)} \in \mathbb{R}^{n\times p_m}$ for $m = 1, \dots, M$ with column-wise mean zero, and we denote the empirical variance and covariances with $\mathbf{S}_{(r,s)} = \frac{1}{n}\mathbf{Y}^{(r)T}\mathbf{Y}^{(s)}$.

2.2.1 Canonical correlation analysis

Given two data tables, canonical correlation analysis (CCA) [92] aims at finding axes for each table such that correlation between the projections of the data tables onto its axis is maximized. As in PCA, subsequent axes are found by the same maximization problem under the constraint of orthogonality with respect to the previous components. The k^{th} canonical vectors are hence given by

$$\begin{aligned} u_k, v_k \in \arg\max_{u,v} \operatorname{cor}(\mathbf{Y}^{(1)}u, \mathbf{Y}^{(2)}v) = \arg\max \frac{u^T\mathbf{S}_{(1,2)}v}{\sqrt{u^T\mathbf{S}_{(1,1)}u}\sqrt{v^T\mathbf{S}_{(2,2)}v}}, \\ \text{s.t. } \mathbf{Y}^{(1)}u \perp \mathbf{Y}^{(1)}u_{k'},\ \mathbf{Y}^{(2)}v \perp \mathbf{Y}^{(2)}v_{k'} \quad \forall k' < k. \end{aligned} \tag{2.12}$$

The solutions can again be found by an eigenvalue problem that is given by

$$\mathbf{S}_{(1,1)}^{-1}\mathbf{S}_{(1,2)}\mathbf{S}_{(2,2)}^{-1}\mathbf{S}_{(2,1)}u = \rho^2 u, \tag{2.13}$$

$$\mathbf{S}_{(2,2)}^{-1}\mathbf{S}_{(2,1)}\mathbf{S}_{(1,1)}^{-1}\mathbf{S}_{(1,2)}v = \rho^2 v, \tag{2.14}$$

where ρ are the canonical correlations and u, v the canonical directions. From this, the canonical variables are obtained as $a = \mathbf{Y}^{(1)}u$, $b = \mathbf{Y}^{(2)}v \in \mathbb{R}^n$.

Like for PCA, sparse formulations of CCA have been suggested with the aim to improve interpretability of canonical vectors and enable applications to high-dimensional data, where the empirical covariance matrix is singular and canonical vectors are not unique [151, 205]. CCA and sparse CCA have been widely used in genomic studies both for relating features from different (omics) technologies [118, 151, 204] and for relating samples from different batches [28]. However, per se it is limited to two data tables. Extensions to multiple data tables have been proposed, that maximize a (weighted) sum of pairwise correlations. The weighting needs to be chosen based on assumptions which tables are connected [182, 204].

2.2.2 Co-inertia analysis

Co-inertia analysis (CIA) is an alternative to CCA that looks at the co-inertia of two data tables and has received particular interest in ecological studies [37, 50, 183]. Inertia and co-inertia are commonly defined in the duality diagram framework [44], where in addition to the data tables distance metrics in the feature space ($\mathbf{Q}$) and sample weights ($\mathbf{D}$) are considered. The inertia is then defined as

$$I = \operatorname{trace}(\mathbf{Y}\mathbf{Q}\mathbf{Y}^T\mathbf{D}). \tag{2.15}$$

Representing $\mathbf{Y}$ as a cloud of n points in the feature space, the inertia thus measures the sum of squared distances of the points to the origin. Introduction of $\mathbf{D}$ enables to up- or down-weight specific points, and $\mathbf{Q}$ determines the importance of different directions in the feature space. The co-inertia of two tables is then defined as

$$CoI = \operatorname{trace}(\mathbf{Y}^{(1)}\mathbf{Q}^{(1)}\mathbf{Y}^{(1)T}\mathbf{D}\mathbf{Y}^{(2)}\mathbf{Q}^{(2)}\mathbf{Y}^{(2)T}\mathbf{D}), \tag{2.16}$$

which provides a measure of the angle between the two inertia matrices according to the inner product given by $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{trace}(\mathbf{A}^T\mathbf{B})$. As a special case, the inertia equals the sum of variances, and the co-inertia the sum of co-variances for centered $\mathbf{Y}$, diagonal $\mathbf{D}$ with entries $1/n$ and Euclidean metric. In this case, the axes that maximize the projected co-inertia are given by

$$u_1, u_2 \in \arg\max \operatorname{cov}(\mathbf{Y}^{(1)}u_1, \mathbf{Y}^{(2)}u_2) = u_1^T \mathbf{Y}^{(1)T}\mathbf{Y}^{(2)}u_2. \tag{2.17}$$

Solutions can be found by a singular value decomposition of $\mathbf{Y}^{(1)T}\mathbf{Y}^{(2)}$. In contrast to CCA, which aims at maximizing correlation, CIA focuses on covariance and thereby finds a balance between the correlation and the variance in two data tables. Subsequent components are found by the same optimization problem with constraints on the orthogonality to previous components, as in CCA.

While less common than CCA, CIA has been applied for omics integration or cross-platform comparisons [42, 119, 137]. Here, often an extension to more than two tables is necessary, and, like for CCA, weights need to be chosen for each table, i. e.

$$u_1, \ldots, u_M, v \in \arg\max \sum_{m=1}^{M} w_m \operatorname{cov}^2(\mathbf{Y}^{(m)}u_m, v), \tag{2.18}$$

where $v \in \mathbb{R}^n$ is the reference score and $w_m \in \mathbb{R}$ the weight of table m. Motivated by improved interpretability, further extensions of CIA have been proposed that employ L_1-penalties to obtain sparse weight vectors [141].

2.2.3 Probabilistic interpretation of CCA

Like for PCA, factor models can provide a probabilistic interpretation of CCA. To this end, Bach & Jordan considered in 2005 [10] the following factor model

$$z_i \overset{iid}{\sim} \mathcal{N}(0, \mathbb{1}_k), \tag{2.19}$$

$$y_i^{(m)}|z_i \overset{iid}{\sim} \mathcal{N}\left(\mu_i^{(m)} + \mathbf{W}^{(m)}z_i, \boldsymbol{\Sigma}^{(m)}\right) \quad m = 1, 2, \tag{2.20}$$

with arbitrary positive definite $\boldsymbol{\Sigma}^{(m)} \in \mathbb{R}^{p_m \times p_m}$ and $k \leq \min(p_1, p_2)$. As the residual covariance can take arbitrary forms, the factors z focus on modelling only variation that is shared across the views, corresponding to the focus on correlation in CCA. Bach & Jordan showed that the maximum likelihood estimates for the weights $\mathbf{W}$ in this model are linked to the canonical directions via

$$\hat{\mathbf{W}}^{(m)}_{\text{MLE}} = \mathbf{S}_{(m,m)}\mathbf{U}^{(m)}\mathbf{M}^{(m)}, \quad m = 1, 2, \tag{2.21}$$

and that the posterior expectations of z are given by

$$\mathbb{E}\left[z_i|y_i^{(m)}\right] = \mathbf{M}^{(m)T}\mathbf{U}^{(m)T}\left(y_i^{(m)} - \mu_i^{(m)}\right), \quad m = 1, 2. \tag{2.22}$$

Here, $\mathbf{U}^{(m)} \in \mathbb{R}^{p_m \times k}$ contain the first k canonical directions and $\mathbf{M}^{(m)} \in \mathbb{R}^{k \times k}$ are arbitrary matrices such that $\mathbf{M}^{(1)}\mathbf{M}^{(2)T}$ yields a diagonal matrix with first k canonical correlations on the diagonal. In particular, the posterior expectations of z span the same subspace as the canonical correlation axes. To infer the parameters in the model an EM-algorithm was proposed.

Based on this probabilistic interpretation, Bayesian models of CCA have been constructed that employ priors on the model parameters [111, 112, 200]. A common choice is an inverse-Wishart prior for the covariance matrices and an ARD prior [131] on the weights that shrinks weights of inactive factors to zero, thereby enabling to determine the number of components in an automated manner. Inference can for example be based on Gibbs sampling [111] or variational methods [200].

2.2.4 Inter-battery factor analysis

The probabilistic CCA model is strongly related to inter-battery factor analysis (IBFA) [192]. Here, not only shared factors but also view-specific factors are included into the model, which is given by

$$y_i^{(m)} = \mathbf{A}^{(m)} z_i + \mathbf{B}^{(m)} z_i^{(m)} + \epsilon_i^{(m)}, \quad \text{for } m = 1, 2, \tag{2.23}$$

with $\epsilon_i^{(m)} \sim \mathcal{N}(0, \boldsymbol{\Psi}^{(m)})$ and $\boldsymbol{\Psi}^{(m)} = \operatorname{diag}(\psi_1^{(m)}, \ldots, \psi_{p_m}^{(m)})$. Here, $A^{(m)} \in \mathbb{R}^{p_m \times k}$ denote the weight matrices of the shared factors $z_i \in \mathbb{R}^k$, and $B^{(m)} \in \mathbb{R}^{p_m \times k_m}$ the weight matrices of the view-specific factors $z_i^{(m)} \in \mathbb{R}^{k_m}$. This model can also be extended to more than two data modalities as in multi-battery factor analysis or the strongly related JIVE model [125], which both decompose variation into common structures present in all data modalities and individual structures.

The probabilistic CCA can be obtained from Equation (2.23) with $z_i^{(m)} \sim \mathcal{N}(0, \mathbb{1}_{k_m})$ when marginalising the view-specific factors out, i. e.

$$y_i^{(m)} | z_i \sim \mathcal{N}\left(\mathbf{A}_i^{(m)} z_i, \mathbf{B}^{(m)} \mathbf{B}^{(m)T} + \boldsymbol{\Psi}^{(m)}\right), \tag{2.24}$$

as long as all view-specific variation (given by $\boldsymbol{\Sigma}^{(m)}$ in Equation (2.20)) is modelled by the view-specific factors.

In its basic formulation in Equation (2.23), IBFA is unidentifiable, as allocation to shared and common factors can be exchanged without changing the likelihood of the model. This can in parts be alleviated by appropriate sparsity assumptions on the joint weight matrix in the concatenated model. For this, we re-write the model as

$$\begin{bmatrix} y_i^{(1)} \\ y_i^{(2)} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{W}\tilde{z}_i, \boldsymbol{\Psi}\right), \qquad \boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi}^{(1)} & 0 \\ 0 & \boldsymbol{\Psi}^{(2)} \end{bmatrix}, \tag{2.25}$$

with latent factors $\tilde{z}_i = [z_i, z_i^{(1)}, z_i^{(2)}] \sim \mathcal{N}(0, \mathbb{1}_{\tilde{k}})$ for $\tilde{k} = k + k_1 + k_2$ and weight matrix

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}^{(1)} \\ \mathbf{W}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{(1)} & \mathbf{B}^{(1)} & 0 \\ \mathbf{A}^{(2)} & 0 & \mathbf{B}^{(2)} \end{bmatrix} \in \mathbb{R}^{(p_1+p_2)\times\tilde{k}}. \tag{2.26}$$

Without the structural constraints on $\mathbf{W}$, this model would reduce to a probabilistic PCA or factor analysis model, depending on the restrictions of the diagonal entries in $\boldsymbol{\Psi}$. However, the specific pattern in $\mathbf{W}$ can be enforced via structured regularization. For example, in the Bayesian setting priors promoting group sparsity on $\mathbf{W}$ can be employed. In [112], the authors suggest using a view-wise ARD prior on the columns of $\mathbf{W}$ to achieve this, i. e. the prior on the k^{th} column of $\mathbf{W}^{(m)}$ is given as

$$\mathbf{W}_{\cdot,k}^{(m)} \sim \mathcal{N}\left(0, \frac{1}{\alpha_k^{(m)}} \mathbb{1}_p\right). \tag{2.27}$$

By allowing for different values for $\alpha_k^{(m)}$ in each view m, the k^{th} component can be specifically inactivated in both, a single or no view as illustrated in Figure 2.1: A small $\alpha_k^{(m)}$ for both $m = 1, 2$ allows for shared components, view-specific components are obtained by learning large values $\alpha_k^{(m)}$ for the inactive view and small values in the active view, and non-essential components are inactivated by learning large values for $\alpha_k^{(m)}$ in both views $m = 1, 2$. In order to make inference on the posterior distributions, we can for example resort to variational methods or Gibbs sampling as in [112].

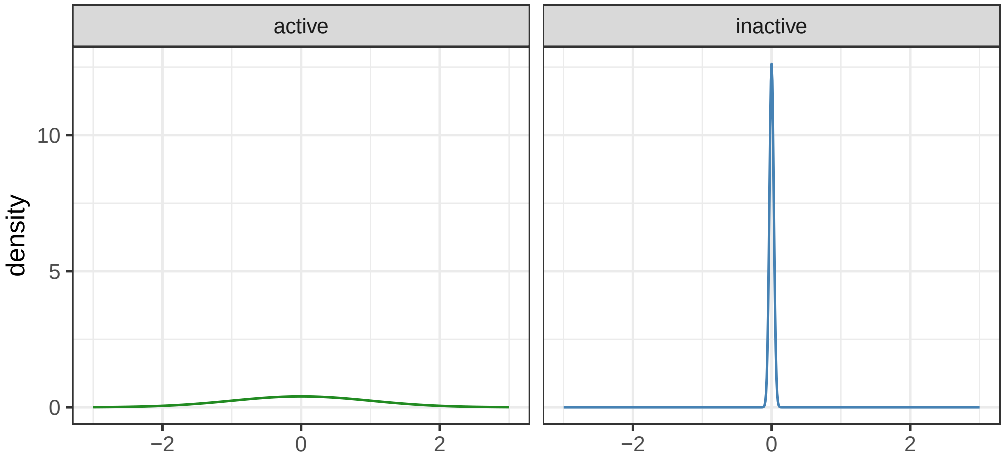


Figure 2.1: Illustration of the ARD prior in Bayesian IBFA. The figure illustrates the effect of the ARD prior on the weight vector $\mathbf{W}^{(m)}_{\cdot,k}$ for factor k. A small value of $\alpha^{(m)}_k$ leads to an active component k (left) as the flat prior poses little constraints on the size of the elements in $\mathbf{W}^{(m)}_{\cdot,k}$. On the other hand, large values of $\alpha^{(m)}_k$ inactivate a component in view m (right), as they result in a very concentrated prior distribution restricting values in $\mathbf{W}^{(m)}_{\cdot,k}$ to lie close to zero.

2.2.5 Group factor analysis

The ideas from the Bayesian IBFA model were extended in the group factor analysis (GFA) framework [197] to multiple views. Here, arbitrary combinations of view-specific factors, fully shared factors and partly shared factors are considered in order to model variation that is present in an arbitrary subset of views, including view-specific and joint variation as special cases. For this purpose, GFA uses the same view-wise regularization on the weight matrix as Bayesian IBFA. By the same principle as above, this can inactivate factors in a subset of views to model variation that is only present in some views. Several extension have been suggest within this framework [27, 113, 122, 160, 214], which differ in the form of $\boldsymbol{\Psi}$ (heteroscedastic or homoscedastic noise), the use of additional low-rank constraints on the factor specificity pattern encoded in $\boldsymbol{\alpha} = (\alpha^{(m)}_k)_{m,k}$ [113], the addition of feature-wise sparsity-promoting priors for improved interpretability of the components [27, 122, 214] and the inference scheme (e. g. Gibbs sampling or variational methods). The GFA framework will be the essential building block for our work in Chapter 4, where we also discuss relationships to different variants of GFA in more detail.

2.3 A glimpse at our contribution

In Chapter 4 we will present a novel method for unsupervised integration of multi-view data tailored to multi-omics data, multi-omics factor analysis (MOFA). This method enables to find the main (shared and unique) sources of variation across samples and gain insights into which axis of variation is important in which omics as well as which molecular processes and markers are underlying the variation. MOFA is part of the GFA framework and adapts it for applications to multi-omics data. In particular, we extended GFA by combining a scalable inference scheme based on variational Bayes (VB), non-Gaussian likelihoods, handling of missing values as well as sparse factor weights. Furthermore, we developed and implemented several downstream analyses tools and demonstrated its application on data from a multi-omics study on leukaemia as well as a single-cell multi-omics data set.

CHAPTER 3

Penalized regression for supervised data integration

Complementary to unsupervised analysis, we are often interested in directly associating the high-dimensional data modalities to a response of interest, such as a phenotypic outcome or trait, by learning a relationship of the form

$$Y = f(X),$$

given a response variable $Y \in \mathbb{R}$ and a vector of predictors $X \in \mathbb{R}^p$. Learning such a relationship can be useful both for prediction of Y based on a new observation of X and to identify which components of X are most relevant for Y. For example, in personalized medicine X could be a vector of molecular markers that we would like to use in order to predict a patient's survival, treatment response or disease risk given by Y. In particular, we again aim at integrative multivariate approaches, which consider all features jointly, as this can be advantageous in order to rule out spurious associations, reduce residual variances and obtain a good predictive model. For this purpose, we consider here (generalized) linear models as a simple yet useful class of models, where f depends on X via a linear combination of the features.

To make it explicit, that X can comprise features from multiple sources of information, such as different omics types, which we want to use jointly for the prediction of Y, we could write

$$Y = f(X^{(1)}, \dots, X^{(M)}),$$

with $X^{(m)} \in \mathbb{R}^{p_m}$ for $m = 1, \dots, M$. In contrast to the setting of unsupervised data integration, Y can now guide the integration and the primary focus is on relationships of Y to the different components of X. Therefore, a straightforward approach to this problem is the application of standard regression techniques on the concatenated vector $X = (X^{(1)}, \dots, X^{(M)}) \in \mathbb{R}^{\sum_l p_l}$. This approach typically leads to a very high-dimensional design matrix and thus requires appropriate regularization techniques. In this chapter, we will introduce some commonly applied methods for penalized regression and eventually discuss ways to account for the heterogeneity of the features in X, which motivates our work in Chapter 5.

3.1 The concepts of penalized regression

Suppose we are given observations $(x_1, y_1), \dots, (x_n, y_n)$ with $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^p$ and denote with $\mathbf{X} = (x_1, \dots, x_n)^T \in \mathbb{R}^{n \times p}$ and $y = (y_i)_{i=1}^n \in \mathbb{R}^n$. We assume that, given x_i, y_i are independent and identically distributed (iid), following a distribution in the exponential family with mean $\mu_i = \mathbb{E}(y_i|x_i)$ described by

$$g(\mu_i) = x_i^T \beta,$$

with a parameter vector $\beta \in \mathbb{R}^p$. This model encompasses both linear regression and logistic regression as special cases. A common way to estimate the coefficients uses a maximum likelihood approach. Here, the coefficients are found by minimizing the negative log-likelihood ℓ of the model, i. e.

$$\hat{\beta} \in \arg\min_{\beta} \ell(\beta). \tag{3.1}$$

For example, in the case of a linear model with normal likelihood this leads to the least-squares approach, which yields an unbiased estimator for β.

Maximum likelihood estimation, however, can encounter several problems with increasing number of predictors. This includes high variance of the estimator, non-identifiability in the high-dimensional setting ($p > n$) and, possibly, the lack of interpretability due to many non-zero coefficients. To alleviate these problems, penalization can be used to reduce the model complexity. At the cost of introducing a bias on the estimator, penalization can decrease the estimator's mean squared error via a reduction in its variance. For this purpose, an additional penalty term is added to the optimization problem in Equation (3.1), i. e.

$$\hat{\beta} \in \arg\min_{\beta} \ell(\beta) + \lambda \text{pen}(\beta) \tag{3.2}$$

Here, $\text{pen}(\cdot)$ is a penalty function that restricts the size of the model coefficients. Common examples are of the form $\text{pen}(\beta) = \sum_j |\beta_j|^q$ for some $q \geq 0$. Two prominent choices of q result in ridge regression ($q = 2$) and Lasso ($q = 1$), which we will discuss in more detail below.

The tuning parameter λ allows to vary the degree of penalization and by this modulates the bias-variance trade-off. With $\lambda = 0$ we recover the maximum likelihood solution, while for $\lambda \to \infty$ all coefficients vanish. Determining a suitable value for λ is important in practice. A very small λ yields a complex model with low bias but highly variable estimators, resulting in over-fitting of the data. On the other hand, a very large λ yields a very simple model, whose estimators have little variance but a strong bias resulting in under-fitting of the data. The relationship of model complexity with the test error and train error is illustrated in Figure 3.1. As the test error is generally not available, in practice cross-validation is commonly used to approximate the test error and to determine a suitable value for λ. For this, the samples are split randomly into k folds of (near) equal size. Then, the model is trained for different values of λ on all except one fold and evaluated on the left-out fold in terms of prediction error. The value with the best average performance across the folds is chosen for λ. Alternatively, it has been suggested to increase this value by one standard error of the cross-validation in order to give preference to the slightly less complex model with comparable performance [66]. While cross-validation provides a useful approach in practice, it is however important to note, that the best model complexity for prediction performance might not be the optimal value for recovery of the true model, e. g. in the context of feature selection [136].

3.2 Ridge regression

Ridge regression was introduced in 1970 by Hoerl & Kennard in order to address the problem of instability in least-squares estimation with non-orthogonal design and the singularity of $\mathbf{X}^T\mathbf{X}$ [90]. With high-dimensional data, ridge regression provides a way to solve the problem of non-identifiability and can lead to an (in terms of mean squared error) improved, albeit biased, estimator with reduced variance. The penalty is given by the Euclidean norm of the coefficients, thus shrinking all coefficients towards zero as illustrated in Figure 3.2.

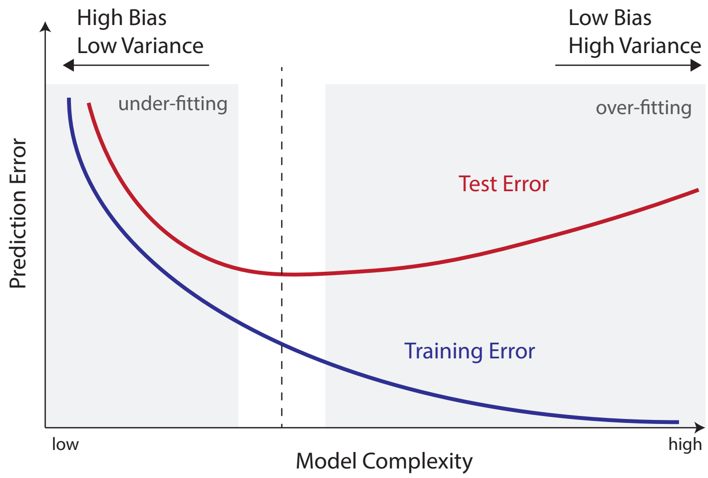


Figure 3.1: Choosing a good model complexity. Red lines indicate the test error, blue lines the training error for varying model complexities. A model with a high model complexity fits the training data perfectly, but has high variance resulting in high prediction error on test data (over-fitting). On the other hand, a low model complexity has high prediction error on both the test and training data as it has a large bias and under-fits the data. An optimal model complexity for prediction purposes would be the point with minimal test error, which can be approximated using cross-validation.

In particular, in case of a linear model the estimate is given by

$$\hat{\beta}_{\text{ridge}} \in \arg\min_{\beta} ||y - \mathbf{X}\beta||_2^2 + \lambda||\beta||_2^2, \tag{3.3}$$

or, equivalently,

$$\hat{\beta}_{\text{ridge}} \in \arg\min_{\beta} ||y - \mathbf{X}\beta||_2^2 \quad \text{s.t. } ||\beta||_2^2 \leq s. \tag{3.4}$$

This optimization problem has a closed-form solution given by

$$\hat{\beta}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbb{1})^{-1}\mathbf{X}^T y =: \mathbf{W}\mathbf{X}^T y, \tag{3.5}$$

and, under the assumption of normal errors with variance σ^2, the distribution of the estimator can easily be found to be

$$\hat{\beta}_{\text{ridge}} \sim \mathcal{N}(\mathbf{W}\mathbf{X}^T\mathbf{X}\beta, \sigma^2\mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W}). \tag{3.6}$$

For large p the estimate can be calculated making use of singular value decomposition of $\mathbf{X}$ [87] or the Woodbury matrix identity [206] to avoid the direct inversion of a $p \times p$ matrix and instead work with $n \times n$ matrices. In general, the solution for different types of likelihoods ℓ can be found using a cyclical coordinate descent algorithm to efficiently calculate the solution path for a sequence of λ values [65].

3.3 Lasso

In 1996, Tibshirani introduced the Lasso [184] which uses an L_1-penalty instead of an L_2-penalty on the model coefficients. Due to the shape of the L_1-norm, the Lasso penalty forces many coefficients to be exactly zero, as illustrated in Figure 3.2. Thereby, the Lasso

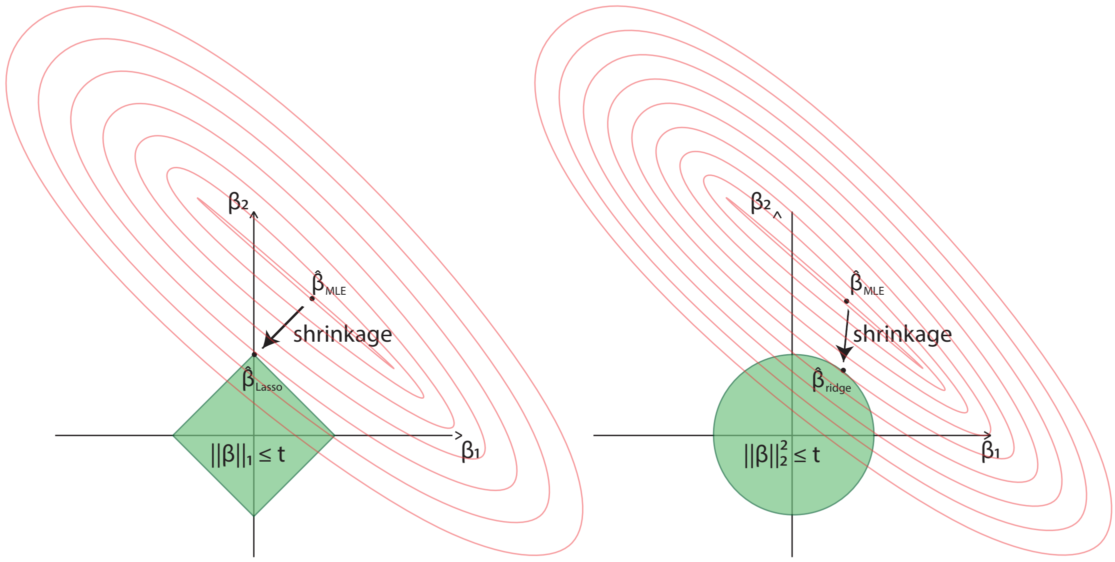


Figure 3.2: Geometry of Lasso and ridge penalties. The figure illustrates in red the contours of the negative log-likelihood and in green the shapes of the constraints implied by the penalty function in Lasso (left) or ridge regression (right). The maximum likelihood estimate is shrunken towards zero to lie within the constrained region.

combines shrinkage with feature selection. A good overview on the theory of the estimator and its application can be found in [26]. The estimate for a linear model is given by

$$\hat{\beta}_{\text{Lasso}} \in \arg\min_{\beta} ||y - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1, \tag{3.7}$$

or, equivalently,

$$\hat{\beta}_{\text{Lasso}} \in \arg\min_{\beta} ||y - \mathbf{X}\beta||_2^2 \quad \text{s.t. } ||\beta||_1 \leq s. \tag{3.8}$$

In contrast to penalties of the form $p(\beta) = \sum_j |\beta_j|^q$ with $q < 1$, this estimate with $q = 1$ can still be found by a convex optimization problem. However, other than for ridge regression, no closed-form solution is available and the distribution of the estimator is complex. Quadratic programming techniques can be employed to solve the convex minimization problem given by the Lasso. By now, efficient ways have been developed to compute the whole Lasso path over a sequence of penalization parameters λ, which in case of a linear model is a piece-wise linear function in λ: In 2004, Efron *et al.* proposed least angle regression [55], whose complexity is linear in the features in high-dimensional settings, i. e. $\mathcal{O}(np \min(n, p))$. A more generally applicable and often faster method is based on a cyclical coordinate-wise descent algorithm proposed in [65].

The Lasso is based on the assumption that the data can be well described by a sparse model, involving only a much smaller number of features than p. If s_0 denotes the number of non-zero coefficients in the model, the prediction error of the Lasso is of order $\frac{s_0 \log p}{n}$ (under the so-called compatibility condition on $\mathbf{X}$ [26]). In particular, compared the ordinary least squares estimator, whose prediction error is of order $\frac{p}{n}$, we are paying a price of $\log p$ for not knowing the truly active variables a priori.

3.4 Extensions of the Lasso

Thanks to its variable selection property, the Lasso method has gained popularity in many applications. However, some potential drawbacks in applications are the following: (i) In case of strongly correlated features the Lasso picks one essentially at random, which can mislead interpretation, (ii) Lasso chooses at most n predictors, (iii) with λ chosen by cross-validation, Lasso aims for best prediction performance and is not consistent for feature selection and (iv) Lasso does not respect group structures in its selection, which might be desirable, e. g. in the presence of categorical predictors.

To address these points, variations on the Lasso penalty have been introduced including elastic net [216], adaptive Lasso [215] or group Lasso [210]. Many other variants have been developed, for details and a comprehensive overview refer e. g. to [26, 66].

3.4.1 Elastic net

To address the first two points Zou & Hastie proposed the elastic net [216], which combines the ridge and the Lasso penalty to maintain the feature selection property of the Lasso, while improving performance in case of correlated features and allowing for a higher number of active predictors. The estimate is given by

$$\hat{\beta}_{\text{eNet}} \in \arg\min_{\beta} \ell(\beta) + \lambda \left(\frac{1-\alpha}{2} ||\beta||_2^2 + \alpha ||\beta||_1 \right), \tag{3.9}$$

where $\alpha \in [0, 1]$ provides the balance between the L_1- and L_2-penalty. In particular, $\alpha = 0$ yields the ridge estimate and $\alpha = 1$ the Lasso. Solution paths in λ can again be found using cyclical coordinate descent [65]. Using a 2-dimensional tuning approach for the two parameters λ and α this method was shown to perform as well as the better of Lasso and ridge approaches in applications to genomic data [198].

3.4.2 Adaptive Lasso

To obtain a version of Lasso that is consistent for feature selection, Zou proposed the adaptive Lasso [215], which consists of a two-step procedure. In a first step, an initial estimate b of the model coefficients is derived, using for example ordinary least squares, Lasso or ridge regression. In a second step, each feature is attributed a penalty factor given by the inverse of its absolute coefficient from the first step, i.e.

$$\hat{\beta}_{\text{ada}} \in \arg\min_{\beta} \ell(\beta) + \lambda \sum_j \frac{1}{|b_j|} |\beta_j|. \tag{3.10}$$

This gives preference to the features that have obtained a high absolute coefficient in the first step, while it penalizes features with a small initial estimate more.

3.4.3 Group Lasso and sparse group Lasso

To account for known groups in the data that should be selected or discarded jointly, Yuan & Lin introduced the group Lasso [210]. Denoting with $\mathcal{G}_g$ the set of indices of all features belonging to a group g, the group Lasso estimate is given by

$$\hat{\beta}_{\text{group}} \in \arg\min_{\beta} \ell(\beta) + \lambda \sum_g m_g ||\beta_{\mathcal{G}_g}||_2. \tag{3.11}$$

The factor m_g usually balances different group sizes and is typically chosen as $m_g = \sqrt{|\mathcal{G}_g|}$. Thus, the penalty is given by the sum of Euclidean norms of the coefficients within each group $\mathcal{G}_g$. As a consequence, the penalization acts like a ridge penalty within groups and like a Lasso penalty at the group level, thereby choosing either all variables within a group or none. For groups of size one, we would recover the ordinary Lasso.

In order to introduce sparsity also within the selected groups, Friedman *et al.* put an additional L_1-penalty on the coefficients, resulting in a method called sparse group Lasso [67]. Here, the estimate is given by

$$\hat{\beta}_{\text{sgl}} \in \arg\min_{\beta} \ell(\beta) + \lambda_1 \sum_g m_g ||\beta_{\mathcal{G}_g}||_2 + \lambda_2 ||\beta||_1. \tag{3.12}$$

3.4.4 Fused Lasso

The fused Lasso [185] is another way to incorporate structural constraints into the penalization. In addition to the usual L_1-penalty, it employs another L_1-penalty on the coefficient differences to enforce piece-wise constant values for neighbouring coefficients, i. e.

$$\hat{\beta}_{\text{fused}} \in \arg\min_{\beta} \ell(\beta) + \lambda_1 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| + \lambda_2 ||\beta||_1. \tag{3.13}$$

This can be useful in applications where a natural ordering of the features exists and we believe that nearby features will have a similar effect. For example, it has been applied for change point detection in copy number profiles from comparative genomic hybridization data [186]. Similar ideas have been applied if features are connected by a known network and smoothness along the graph is assumed [124].

3.5 A Bayesian view on penalized regression

In parallel to penalized regression methods, Bayesian regression methods have been developed to cope with high-dimensional data. These use prior distributions on the model coefficients as a form of regularization. This approach shows clear parallels to the frequentist methods above. For example, the Lasso or ridge estimate can be equivalently characterized as a maximum posterior estimate in a Bayesian model:

Proposition 1 *Given* $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^{p+1}$ *and assume* $y_i \sim \mathcal{N}(x_i^T\beta, \sigma^2)$ *with* σ^2 *known. If the priors on* β *are chosen as*

$$\beta_j \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{\lambda}\right) \quad or \quad \beta_j \overset{iid}{\sim} \text{Laplace}\left(0, \frac{\sigma^2}{\lambda}\right) \quad \forall\, j = 1, \dots, p,$$

the maximum-a-posterior estimate $\hat{\beta}_{MAP} \in \arg\max_\beta p(\beta|y)$ *corresponds to the ridge and Lasso estimate in a linear regression model with penalty parameter* λ*, respectively.*

Methods using a normal or Laplace prior are therefore also referred to as Bayesian ridge regression [95] or Bayesian Lasso [150]. In practice, the error variance it typically unknown and additional priors are used for σ^2, e. g. a conjugate inverse-Gamma prior.

In a Bayesian approach the parameters of the prior distribution determine the model complexity, corresponding to the penalty parameter λ in penalized regression. As examples, let's consider the case of a normal or Laplace prior on β, i. e. $\beta \sim \mathcal{N}(0, \lambda^{-1})$ or $\beta \sim \text{Laplace}(0, \lambda^{-1})$. Here, the variance or scale parameter λ^{-1} determines the width of the prior distribution, as already illustrated in the context of unsupervised methods in Figure 2.1. For a high value of λ^{-1} the prior distribution is very wide and the regularization is weak, resulting in higher model complexity. On the other hand, a very low value of λ^{-1} leads to a very narrow prior distribution and hence a strong regularization. By employing hyper-priors on the parameter λ, the Bayesian setting opens up alternative ways to choose a suitable model complexity in an adaptive manner, without the need to refer to cross-validation. A common choice is a conjugate Gamma prior on λ (in Bayesian ridge regression) or λ^2 (in the Bayesian Lasso).

Apart from the priors used in Bayesian ridge regression or Bayesian Lasso, many alternative choices have been suggested. Examples include the spike-and-slab prior [143], the automatic relevance determination (ARD) prior [131] or the Horseshoe prior [33]. Furthermore, structural information can naturally be incorporated into the priors, in a similar spirit to the group or fused Lasso. Examples include temporal or spatial relationships and group dependencies (e.g. [5, 56, 88, 159, 163, 207, 208] and references therein), which are incorporated via multivariate or non-exchangeable priors on the coefficients.

Although the correspondence in Proposition 1 is based on the maximum-a-posterior estimate, Bayesian inference is usually based on the whole posterior distribution and point estimates are often obtained by the posterior mean. In simple examples, the posterior can be derived analytically, but typically one needs to refer to sampling-based or deterministic approximations such as Markov chain Monte Carlo (MCMC) methods, Laplace's method or Integrated Nested Laplace Approximations (INLA) [165], Variational Bayes [21] or Expectation Propagation [142]. In addition to point estimates, the posterior distribution can yield uncertainty measures, predictive distributions and feature inclusion probabilities.

3.6 Penalized regression with heterogeneous data modalities

Up to here, we have mainly considered the general regression setting given a single design matrix. In our original problem, however, the independent variable X had components from very different data modalities. While the penalization accounts for the high-dimensionality of X, we so far have not explicitly addressed its heterogeneity and simply concatenated all features. This enabled us to investigate the relationships of individual features to the response in a joint model including features from all data modalities. However, concatenation makes it harder to find relationships between whole data modalities and the response. For example, often we are interested in which modalities are most important for an accurate prediction of Y, which might not be obvious from the concatenated regression model. In addition, concatenation ignores all available information on the source of the features in X, which could help to make the regression more powerful, e. g., if two data modalities are of very different quality. For these reasons, it can be desirable to take the structure of X into account. Here, we outline some possible avenues based on the methods introduced above.

Using the group Lasso We have already seen one way to account for the heterogeneity with the group Lasso in Section 3.4. By taking data modalities as groups, we can select or discard entire modalities from the regression problem. As typically we do not want to include a complete modality as a whole, the sparse group Lasso can provide a means of selection both at the level of data modalities and individual features. While useful in many applications, drawbacks of this approach include that the amount of regularization towards sparsity is the same for all groups and that in many applications a hard in- or exclusion of modalities might be too drastic, calling for a softer weighting scheme of different modalities.

Combining modality-wise models Another way of dealing with multiple data modalities consists in fitting separate models to each single modality in a first step and combining them in a second step. This can provide a softer means of down- and up-weighting modalities in the joint regression model. For example, the first step could be used to select a set of features from each data modality, e. g. using Lasso, and the second step could consider a joint regression model with all selected features, as e. g. in [213]. Other approaches are based on model averaging in the second step. A drawback of such approaches is that the selection or model obtained from the first step can be misleading or sub-optimal, because it cannot take into account the variables from other modalities. Collaborative regression [78] shares a similar idea to two-step approaches by making individual predictions for each modality. Other than two-step approaches, this methods simultaneously minimizes the sum of modality-wise prediction errors and the discrepancies between modality-wise predictions. However, this was seen to be unsuited for prediction tasks and mainly focused on finding common patterns [78].

Based on latent components Instead of directly using the features as predictors, data modalities can be summarized into latent factors, which can then be used in the regression model. Many approaches in this context are based on two separate steps. First, an unsupervised dimensionality reduction is employed as discussed in Chapter 2, e.g. using PCA per modality or multi-table methods. In a second step, the inferred components are then jointly used in a regression model. In practice, such approaches can help to reduce the dimensionality and noise in the joint model and - if multi-table methods are applied - borrow strength across modalities. However, as the selection of latent factors is de-coupled from the response variable, this can be underpowered as the major axes of (co-)variation might not correspond to the most relevant predictors for the response. Therefore, approaches have been developed that simultaneously construct the latent factors explaining variation in each modality and optimize them for explaining the response Y, e. g. based on partial least squares (PLS) [171] or sparse factor models [179]. This can provide a compromise between discovery of biologically relevant connections between data modalities and the prediction or classification performance. However, such approaches often seem to be less powerful compared to concatenation based regression approaches in terms of prediction performance [171].

Differential penalization using penalty factors Differential penalization provides an alternative way to account for heterogeneity of the features. This can provide a softer weighting scheme of different feature groups compared to the group Lasso. In the most general form, we can allow for a different penalty factor $s_j \in \mathbb{R}_{\geq 0}$ per feature, i. e.

$$\hat{\beta} \in \arg\min_{\beta} \ell(\beta) + \lambda \sum_{j=1}^{p} s_j q(\beta_j), \tag{3.14}$$

e. g. with $q(\beta_j) = |\beta_j|$ or $q(\beta_j) = \beta_j^2$. By using a common $s^{(l)} = s_j$ for all features j from a modality l, we jointly up- or down-weight features from the same data modality. While this provides a very general approach, the main problem here consists in finding a good set of penalty factors adaptively. For example, cross-validation soon becomes prohibitive as it would require re-fitting the model over a grid exponential in the number of distinct penalty factors. Therefore, alternative approaches are required, and we will come back to this problem in Chapter 5.

Using Bayesian approaches Taking a Bayesian view on penalized regression, feature heterogeneity can be incorporated via the prior on the model coefficients. For example, the introduction of differential penalization in Equation (3.14) corresponds to non-exchangeable coefficients with univariate priors that have different scale parameters. As mentioned in Section 3.5, multivariate or non-exchangeable priors can be employed to account for known structure in the design matrix (e. g. [5, 56, 88, 159, 163, 207, 208]), including Bayesian versions of the group Lasso. This provides a flexible way of taking heterogeneities of the features into account. A major challenge that is critical for the applicability of Bayesian methods are scalable inference schemes.

3.7 A glimpse at our contribution

In Chapter 5 we will follow up on the idea of using differential penalization of features to account for feature heterogeneity in penalized regression. In particular, we want to use additional information on the features in order to guide the relative strength of penalization across different feature groups in a scalable and adaptive manner. Such additional information is available in most applications: While the setting of different data modalities is one motivating example, other types of annotations could provide useful information on a feature. Examples include quality metrics, insights from prior studies and the features' empirical variance or frequency. We developed a method to incorporate such information by adapting the relative strength of penalization across features in a data-driven manner and demonstrated performance gains on simulated data. Furthermore, we investigated the use of important covariates in applications to data from genome biology, such as the omics or tissue type.

CHAPTER 4

Multi-Omics Factor Analysis - a framework for unsupervised integration of multi-omics data sets

This chapter is a slightly modified version of the peer-reviewed article published under [8]. The original paper is based on joint work with Ricard Argelaguet, who is shared first-author on this paper. The manuscript was jointly written by Ricard Argelaguet, Florian Buettner, Oliver Stegle, Wolfgang Huber and me. All the analyses were carried out by Ricard and me with focus on the CLL study on my side and the simulation and single cell study on Ricard's side but contributions vice-versa. The method was implemented by Ricard, Damien Arnold and me, where I in particular contributed updates for non-Gaussian likelihoods and methods for down-stream analyses.

Multi-omics studies promise the improved characterization of biological processes across molecular layers. However, methods for the unsupervised integration of the resulting heterogeneous datasets are lacking. We present Multi-Omics Factor Analysis (MOFA), a computational method for discovering the principal sources of variation in multi-omics datasets. MOFA infers a set of (hidden) factors that capture biological and technical sources of variability. It disentangles axes of heterogeneity that are shared across multiple modalities and those specific to individual data modalities. The learnt factors enable a variety of downstream analyses, including identification of sample subgroups, data imputation, and the detection of outlier samples. We applied MOFA to a cohort of 200 patient samples of chronic lymphocytic leukaemia, profiled for somatic mutations, RNA expression, DNA methylation and ex-vivo drug responses. MOFA identified major dimensions of disease heterogeneity, including immunoglobulin heavy chain variable region status, trisomy of chromosome 12 and previously underappreciated drivers, such as response to oxidative stress. In a second application, we used MOFA to analyse single-cell multi-omics data, identifying coordinated transcriptional and epigenetic changes along cell differentiation.

4.1 Introduction

Technological advances increasingly enable multiple biological layers to be probed in parallel, ranging from genome, epigenome, transcriptome, proteome and metabolome to phenome profiling [85]. Integrative analyses that use information across these data modalities promise to deliver more comprehensive insights into the biological systems under study. Motivated by this, multi-omics profiling is increasingly applied across biological domains, including cancer biology [29, 77, 99, 139], regulatory genomics [34], microbiology [110] or host-pathogen biology [173]. Most recent technological advances have also enabled performing multi-omics analyses at the single cell level [7, 38, 40, 81, 130]. A common aim of such applications is to characterize heterogeneity between samples, as manifested in one or several of the omics data types [162]. Multi-omics profiling is particularly appealing if the relevant axes of variation are not known a priori, and hence may be missed by studies that consider a single data modality or targeted approaches.

A basic strategy for the integration of omics data is testing for marginal associations between different data modalities. A prominent example is molecular quantitative trait locus mapping, where large numbers of association tests are performed between individual genetic variants and gene expression levels [79] or epigenetic marks [34]. While eminently useful for variant annotation, such association studies are inherently local and do not provide a coherent global map of the molecular differences between samples. A second strategy is the use of kernel- or graph-based methods to combine different data types into a common similarity network between samples [116, 199]; however, it is difficult to pinpoint the molecular determinants of the resulting graph structure. Related to this, there exist generalizations of other clustering methods to reconstruct discrete groups of samples based on multiple data modalities [144, 170].

A key challenge that is not sufficiently addressed by these approaches is interpretability. In particular, it would be desirable to reconstruct the underlying factors that drive the observed variation across samples. These could be continuous gradients, discrete clusters, or combinations thereof. Such factors would also help in establishing or explaining associations with external data such as phenotypes or clinical covariates. Although factor models that aim to address this have previously been proposed, e.g., [137, 138, 171, 181], these methods either lack sparsity, which can reduce interpretability, or they require a substantial number of parameters to be determined using computationally demanding cross-validation or post hoc. Further challenges faced by existing methods are computational scalability to larger datasets, handling of missing values and non-Gaussian data modalities, such as binary readouts or count-based traits.

4.2 Results

We present MOFA, a statistical method for integrating multiple modalities of omics data in an unsupervised fashion. Intuitively, MOFA can be viewed as a versatile and statistically rigorous generalization of PCA to multi-omics data. Given several data matrices with measurements of multiple omics data types on the same or on partially overlapping sets of samples, MOFA infers an interpretable low-dimensional data representation in terms of (hidden) factors (Figure 4.1a). These learnt factors capture major sources of variation across data modalities, thus facilitating the identification of continuous molecular gradients or discrete subgroups of samples. The inferred factor loadings can be sparse, thereby facilitating the linkage between the factors and the most relevant molecular features. Importantly, MOFA disentangles to what extent each factor is unique to a single data modality

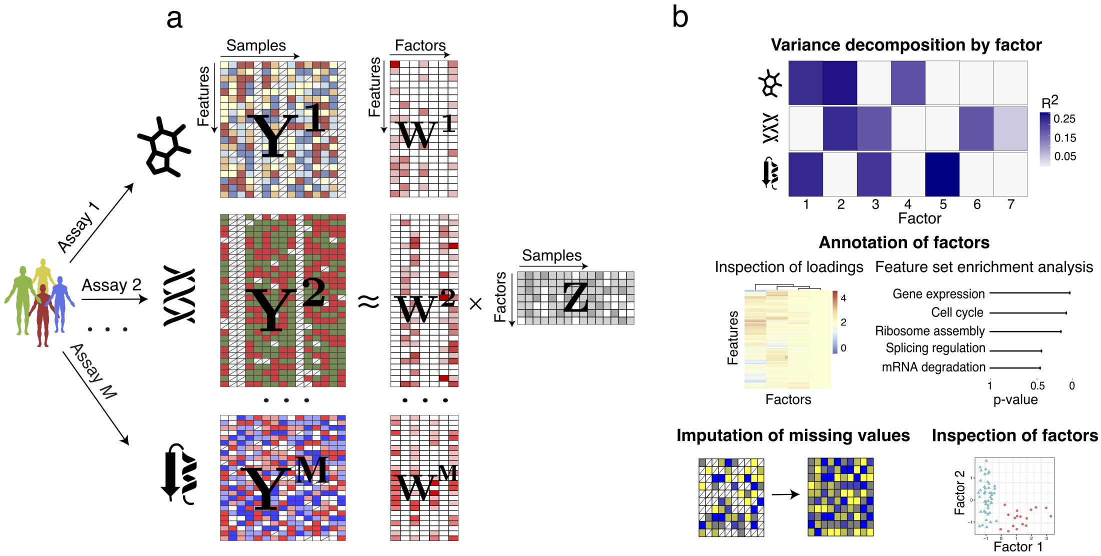


Figure 4.1: Multi-Omics Factor Analysis: model overview and downstream analyses. (a) Model overview: MOFA takes an arbitrary number of M data matrices as input ($\mathbf{Y}^1, \ldots, \mathbf{Y}^M$), one or more from each data modality, with co-occurrent samples but features that are in general unrelated and that differ in numbers. MOFA decomposes these matrices into a matrix of factors, $\mathbf{Z}$, with samples in columns and M weight matrices, one for each data modality (loadings $\mathbf{W}^1, \ldots, \mathbf{W}^M$). White cells in the weight matrices correspond to zeros, i.e. inactive features, whereas the cross symbol in the data matrices denote missing values. **(b)** The fitted MOFA model can be queried for different downstream analyses, including (i) variance decomposition, assessing the proportion of variance explained by each factor in each data modality, (ii) semi-automated factor annotation based on the inspection of loadings and gene set enrichment analysis, (iii) visualization of the samples in the factor space and (iv) imputation of missing values, including missing assays.

or is manifested in multiple modalities (Figure 4.1b), thereby revealing shared axes of variation between the different omics layers. Once trained, the model output can be used for a range of downstream analyses, including visualisation, clustering and classification of samples in the low-dimensional space(s) spanned by the factors, as well as the identification of outlier samples and the imputation of missing values (Figure 4.1b).

Technically, MOFA builds upon the statistical framework of group factor analysis [27, 109, 113, 122, 197, 214], which we have adapted to the requirements of multi-omics studies by combining: (i) fast inference based on a variational approximation, (ii) inference of sparse solutions facilitating interpretation, (iii) efficient handling of missing values, and (iv) flexible combination of different likelihood models for each data modality, which enables integrating diverse data types such as binary-, count- and continuous-valued data. The relationship of MOFA to previous approaches is discussed in Methods Section 4.4.5 and Appendix Table 4.A.3.

MOFA is implemented as well-documented open-source software and comes with tutorials and example workflows for different application domains (Methods Section 4.4.9). Taken together, these functionalities provide a powerful and versatile tool for disentangling sources of variation in multi-omics studies.

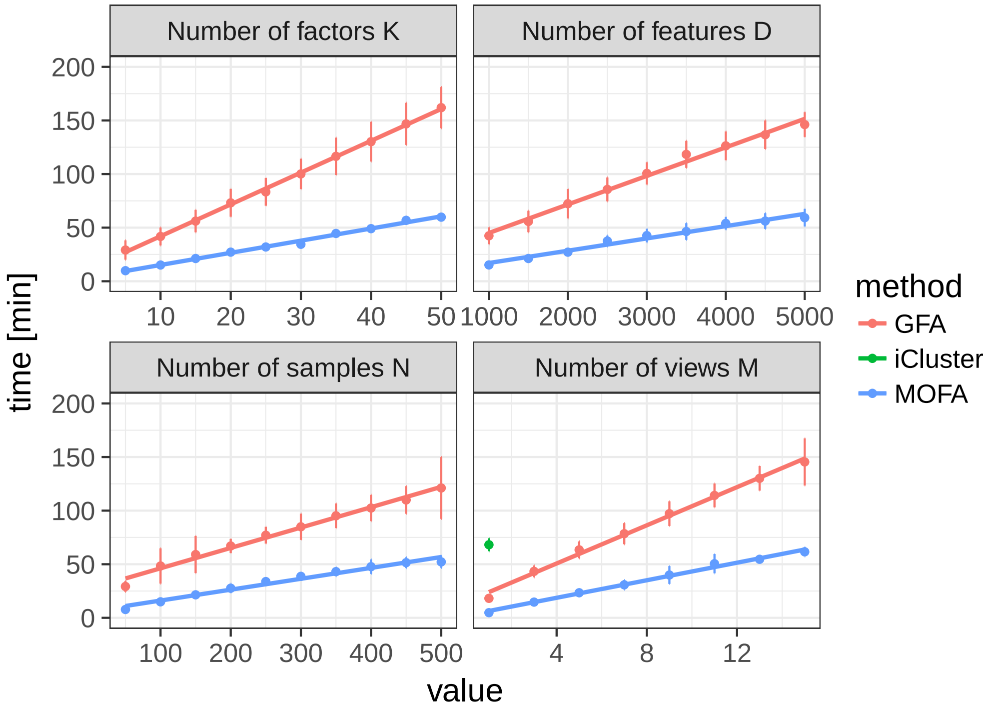


Figure 4.2: Scalability of MOFA, GFA and iCluster. Time required for model training for GFA (red), MOFA (blue) and iCluster (green) as a function of number of factors K, number of features D, number of samples N and number of views M. Baseline parameters were $M = 3$, $K = 10$, $D = 1000$ and $N = 100$ and 5% missing values. Shown are average time across 10 trials, error bars denote standard deviation. iCluster is only shown for the lowest M value as all other settings required on average more than 200 minutes for training.

4.2.1 Model validation and comparison on simulated data

First, to validate MOFA, we simulated data from its generative model, varying the number of views, the likelihood models, the number of latent factors and other parameters (Methods Section 4.4.6, Appendix Table 4.A.1). We found that MOFA was able to accurately reconstruct the latent dimension, except in settings with large numbers of factors or high proportions of missing values (Appendix Figure 4.A.1). We also found that models that account for non-Gaussian observations improved the fit when simulating binary or count data (Appendix Figures 4.A.2 and 4.A.3).

We also compared MOFA to two previously reported latent variable models for multi-omics integration: GFA [122] and iCluster [144]. Over a range of simulations, we observed that GFA and iCluster tended to infer redundant factors (Appendix Figure 4.A.4) and were less accurate in recovering patterns of shared factor activity across views (Appendix Figure 4.A.5). MOFA is also computationally more efficient than these existing methods (Figure 4.2). For example, the training on the chronic lymphocytic leukaemia (CLL) data, which we consider next, required 45 minutes using MOFA vs. 34 hours with GFA and 5-6 days with iCluster.

4.2.2 Application to chronic lymphocytic leukaemia

We applied MOFA to a study of chronic lymphocytic leukaemia (CLL), which combined ex-vivo drug response measurements with somatic mutation status, transcriptome profiling and DNA methylation assays [47] (Figure 4.3a). Notably, nearly 40% of the 200 samples were profiled with some but not all omics types; such a missing value scenario is not uncommon in large cohort studies, and MOFA is designed to cope with it (Appendix Section 4.A.4; Appendix Figure 4.A.1). MOFA was configured to combine different likelihood models in order to accommodate the combination of continuous and discrete data types in this study.

MOFA identified 10 factors (minimum explained variance 2% in at least one data type; Methods Section 4.4.7). These were robust to algorithm initialisation as well as subsampling of the data (Appendix Figures 4.A.6, 4.A.7). The factors were largely orthogonal, capturing independent sources of variation (Appendix Figure 4.A.6). Among these, Factors 1 and 2 were active in most assays, indicating broad roles in multiple molecular layers (Figure 4.3b). In contrast, other factors such as Factor 3 or Factor 5 were specific to two data modalities, and Factor 4 was active in a single data modality only. Cumulatively, the 10 factors explained 41% of variation in the drug response data, 38% in the mRNA data, 24% in the DNA methylation data and 24% in the mutation data (Figure 4.3c).

We also trained MOFA when excluding individual data modalities to probe their redundancy, finding that factors that were active in multiple data modalities could still be recovered, while the identification of others was dependent on a specific data type (Appendix Figure 4.A.8). In comparison to GFA [122] and iCluster [144], MOFA was more consistent in identifying factors across multiple model instances (Appendix Figure 4.A.9).

MOFA identifies important clinical markers in CLL and reveals an underappreciated axis of variation attributed to oxidative stress.

As part of the downstream pipeline, MOFA provides different strategies to use the loadings of the features on each factor to identify their aetiology (Figure 4.1b). For example, based on the top weights in the mutation data, Factor 1 was aligned with the somatic mutation status of the immunoglobulin heavy chain variable region gene (IGHV), while Factor 2 aligned with trisomy of chromosome 12 (Figure 4.3d,e). Thus, MOFA correctly identified two major axes of molecular disease heterogeneity and aligned them with two of the most important clinical markers in CLL [58, 211] (Figure 4.3d,e).

IGHV status, the marker associated with Factor 1, is a surrogate of the differentiation state of the tumour's cell of origin and the level of activation of the B-cell receptor. While in clinical practice this axis of variation is generally considered binary [58], our results indicate a more complex substructure (Figure 4.4a, Appendix Figure 4.A.10). At the current resolution, this factor was consistent with three subgroup models such as proposed by [148, 158] (Appendix Figure 4.A.11), although there is suggestive evidence for an underlying continuum. MOFA connected this factor to multiple molecular layers (Appendix Figures 4.A.12, 4.A.13), including changes in the expression of genes previously linked to IGHV status [51, 133, 146, 155, 190, 195] (Figure 4.4b,c) and with drugs that target kinases in or downstream of the B-cell receptor pathway (Figure 4.4d,e).

Despite their clinical importance, the IGHV and the trisomy 12 factors accounted for less than 20% of the variance explained by MOFA, suggesting the existence of other sources of heterogeneity. One example is Factor 5, which was active in the mRNA and drug response data. Analysis of the weights in the mRNA revealed that this factor tagged a set of genes enriched for oxidative stress and senescence pathways (Figure 4.3f, Figure 4.5), with the

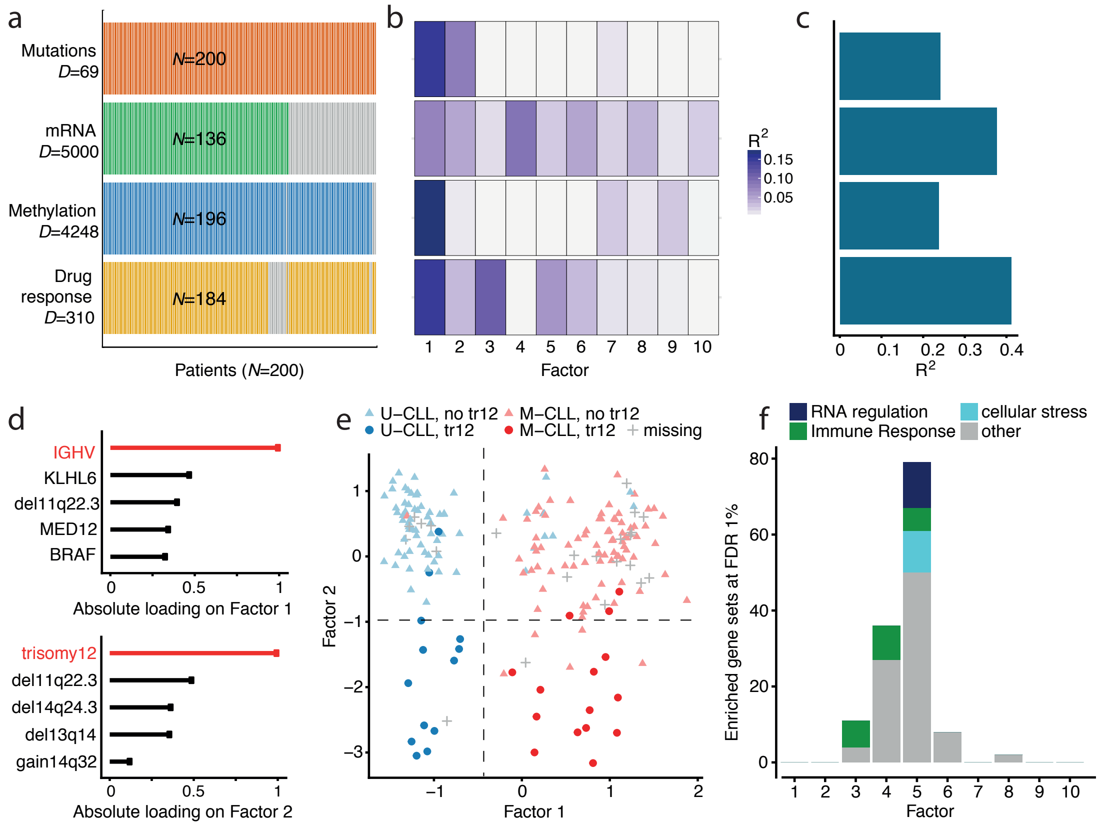


Figure 4.3: Application of MOFA to a study of chronic lymphocytic leukaemia. **(a)** Study overview and data types. Data modalities are shown in different rows (D = number of features) and samples in columns, with missing samples shown using grey bars. **(b)** Proportion of total variance explained (R^2) by individual factors for each assay and **(c)** cumulative proportion of total variance explained. **(d)** Absolute loadings of the top features of Factors 1 and 2 in the mutations data. **(e)** Visualisation of samples using Factors 1 and 2. The colours denote the IGHV status of the tumours; symbol shape and colour tone indicate chromosome 12 trisomy status. **(f)** Number of enriched Reactome gene sets per factor based on the gene expression data (FDR $<$ 1%). The colours denote categories of related pathways defined as in Appendix Table 4.A.2.

top weights corresponding to heat shock proteins (HSPs) (Figure 4.5b,c), genes that are essential for protein folding and are up-regulated upon stress conditions [2, 175]. Although genes in HSP pathways are up-regulated in some cancers and have known roles in tumour cell survival [189], thus far this gene family has received little attention in the context of CLL. Consistent with this annotation based on the mRNA data, we observed that the drugs with the strongest weights on Factor 5 were associated with response to oxidative stress, such as target reactive oxygen species (ROS), DNA damage response and apoptosis (Figure 4.5d,e).

Factor 4 captured 9% of variation in the mRNA data, and gene set enrichment analysis on the mRNA loadings suggested aetiologies related to immune response pathways and T-cell receptor signalling (Figure 4.3f), likely due to differences in cell type composition between samples: While the samples are comprised mainly of B-cells, Factor 4 revealed a possible contamination with other cell types such as T-cells and monocytes (Appendix Figure 4.A.14). Factor 3 explained 11% of variation in the drug response data, capturing heterogeneity in the samples' general level of drug sensitivity [75] (Appendix Figure 4.A.15).

MOFA identifies outlier samples and accurately imputes missing values

Next, we explored the relationship between inferred factors and clinical annotations, which can be missing, mis-annotated or inaccurate, since they are frequently based on single markers or imperfect surrogates [202]. Since IGHV status is the major biomarker impacting on clinical care, we assessed the consistency between the inferred continuous Factor 1 and this binary marker. For 176 out of 200 patients, the MOFA factor was in agreement with the clinical IGHV status, and MOFA further allowed for classifying 12 patients that lacked clinically measured IGHV status (Figure 4.6a,b). Interestingly, MOFA assigned 12 patients to a different group than suggested by their clinical IGHV label. Upon inspection of the underlying molecular data, nine of these cases showed intermediate molecular signatures, suggesting that they are borderline cases that are not well captured by the binary classification; the remaining three cases were clearly discordant (Figure 4.6c,d). Additional independent drug response assays as well as whole exome sequencing data confirmed that these cases are outliers within their IGHV group (Figure 4.6e,f).

As incomplete data is a common problem in studies that combine multiple high-throughput assays, we assessed the ability of MOFA to fill in missing values within assays as well as when entire data modalities are missing for some of the samples. For both imputation tasks, MOFA yielded more accurate predictions than other established imputation strategies, including imputation by feature-wise mean, SoftImpute [134] and a k-nearest neighbour (kNN) method [191] (Figure 4.7, Appendix Figure 4.A.16), and MOFA was more robust than GFA, especially in the case of missing assays (Appendix Figure 4.A.17).

Latent factors inferred by MOFA are predictive of clinical outcomes

Finally, we explored the utility of the latent factors inferred by MOFA as predictors in models of clinical outcomes. Three of the 10 factors identified by MOFA were significantly associated with time to next treatment (Cox regression, Methods Section 4.4.7, FDR $<$ 1%, Figure 4.8a,b): Factor 1, related to the B cell of origin, and two Factors, 7 and 8, associated with chemo-immunotherapy treatment prior to sample collection ($P < 0.01$, t-test). In particular, Factor 7 captured del17p and TP53 mutations as well as differences in methylation patterns of oncogenes [64, 73] (Appendix Figure 4.A.18), while Factor 8 was associated with WNT signalling (Appendix Figure 4.A.19).

We also assessed the prediction performance when combining the 10 MOFA factors in a

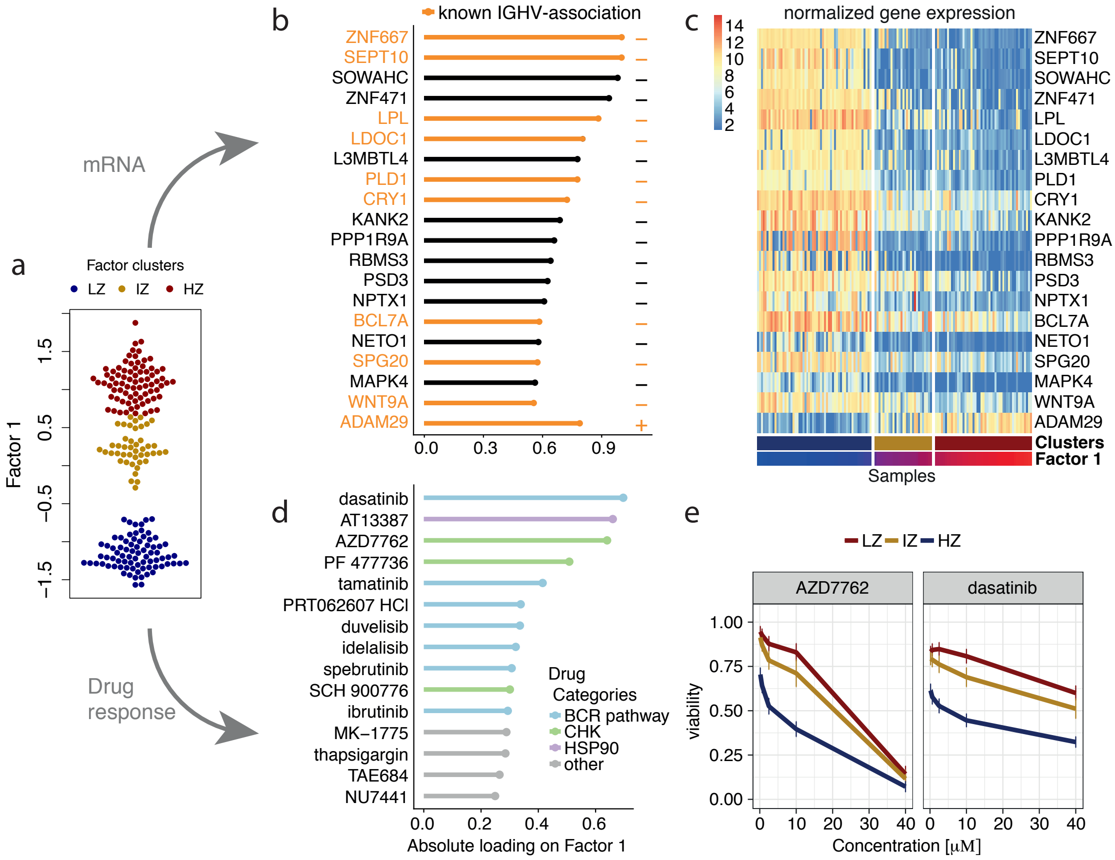


Figure 4.4: Characterization of Factor 1 (associated to the differentiation state of the cell of origin). (a) Beeswarm plot with Factor 1 values for each sample with colours corresponding to three clusters found by 3-means clustering with low factor values (LZ), intermediate factor values (IZ) and high factor values (HZ). **(b)** Absolute loadings for the genes with the largest absolute weights in the mRNA data. Plus or minus symbols on the right indicate the sign of the loading. Genes highlighted in orange were previously described as prognostic markers in CLL and associated with IGHV status [51, 133, 146, 155, 190, 195]. **(c)** Heatmap of gene expression values for genes with the largest absolute weights as in **(b)**. **(d)** Absolute loadings of the drugs with the largest absolute weights, annotated by target category. **(e)** Drug response curves for two of the drugs, stratified by the clusters as in **(a)**.

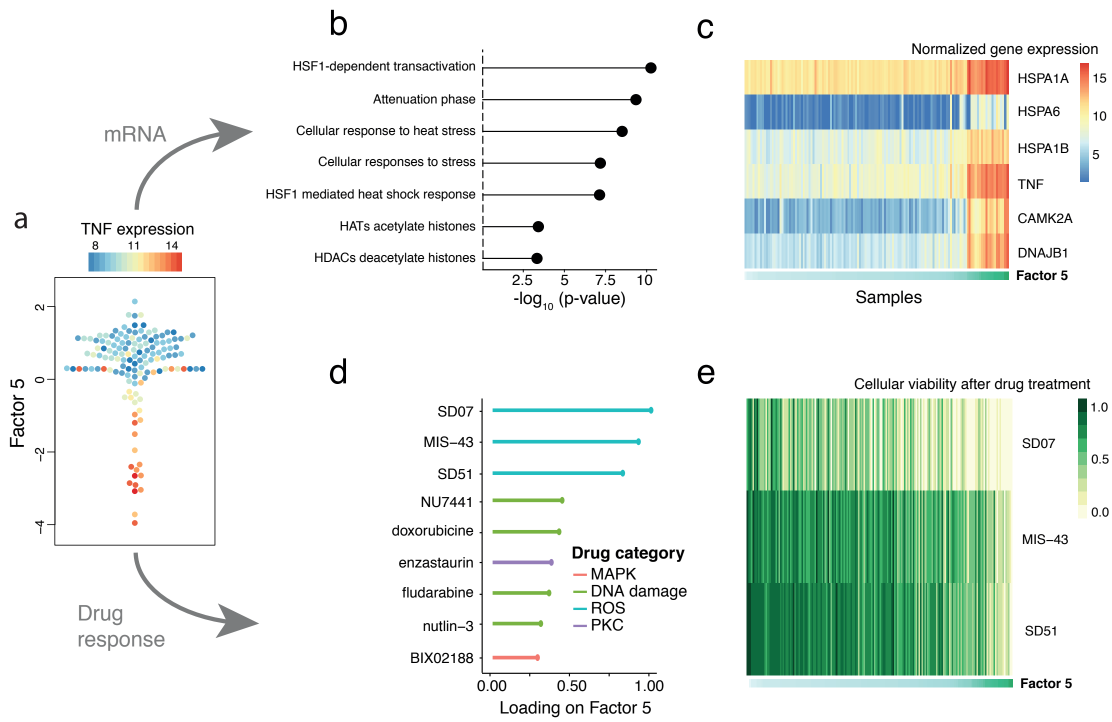


Figure 4.5: Characterization of Factor 5 (oxidative stress response factor) in the CLL data. (a) Beeswarm plot of Factor 5. Colours denote the expression of TNF, an inflammatory stress marker. **(b)** Gene set enrichment for the top Reactome pathways in the mRNA data (t-test, Methods Section 4.4.7). **(c)** Heatmap of gene expression values for the six genes with largest absolute loadings. Samples are ordered by their factor values. **(d)** Absolute loadings for the top drugs with largest absolute loading in the drug response data, annotated by target category. **(e)** Heatmap of drug response values for the top three drugs with largest absolute loading. Samples are ordered by their factor values.

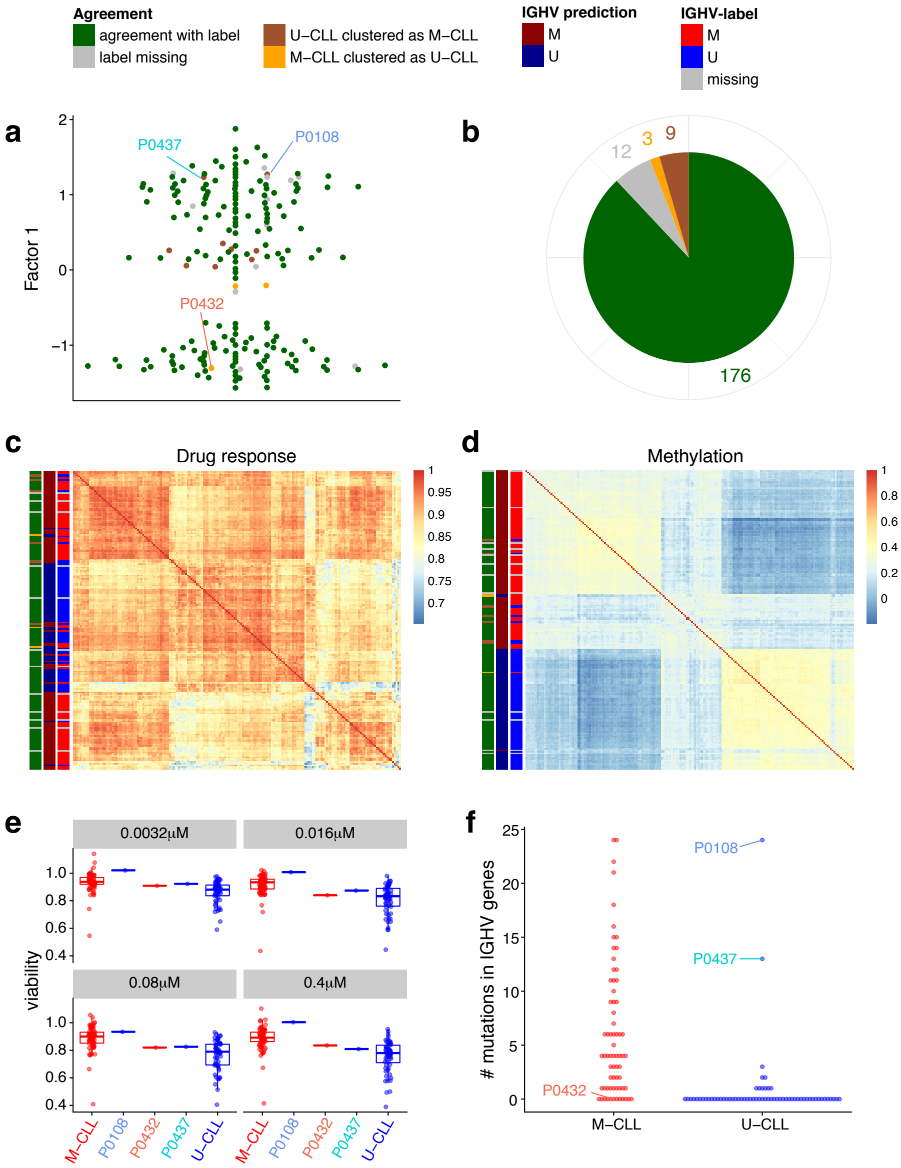


Figure 4.6: Prediction of IGHV status based on Factor 1 in the CLL data and validation of outlier cases on independent assays. **(a)** Beeswarm plot of Factor 1 with colours denoting agreement between predicted and clinical labels as in **(b)**. **(b)** Pie chart showing total numbers for agreement of imputed labels with clinical label. **(c)** Sample-to-sample correlation matrix based on drug response data. **(d)** Sample-to-sample correlation matrix based on methylation data. **(e)** Drug response to ONO-4509 (not included in the training data): Boxplots for the viability values in response to ONO-4509. The three outlier samples are shown in the middle, on the left and right the viabilities of the other M-CLL and U-CLL samples are shown, respectively. The panels show different drug concentrations tested. **(f)** Whole exome sequencing data on IGHV genes (not included in the training data): the number of mutations found on IGHV genes using whole exome sequencing is shown on the y-axis, separately for U-CLL and M-CLL samples. The three outlier samples are labelled.

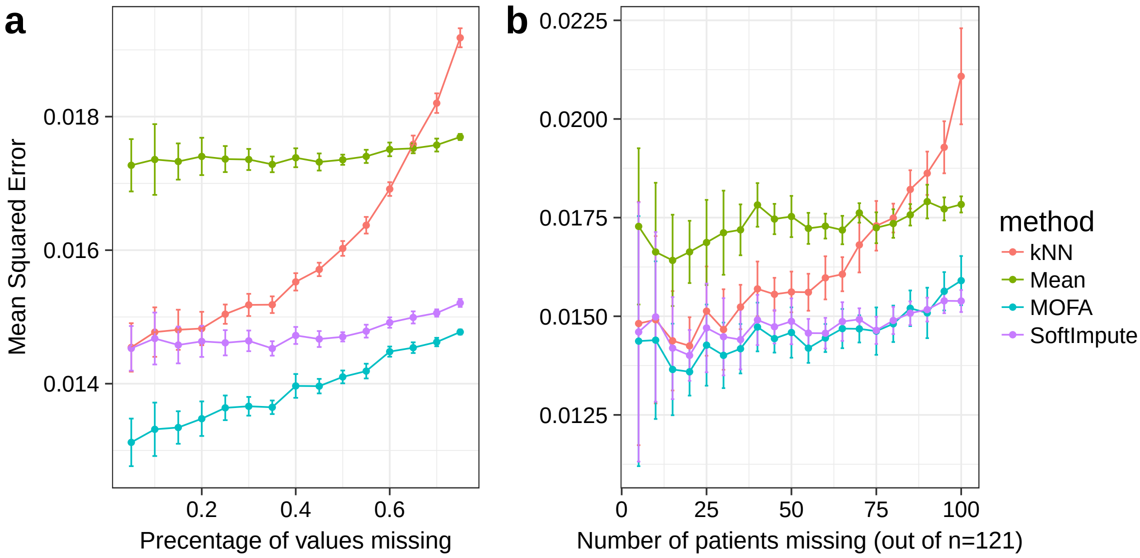


Figure 4.7: Imputation of missing values in the drug response assay of the CLL data. Considered were MOFA, SoftImpute, imputation by feature-wise mean (Mean) and k-nearest neighbour (kNN). Shown are averages of the mean squared error (MSE) across 15 imputation experiments for increasing fractions of missing data, considering **(a)** values missing at random and **(b)** entire assay missing for samples at random. Error bars denote plus or minus two standard error.

multivariate Cox regression model. Notably, this model yielded higher prediction accuracy than models using components derived from conventional PCA (Figure 4.8c), individual molecular features (Appendix Figure 4.A.20) or MOFA factors derived from only a subset of the available data modalities (Appendix Figure 4.A.8b,d) (assessed using cross-validation, Methods Section 4.4.7). The predictive value of MOFA factors was similar to clinical co-variates (such as lymphocyte doubling time) that are used to guide treatment decisions (Appendix Figure 4.A.21).

4.2.3 Application to a single-cell multi-omics study

As multi-omics approaches are also beginning to emerge in single cell biology [7, 38, 40, 81, 130], we investigated the potential of MOFA to disentangle the heterogeneity observed in such studies. We applied MOFA to a data set of 87 mouse embryonic stem cells (mESCs), comprising of 16 cells cultured in ‘2i’ media, which induces a naive pluripotency state, and 71 serum-grown cells, which commits cells to a primed pluripotency state poised for cellular differentiation [7]. All cells were profiled using single-cell methylation and transcriptome sequencing, which provides parallel information of these two molecular layers (Figure 4.9a). We applied MOFA to disentangle the observed heterogeneity in the transcriptome and the CpG methylation at three different genomic contexts: promoters, CpG islands and enhancers.

MOFA identified three major factors driving cell-cell heterogeneity (minimum explained variance of 2%, Methods Section 4.4.8): While Factor 1 is shared across all data modalities (7% variance explained in the RNA data and between 53% and 72% in the methylation data sets), Factors 2 and 3 are active primarily in the RNA data (Figure 4.9b,c). Gene loadings revealed that Factor 1 captured the cells’ transition from naive to primed pluripotent states, pinpointing pluripotency markers such as Rex1/Zpf42, Tbx3, Fbxo15 and Essrb

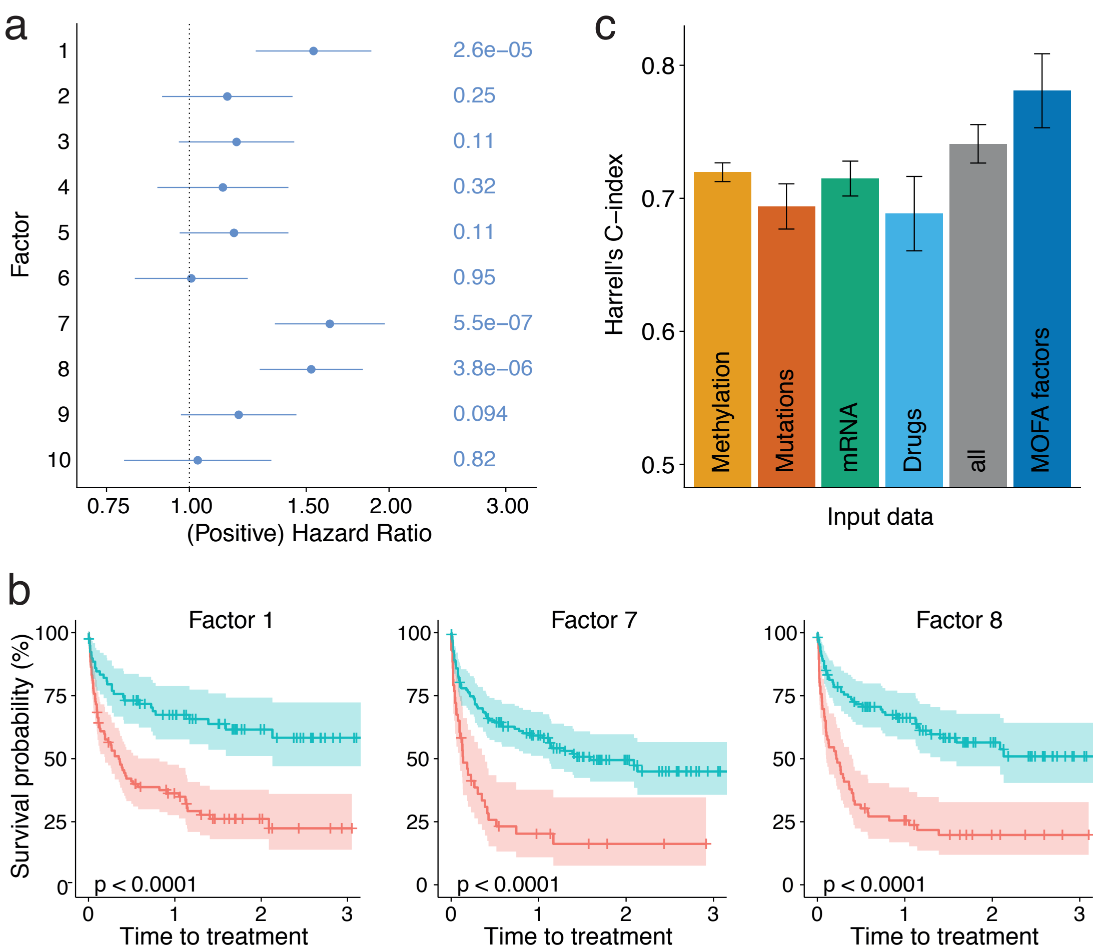


Figure 4.8: Relationship between clinical data and latent factors. (a) Association of MOFA factors to time to next treatment using univariate Cox models. Error bars denote 95% confidence intervals. Numbers on the right denote p-values for each predictor. **(b)** Kaplan-Meier plots measuring time to next treatment for the individual MOFA factors. The cut-points on each factor were chosen using maximally selected rank statistics [94], and p-values were calculated using a log-rank test on the resulting groups. **(c)** Prediction accuracy of time to treatment using multivariate Cox regression trained using the 10 factors derived using MOFA, as well using the first 10 components obtained from PCA applied to the corresponding single data modalities and the full dataset (assessed on hold-out data). Shown are average values of Harrell's C-index from 5-fold cross-validation. Error bars denote standard error of the mean.

[145] (Figure 4.9d and Figure 4.10a). MOFA connected these transcriptomic changes to coordinated changes of the genome-wide DNA methylation rate across all genomic contexts (Figure 4.10b), as previously described both in vitro [7] and in vivo [9]. Factor 2 captured a second axis of differentiation from the primed pluripotency state to a differentiated state with highest RNA loadings for known differentiation markers such as keratins and annexins [70] (Figure 4.9d and Figure 4.10c). Finally, Factor 3 captured the cellular detection rate, a known technical covariate associated with cell quality and mRNA content [63] (Appendix Figure 4.A.22).

Jointly, Factors 1 and 2 captured the entire differentiation trajectory from naive pluripotent cells via primed pluripotent cells to differentiated cells (Figure 4.9e), illustrating the importance of learning continuous latent factors rather than discrete sample assignments. Multi-omics clustering algorithms such as similarity network fusion (SNF) [199] or iCluster [144, 170] were only capable of distinguishing cellular subpopulations, but not of recovering continuous processes such as cell differentiation (Appendix Figure 4.A.23).

4.3 Discussion

Multi-Omics Factor Analysis (MOFA) is an unsupervised method for decomposing the sources of heterogeneity in multi-omics data sets. We applied MOFA to high-dimensional and incomplete multi-omics profiles collected from patient-derived tumour samples and to a single-cell study of mESCs.

First, in the CLL study, we demonstrated that our method is able to identify major drivers of variation in a clinically and biologically heterogeneous disease. Most notably, our model identified previously known clinical markers as well as novel putative molecular drivers of heterogeneity, some of which were predictive of clinical outcome. Additionally, since MOFA factors capture variations of multiple features and data modalities, inferred factors can help to mitigate assay noise, thereby increasing the sensitivity for identifying molecular signatures compared to using individual features or assays. Our results also demonstrate that MOFA can leverage information from multiple omics layers to accurately impute missing values from sparse profiling datasets and guide the detection of outliers, e.g. due to mislabelled samples or sample swaps.

In a second application, we used MOFA for the analysis of single-cell multi-omics data. This use case illustrates the advantage of learning continuous factors, rather than discrete groups, enabling MOFA to recover a differentiation trajectory by combining information from two sparsely profiled molecular layers.

While applications of factor models for integrating different data types were reported previously [1, 116, 144, 170], MOFA provides unique features (Methods Section 4.4.5, Appendix Table 4.A.3) that enable the interpretable reconstruction of the underlying factors and accommodating different data types as well as different patterns of missing data. MOFA is available as open source software and includes semi-automated analysis pipelines allowing for in-depth characterisations of inferred factors. Taken together, this will foster the accessibility of interpretable factor models for a wide range of multi-omics studies.

Although we have addressed important challenges for multi-omics applications, MOFA is not free of limitations. The model is linear, which means that it can miss strongly non-linear relationships between features within and across assays [25]. Non-linear extensions of MOFA may address this, although, as with any models in high-dimensional spaces, there will be trade-offs between model complexity, computational efficiency and interpretability [43]. A related area of work is to incorporate prior information on the relationships between individual features. For example, future extensions could make use of pathway databases

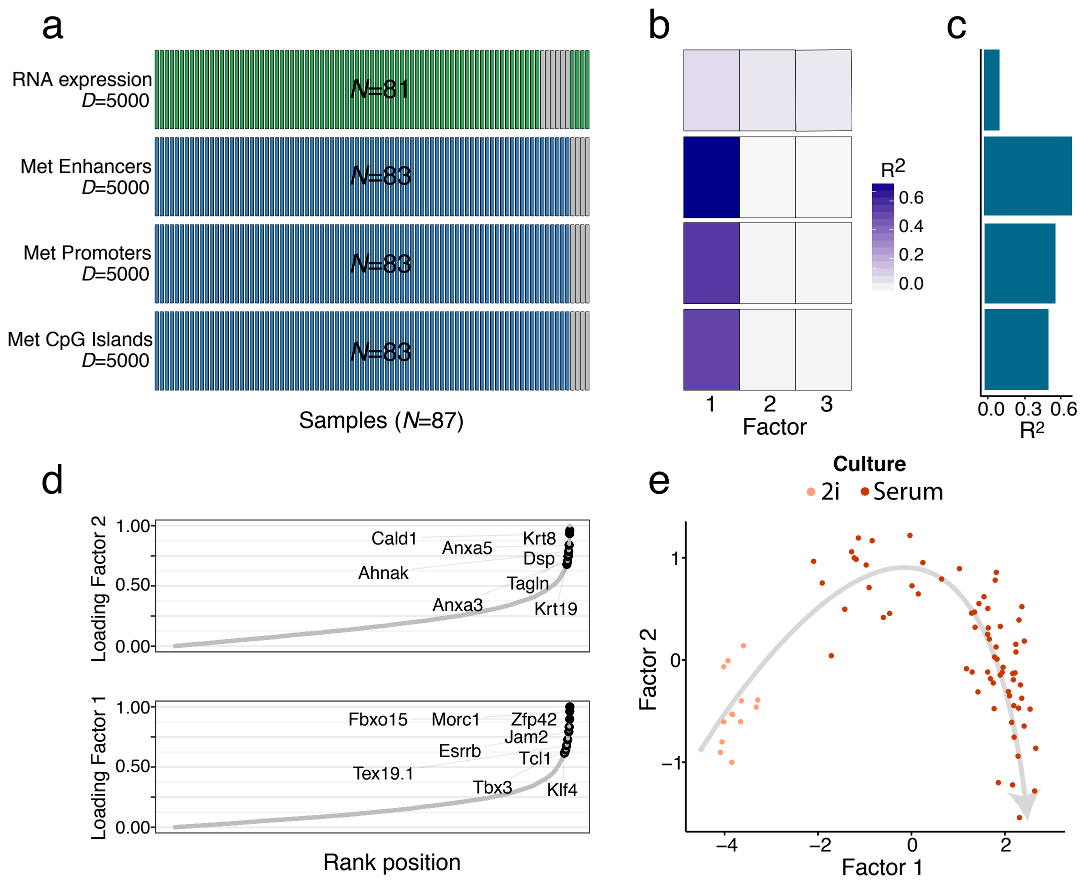


Figure 4.9: Application of MOFA to a single-cell multi-omics study. (a) Study overview and data types. Data modalities are shown in different rows (D = number of features) and samples in columns, with missing samples shown using grey bars. **(b)** Fraction of the total variance explained (R^2) by individual factors for each data modality and **(c)** cumulative proportion of total variance explained. **(d)** Absolute loadings of Factor 1 (bottom) and Factor 2 (top) in the mRNA data. Labelled genes in Factor 1 are known markers of pluripotency [145] and genes labelled in Factor 2 are known differentiation markers [70]. **(e)** Scatterplot of Factors 1 and 2. Colours denote culture conditions. The grey arrow illustrates the differentiation trajectory from naive pluripotent cells via primed pluripotent cells to differentiated cells.

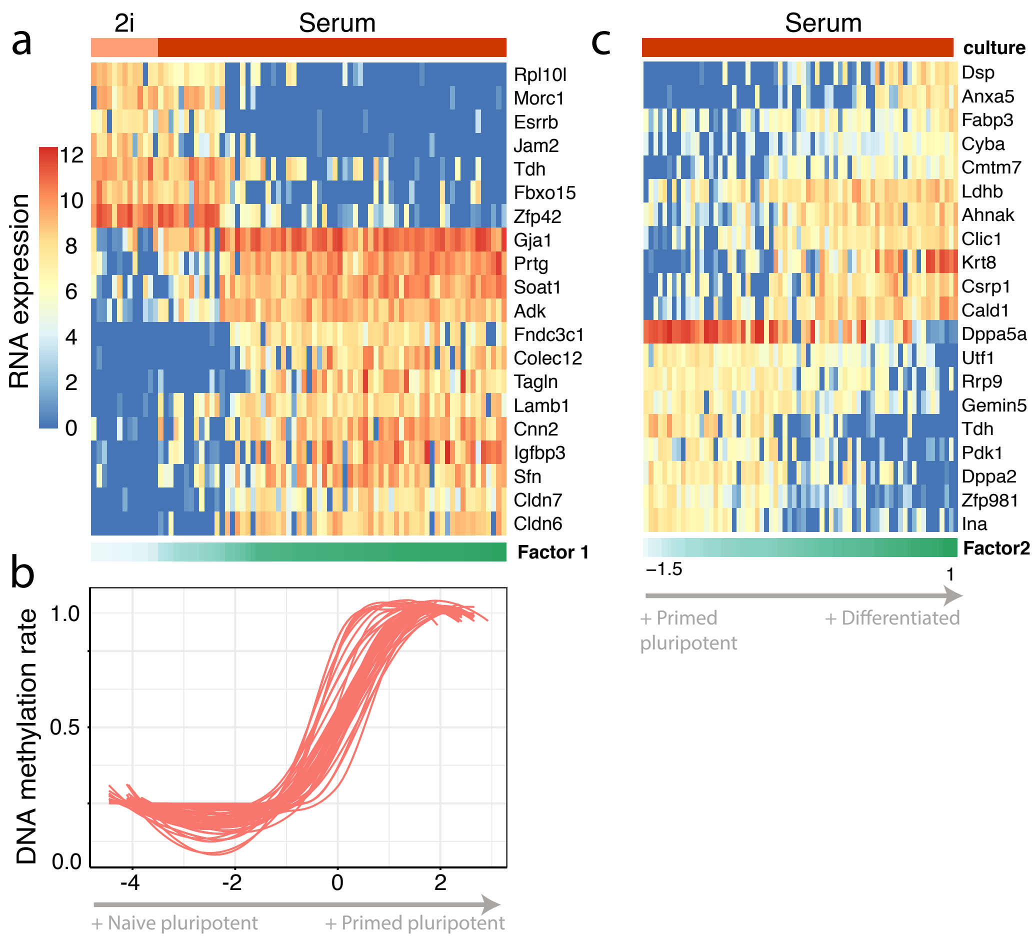


Figure 4.10: Transcriptomic and epigenetic changes associated with Factor 1 in the single cell data. **(a)** RNA expression changes for the top 20 genes with largest absolute weight on Factor 1. **(b)** DNA methylation rate changes for the top 20 CpG sites with largest absolute weight. Shown is a non-linear loess regression model fit per CpG site. **(c)** RNA expression changes for the top 20 genes with largest absolute weight on Factor 2.

within each omics type [24] or priors that reflect relationships given by the 'dogma of molecular biology'. In addition, new likelihoods and noise models could expand the value of MOFA in data sets with specific statistical properties that hamper the application of traditional statistical methods, including zero-inflated data (i.e. scRNA-seq [154]) or binomial distributed data (i.e. splicing events [96]). Finally, while here we focus our attention on the point estimates of inferred factors, future extensions could attempt a more comprehensive Bayesian treatment that propagates evidence strength and estimation uncertainties to diagnostics and downstream analyses.

4.4 Methods

4.4.1 Multi-Omic Factor Analysis model

Starting from M data matrices $\mathbf{Y}^1, .., \mathbf{Y}^M$ of dimensions $N \times D_m$, where N is the number of samples and D_m the number of features in data matrix m, MOFA decomposes these matrices as

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\epsilon}^m \qquad m = 1, \ldots, M. \tag{4.1}$$

Here, $\mathbf{Z}$ denotes the factor matrix (common for all data matrices) and $\mathbf{W}^m$ denote the weight matrices for each data matrix m (also referred to as view m in the following). $\boldsymbol{\epsilon}^m$ denotes the view-specific residual noise term, with its form depending on the specifics of the data-type (see *Noise model*).

The model is formulated in a probabilistic Bayesian framework, where we place prior distributions on all unobserved variables of the model (see plate diagram in Figure 4.11), i.e. the factors $\mathbf{Z}$, the weight matrices $\mathbf{W}^m$ and the parameters of the residual noise term. In particular, we use a standard normal (or Gaussian) prior for the factors $\mathbf{Z}$ and employ sparsity priors for the weight matrices as described next.

Model regularization

An appropriate regularization of the weight matrices is essential for the model's ability to disentangle variation across data sets and to yield interpretable factors. MOFA uses a two-level regularization: The first level encourages view- and factor-wise sparsity, thereby allowing to directly identify which factor is active in which view. The second level encourages feature-wise sparsity, thereby typically resulting in a small number of features with active weights. To encode these sparsity levels we combine an ARD prior for the first type of the sparsity with a spike-and-slab prior for the second. For amenable inference we model the spike-and-slab prior by parametrizing the weights as a product of a Bernoulli random variable and a Gaussian random variable: $\mathbf{W}^m = \mathbf{S}^m\hat{\mathbf{W}}^m$, where $s^m_{dk} \sim \text{Ber}(\theta^m_k)$ and $\hat{w}^m_{dk} \sim \mathcal{N}(0, \frac{1}{\alpha^m_k})$. To automatically learn the appropriate level of regularization for each factor and view, we use uninformative conjugate priors on α^m_k, which controls the strength of factor k in view m, and on θ^m_k, which determines the feature-wise sparsity level of factor k in view m (see Appendix Section 4.A.1 for details).

Noise model

MOFA supports the combination of different noise models to integrate diverse data types, including continuous, binary and count data. A standard noise model for continuous data is the Gaussian noise model assuming independent and identically distributed (iid) residuals $\boldsymbol{\epsilon}^m$ that are heteroscedastic across features, i.e. $\epsilon^m_{nd} \sim \mathcal{N}(0, \frac{1}{\tau^m_d})$, with a gamma prior on the

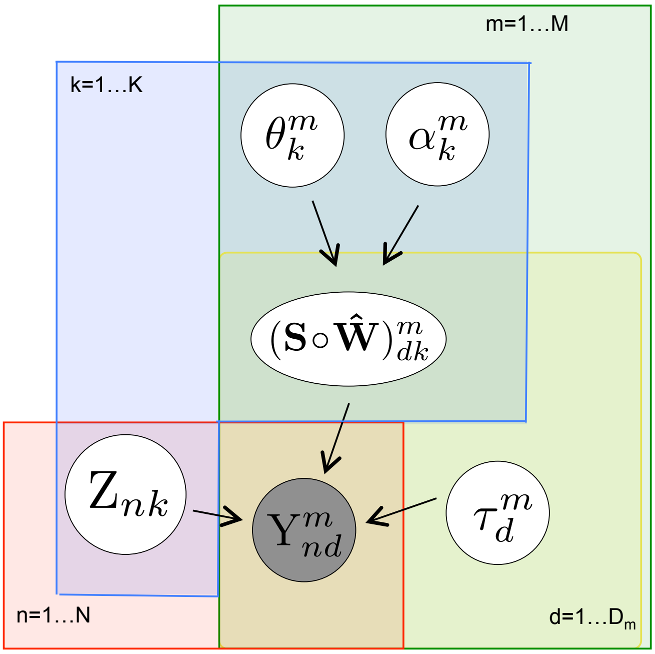


Figure 4.11: Graphical model representation of MOFA. Grey-filled nodes denote observed variables whereas white-filled nodes denote unobserved variables that are inferred by the model. $\mathbf{Y}$ denotes the observed data matrices, $\mathbf{Z}$ denotes latent factors and $\mathbf{S}\hat{\mathbf{W}}$ denotes the model weights with a spike-and-slab prior, implemented as the product of a Bernoulli variable $\mathbf{S}$ and a Gaussian variable $\hat{\mathbf{W}}$. $\boldsymbol{\theta}$ represents the sparsity parameters of the spike-and-slab prior and $\boldsymbol{\alpha}$ corresponds to the view- and factor-wise parameters of the automatic relevance determination prior. $\boldsymbol{\tau}$ represents the precision of the Gaussian noise. N is the number of samples, M is the number of views, D_m is the number of features in the m^{th} view and K is the number of latent factors.

precision parameters τ_d^m. MOFA further supports noise models for binary and count data that are not appropriately modelled using a Gaussian likelihood. In the current version, MOFA models count data using a Poisson model and binary data using a Bernoulli model. Here, the model likelihood is given by $y_{nd}^m \sim \text{Poi}(\lambda(z_{n:}w_{d:}^T))$ and $y_{nd}^m \sim \text{Ber}(\sigma(z_{n:}w_{d:}^T))$, respectively, where $\lambda(x) = \log(1 + e^x)$ and σ denotes the logistic function $\sigma(x) = (1 + e^{-x})^{-1}$.

4.4.2 Parameter inference

For scalability, we make use of a variational Bayesian framework, which uses a mean field approximation for inference [21]. The key idea is to approximate the intractable posterior distribution using a simpler class of distributions by minimizing the Kullback-Leibler divergence to the exact posterior, or equivalently, maximizing the evidence lower bound (ELBO). Convergence of the algorithm can be monitored based on the ELBO. An overview of variational inference and details on the specific implementation for MOFA can be found in Appendix Section 4.A.2. To enable an efficient inference for non-Gaussian likelihoods we employ variational lower bounds on the likelihood [101, 168] (see Appendix Section 4.A.3 for details).

4.4.3 Model selection

An important part of the training is the determination of the number of factors. Factors are automatically inactivated by the ARD prior of the model as described in *Model regularization* in Section 4.4.1. In practice, factors are pruned during training using a minimum fraction of variance explained threshold that needs to be specified by the user. Alternatively, the user can fix the number of factors and the minimum variance criterion is ignored. In the analyses presented we initialised the models with $K = 25$ factors and they were pruned during training using a threshold of variance explained of 2%. For details on the implementation as well as practical considerations for training and choice of the threshold parameter refer to Appendix Section 4.A.4.

While the inferred factors are robust under different initializations (e.g. Appendix Figure 4.A.6c,d) the optimization landscape is non-convex and hence the algorithm is not guaranteed to converge to a global optimum. Results presented here are based on 10-25 random restarts, selecting the model with the highest ELBO (e.g. Appendix Figure 4.A.6b).

4.4.4 Downstream analysis for factor interpretation and annotation

As part of MOFA we provide the R package MOFAtools, which provides a semi-automated pipeline for the characterisation and interpretation of the latent factors. In all downstream analyses we use the expectations of the model variables under the posterior distributions inferred by the variational framework.

The first step, after a model has been trained, is to disentangle the variation explained by each factor in each view. To this end, we compute the fraction of the variance explained (R^2) by factor k in view m as

$$R^2_{m,k} = 1 - \frac{\sum_{n,d}(y^m_{nd} - z_{nk}w^m_{kd} - \mu^m_d)^2}{\sum_{n,d}(y^m_{nd} - \mu^m_d)^2} \tag{4.2}$$

as well as the fraction of variance explained per view taking into account all factors

$$R^2_{m} = 1 - \frac{\sum_{n,d}(y^m_{nd} - \sum_k z_{nk}w^m_{kd} - \mu^m_d)^2}{\sum_{n,d}(y^m_{nd} - \mu^m_d)^2}. \tag{4.3}$$

Here, μ^m_d denotes the feature-wise mean. Subsequently, each factor is characterised by three complementary analyses:

1. *Ordination of the samples in factor space:* Visualise a low-dimensional representation of the main drivers of sample heterogeneity.

2. *Inspection of top features with largest weight:* The loadings can give insights into the biological process underlying the heterogeneity captured by a latent factor. Due to scale differences between assays, the weights of different views are not directly comparable. For simplicity, we scale each weight vector by its maximal absolute value.

3. *Feature set enrichment analysis:* We combine the signal from functionally related sets of features (e.g., gene sets) to derive a feature-set based annotation. By default, we use a parametric t-test comparing the means of the foreground set (the weights of features that belong to a set G) and the background set (the weights of features that do not belong to the set G), similar to the approach described in [69].

4.4.5 Relationship to existing methods

MOFA builds upon the statistical framework of group factor analysis [27, 109, 113, 122, 197, 214] and is in part also related to the iCluster methods [144, 170] as shown in Appendix Table 4.A.3. Here we describe these connections in further detail:

(i) **iCluster**: In contrast to MOFA, iCluster uses in each view the same extent of regularization for all factors, which may be sufficient for the purpose of clustering (the primary application of iCluster), however it results in a reduced ability for distinguishing factors that drive variation in distinct subsets of views (Appendix Figure 4.A.5). Additionally, unlike MOFA and GFA, iCluster does not handle missing values and is computationally demanding (Figure 4.2), as it requires re-fitting the model for a large range of different penalty parameters and choices of the model dimension.

(ii) **Group factor analysis**: While the underlying model of MOFA is closely connected to the most recent GFA implementation [122], GFA is restricted to Gaussian observation noise. In terms of the algorithmic implementation, MOFA uses an additional burn-in period during training in which the sparsity constraints are deactivated ($\theta = 1$) to avoid early splitting of a common signal onto distinct factors to meet sparsity assumptions, and it actively drops factors below a predefined variance threshold (see Section 4.4.3). In contrast, GFA directly uses sparsity constraints throughout training and also maintains factors that have near-zero relevance. In terms of inference, MOFA is implemented using a variational approximate Bayesian inference, whereas GFA is based on a Gibbs sampler. In terms of computational scalability (Figure 4.2), both methods are linear in the model's parameters, although GFA is computationally more expensive in absolute terms. This difference is particularly pronounced for datasets with missing data. This, together with the inability to deactivate factors during inference (Appendix Figure 4.A.4) renders GFA considerably slower in applications to real data.

4.4.6 Details on the simulation studies

Model validation

To validate MOFA we simulated data from the generative model for a varying number of views ($M = 1, 3, \ldots, 21$), features ($D = 100, 500, \ldots, 10000$), factors ($K = 5, 10, \ldots, 60$), missing values (from 0% to 90%) as well as for non-Gaussian likelihoods (Poisson, Bernoulli) (see Appendix Table 4.A.1 for simulation parameters). We assessed the ability of MOFA to recover the true simulated number of factors in the different settings, where we considered 10 repeat experiments for every configuration. All trials were started with a high number of factors ($K = 100$), and inactive factors were pruned as described in Section 4.4.3.

Model comparison

To compare MOFA to GFA, we simulated data from the underlying generative model with $K_{true} = 10$ factors, $M = 3$ views, $N = 100$ samples, $D = 5,000$ features per view and 5% missing values (missing at random). For each of the three views we used a different likelihood model: continuous data was simulated with a Gaussian distribution, binary data with a Bernoulli distribution and count data with a Poisson distribution. Except for the non-Gaussian likelihood extension, both methods share the same underlying generative model, thus allowing for a meaningful comparison. We fit ten realizations of the MOFA and GFA models with $K_{initial} = 20$ factors and let the method determine the most likely number of factors. To assess scalability, we considered the same base parameter settings, varying one of the simulation parameters at a time (number of factors K, number of features D, number of samples N and number of views M, all Gaussian). To assess the ability to reconstruct factor activity patterns we simulated data from the generative model for $K_{true} = 10$ and $K_{true} = 15$ factors (M, N, D as before, no missing values, only Gaussian views), where factors were set to either active or inactive in a specific view by sampling the parameter α_k^m from $\{1, 10^3\}$. Appendix Table 4.A.1 shows in more detail the simulation parameters used in each setting.

4.4.7 Details on the CLL analysis

Data processing

The data were taken from [47], where details on the data generation and processing can be found. Briefly, this dataset consists of somatic mutations (combination of targeted and whole exome sequencing), RNA expression (RNA-Seq), DNA methylation (Illumina arrays) and ex-vivo drug response screens (ATP-based CellTiter Glo assay). For the training of MOFA we included 62 drug response measurements (excluding NSC 74859 and bortezomib due to bad quality) at five concentrations each ($D = 310$) with a threshold at 1.1 to remove outliers. Mutations were considered if present in at least 3 samples ($D = 69$). Low counts from RNAseq data were filtered out and the data were normalized using the *estimateSizeFactors* and *varianceStabilizingTransformation* function of DESeq2 [127]. For training we considered the top $D = 5,000$ most variable mRNAs after exclusion of genes from the Y chromosome. Methylation data were transformed to M-values and we extracted the top 1% most variable CpG sites excluding sex chromosomes ($D = 4,248$). We included patients diagnosed with CLL and having data in at least two views into the MOFA model leading to a total of $N = 200$ samples.

Model training and selection

We trained MOFA using 25 random initializations with a variance threshold of 2% and selected the model with the best fit for downstream analysis (see Section 4.4.3).

Gene set enrichment analysis

Gene set enrichment analysis was performed based on Reactome gene sets [59] as described above. Resulting p-values were adjusted for multiple testing for each factor using the Benjamini-Hochberg procedure [16]. Significant enrichments were at a false discovery rate of 1%.

Imputation

To compare imputation performance, we trained MOFA on the subset of samples with all measurements ($N = 121$) and masked at random either single values or all measurements for random samples in the drug response. After model training the masked values were imputed directly from the model equation (4.1) and the accuracy was assessed in terms of mean squared error on the true (masked) values. For both settings we fixed the number of factors in MOFA to $K = 10$. To investigate the dependence on K for imputation and to compare MOFA to GFA we re-ran the same masking experiments varying $K = 1, \ldots, 20$ (Appendix Figure 4.A.17).

Survival Analysis

Associations between the inferred factors and clinical covariates were assessed using the patients' time to next treatment as response variable in a Cox model ($N = 174$ samples with treatment information, 96 of which were uncensored cases). For univariate association tests (as shown in Figure 4.8, Appendix Figure 4.A.21) we scaled all predictors to ensure comparability of the hazard ratios and we rotated factors, which are rotational invariant, such that their hazard ratio is greater or equal to 1. To investigate the predictive power of different datasets, we used a multivariate Cox model and compared Harrell's C-index of predictions in a stratified 5-fold cross-validation scheme. As predictors we included the top 10 principal components calculated on the data for each single view, a concatenated data set ('all') as well as the ten MOFA factors. Missing values in a view were set to the feature-wise mean. In a second set of models we used the complete set of all features in a view with a ridge penalty in the Cox model as implemented in the R package glmnet. For the Kaplan-Meier plots an optimal cut-point on each factor was determined to define the two groups using the maximally selected rank statistics as implemented in the R package survminer with p-values based on a log-rank test between the resulting groups.

4.4.8 Details on the single cell analysis

The data were obtained from [7], where details on the data generation and pre-processing can be found. Briefly, for each CpG site we calculated a binary methylation rate from the ratio of methylated read counts to total read counts. RNA expression data were normalized following [128]. To fit MOFA, we considered the top 5,000 most variable genes with a maximum dropout of 90%, and the top 5,000 most variable CpG sites with a minimum coverage of 10% across cells. Model selection was performed as described for the CLL data and factors were inactivated below a minimum explained variance of 2%. For the clustering analysis using SNF and iCluster, the optimal number of clusters was selected using the Bayesian information criterion (BIC) criterion.

4.4.9 Software and data availability

An open source implementation of MOFA in R and Python is available from `https://github.com/bioFAM/MOFA`. Code to reproduce all the analyses presented is available at `https://github.com/bioFAM/MOFA_analysis`.

The CLL data were obtained from [47] and are available at the European Genome-phenome Archive under accession EGAS00001001746 and data tables as R objects can be downloaded from `http://pace.embl.de/`. The single-cell data were obtained from [7] and are available in the Gene Expression Omnibus under accession GSE74535. All data used are contained within the MOFA vignettes and can be downloaded from `https://github.com/bioFAM/MOFA`.

4.A Appendix

Multi-Omics Factor Analysis (MOFA) is a statistical model aimed at disentangling sources of variation in multi-omics data. Here, we introduce the statistical model (Section 4.A.1) and its inference procedure in more detail, both in case of Gaussian data (Section 4.A.2) and non-Gaussian data (Section 4.A.3). In addition, we provide practical considerations for training (Section 4.A.4).

Mathematical notation

- Matrices are denoted with bold capital letters, e. g. $\mathbf{W}$.
- Vectors are denoted with bold non-capital letters. If the vector comes from a matrix, two indices separated by a comma will always be shown at the bottom: the first one corresponding to the row and the second one to the column. The symbol ':' denotes the entire row/column. For instance, $\mathbf{w}_{j,:}$ refers to the entire j^{th} row of a matrix $\mathbf{W}$.
- Scalars are denoted with non-bold non-capital letters. If the value comes from a matrix, two indices will always be shown at the bottom: the first one corresponding to the row and the second one to the column. For instance, w_{jk} refers to the value coming from the j^{th} row and the k^{th} column of a matrix $\mathbf{W}$.
- $\mathbb{E}_q[x]$ denotes the expectation of x under the distribution q. Sometimes, when the expectations are taken with respect to the same distribution many times, we will use $\langle x \rangle$ to avoid cluttered notation.
- $\mathcal{F}(x|a)$ denotes the probability density function of a distribution $\mathcal{F}$ in x with parameters a. For example, $\mathcal{N}\left(x \,|\, \mu, \sigma^2\right)$ denotes the density of a univariate normal (or Gaussian) distribution with mean μ and variance σ^2, and $\text{Gamma}\left(x \,|\, a, b\right)$ the density of a gamma distribution with parameters a and b.
- $\mathrm{B}(a, b)$ denotes the beta function.
- $\Gamma(a)$ denotes the gamma function.

4.A.1 Details on the Multi-Omics Factor Analysis model

Factor analysis models, also called latent variable models, are a probabilistic modelling approach which aim to reduce the dimensionality of a (big) dataset into a small set of variables which are easier to interpret and visualise. More formally, given a dataset $\mathbf{Y}$ of N samples and D features, latent variable models attempt to explain dependencies between the features by means of a potentially smaller set of K unobserved (latent) factors. MOFA is a generalisation of traditional factor analysis where the input data consists of M matrices $\mathbf{Y}^m = [y^m_{nd}] \in \mathbb{R}^{N \times D_m}$ where each matrix m is called a view. Each view consists of non-overlapping features which usually, but not necessarily, represent different assays. The input data is then factorised as

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\epsilon}^m, \tag{4.4}$$

where $\mathbf{Z} = [z_{nk}] \in \mathbb{R}^{N \times K}$ is a single matrix that contains the low-dimensional latent variables, $\mathbf{W}^m = [w^m_{dk}] \in \mathbb{R}^{D_m \times K}$ are loading matrices that relate the high-dimensional space to the low dimensional representation, and $\boldsymbol{\epsilon}^m = [\epsilon^m_{nd}] \in \mathbb{R}^{N \times D_m}$ denotes residual noise. We start by assuming independent Gaussian residuals $\boldsymbol{\epsilon}^m$, similar to standard (group) factor analysis models, while allowing for heteroscedasticity across features, i. e. for $n = 1, \ldots, N$

$$p(\epsilon^m_{nd}|\tau^m_d) = \mathcal{N}\left(\epsilon^m_{nd} \,|\, 0, 1/\tau^m_d\right). \tag{4.5}$$

This results in the following Gaussian likelihood (for extensions to non-Gaussian settings see Section 4.A.3):

$$p(y_{nd}^m|\mathbf{z}_{n,:}, \mathbf{w}_{d,:}^m, \tau_d^m) = \mathcal{N}\left(y_{nd}^m \,|\, \mathbf{z}_{n,:}\mathbf{w}_{d,:}^{mT}, 1/\tau_d^m\right), \tag{4.6}$$

where $\mathbf{w}_{d,:}^m$ denotes the d^{th} row of the loading matrix $\mathbf{W}^m$ and $\mathbf{z}_{n,:}$ the n^{th} row of the latent factor matrix $\mathbf{Z}$. For a fully probabilistic treatment we place prior distributions on the weights $\mathbf{W}^m$, the latent variables $\mathbf{Z}$ as well as on the precision of the noise $\boldsymbol{\tau}^m$. We use a standard Gaussian prior on the latent variables and a conjugate Gamma prior for the precision, i. e.

$$p(z_{nk}) = \mathcal{N}\left(z_{nk} \,|\, 0, 1\right), \tag{4.7}$$

$$p(\tau_d^m) = \text{Gamma}\left(\tau_d^m \,|\, a_0^\tau, b_0^\tau\right), \tag{4.8}$$

with $a_0^\tau, b_0^\tau = 10^{-14}$ to obtain uninformative priors.

A key determinant of the model is the regularization used on the weights $\mathbf{W}^m$. MOFA encodes two levels of sparsity: a view- and factor-wise sparsity and a feature-wise sparsity. The aim of the factor- and view-wise sparsity is to identify which factors are active in which view, such that the weight vector $\mathbf{w}_{:,k}^m$ is shrunk to zero if the factor k does not drive any variation in view m. This is the general property that allows the model to disentangle the sources of variability between different assays.

In addition, we place a second layer of feature-wise sparsity which puts zero weights on individual features from active factors. This relies on the assumption that biological sources of variability are typically sparse, i.e. only a small number of features are 'active', i. e., have non-zero weight. We achieve both levels of sparsity by placing appropriate priors on the weight matrices.

Specifically, we combine an automatic relevance determination (ARD) prior [131] for the view- and factor-wise sparsity with a spike-and-slab prior [143] for the feature-wise sparsity, similar to [109]. However, as the spike-and-slab prior

$$p(w) = (1-\theta)\mathbb{1}_0(w) + \theta\mathcal{N}(w \,|\, 0, 1/\alpha) \tag{4.9}$$

contains a Dirac delta function, which makes the inference troublesome, here we use a re-parametrization of the weights w as a product of a Gaussian random variable $\hat{w}$ and a Bernoulli random variable s [24, 188], resulting in the following prior:

$$p(\hat{w}_{dk}^m, s_{dk}^m|\theta_k^m, \alpha_k^m) = \mathcal{N}(\hat{w}_{dk}^m \,|\, 0, 1/\alpha_k^m)\,\text{Ber}(s_{dk}^m \,|\, \theta_k^m) \tag{4.10}$$

In this formulation α_k^m controls the strength of factor k in view m and θ_k^m controls the degree of contribution from the spike term, determining the overall feature-wise sparsity levels of factor k in view m. In order to automatically learn these parameters we use the following conjugate priors

$$p(\theta_k^m) = \text{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right), \tag{4.11}$$

$$p(\alpha_k^m) = \text{Gamma}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right), \tag{4.12}$$

with hyper-parameters $a_0^\theta, b_0^\theta = 1$ and $a_0^\alpha, b_0^\alpha = 10^{-14}$ to get uninformative priors. A value of θ_k^m close to 0 implies that most of the weights of factor k in view m are shrunk to 0, yielding a sparse factor. In contrast, a value of θ_k^m close to 1 implies that most of the weights are non-zero, yielding a non-sparse factor.

In practice, the ARD prior yields a matrix $\boldsymbol{\alpha} \in \mathbb{R}^{M\times K}$ that defines four different types of factors:

- Factors that do not explain variation in any data set (inactive factors): all values in the corresponding columns of $\boldsymbol{\alpha}$ are large. These factors are actively removed from the model during training.
- Factors that explain variation in all data sets (fully shared factors): all M values in the corresponding columns of $\boldsymbol{\alpha}$ are small.
- Factors that explain variation in a single data set (unique factors): all values in the corresponding columns of $\boldsymbol{\alpha}$ are very large, except one.
- Factors that explain variation in a subset of data sets (partially shared factors): some values in the corresponding columns of $\boldsymbol{\alpha}$ are very large whereas others are small.

Using these prior distributions, the joint probability density function is given by

$$
\begin{aligned}
p(\mathbf{Y}, \mathbf{Z}, \hat{\mathbf{W}}, \mathbf{S}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) = & \prod_{m=1}^{M} \prod_{n=1}^{N} \prod_{d=1}^{D_m} \mathcal{N}\left(y_{nd}^m \,|\, \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk}, 1/\tau_d\right) \\
& \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \\
& \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{k=1}^{K} \mathcal{N}\left(\hat{w}_{dk}^m \,|\, 0, 1/\alpha_k^m\right) \mathrm{Ber}(s_{dk}^m | \theta_k^m) \\
& \prod_{m=1}^{M} \prod_{k=1}^{K} \mathrm{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right) \\
& \prod_{m=1}^{M} \prod_{k=1}^{K} \mathrm{Gamma}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \\
& \prod_{m=1}^{M} \prod_{d=1}^{D_m} \mathrm{Gamma}\left(\tau_d^m \,|\, a_0^\tau, b_0^\tau\right).
\end{aligned}
\tag{4.13}
$$

This completes the definition of the model, which is graphically in Figure 4.11.

4.A.2 Details on model inference

4.A.2.1 Introduction to variational Bayes inference

To ensure scalable inference we use a variational approach with a mean-field approximation [20]. Briefly, in variational inference the true intractable posterior distribution of the unobserved variables $p(\mathbf{X}|\mathbf{Y})$ is approximated by a simpler distribution of factorized form $q(\mathbf{X}) = \prod_i q(\mathbf{X}_i)$ that leads to an efficient inference scheme. Here, $\mathbf{X}$ denotes all the hidden variables (including parameters) and $\mathbf{Y}$ denotes all the observed variables.

Under this approximation, the true log marginal likelihood $\log p(\mathbf{Y})$ is lower bounded by

$$
\begin{aligned}
\mathcal{L}(\mathbf{X}) &= \int q(\mathbf{X}) \Big(\log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} + \log p(\mathbf{Y}) \Big) d\mathbf{X} \\
&= \log p(\mathbf{Y}) - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{X}) || p(\mathbf{X}|\mathbf{Y})) \\
&\leq \log p(\mathbf{Y}),
\end{aligned}
\tag{4.14}
$$

where D_{KL} denotes the Kullback-Leibler (KL) divergence.

$\mathcal{L}(\mathbf{X})$ is called the evidence lower bound (ELBO), which is equal to the sum of the model evidence and the negative KL-divergence between the true posterior and the variational distribution. The key observation here is that increasing the ELBO is equivalent to decreasing the KL-divergence between the two distributions.

Variational learning involves optimising the functional $\mathcal{L}(\mathbf{X})$ with respect to the distribution $q(\mathbf{X})$. If we allow any possible choice of $q(\mathbf{X})$, then the maximum of the lower bound $\mathcal{L}(\mathbf{X})$ will occur when the KL-divergence vanishes, which occurs when $q(\mathbf{X})$ equals the true posterior distribution $p(\mathbf{X}|\mathbf{Y})$. Nevertheless, since the true posterior is intractable, this does not lead to any simplification of the problem. Instead, it is necessary to consider a restricted family of variational distributions that are tractable to compute and then seek the member of this family for which the KL-divergence is minimised [15].

Mean-field approximation The most common type of variational Bayes, known as mean-field approach, assumes that the variational distribution factorises over L disjoint groups of variables, i. e.

$$q(\mathbf{X}) = \prod_{i=1}^{L} q(\mathbf{X}_i).$$

Evidently, this family of distributions does not usually contain the true posterior because the unobserved variables have dependencies, but this assumption allows the derivation of an analytical inference scheme [15]. It follows that the optimal distribution q_i^* that maximises the lower bound $\mathcal{L}(\mathbf{X})$, for each variable $\mathbf{X}_i$, can be calculated as

$$\log q_i^*(\mathbf{X}_i) = \mathbb{E}_{-i}[\log p(\mathbf{Y}, \mathbf{X})] + \text{const}, \tag{4.15}$$

where $\mathbb{E}_{-i}$ denotes an expectation with respect to the q distributions over all variables $\mathbf{X}_j$ except for $\mathbf{X}_i$. The additive constant is set by normalizing the distribution $q_i^*(\mathbf{X}_i)$, i. e.

$$q_i^*(\mathbf{X}_i) = \frac{\exp(\mathbb{E}_{-i}[\log p(\mathbf{Y}, \mathbf{X})])}{\int \exp(\mathbb{E}_{-i}[\log p(\mathbf{Y}, \mathbf{X})]) d\mathbf{X}_i}.$$

This provides the general expression which yields the set of variational distributions that maximise the lower bound of the log marginal likelihood, subject to the factorisation constraint. Or equivalently, the set of distributions that minimise the KL-divergence between the $q(\mathbf{X})$ distribution and the true posterior $p(\mathbf{X}|\mathbf{Y})$. For MOFA we adopt the following mean field approximation, which factorizes in all model variables except for $\hat{w}_{dk}^m, s_{dk}^m$, which are strongly connected by the re-parametrization $w_{dk}^m = \hat{w}_{dk}^m s_{dk}^m$, i. e. we assume

$$\begin{aligned} q(\mathbf{Z}, \hat{\mathbf{W}}, \mathbf{S}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) &= q(\mathbf{Z}) q(\hat{\mathbf{W}}, \mathbf{S}) q(\boldsymbol{\theta}) q(\boldsymbol{\alpha}) q(\boldsymbol{\tau}) \\ &= \prod_{n=1}^{N} \prod_{k=1}^{K} q(z_{nk}) \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{k=1}^{K} q(\hat{w}_{dk}^m, s_{dk}^m) \\ &\quad \prod_{m=1}^{M} \prod_{k=1}^{K} q(\theta_k^m) q(\alpha_k^m) \prod_{m=1}^{M} \prod_{d=1}^{D_m} q(\tau_d^m). \end{aligned}$$

Variational Bayes expectation maximization algorithm Note that in Equation (4.15), for a given variable $\mathbf{X}_i$, the expectation on the right-hand side is taken with respect to the other variables' variational distribution $q_j(\mathbf{X}_j)$ for $j \neq i$. Therefore, there are circular dependencies between the different equations and there is no analytical solution for the parameters of the variational distribution. This naturally suggests an iterative algorithm similar to the expectation-maximization (EM) algorithm. In each step we update the moments and parameters of the variational distribution of the latent variables $q_j(\mathbf{X}_j)$ using the current estimates of the variational distributions of the parameters $q_{-j}(\mathbf{X}_{-j})$ [15]. The algorithm is stopped when the change in the ELBO is small enough.

4.A.2.2 Update equations for Gaussian data

Latent variables (Z) The variational distribution of $\mathbf{Z}$ is given by

$$q(\mathbf{Z}) = \prod_{k=1}^{K}\prod_{n=1}^{N} q(z_{nk}) = \prod_{k=1}^{K}\prod_{n=1}^{N} \mathcal{N}\left(z_{nk} \,|\, \mu_{z_{nk}}, \sigma^2_{z_{nk}}\right),$$

where

$$\sigma^2_{z_{nk}} = \Big(\sum_{m=1}^{M}\sum_{d=1}^{D_m}\langle\tau_d^m\rangle\langle(s_{dk}^m\hat{w}_{dk}^m)^2\rangle + 1\Big)^{-1},$$

$$\mu_{z_{nk}} = \sigma^2_{z_{nk}}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\langle\tau_d^m\rangle\langle s_{dk}^m\hat{w}_{dk}^m\rangle\Big(y_{nd}^m - \sum_{j\neq k}\langle s_{dj}^m\hat{w}_{dj}^m\rangle\langle z_{nj}\rangle\Big).$$

Spike and slab weights ($\mathbf{W} = \hat{\mathbf{W}}\mathbf{S}$) The variational distribution of $\hat{\mathbf{W}}$, $\mathbf{S}$ is given by

$$q(\hat{\mathbf{W}}, \mathbf{S}) = \prod_{m=1}^{M}\prod_{d=1}^{D_m}\prod_{k=1}^{K} q(\hat{w}_{dk}^m, s_{dk}^m) = \prod_{m=1}^{M}\prod_{d=1}^{D_m}\prod_{k=1}^{K} q(\hat{w}_{dk}^m|s_{dk}^m)q(s_{dk}^m).$$

The update for $q(s_{dk}^m)$ is obtained from

$$\gamma_{dk}^m = q(s_{dk} = 1) = \frac{1}{1+\exp(-\lambda_{dk}^m)},$$

where

$$\lambda_{dk}^m = \langle\log\frac{\theta_k^m}{1-\theta_k^m}\rangle + 0.5\log\frac{\langle\alpha_k^m\rangle}{\langle\tau_d^m\rangle} - 0.5\log\Big(\sum_{n=1}^{N}\langle z_{nk}^2\rangle + \frac{\langle\alpha_k^m\rangle}{\langle\tau_d^m\rangle}\Big)$$
$$+\frac{\langle\tau_d^m\rangle}{2}\frac{\Big(\sum_{n=1}^N y_{nd}^m\langle z_{nk}\rangle - \sum_{j\neq k}\langle s_{dj}^m\hat{w}_{dj}^m\rangle\sum_{n=1}^N\langle z_{nk}\rangle\langle z_{nj}\rangle\Big)^2}{\sum_{n=1}^N\langle z_{nk}^2\rangle + \frac{\langle\alpha_k^m\rangle}{\langle\tau_d^m\rangle}},$$

and the update for $q(\hat{w}_{dk}^m|s_{dk}^m)$ as

$$q(\hat{w}_{dk}^m|s_{dk}^m = 0) = \mathcal{N}\left(\hat{w}_{dk}^m \,|\, 0, 1/\alpha_k^m\right),$$
$$q(\hat{w}_{dk}^m|s_{dk}^m = 1) = \mathcal{N}\left(\hat{w}_{dk}^m \,|\, \mu_{w_{dk}^m}, \sigma^2_{w_{dk}^m}\right),$$

where

$$\mu_{w_{dk}^m} = \frac{\sum_{n=1}^N y_{nd}^m\langle z_{nk}\rangle - \sum_{j\neq k}\langle s_{dj}^m\hat{w}_{dj}^m\rangle\sum_{n=1}^N\langle z_{nk}\rangle\langle z_{nj}\rangle}{\sum_{n=1}^N\langle z_{nk}^2\rangle + \frac{\langle\alpha_k^m\rangle}{\langle\tau_d^m\rangle}},$$

$$\sigma_{w_{dk}^m} = \frac{\langle\tau_d^m\rangle^{-1}}{\sum_{n=1}^N\langle z_{nk}^2\rangle + \frac{\langle\alpha_k^m\rangle}{\langle\tau_d^m\rangle}}.$$

Taken together this means that we can update $q(\hat{w}_{dk}^m, s_{dk}^m)$ using

$$q(\hat{w}_{dk}^m|s_{dk}^m)q(s_{dk}^m) = \mathcal{N}\left(\hat{w}_{dk}^m \,|\, s_{dk}^m\mu_{w_{dk}^m}, s_{dk}^m\sigma^2_{w_{dk}^m} + (1-s_{dk}^m)/\alpha_k^m\right)$$
$$(\gamma_{dk}^m)^{s_{dk}^m}(1-\gamma_{dk}^m)^{1-s_{dk}}.$$

ARD precision ($\boldsymbol{\alpha}$) The variational distribution of $\boldsymbol{\alpha}$ is given by

$$q(\boldsymbol{\alpha}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \text{Gamma}(\alpha_k^m | \hat{a}_{mk}^{\alpha}, \hat{b}_{mk}^{\alpha}),$$

where

$$\hat{a}_{mk}^{\alpha} = a_0^{\alpha} + \frac{D_m}{2},$$
$$\hat{b}_{mk}^{\alpha} = b_0^{\alpha} + \frac{\sum_{d=1}^{D_m} \langle (\hat{w}_{dk}^m)^2 \rangle}{2}.$$

Noise precision ($\boldsymbol{\tau}$) The variational distribution of $\boldsymbol{\tau}$ is given by

$$q(\boldsymbol{\tau}) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} q(\tau_d^m) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} \text{Gamma}(\tau_d^m | \hat{a}_{md}^{\tau}, \hat{b}_{md}^{\tau}),$$

where

$$\hat{a}_{md}^{\tau} = a_0^{\tau} + \frac{N}{2},$$
$$\hat{b}_{md}^{\tau} = b_0^{\tau} + \frac{1}{2} \sum_{n=1}^{N} \langle (y_{nd}^m - \sum_{k}^{K} \hat{w}_{dk}^m s_{dk}^m z_{nk})^2 \rangle.$$

Spike and slab sparsity parameter ($\boldsymbol{\theta}$) The variational distribution of $\boldsymbol{\theta}$ is given by

$$q(\boldsymbol{\theta}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \text{Beta}(\theta_k^m | \hat{a}_{mk}^{\theta}, \hat{b}_{mk}^{\theta}),$$

where

$$\hat{a}_{mk}^{\theta} = \sum_{d=1}^{D_m} \langle s_{dk}^m \rangle + a_0^{\theta},$$
$$\hat{b}_{mk}^{\theta} = b_0^{\theta} - \sum_{d=1}^{D_m} \langle s_{dk}^m \rangle + D_m.$$

Evidence lower bound (ELBO) In order to monitor training and assess convergence we calculate the ELBO alongside with the other updates. The ELBO can be decomposed into a likelihood term and terms for each model variable $\mathbf{X}_i$ as

$$\begin{aligned}\mathcal{L}(\mathbf{X}) &= \int q(\mathbf{X}) \Big(\log \frac{p(\mathbf{X}, \mathbf{Y})}{q(\mathbf{X})} \Big) d\mathbf{X} \\ &= \mathbb{E}_q \log p(\mathbf{Y}|\mathbf{X})\rangle + \sum_i \left(\mathbb{E}_q \log p(\mathbf{X}_i) - \mathbb{E}_q \log q(\mathbf{X}_i) \right),\end{aligned}$$

where the expectation is under the variational distribution of the current step. Each of the terms from the above is computed as follows:

- *Likelihood term:* If using the Gaussian likelihood, this is given by

$$-\sum_{m=1}^{M}\left(\frac{ND_m}{2}\log(2\pi)+\frac{N}{2}\sum_{d=1}^{D_m}\log(\langle\tau_d^m\rangle)-\sum_{d=1}^{D_m}\frac{\langle\tau_d^m\rangle}{2}\sum_{n=1}^{N}\left(y_{nd}^m-\sum_{k=1}^{K}\langle s_{dk}^m\hat{w}_{dk}^m\rangle\langle z_{nk}\rangle\right)^2\right)$$

  Otherwise, this expression is replaced by the corresponding likelihood or bound.

- $\hat{\mathbf{W}}$ *and* $\mathbf{S}$ *terms:*

$$\mathbb{E}_q[\log p(\hat{\mathbf{W}},\mathbf{S})]=-\sum_{m=1}^{M}\left(\frac{KD_m}{2}\log(2\pi)+\frac{D_m}{2}\sum_{k=1}^{K}\log(\alpha_k^m)-\frac{\alpha_k^m}{2}\sum_{d=1}^{D_m}\sum_{k=1}^{K}\langle(\hat{w}_{dk}^m)^2\rangle\right.$$
$$\left.+\sum_{d=1}^{D_m}\sum_{k=1}^{K}\langle\log(\theta_k^m)\rangle\langle s_{dk}^m\rangle+\sum_{d=1}^{D_m}\sum_{k=1}^{K}\langle\log(1-\theta_k^m)\rangle(1-\langle s_{dk}^m\rangle)\right)$$

$$\mathbb{E}_q[\log q(\hat{\mathbf{W}},\mathbf{S})]=-\sum_{m=1}^{M}\left(\frac{KD_m}{2}\log(2\pi)+\frac{1}{2}\sum_{d=1}^{D_m}\sum_{k=1}^{K}\log\left(\langle s_{dk}^m\rangle\sigma_{w_{dk}^m}^2+\frac{1-\langle s_{dk}^m\rangle}{\alpha_k^m}\right)\right.$$
$$\left.+\sum_{d=1}^{D_m}\sum_{k=1}^{K}(1-\langle s_{dk}^m\rangle)\log(1-\langle s_{dk}^m\rangle)-\langle s_{dk}^m\rangle\log\langle s_{dk}^m\rangle\right)$$

- $\mathbf{Z}$ *terms:*

$$\mathbb{E}_q[\log p(\mathbf{Z})]=-\frac{NK}{2}\log(2\pi)-\frac{1}{2}\sum_{n=1}^{N}\langle z_{nk}^2\rangle$$
$$\mathbb{E}_q[\log q(\mathbf{Z})]=-\frac{NK}{2}(1+\log(2\pi))-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\log(\sigma_{z_{nk}}^2)$$

- $\boldsymbol{\alpha}$ *terms:*

$$\mathbb{E}_q[\log p(\boldsymbol{\alpha})]=\sum_{m=1}^{M}\sum_{k=1}^{K}\left(a_0^\alpha\log b_0^\alpha+(a_0^\alpha-1)\langle\log\alpha_k\rangle-b_0^\alpha\langle\alpha_k\rangle-\log\Gamma(a_0^\alpha)\right)$$
$$\mathbb{E}_q[\log q(\boldsymbol{\alpha})]=\sum_{m=1}^{M}\sum_{k=1}^{K}\left(\hat{a}_{mk}^\alpha\log\hat{b}_{mk}^\alpha+(\hat{a}_{mk}^\alpha-1)\langle\log\alpha_k\rangle-\hat{b}_{mk}^\alpha\langle\alpha_k\rangle-\log\Gamma(\hat{a}_{mk}^\alpha)\right)$$

- $\boldsymbol{\tau}$ *terms:*

$$\mathbb{E}_q[\log p(\tau)]=\sum_{m=1}^{M}\left(D_m a_0^\tau\log b_0^\tau+\sum_{d=1}^{D_m}(a_0^\tau-1)\langle\log\tau_d^m\rangle\right.$$
$$\left.-\sum_{d=1}^{D_m}b_0^\tau\langle\tau_d^m\rangle-D_m\log\Gamma(a_0^\tau)\right)$$
$$\mathbb{E}_q[\log q(\boldsymbol{\tau})]=\sum_{m=1}^{M}\sum_{d=1}^{D_m}\left(\hat{a}_{md}^\tau\log\hat{b}_{md}^\tau+(\hat{a}_{md}^\tau-1)\langle\log\tau_d^m\rangle-\hat{b}_{md}^\tau\langle\tau_d^m\rangle\right.$$
$$\left.-\log\ \Gamma(\hat{a}_{md}^\tau)\right)$$

- **$\boldsymbol{\theta}$** *terms:*

$$\mathbb{E}_q\left[\log p(\theta)\right] = \sum_{m=1}^{M}\sum_{k=1}^{K}\Big((a_0^\theta - 1)\langle\log(\theta_k^m)\rangle + (b_0^\theta - 1)\langle\log(1-\theta_k^m)\rangle - \log(\mathrm{B}(a_0^\theta, b_0^\theta))\Big)$$

$$\mathbb{E}_q\left[\log q(\theta)\right] = \sum_{m=1}^{M}\sum_{k=1}^{K}\Big((\hat{a}_{mk}^\theta - 1)\langle\log(\theta_k^m)\rangle + (\hat{b}_{mk}^\theta - 1)\langle\log(1-\theta_k^m)\rangle - \log(\mathrm{B}(\hat{a}_{mk}^\theta, \hat{b}_{mk}^\theta))\Big)$$

4.A.3 Modelling and inference with non-Gaussian data

To implement efficient variational inference in conjunction with a non-Gaussian likelihood we adapt prior work from [168] using local variational bounds. The key idea is to dynamically approximate non-Gaussian data by Gaussian pseudo-data based on a second-order Taylor expansion. To make the approximation justifiable we need to introduce variational parameters that are adjusted alongside the updates to improve the fit. Denoting the parameters in the MOFA model as $\mathbf{X} = (\mathbf{Z}, \mathbf{W}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau})$, recall that the variational framework approximates the posterior $p(\mathbf{X}|\mathbf{Y})$ with a distribution $q(\mathbf{X})$, which is indirectly optimised by optimising a lower bound of the log model evidence. The resulting optimization problem can be re-written from Equation (4.14) as

$$\min_{q(\mathbf{X})} -\mathcal{L}(\mathbf{X}) = \min_{q(\mathbf{X})} \mathbb{E}_q\big[-\log p(\mathbf{Y}|\mathbf{X})\big] + \mathrm{D}_{\mathrm{KL}}[q(\mathbf{X})||p(\mathbf{X})].$$

Expanding the MOFA model to non-Gaussian likelihoods we now assume a general likelihood of the form $p(\mathbf{Y}|\mathbf{X}) = p(\mathbf{Y}|\mathbf{C})$ with $\mathbf{C} = \mathbf{Z}\mathbf{W}^T$, that can be written as

$$-\log p(\mathbf{Y}|\mathbf{X}) = \sum_{n=1}^{N}\sum_{d=1}^{D} f_{nd}(c_{nd}),$$

with $f_{nd}(c_{nd}) = -\log p(y_{nd}|c_{nd})$. Note that we dropped the view index m to keep notation uncluttered.

Extending [168] to our heteroscedastic noise model, we require $f_{nd}(c_{nd})$ to be twice differentiable and bounded by κ_d, such that $f''_{nd}(c_{nd}) \leq \kappa_d \,\forall n, d$. This holds true in many important models as for example the Bernoulli and Poisson case. Under this assumption a lower bound on the log likelihood can be constructed using Taylor expansion,

$$f_{nd}(c_{nd}) \leq \frac{\kappa_d}{2}(c_{nd} - \zeta_{nd})^2 + f'(\zeta_{nd})(c_{nd} - \zeta_{nd}) + f_{nd}(\zeta_{nd}) =: q_{nd}(c_{nd}, \zeta_{nd}),$$

where $\boldsymbol{\zeta} = \zeta_{nd}$ are additional variational parameters that determine the location of the Taylor expansion and have to be optimised to make the lower bound as tight as possible. Plugging the bounds into above optimization problem, we obtain

$$\min_{q(\mathbf{X}),\boldsymbol{\zeta}} \quad \sum_{d=1}^{D}\sum_{n=1}^{N} \mathbb{E}_q[q_{nd}(c_{nd}, \zeta_{nd})] + \mathrm{D}_{\mathrm{KL}}[q(\mathbf{X})||p(\mathbf{X})].$$

The algorithm proposed in [168] then alternates between updates of $\boldsymbol{\zeta}$ and $q(\mathbf{X})$. The update for $\boldsymbol{\zeta}$ is given by

$$\boldsymbol{\zeta} \leftarrow \mathbb{E}[\mathbf{W}]\mathbb{E}[\mathbf{Z}]^T,$$

where the expectations are taken with respect to the corresponding q distributions. On the other hand, the updates for $q(\mathbf{X})$ can be shown to be identical to the variational Bayesian updates with a conjugate Gaussian likelihood when replacing the observed data $\mathbf{Y}$ by a pseudo-data $\hat{\mathbf{Y}}$ and the precisions τ_{nd} (which were treated as random variables) by the constant terms κ_d introduced above. The pseudo-data is given by

$$\hat{y}_{nd} = \zeta_{nd} - f'(\zeta_{nd})/\kappa_d.$$

Depending on the log likelihood $f(\cdot)$, different κ_d are used resulting in different pseudo-data updates. Two special cases implemented in MOFA are the Poisson and Bernoulli likelihood described in the following.

4.A.3.1 Bernoulli likelihood for binary data

When the observations are binary, $y \in \{0, 1\}$, they can be modelled using a Bernoulli likelihood, i. e.

$$\mathbf{Y}|\mathbf{Z}, \mathbf{W} \sim \mathrm{Ber}(\sigma(\mathbf{Z}\mathbf{W}^T)),$$

where $\sigma(a) = (1 + e^{-a})^{-1}$ is the logistic link function and $\mathbf{Z}$ and $\mathbf{W}$ are the latent factors and weights in our model, respectively. In order to make the variational inference efficient and explicit as in the Gaussian case, we aim to approximate the Bernoulli data by a Gaussian pseudo-data as proposed in [168] and described above. This allows to recycle all the updates from the model with Gaussian views. While [168] assumes a homoscedastic approximation with a spherical Gaussian, we adopt an approach following [101], which allows for heteroscedasticity and provides a tighter bound on the Bernoulli likelihood.

Denoting $c_{nd} = (\mathbf{Z}\mathbf{W}^T)_{nd}$ the Jaakkola upper bound [101] on the negative log-likelihood is given by

$$\begin{aligned}
-\log\left(p(y_{nd}|c_{nd})\right) &= -\log\left(\sigma\left((2y_{nd}-1)c_{nd}\right)\right) \\
&\leq -\log(\zeta_{nd}) - \frac{(2y_{nd}-1)c_{nd} - \zeta_{nd})}{2} + \lambda(\zeta_{nd})\left(c_{nd}^2 - \zeta_{nd}^2\right) \\
&=: b_J(\zeta_{nd}, c_{nd}, y_{nd})
\end{aligned}$$

with λ given by $\lambda(\zeta) = \frac{1}{4\zeta}\tanh\left(\frac{\zeta}{2}\right)$. This can be derived from a first-order Taylor expansion on the function $f(x) = -\log(e^{\frac{x}{2}} + e^{-\frac{x}{2}}) = \frac{x}{2} - \log(\sigma(x))$ in x^2 and by the convexity of f in x^2 this bound is global as discussed in [101].

In order to make use of this tighter bound but still be able to re-use the variational updates from the Gaussian case we re-formulate the bound as a Gaussian likelihood on pseudo-data $\hat{\mathbf{Y}}$.

As above we can plug this bound on the negative log-likelihood into the variational optimization problem to obtain

$$\min_{q(\mathbf{X}),\boldsymbol{\zeta}} \quad \sum_{d=1}^{D}\sum_{n=1}^{N} \mathbb{E}_q b_J(\zeta_{nd}, c_{nd}, y_{nd}) + \mathrm{D_{KL}}[q(\mathbf{X})||p(\mathbf{X})].$$

This is minimized iteratively in the variational parameter ζ_{nd} and the variational distribution of $\mathbf{Z}, \mathbf{W}$: Minimizing in the variational parameter $\boldsymbol{\zeta}$ leads to the updates given by

$$\zeta_{nd}^2 = \mathbb{E}[c_{nd}^2]$$

as described in [101], [20]. For the variational distribution $q(\mathbf{Z}, \mathbf{W})$ we observe that the Jaakkola bound can be re-written as

$$b_J(\zeta_{nd}, c_{nd}, y_{nd}) = -\log\left(\varphi\left(\hat{y}_{nd}; c_{nd}, \frac{1}{2\lambda(\zeta_{nd})}\right)\right) + \gamma(\zeta_{nd}),$$

where $\varphi(\cdot; \mu, \sigma^2)$ denotes the density function of a normal distribution with mean μ and variance σ^2 and γ is a term only depending on $\boldsymbol{\zeta}$. This allows us to re-use the updates for $\mathbf{Z}$ and $\mathbf{W}$ from a setting with Gaussian likelihood by considering the Gaussian pseudo-data

$$\hat{y}_{nd} = \frac{2y_{nd} - 1}{4\lambda(\zeta_{nd})},$$

updating the data precision as $\tau_{nd} = 2\lambda(\zeta_{nd})$ using updates generalized for sample- and feature-wise precision parameters on the data.

4.A.3.2 Poisson likelihood for count data

When observations are natural numbers, such as count data $y \in \mathbb{N}_0$, they can be modelled using a Poisson likelihood, i. e.

$$p(y|c) = \lambda(c)^y e^{-\lambda(c)},$$

where $\lambda(c) > 0$ is the rate function and has to be convex and log-concave in order to ensure that the likelihood is log-concave. Following [168], here we choose the rate function

$$\lambda(c) = \log(1 + e^c).$$

Then, an upper bound of the second derivative of the log-likelihood is given by

$$f''_{nd}(c_{nd}) \leq \kappa_d = 1/4 + 0.17 \max(\mathbf{y}_{:,d}),$$

and the pseudo-data updates are given by

$$\hat{y}_{nd} = \zeta_{nd} - \frac{\sigma(\zeta_{nd})(1 - y_{nd}/\lambda(\zeta_{nd}))}{\kappa_d}.$$

4.A.4 Implementation and practical considerations for training

4.A.4.1 Monitoring convergence

In contrast to sampling methods, variational approximations have the appealing property that convergence is easily monitored by changes in the ELBO, which is required to increase monotonically [20]. In practice, we set a default threshold for convergence corresponding to a change in ELBO smaller than 0.1%.

4.A.4.2 Handling of missing values

The model naturally accounts for missing values and no prior imputation is required. Non-observed data points do not intervene in the likelihood and are ignored in the update equations. In practice, we use a binary mask $\mathcal{O}^m \in \mathbb{R}^{N \times D_m}$ for each view m, such that $\mathcal{O}_{n,d} = 1$ when feature d is observed for sample n, 0 otherwise.

4.A.4.3 Data pre-processing

MOFA does not require the data to be centered or scaled. The first property is achieved by incorporating a constant factor of ones that will capture any feature-wise intercept effect. This ensures that the rest of the factors capture variation independent of the feature-wise means. The second property is achieved by the factor- and view-wise ARD prior, which allows different scales of the weights for each view. However, when using the Gaussian noise model, it is recommended to use methods for normalization and variance stabilisation (e.g. as implemented in [127] for RNAseq data) prior to model training. This makes the normality assumption of the model residuals more appropriate.

4.A.4.4 Consistency across random initializations

The variational Bayes algorithm is not guaranteed to find the optimal solution [20] and the estimates will depend on the parameter initialization. We suggest to adopt common practice [91] and assess the consistency of factors by running MOFA multiple times (e.g. 10 trials) under different initialisations. Subsequently, a single model with the highest ELBO should be selected for downstream analysis. Appropriate functions for model selection are provided in the R package MOFAtools.

4.A.4.5 Determining the number of factors

The model can automatically learn the number of factors by removing inactive factors during training if they do not explain significant variation in any view. This is achieved by the view- and factor-wise ARD prior (Eq. (4.10)). In practice, factors are pruned during training using a minimum fraction of variance explained threshold that needs to be specified by the user. Alternatively, the user can fix the number of factors and the minimum variance criterion is ignored.

4.A.4.6 Rotational invariance

An important consequence of the definition of MOFA (and most factor analysis models [14, 197]) is their unidentifiability due to rotational and scaling invariance. This means that the factors and corresponding loadings can only be identified up to an orthogonal rotation. In practice, this property implies that the actual factor and weight values need to be interpreted in a relative manner, always within the same model instance.

4.A.5 Supplementary Figures

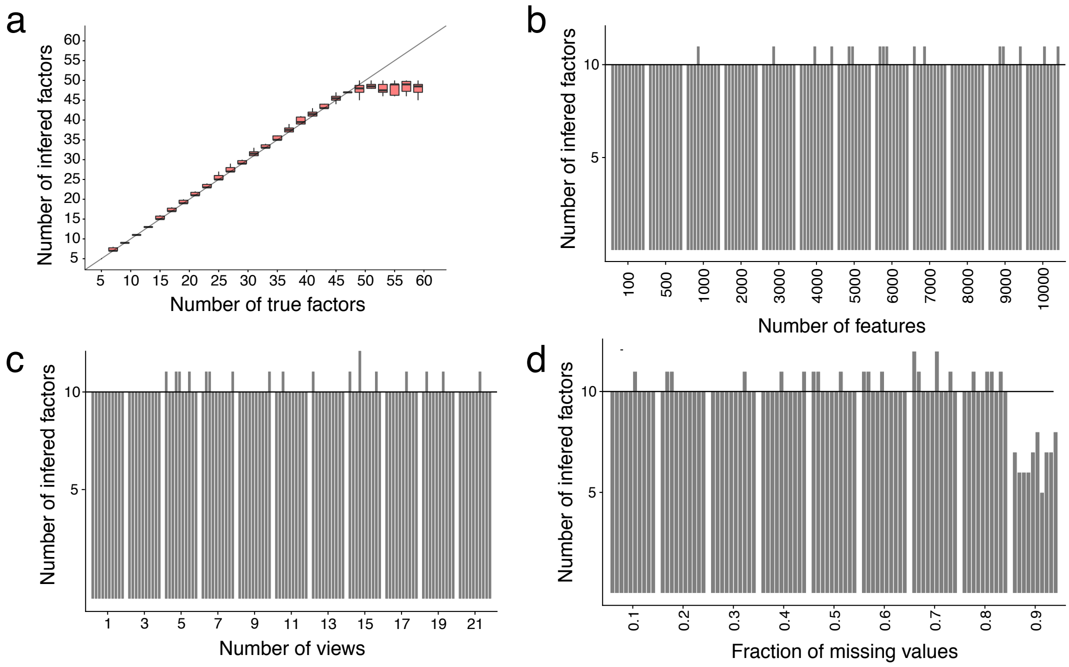


Figure 4.A.1: Model validation of MOFA using simulated data. (a) Comparison of the number of simulated and estimated factors. Boxplots show the distribution across 10 model instances. **(b-d)** Recovery of the true number of latent factors ($K = 10$) under different number of **(b)** features, **(c)** views and **(d)** fraction of missing values. Individual bars correspond to different model instances.

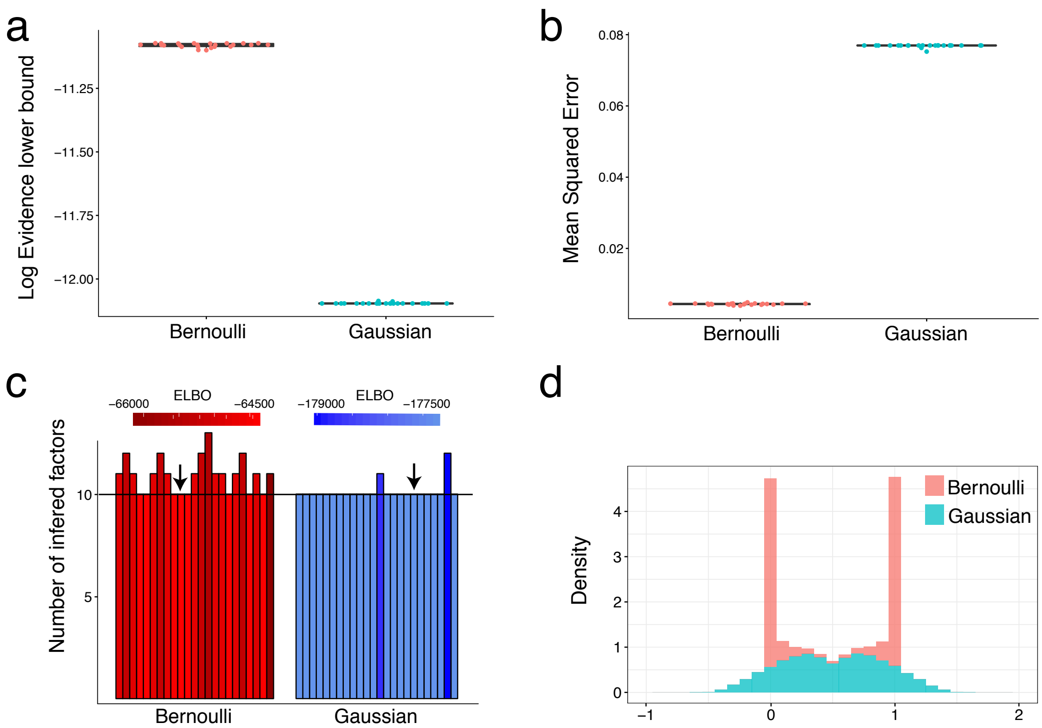


Figure 4.A.2: Validation of the Bernoulli likelihood model. On simulated binary data 25 instances of a MOFA model were trained considering a Bernoulli (red) or Gaussian (blue) likelihood, respectively. **(a)** Variational evidence lower bound (ELBO) for each model instance. **(b)** Reconstruction error for each model instance. **(c)** Number of estimated factors. The horizontal line denotes the true number of factors ($K = 10$). Individual model instances are coloured based on their respective ELBO value. The arrows mark the models with the highest ELBO that would be selected for downstream analysis. **(d)** Distribution of the reconstructed data, with the Bernoulli (red) or Gaussian (blue) likelihood model, respectively.

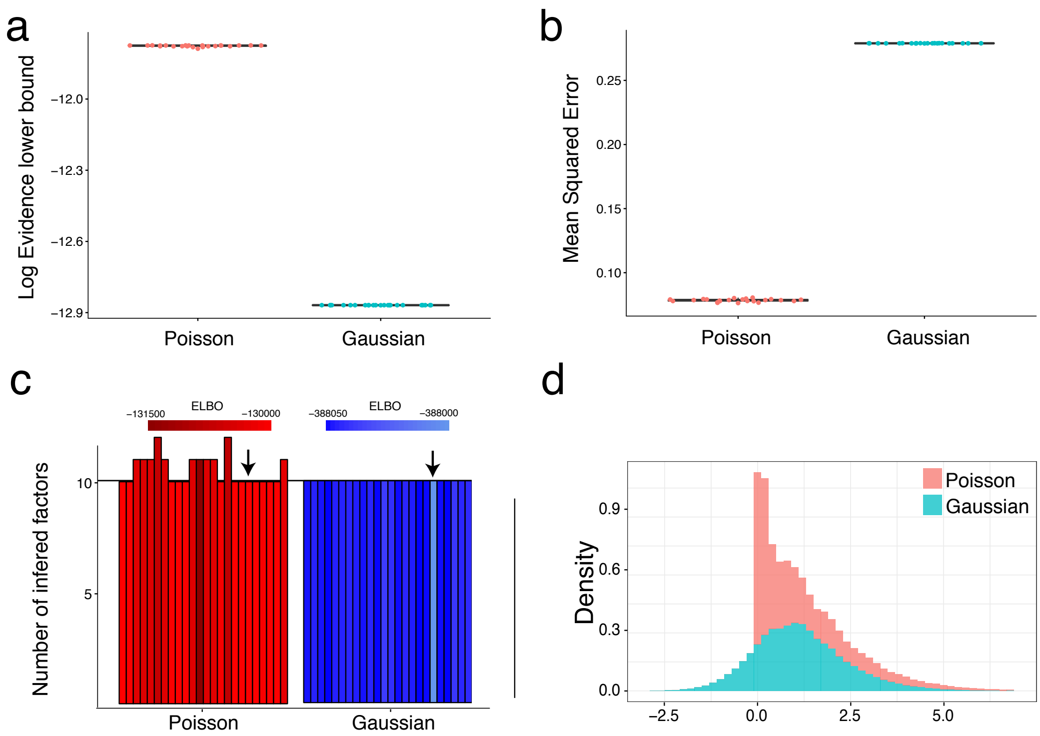


Figure 4.A.3: Validation of the Poisson likelihood model. On simulated count data 25 instances of a MOFA model were trained considering a Poisson (red) or Gaussian (blue) likelihood, respectively. **(a)** Variational evidence lower bound (ELBO) for each model instance. **(b)** Reconstruction error for each model instance. **(c)** Number of estimated factors. The horizontal line denotes the true number of factors ($K = 10$). Individual model instances are coloured based on their respective ELBO value. The arrows mark the models with the highest ELBO that would be selected for downstream analysis. **(d)** Distribution of the reconstructed data, with the Poisson (red) or Gaussian (blue) likelihood model, respectively.

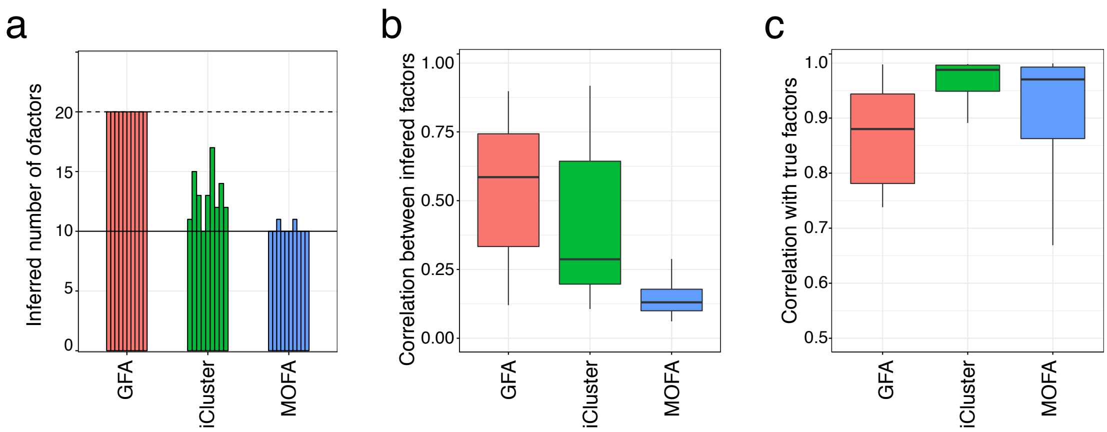


Figure 4.A.4: Comparison of MOFA, GFA and iCluster on simulated data. (a) Estimated number of factors. The solid horizontal line denotes the true number of simulated factors ($K = 10$) and the dashed horizontal line indicates the initial number of factors ($K = 20$). Each bar represents a different model realization of the simulated data. **(b)** Pearson correlation coefficient between pairs of inferred latent factors for individual trials. For each factor, the maximum correlation coefficient with any of the remaining factors is shown. Factors were simulated to be uncorrelated. **(c)** Pearson correlation coefficient between true and inferred factors (for the top ten factors in each fit). For each factor, shown is the maximum correlation coefficient with any of the true factors.

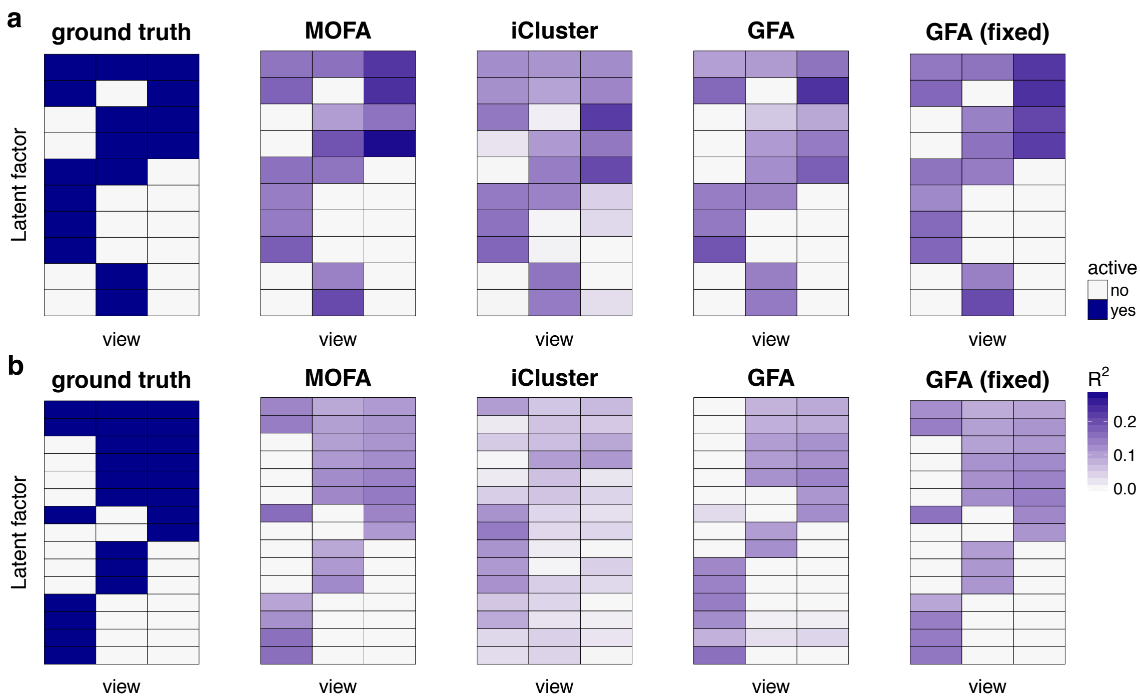


Figure 4.A.5: Assessment of MOFA, iCluster and GFA in terms of recovering the pattern of factor activity across views. Data were simulated using a Gaussian likelihood based on the true activity pattern of factors per view displayed on the left. **(a)** Number of true underlying factors $K = 10$. **(b)** Number of true underlying factors $K = 15$. In both cases, MOFA and GFA were trained starting with $K = 25$ factors whereas for iCluster and GFA (fixed) the true number of factors was used. Shown is the fraction of variance explained for each factor and view.

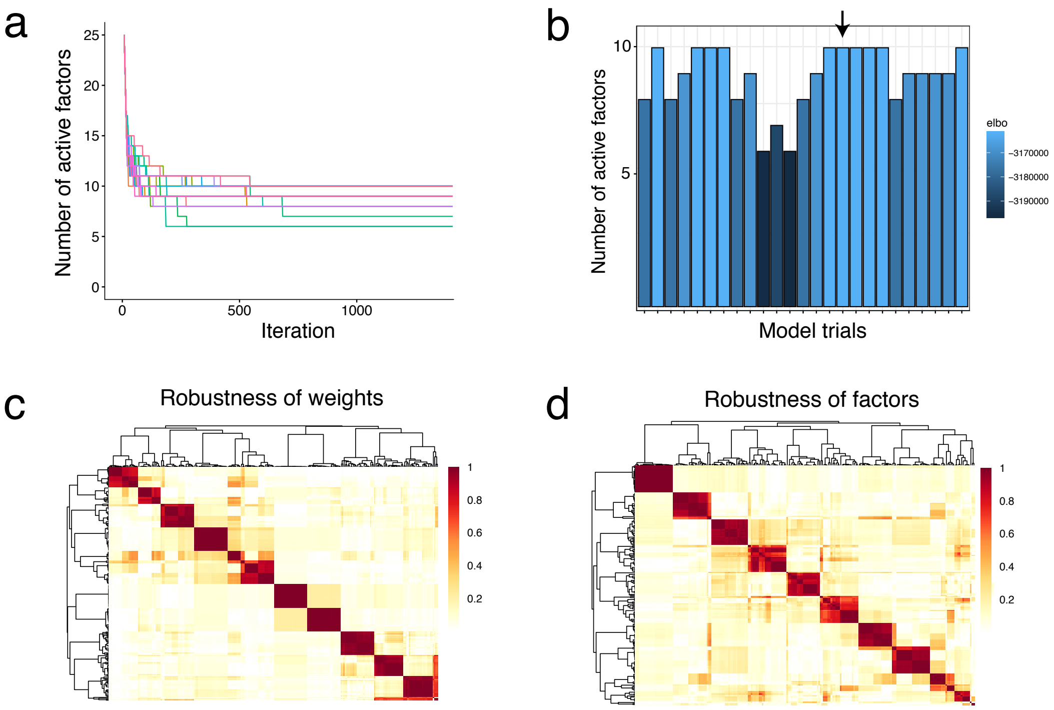


Figure 4.A.6: Assessment of model consistency across different trials. 25 MOFA instances were trained on the CLL data. **(a)** The training curve for the number of active factors. **(b)** The number of estimated factors for each trial, coloured by the corresponding evidence lower bound (ELBO). The arrow indicates the model with highest ELBO that was selected for downstream analysis. **(c)** Absolute value of the Pearson correlation coefficient between the weights of the mRNA data. Each block in the diagonal captures a weight vector consistently learnt across multiple trials. **(d)** Absolute value of the Pearson correlation coefficient between the factors. Each block in the diagonal captures a latent factor consistently learnt across multiple trials.

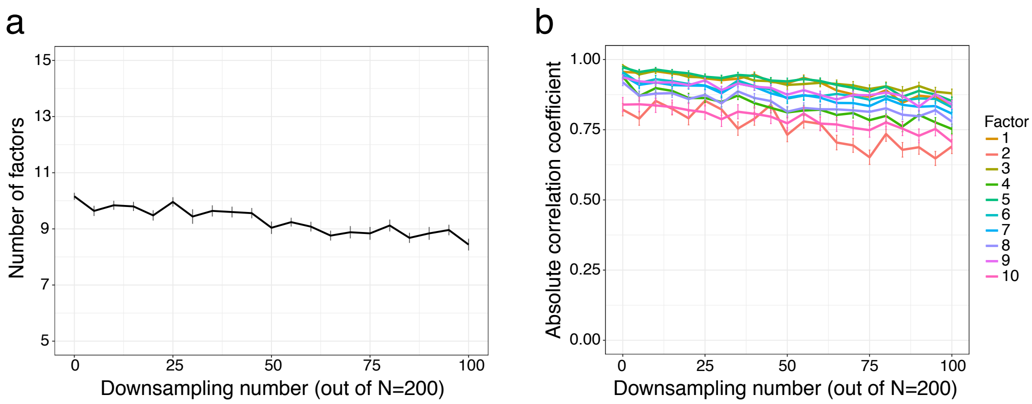


Figure 4.A.7: Model robustness on the CLL data assessed using downsampling of samples. Shown is **(a)** the number of factors and **(b)** the absolute Pearson correlation coefficient between factors estimated on downsampled data and the factors estimated on the full dataset. Shown are averages across 25 trials. Error bars denote plus or minus one standard error.

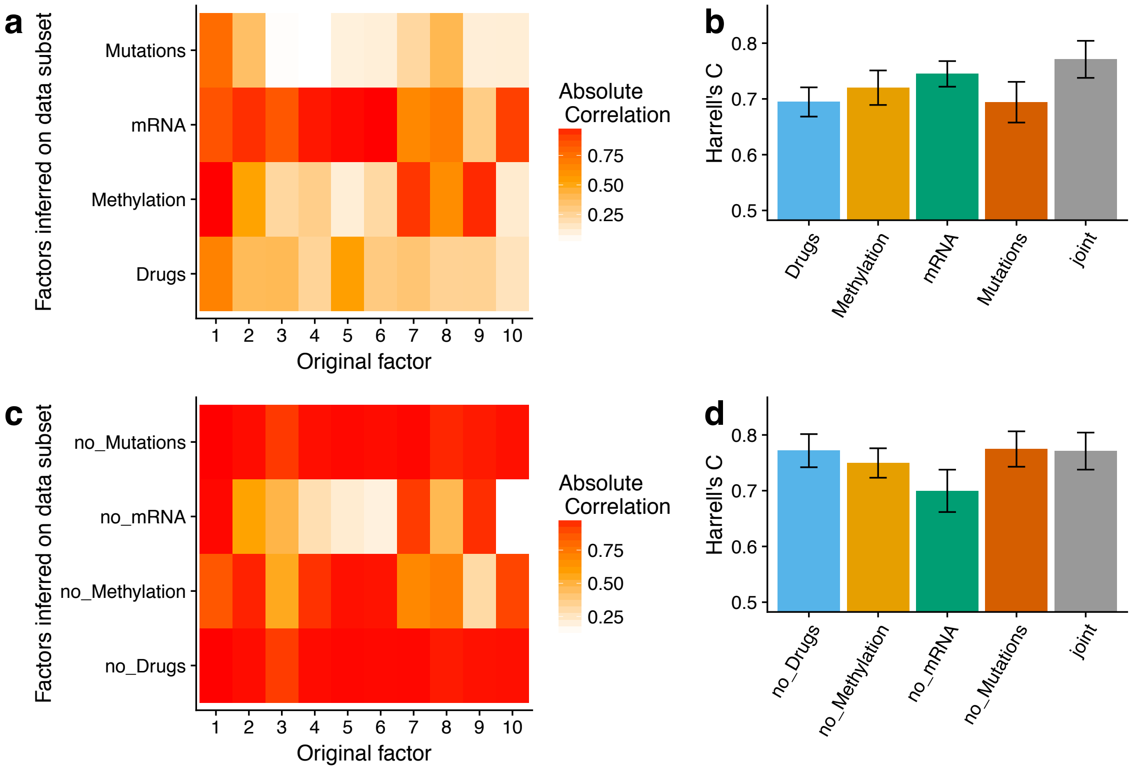


Figure 4.A.8: MOFA trained on a subset of the available assays. **(a)** Absolute correlation between the MOFA factors recovered on the full data sets (x-axis) with the most associated factor recovered when using only one data modality (y-axis). Correlation is calculated on the $N = 121$ samples that were profiled in all assays. **(b)** Harrell's C-index for prediction of time to next treatment for the $N = 121$ samples with data in all modalities using 10 factors obtained from MOFA on each single data modality as well as the full data. **(c)** Same as in **(a)** for MOFA trained on all assays except one. **(d)** Same as in **(b)** for MOFA trained on all assays except one.

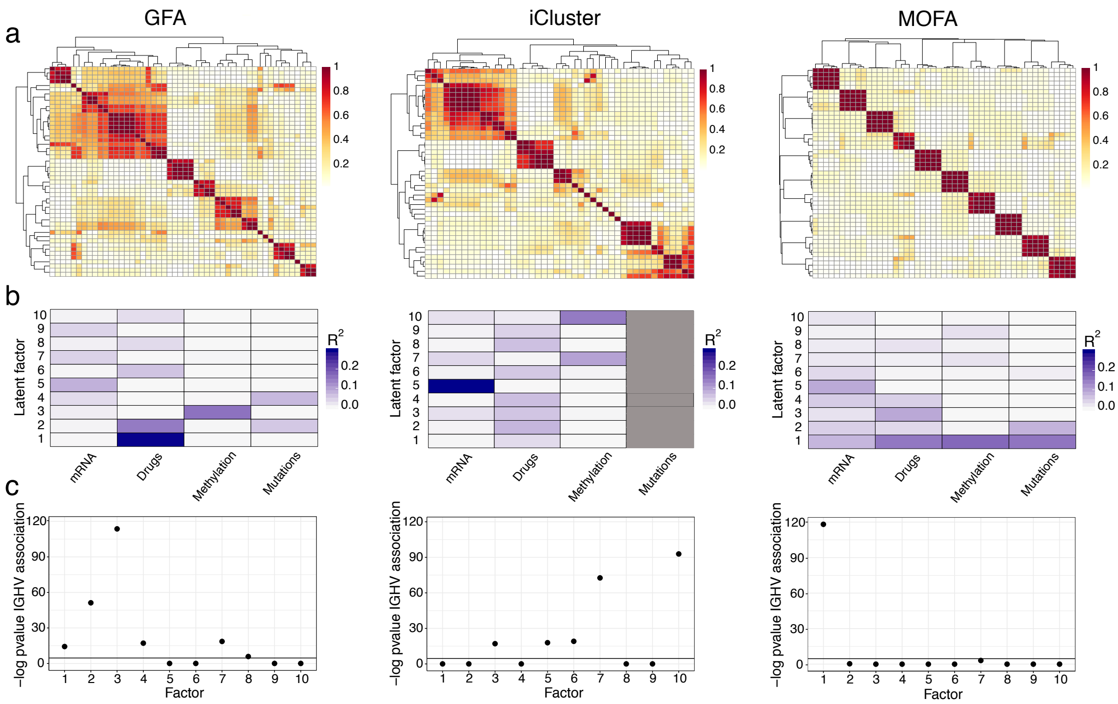


Figure 4.A.9: Performance of MOFA, GFA and iCluster on the CLL data. **(a)** Consistency of inferred factors across multiple trials. Shown are absolute Pearson correlation coefficients between pairs of factors in different trials. Each diagonal block captures a factor that is consistently learnt across multiple trials. For the first trial of each model, we show: **(b)** Fraction of variance explained (R^2) by individual factors for each view. No variance measure can be estimated in the (binary) mutation data by the iCluster method. **(c)** Negative log FDR-adjusted p-values from the association analysis (t-test) between individual factors and IGHV status. The line denotes the statistical significance threshold of 1% FDR.

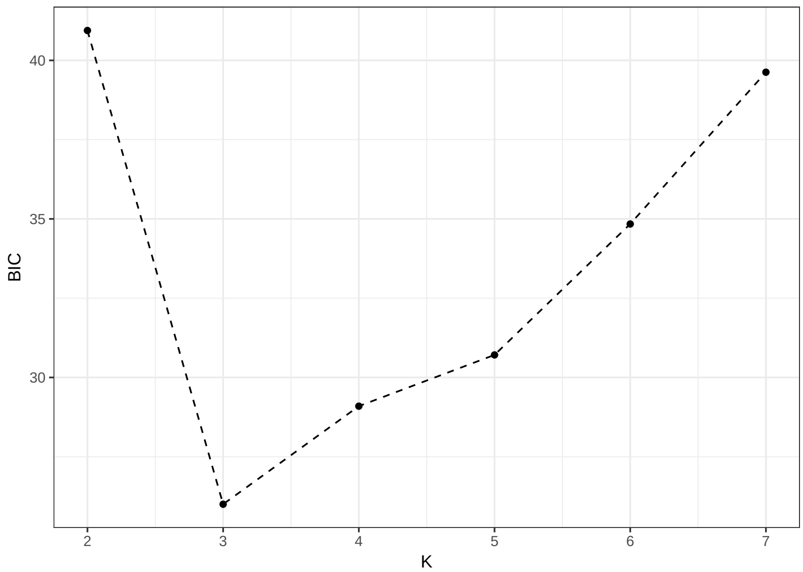
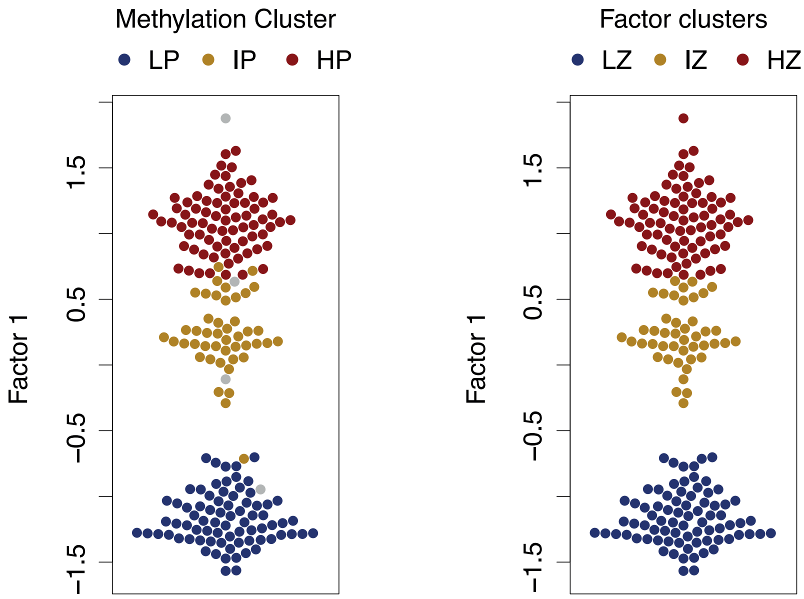


Figure 4.A.10: BIC for the K-Means clustering on Factor 1 in the CLL data. Values of the Bayesian information criterion (BIC) for different values of K in the K-means clustering on Factor 1. A minimum is obtained for $K = 3$.



	low Factor 1 (LZ)	intermediate Factor 1 (IZ)	high Factor 1 (HZ)
low-programmed (LP)	76	0	0
intermediate-programmed (IP)	1	42	2
high-programmed (HP)	0	0	75
missing	1	2	1
total	**78**	**44**	**78**
Drug response data present	73 (93.5%)	41 (93.1%)	70 (89.7%)
Methylation data present	77 (98.7%)	43 (97.7%)	76 (97.4%)
Mutation data present	78 (100%)	44 (100%)	78 (100%)
RNAseq data present	54 (69.2 %)	28 (63.6%)	54 (69.2%)

Figure 4.A.11: Correspondence of patient clusters on Factor 1 with previously described CLL subgroups. Beeswarm plots of Factor 1 coloured by previously described CLL subgroups [148] (left) and 3-means clusters (right) as in Figure 4.4a. The table below indicates the sample numbers that fall in each of these clusters as well as the number of samples in each cluster with data available for the given omics type.

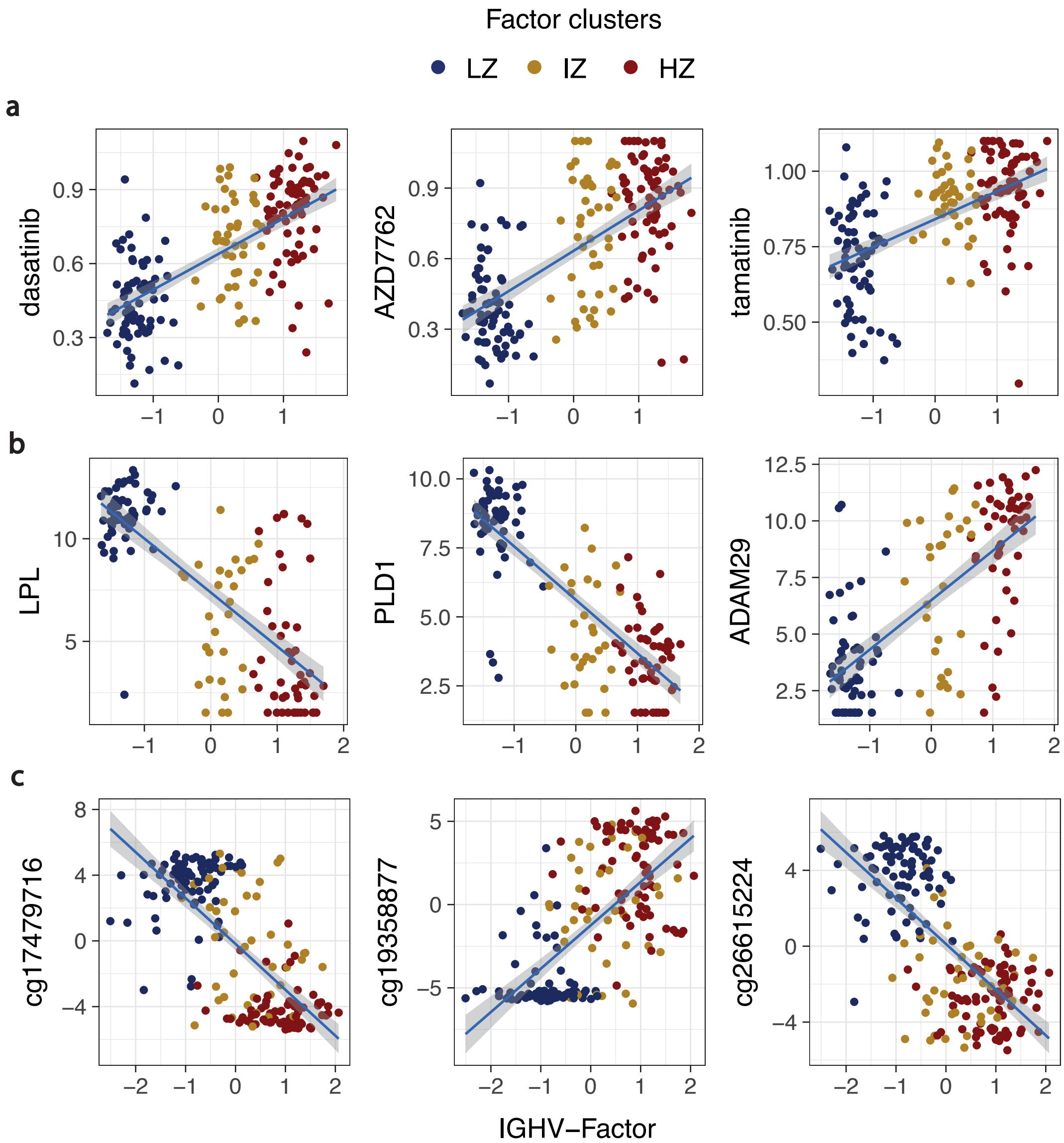


Figure 4.A.12: Correlation between the continuous IGHV state inferred by MOFA (Factor 1) and individual molecular features. Scatterplots showing the correlation between the continuous Factor 1 inferred by MOFA and molecular features. To avoid circularity, models were re-trained holding out different data modalities in turn: **(a)** drug response, **(b)** gene expression and **(c)** methylation. Colours denote cluster assignments using the factor obtained from the full data set. Displayed are representative features with high absolute loading on the Factor 1 from the full model.

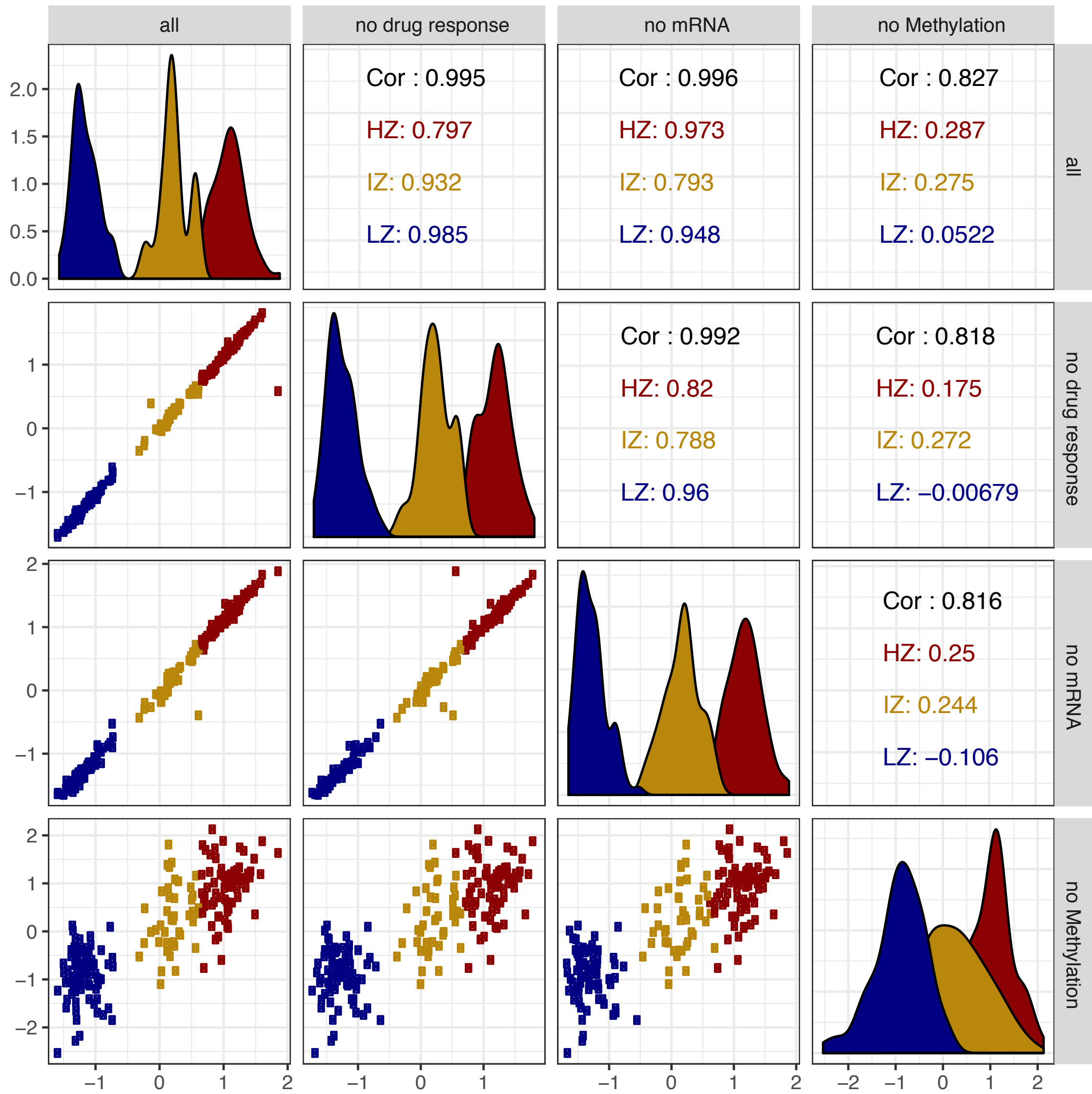


Figure 4.A.13: Correlation between the continuous IGHV state inferred by MOFA (Factor 1) trained on different subset of data modalities. Scatterplots on the lower panels show the pairwise correlation between the continuous Factor 1 inferred by MOFA when training on all assays, without the drug response assay, mRNA assay and methylation assay, respectively. Colours are based on the clusters on Factor 1 inferred by the full model (trained on all data modalities). The panels on the diagonal show the densities of factor values of the 3 different clusters in each setting and the upper panels denote the overall and within-cluster correlation.

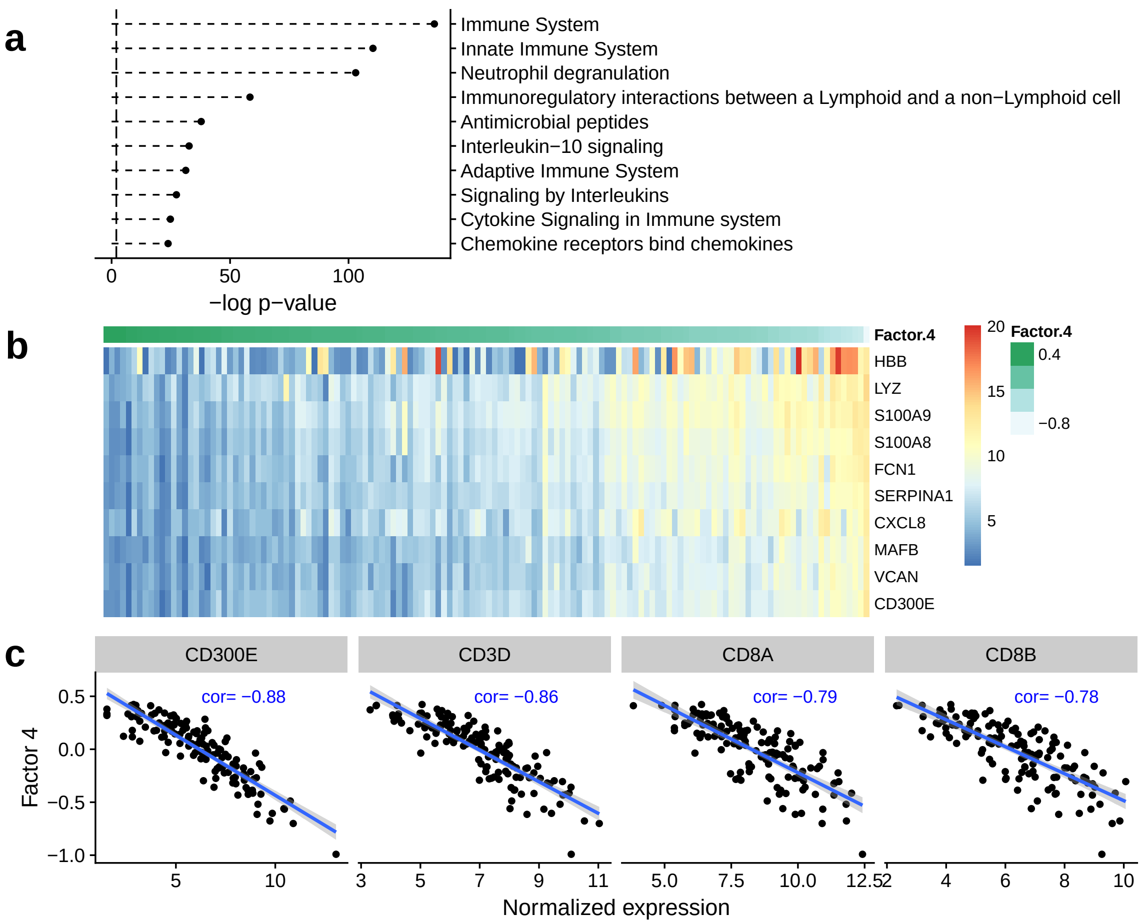


Figure 4.A.14: Characterization of Factor 4 in the CLL data. (a) Gene sets of the Reactome pathways enriched in the loadings of the mRNA data in Factor 4 (t-test, Methods Section 4.4.4, dashed line represents a FDR of 1%). **(b)** Heatmap of the mRNA data in the top ten features for Factor 4. Samples are ordered along their value on Factor 4 as shown on top of the heatmap. **(c)** Scatterplot of the normalized expression of important surface markers of T-cells (CD8A, CD8B, CD3D) and monocytes (CD300E) versus the values on Factor 4.

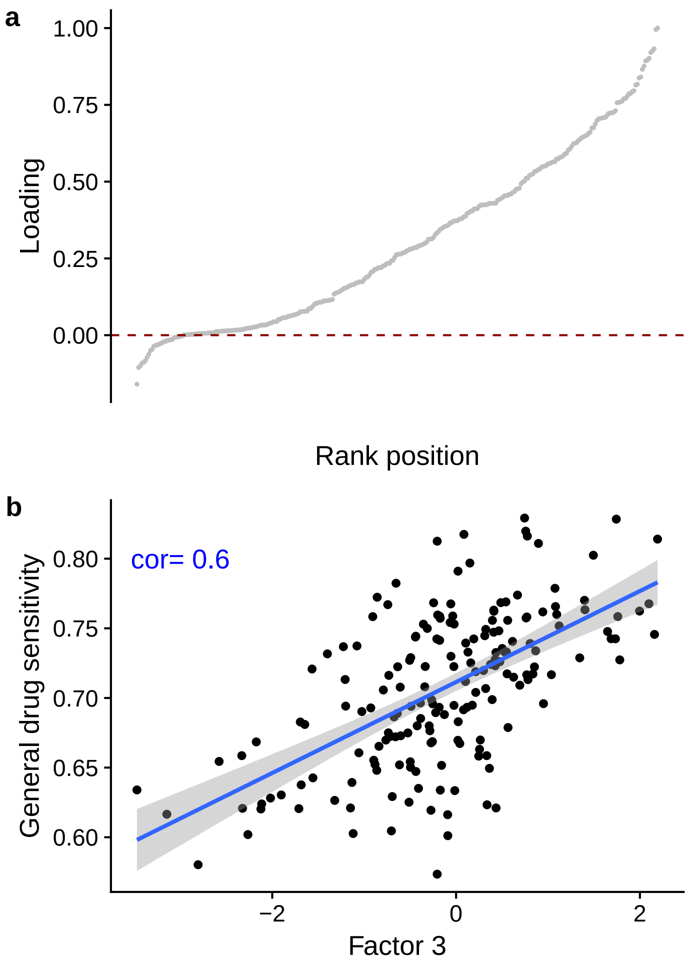


Figure 4.A.15: Characterization of Factor 3 in the CLL data. (a) Loadings of all drugs and concentrations on Factor 3. **(b)** Scatterplot of Factor 3 versus a general level of drug sensitivity calculated as the mean viability of a sample across all drugs and concentrations.

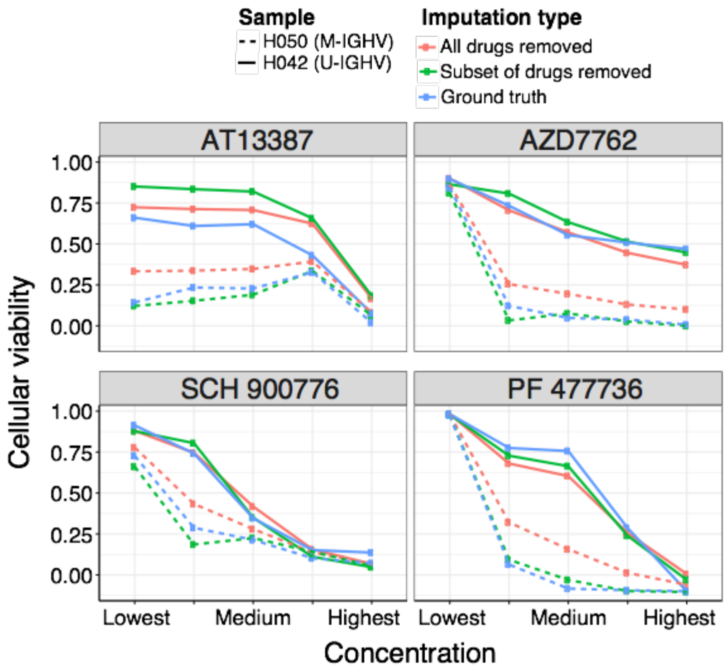


Figure 4.A.16: Prediction of drug response curves in the CLL data. Prediction of drug response curves for two samples clinically annotated as M-CLL (M-IGHV, H050, dashed line) and U-CLL (U-IGHV, H052, solid line), respectively, for four representative drugs known to be affected by IGHV status. Scatterplots show the predicted drug response curve as cellular viability versus concentration when training a MOFA model removing all drugs from the corresponding patients (red) and removing only the four drugs (green). The true response curve is shown in blue.

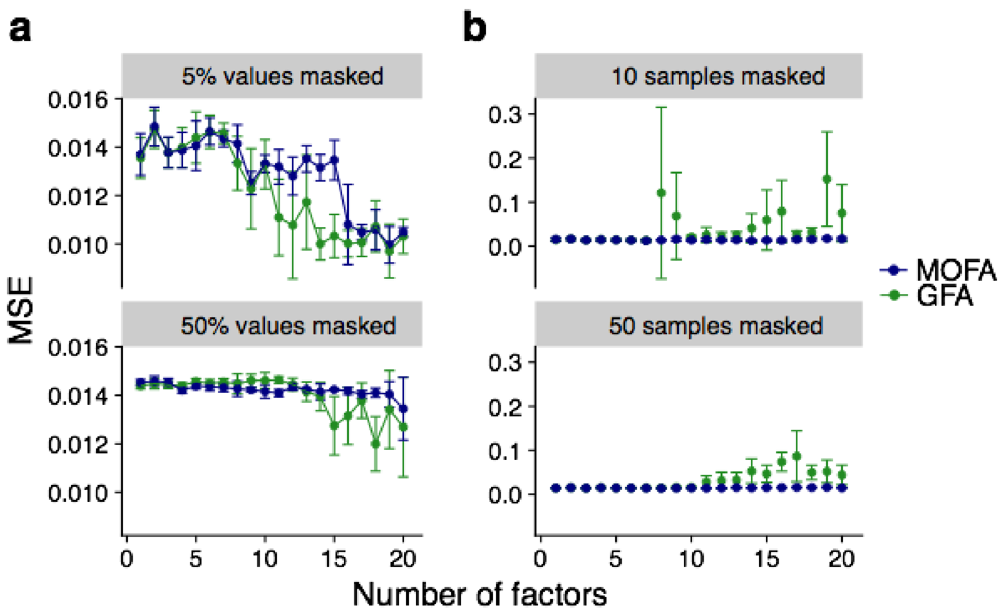


Figure 4.A.17: Comparison of the accuracy of MOFA and GFA for imputing missing values in the drug response assay of the CLL data. GFA and MOFA models were trained with different numbers of factors. Shown are averages of the mean squared error (MSE) across 5 imputation experiments for different fractions of missing data, considering **(a)** values missing at random (top panel: 5%; bottom panel: 50%) and **(b)** entire assay missing for samples at random (top panel: $N = 10$; bottom panel: $N = 50$). Error bars denote plus or minus two standard error.

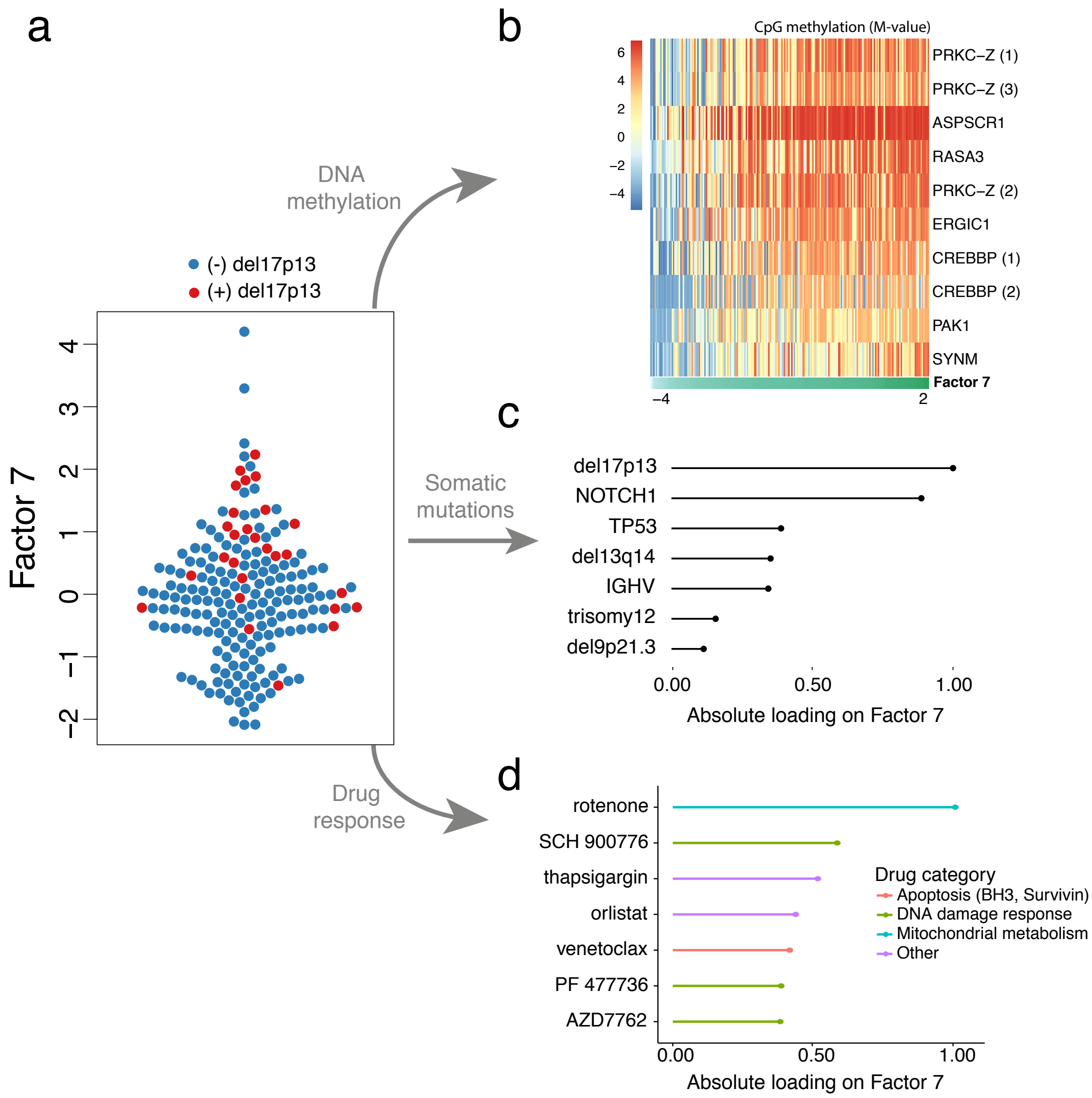


Figure 4.A.18: Characterisation of Factor 7 in the CLL data. **(a)** Beeswarm plot with Factor 7 values for each sample. Colours denote the presence or absence of the deletion del17p13. **(b)** Heatmap of methylation (M-value) for CpG sites with the largest absolute loading (matched to overlapping genes). **(c)** Absolute loadings of top features in the somatic mutation data. **(d)** Absolute loadings of the top features in the drug response data.

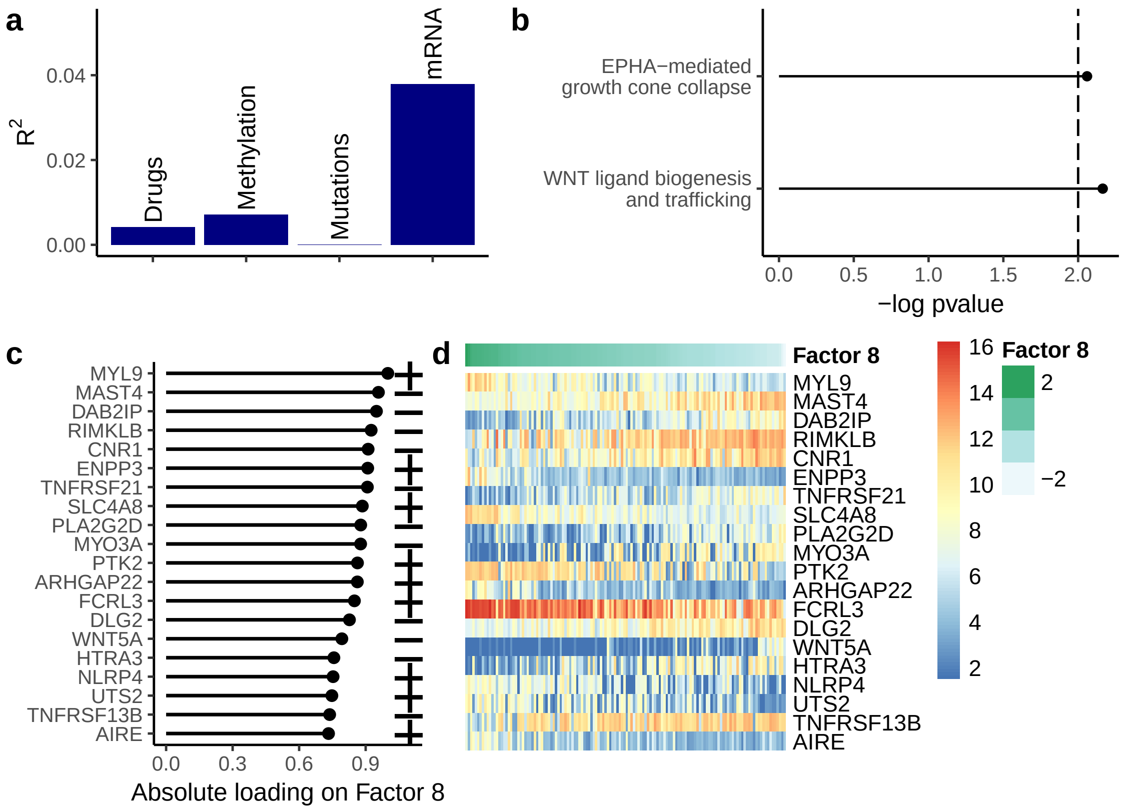


Figure 4.A.19: Characterization of Factor 8 in the CLL data. (a) Variance explained by Factor 8 in the four assays. **(b)** Gene sets enriched for the Reactome pathways in the loadings of the mRNA data at a FDR of 1% (t-test, Methods Section 4.4.4). **(c)** Absolute values for the weights of top 20 genes in the mRNA data, sign indicating the direction of their effect. **(d)** Heatmap of the normalized expression values for the genes shown in **(c)**, samples are ordered along their values on Factor 8.

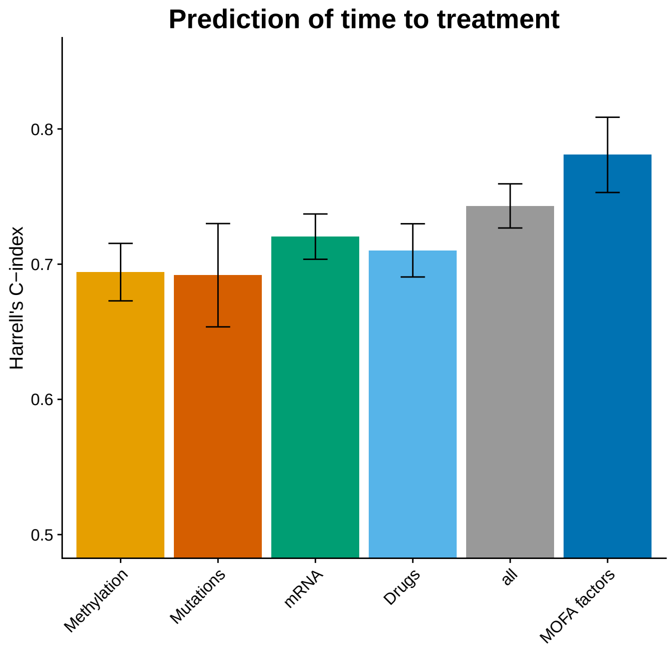


Figure 4.A.20: Prediction accuracy of time to next treatment using MOFA factors and individual features of the assays in the CLL data. Considered are L_2-penalized Cox models trained on the features of individual assays as well as their superset (all). For comparison the result using a Cox model trained on the 10 MOFA factors is shown as in Figure 4.8. The y-axis shows Harrell's C-index as a measure of prediction performance. The average value over 5-fold cross-validation is shown, with error bars indicating the standard error. Assays with missing values were imputed using the feature-wise mean.

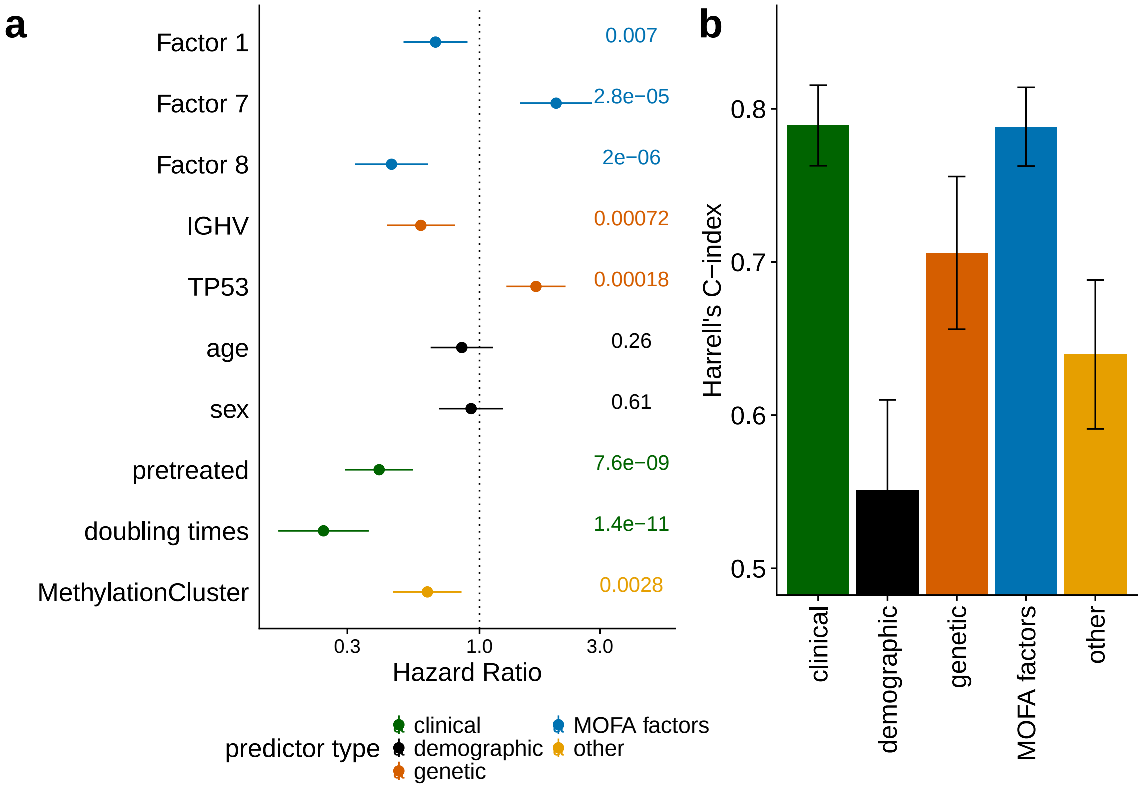


Figure 4.A.21: Comparison of MOFA factors with clinical covariates in the CLL data. **(a)** Association of MOFA factors and clinical covariates with time to next treatment using a univariate Cox models for $N = 76$ samples, for which the clinical information was available. Error bars denote 95% confidence intervals. Numbers on the right denote p-values for each predictor. 'Doubling times' refers to the clinically measured doubling time of lymphocytes, 'pretreated' to whether the patient was treated by chemo-immunotherapy prior to sample collection and 'MethylationCluster' to the previously described CLL subgroups as in Figure 4.A.11. **(b)** Prediction accuracy of time to treatment using multivariate Cox regression trained using the 10 factors derived using MOFA as well as the selected clinical predictors in panel **(a)**. Shown are average values of Harrell's C-index from 5-fold cross-validation. Error bars denote standard error of the mean.

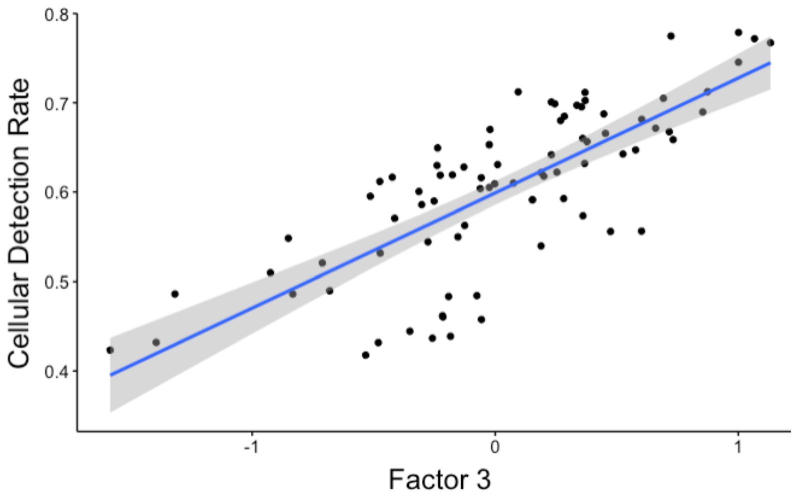


Figure 4.A.22: Characterisation of Factor 3 in the single cell data. Scatterplot depicting the correlation between Factor 3 and the cellular detection rate, a known technical factor in single-cell RNA-seq data that corresponds to the fraction of expressed genes.

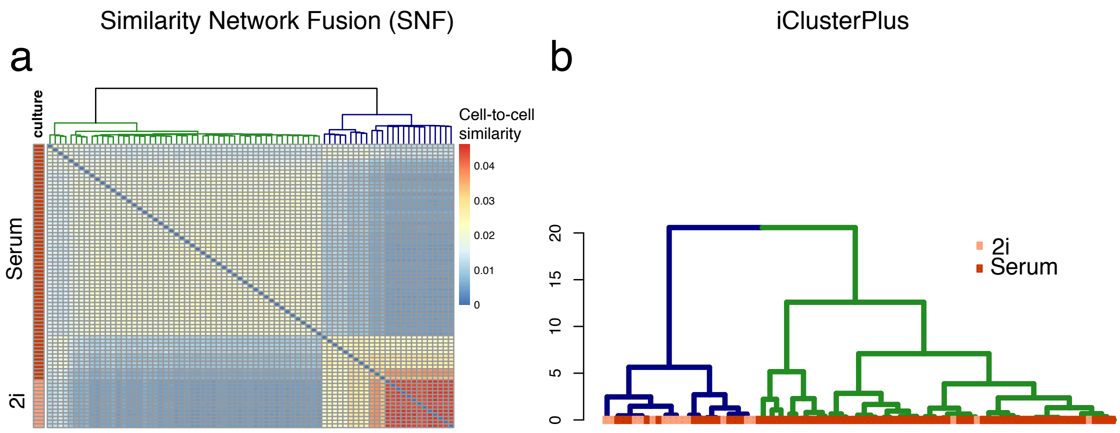


Figure 4.A.23: Multi-omics clustering applied to the single cell data. (a) Similarity matrix and dendrogram obtained using Similarity Network Fusion [199]. **(b)** Dendrogram obtained using iClusterPlus [144] with two clusters ($K = 1$, model selection by optimal BIC). The cells at the leaves are coloured by culture condition.

4.A.6 Supplementary Tables

Table 4.A.1: Parameters for simulation settings

(a) Simulation settings to validate the ability to learn the number of active factors (Appendix Figure 4.A.1)

likelihood	# factors	# features	# views	# samples	% missing
Gaussian	5,10,...,60	5000	3	100	0
Gaussian	10	100,500,...,10000	3	100	0
Gaussian	10	5000	1,3,...,21	100	0
Gaussian	10	5000	3	100	0,5,...,90

(b) Simulation settings to validate non-Gaussian likelihoods (Appendix Figures 4.A.2 and 4.A.3)

likelihood	# factors	# features	# views	# samples	% missing
Bernoulli	10	5000	3	100	0
Poisson	10	5000	3	100	0

(c) Simulations settings for the GFA and iCluster comparison (Appendix Figures 4.A.4 and 4.A.5)

likelihood	# factors	# features	# views	# samples	% missing
Gaussian, Bernoulli, Poisson	10	5000	3	100	5

Table 4.A.2: Coarse-grain categories of gene sets used in Figure 4.3

Immune Response
Interleukin-6 signalling
Interleukin-7 signalling
Cytokine Signalling in Immune system
Adaptive Immune System
Innate Immune System
Immune System
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
TCR signalling
Downstream TCR signalling
Phosphorylation of CD3 and TCR zeta chains
Translocation of ZAP-70 to Immunological synapse
Interleukin-1 signalling
Signalling by Interleukins
Interleukin-2 signalling
Interleukin-3, 5 and GM-CSF signalling
Diseases of Immune System
Interleukin-6 family signalling
Interleukin-10 signalling
Interleukin-4 and 13 signalling
IL-6-type cytokine receptor ligand interactions
Interleukin receptor SHC signalling

Cellular Stress/Senescence
Telomere Maintenance
Packaging Of Telomere Ends
Polymerase switching on the C-strand of the telomere
Processive synthesis on the C-strand of the telomere
Telomere C-strand (Lagging Strand) Synthesis
Activation of ATR in response to replication stress
Extension of Telomeres
Cellular responses to stress
Oxidative Stress Induced Senescence
Senescence-Associated Secretory Phenotype (SASP)
Cellular Senescence
Formation of Senescence-Associated Heterochromatin Foci (SAHF)
Oncogene Induced Senescence
DNA Damage/Telomere Stress Induced Senescence
Regulation of HSF1-mediated heat shock response
HSF1 activation
Cellular response to heat stress
Attenuation phase
HSF1-dependent transactivation

RNA regulation
RNA Polymerase II HIV Promoter Escape
SIRT1 negatively regulates rRNA Expression
ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression
NoRC negatively regulates rRNA expression
Regulation of mRNA stability by proteins that bind AU-rich elements
RNA Polymerase I, RNA Polymerase III, and Mitochondrial Transcription
Positive epigenetic regulation of rRNA expression
B-WICH complex positively regulates rRNA expression

Negative epigenetic regulation of rRNA expression
Transcriptional regulation by small RNAs
RNA Polymerase II Pre-transcription Events
RNA Polymerase I Promoter Opening
RNA Polymerase I Transcription Initiation
RNA Polymerase I Promoter Escape
RNA Polymerase II Promoter Escape
RNA Polymerase I Chain Elongation
RNA Polymerase II Transcription Pre-Initiation And Promoter Opening
RNA Polymerase III Chain Elongation
RNA Polymerase I Promoter Clearance
RNA Polymerase II Transcription Termination
RNA Polymerase II Transcription
RNA Polymerase I Transcription Termination
RNA Polymerase I Transcription
RNA Polymerase III Transcription Termination
RNA Polymerase III Transcription
RNA Polymerase III Abortive And Retractive Initiation
RNA Polymerase II Transcription Initiation
RNA Polymerase II Transcription Elongation
RNA Polymerase II Transcription Initiation And Promoter Clearance
RNA Polymerase III Transcription Initiation
RNA Polymerase III Transcription Initiation From Type 1 Promoter
RNA Polymerase III Transcription Initiation From Type 2 Promoter
RNA Polymerase III Transcription Initiation From Type 3 Promoter

Table 4.A.3: Overview of GFA and iCluster methods.
Abbreviations used in the table: variational Bayes (VB), Gibbs (Gibbs sampling), automatic relevance determination (ARD), expectation-maximization (EM)

Publication	Inference	Group-sparsity	Feature-sparsity	Missing values	Likelihood	Noise model
Shen *et al.* 2009 [170]	EM, grid search	different L_1-penalties	L_1-penalty	No	Gaussian	Heteroscedastic
Mo *et al.* 2013 [144]	EM, grid search	different L_1-penalties	L_1-penalty	No	Gaussian, Poisson, Bernoulli, Multinomial	Heteroscedastic
Virtanen *et al.* 2012 [197]	VB	ARD	None	No	Gaussian	Homoscedastic
Klami *et al.* 2015 [113]	VB	ARD	None	No	Gaussian	Homoscedastic
Bunte *et al.* 2016 [27]	Gibbs	ARD	Spike & Slab	No	Gaussian	Homoscedastic
Hore *et al.* 2016 [91]	VB	None	Spike & Slab	Yes	Gaussian	Heteroscedastic
Remes *et al.* 2015 [160]	VB	ARD	None	No	Gaussian	Homoscedastic
Zhao *et al.* 2016 [214]	Gibbs	ARD	Three-parameter beta prior	No	Gaussian	Heteroscedastic
Leppäaho *et al.* 2017 [122]	Gibbs	ARD	Spike & Slab	Yes	Gaussian	Homoscedastic
MOFA	VB	ARD	Spike & Slab	Yes	Gaussian, Poisson, Bernoulli	Heteroscedastic

CHAPTER 5

Adaptive penalization in high-dimensional regression and classification with external covariates

This chapter is a slightly modified version of a pre-print available under `https://arxiv.org/abs/1811.02962`.

Penalization schemes like Lasso or ridge regression are routinely used to regress a response of interest on a high-dimensional set of potential predictors. Despite being decisive, the question of the relative strength of penalization is often glossed over and only implicitly determined by the scale of individual predictors. At the same time, additional information on the predictors is available in many applications but left unused. Here, we propose to make use of such external covariates to adapt the penalization in a data-driven manner. We present a method that differentially penalizes feature groups defined by the covariates and adapts the relative strength of penalization to the information content of each group. Using techniques from the Bayesian tool-set our procedure combines shrinkage with feature selection and provides a scalable optimization scheme.

We demonstrate in simulations that the method accurately recovers the true effect sizes and sparsity patterns per feature group. Furthermore, it leads to an improved prediction performance in situations where the groups have strong differences in dynamic range. In applications to data from high-throughput biology, the method enables re-weighting the importance of feature groups from different assays. Overall, using available covariates extends the range of applications of penalized regression, improves model interpretability and can improve prediction performance.

Code Availability The software is freely available as an R package `https://git.embl.de/bvelten/graper`, scripts for the analyses contained in this paper can be found at `https://git.embl.de/bvelten/graper_analyses`.

5.1 Introduction

We are interested in the setup where we observe a continuous or categorical response Y together with a vector of potential predictors, or features, $X \in \mathbb{R}^p$ and aim to find a relationship of the form

$$Y = f(X).$$

Two main questions are of potential interest in this setting. First, we want to obtain an f that yields good predictions for Y given a new observation X. Second, we aim at finding which components in X are the 'important ones' for the prediction.

A common and useful approach to this end are (generalized) linear regression methods, which assume that the distribution of $Y|X$ depends on X via a linear term $X^T\beta$. In order to cope with high-dimensionality of X and avoid over-fitting, penalization on β is employed, e.g. in ridge regression [90], Lasso [184] or elastic net [216]. By constraining the values of β, the complexity of the model is restricted, resulting in biased but less variable estimates and improved prediction performance. In addition, some choices of the penalty yield estimates with a relatively small number of non-zero components, thereby facilitating feature selection. An example is the L_1-penalty employed in Lasso or elastic net.

Commonly, penalization methods apply a penalty that is symmetric in the model coefficients. Real data, however, often consists of a collection of heterogeneous features, which such an approach does not account for. In particular, it ignores any additional information or structural differences that may be present in the features. Often we encounter X whose components comprise multiple data modalities and data qualities, e.g., measurement values from different assays. Other side-information on individual features could include temporal or spatial information, quality metrics associated to each measurement or the features' sample variance, frequency or signal-to-noise ratio. It has already been observed in multiple testing that the power of the analysis can be improved by making use of such external information (e.g. [48, 62, 97, 121, 123]). However, in current penalized regression models this information is frequently ignored. Making use of it could on one hand improve prediction performance. On the other hand, it might yield important insight into the relationship of external covariates to the features' importance. For example, if the covariate encodes different data modalities, insights into their relative importance could help cutting costs by reducing future assays to the essential data modalities.

As a motivating example we consider applications in molecular biology and precision medicine. Here, the aim is to predict phenotypic outcomes, such as treatment response, and identify reliable disease markers based on molecular data. Nowadays, different high-throughput technologies can be combined to jointly measure thousands of molecular features from different biological layers [85, 162]. Examples include genetic alterations, gene expression, methylation patterns, protein abundances or microbiome occurrences. However, despite the increasing availability of molecular and clinical data, outcome prediction remains challenging [4, 35, 82]. Common applications of penalized regression only make use of parts of the available data. For example, different assay types are simply concatenated or analysed separately. In addition, available annotations on individual features are left unused, such as their chromosomal location or gene set and pathway membership. Incorporating side-information on the assay type and spatial or functional annotations could help to improve prediction performance. Furthermore, it could help prioritizing feature groups, such as different assays or gene sets.

Here, we propose a method that incorporates external covariates in order to guide penalization and can learn relationships of the covariate to the feature's effect size in a data-driven way. We introduce the method for linear models and extend it to classification purposes. We demonstrate that this can improve prediction performance and yields insights into the relative importance of different feature sets, both on simulated data and applications in high-throughput biology.

5.2 Methods

5.2.1 Problem statement

Assume we are given observations $(x_1, y_1), \ldots, (x_n, y_n)$ with $y_i \in \mathcal{Y} \subseteq \mathbb{R}$, $x_i \in \mathbb{R}^p$ (possibly $n \ll p$) from a linear model, i.e.

$$y_i = x_i^T \beta + \epsilon_i \tag{5.1}$$

with $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. In addition, we suppose that we have access to a covariate $\zeta_j \in \mathcal{Z} \subseteq \mathbb{R}^k$ for each predictor $j = 1, \ldots, p$. We hope, loosely speaking, that ζ_j contains some sort of information on the magnitude of β_j. The question we want to address is: Can we use the information from ζ to improve upon estimation of β and prediction of Y?

In order to estimate β from a finite sample $y = (y_i)_{i=1}^n \in \mathbb{R}^n$ and $\mathbf{X} = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n \times p}$ we can employ penalization on the negative log-likelihood of the model, i.e.

$$\hat{\beta}(\lambda) \in \arg\min_\beta \frac{1}{n} ||y - \mathbf{X}\beta||_2^2 + \lambda \text{pen}(\beta), \tag{5.2}$$

where $\text{pen}(\cdot)$ denotes a penalty function on the model coefficients. For example, the choice $\text{pen}(\beta) = \sum_j |\beta_j|^q$ leads to Lasso ($q = 1$) or ridge regression ($q = 2$). The parameter λ controls the amount of penalization and thereby the model complexity. Ideally, we would like to choose an optimal λ. For estimation this means minimizing the mean squared error $\text{MSE}(\hat{\beta}(\lambda)) = \mathbb{E}||\hat{\beta}(\lambda) - \beta||_2^2$; for prediction this means minimizing the expected prediction error. In practice, λ is often chosen to minimize the cross-validated error.

In most applications, the penalization is symmetric, i.e. for any permutation π we have $\lambda\text{pen}(\beta_1, \ldots, \beta_p) = \lambda\text{pen}(\beta_{\pi(1)}, \ldots, \beta_{\pi(p)})$. However, as we have external information on each feature given by ζ we want to allow for differential penalization guided by ζ. For this, we will consider the following non-symmetric generalization, which still leads to a convex optimization problem in β for a convex penalty, such as $\text{pen}(x) = |x|$ or $\text{pen}(x) = x^2$:

$$\hat{\beta}(\lambda) \in \arg\min_\beta \frac{1}{n} ||y - \mathbf{X}\beta||_2^2 + \sum_j \lambda(\zeta_j)\text{pen}(\beta_j). \tag{5.3}$$

Instead of a constant λ, here $\lambda : \mathcal{Z} \to \mathbb{R}_{\geq 0}$ provides a mapping from the covariate ζ to a non-negative penalty factor $\lambda(\zeta)$. This additional flexibility compared to a single penalty parameter can be helpful if ζ contains information on β. For example, in the simple case of ridge regression with deterministic orthonormal design matrix, known noise variance σ^2 and 'oracle covariate' $\zeta_j = \beta_j$ the optimal λ is seen to be $\lambda^*(\zeta_j) = \frac{\sigma^2}{\zeta_j^2}$. However, in practice the information in ζ is not that explicit and hence we do not know which λ is optimal.

If λ takes values in a small set of discrete values, e.g. for categorical covariates ζ, cross-validation could be used to determine a suitable set of function values. This approach is employed in the IPF-Lasso [22], where categorical covariates encode different data modalities. However, cross-validation soon becomes prohibitive, as it requires a grid search exponential in the number of categories defined by ζ. Similarly, cross-validation can be employed with λ parametrized by a small number of tuning parameters using domain knowledge to come up with a suitable parametric form for λ [17, 196]. However, such an explicit form is often not available. In many situations it is a major problem itself to come up with a helpful relationship between ζ and β and thereby knowledge of which values of a covariate would require more or less penalization. Therefore, we aim at finding λ in a data-driven manner and with improved scalability compared to cross-validation.

5.2.2 Problem statement from a Bayesian perspective

There is a direct correspondence between estimates obtained from penalized regression and a Bayesian estimate with penalization via corresponding priors on the coefficients. For example, the ridge estimate corresponds to the maximum a posterior estimate (MAP) in a Bayesian regression model with normal prior on β and the Lasso estimate to a MAP with a Laplace prior on β. This correspondence opens up alternative strategies using tools from the Bayesian mindset to approach the problem outlined above: Differential penalization translates to introducing different priors on the components of β. Our belief that ζ carries information on β can be incorporated by using prior distributions whose parameters depend on ζ. In GRridge [203], the authors used this idea to derive an Empirical Bayes approach for finding group-wise penalty parameters in ridge regression. However, this approach does not obviously generalize to other penalties such as the Lasso.

Moving completely into the Bayesian mindset we instead turn to explicit specification of priors to implement the penalization task. Different priors have been suggested [33, 131, 143, 150] and structural knowledge was incorporated into the penalization by employing multivariate priors that encode the structure in the covariance or non-exchangeable priors with different hyper-parameters (e.g. [5, 56, 88, 163, 207, 208] and references therein). Despite the possible gains in prediction performance when incorporating such structural knowledge, these methods have not been widely applied. A limiting factor has often been the lack of scalability to large data sets.

5.2.3 Setup and notation

From the linear model assumption we have

$$y_i = x_i^T \beta + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \tau^{-1}), \tag{5.4}$$

where τ denotes the precision of the noise. Based on the external covariate ζ we define a partition of the p predictors into G groups:

$$g_\zeta = g : \{1, \ldots, p\} \to \{1, \ldots, G\}. \tag{5.5}$$

For instance, categorical covariates ζ, such as different assay types, naturally define such a partition. For continuous covariates g_ζ can be defined based on suitable binning or clustering.

To achieve penalization in dependence of ζ we consider a spike and slab prior [143] on the model coefficients β with a different slab precision γ and mixing parameter π for each group. We re-parametrize β as $\beta_j = s_j b_j$ with

$$b_j | \gamma_{g_\zeta(j)} \sim \mathcal{N}\left(0, \gamma^{-1}_{g_\zeta(j)}\right), \tag{5.6}$$

$$s_j | \pi_{g_\zeta(j)} \sim \mathrm{Ber}(\pi_{g_\zeta(j)}). \tag{5.7}$$

In the special case of $\pi = 1$ this yields a normal prior as in [131] corresponding to ridge regression. With $\pi < 1$ we additionally promote sparsity on the coefficients, and the value of π controls the number of active predictors in each group. The value of γ controls the overall shrinkage per group. To learn the model hyperparameters γ, π and the noise precision τ, we choose the following conjugate priors

$$\tau \sim \text{Gamma}(r_\tau, d_\tau), \tag{5.8}$$

and for each group $k \in \{1, \dots, G\}$

$$\gamma_k \sim \text{Gamma}(r_\gamma, d_\gamma), \tag{5.9}$$
$$\pi_k \sim \text{Beta}(d_\pi, r_\pi), \tag{5.10}$$

with $d_\tau, r_\tau, d_\gamma, r_\gamma = 0.001$ and $r_\pi, d_\pi = 1$. Hence, the joint probability of the model is given by

$$p(y, b, s, \gamma, \pi, \tau) = p(y|b, s, \tau)p(b, s|\pi, \gamma)p(\gamma)p(\pi)p(\tau). \tag{5.11}$$

5.2.4 Inference using variational Bayes

The challenge now lies in inferring the posterior of the model parameters from the observed data $\mathbf{X}, y$ and the covariate ζ. While Markov chain Monte Carlo methods are frequently used for this purpose they do not scale well to large data sets. Here, we adopt a variational inference framework [20, 21] that has been used (in combination with importance sampling) for variable selection with exchangeable priors (e. g. varbvs [31, 32]). Denoting all unobserved model components by $\theta = (b, s, \gamma, \pi, \tau)$, we approximate the posterior $p(\theta|\mathbf{X}, y)$ by a distribution $q(\theta)$ from a restricted class of distributions $\mathcal{Q}$, where the goodness of the approximation is measured in terms of the Kullback-Leibler (KL) divergence, i.e.

$$q \in \arg\min_{q \in \mathcal{Q}} D_{\text{KL}}(q \,||\, p(\theta|\mathbf{X}, y)). \tag{5.12}$$

A common and useful choice for distributions in class $\mathcal{Q}$ is the mean-field approximation, i.e. assuming that a distribution in $\mathcal{Q}$ factorizes in its parameters. We consider

$$q(\theta) = q(b, s, \gamma, \pi, \tau) = \prod_{j=1}^{p} q(b_j, s_j)q(\gamma)q(\pi)q(\tau), \tag{5.13}$$

where b_j and s_j are not factorised due to their strong dependencies [188].

The variational approach leads to an iterative inference algorithm [21] by observing that minimizing the KL-divergence is equivalent to maximizing the evidence lower bound $\mathcal{L}$ defined by

$$\log(p(y)) = \mathcal{L}(q) + D_{\text{KL}}(q \,||\, p(\theta \,|\, \mathbf{X}, y)). \tag{5.14}$$

From this, we have

$$\mathcal{L}(q) = \int \log \frac{p(y, \theta)}{q(\theta)} \, q(\theta) \, d\theta \tag{5.15}$$
$$= \int \log p(y, \theta) \, q(\theta) \, d\theta + H(q(\theta)), \tag{5.16}$$

with $H(q) = \int -q(\theta) \, \log q(\theta) \, d\theta$ denoting the differential entropy.

Variational methods are based on maximisation of the functional $\mathcal{L}$ with respect to q in order to obtain a tight lower bound on the log model evidence and minimize the KL-distance between the density q and the true (intractable) posterior. Under a mean-field assumption $q(\theta) = \prod_j q(\theta_j)$, the optimal q_j keeping all other factors fixed is given by

$$\log(q_j^*)(\theta_j) = \mathbb{E}_{-j}(\log(p(y, \theta))) - \text{const}. \tag{5.17}$$

Iterative optimization of each factor results in Algorithm 1. Details on the variational inference and the updates can be found in Appendix 5.A.1 and 5.A.2. The method is implemented in the freely available R package graper. From the obtained approximation q of the posterior distribution, we obtain point estimates for the model parameters. In particular, we will use the posterior means $\hat{\beta} = \int \beta\, q(\beta)\, d\beta$, $\hat{\gamma} = \int \gamma\, q(\gamma)\, d\gamma$ and $\hat{\pi} = \int \pi\, q(\pi)\, d\pi$.

Algorithm 1 Inference algorithm

1: Input: $\mathbf{X}, y, \bigsqcup_{k=1}^{G} \mathcal{G}_k = \{1, \ldots, p\}$

2: Initialize $\mathbb{E}s_j = 1$, $\mathbb{E}\beta_j$ sampled from $\mathcal{N}(0,1)$, $\mathbb{E}\tau = \mathbb{E}\gamma_k = 1$

3: **while** $\mathcal{L}(q)$ has not converged **do**:

4: **for** $k = 1, \ldots G$ **do**

5: Set $q(\pi_k) = \mathrm{Beta}(\pi_k|\alpha_k^\pi, \beta_k^\pi)$ with

$$\alpha_k^\pi = d_\pi + \sum_{j \in \mathcal{G}_k} \mathbb{E}s_j \text{ and } \beta_k^\pi = r_\pi + \sum_{j \in \mathcal{G}_k} (1 - \mathbb{E}s_j)$$

6: **for** $j = 1, \ldots p$ **do**

7: Set $q(s_j) = \mathrm{Ber}(s_j|\psi_j)$, $q(b_j|s_j = 1) = \mathcal{N}(b_j|\mu_j, \sigma_j^2)$ and $q(b_j|s_j = 0) = \mathcal{N}(b_j|0, (\mathbb{E}\gamma_{g(j)})^{-1})$ with

$$\sigma_j^2 = (\mathbb{E}\tau ||\mathbf{X}_{\cdot,j}||_2^2 + \mathbb{E}\gamma_{g(j)})^{-1}$$

$$\mu_j = \sigma_j^2 \mathbb{E}\tau \left(- \sum_{i=1}^{n} \sum_{l \neq j}^{p} \mathbf{X}_{ij} \mathbf{X}_{il} \mathbb{E}(\beta_l) + \mathbf{X}_{\cdot,j}^T y \right)$$

$$\mathrm{logit}(\psi_j) = \mathbb{E} \log \frac{\pi_{g(j)}}{1 - \pi_{g(j)}} + \frac{1}{2} \log(\mathbb{E}\gamma_{g(j)}) + \frac{1}{2} \log(\sigma_j^2) + \frac{1}{2} \frac{\mu_j^2}{\sigma_j^2}$$

8: Set $q(\tau) = \mathrm{Gamma}(\tau|\alpha^\tau, \beta^\tau)$ with

$$\alpha^\tau = r_\tau + \frac{n}{2} \text{ and } \beta^\tau = d_\tau + \frac{1}{2} \mathbb{E}||y - \mathbf{X}\beta||_2^2$$

9: **for** $k = 1, \ldots G$ **do**

10: Set $q(\gamma_k) = \mathrm{Gamma}(\gamma_k|\alpha_k^\gamma, \beta_k^\gamma)$ with

$$\alpha_k^\gamma = r_\gamma + \frac{1}{2}|\mathcal{G}_k| \text{ and } \beta_k^\gamma = d_\gamma + \frac{1}{2} \sum_{j \in \mathcal{G}_k} \mathbb{E}b_j^2$$

11: Calculate $\mathcal{L}(q) = \mathbb{E} \log p(y, b, s, \gamma, \pi, \tau) + H(q)$

Notes: *The expectations are taken under the current variational distribution q, and* $H(q) = \int -q(\theta) \log q(\theta) d\theta$ *denotes the differential entropy. We use* $\mathcal{F}(x|a)$ *to denote the probability density function in x of a distribution* $\mathcal{F}$ *with parameters a, e.g.* $\mathrm{Beta}(x|\alpha, \beta)$*. In step 7 it is important to keep track of* $v = \mathbf{X}\, \mathbb{E}\beta$ *in the implementation to obtain linear computational complexity in p. We set* $r_\tau = r_\gamma = d_\tau = d_\gamma = 0.001$ *and* $d_\pi = r_\pi = 1$.

Remark on the choice of the mean-field assumption An interesting deviation from the standard fully factorized mean-field assumption in Equation (5.13) is taking a multivariate variational distribution for the model coefficients. This is easily possible for the dense model

$(\pi = 1, s = 1, b = \beta)$, where we can consider the factorization

$$q(\beta, \gamma, \tau) = q(\beta)q(\gamma)q(\tau).$$

In particular, a multivariate distribution is kept for the model coefficients β instead of factorizing $q(\beta) = \prod_j q(\beta_j)$. Thereby, this approach allows to capture dependencies between model coefficients in the inferred posterior and is less approximative. We will show below that this can improve the prediction results. However, a drawback of this approach is its computational complexity, as it requires the calculation and inversion of a $p \times p$ covariance matrix in each step. While this can be reduced to a quadratic complexity as described in Appendix 5.A.2.1, this is still prohibitive for many applications. Therefore, we concentrate in the following on the fully factorized mean-field assumption but include comparisons to the multivariate approach in the Results section.

5.2.5 Extension to logistic regression

The model of Section 5.2.3 can be flexibly adapted to other types of generalised linear regression setups with suitable link functions and likelihoods. However, the inference framework needs to be adapted due to loss of conjugacy. Here, we extend the model to the framework of logistic regression with a binary response variable, where we assume that the response follows a Bernoulli likelihood with a logistic link function

$$y_i|\beta \sim \mathrm{Ber}(\sigma(x_i^T\beta)) \quad \text{with} \quad \sigma(z) = \frac{1}{1+\exp(-z)}. \tag{5.18}$$

While the prior structure and core of the variational inference are identical to the case of a linear model, additional approximations are necessary. For this purpose we adopt [101] and approximate the likelihood using a lower bound on the logistic function. For an arbitrary $\xi \in \mathbb{R}$ we have

$$\sigma(z) \geq \sigma(\xi) \exp\left(\frac{1}{2}(z-\xi) - \eta(\xi)(z^2-\xi^2)\right) \tag{5.19}$$

with $\eta(\xi) = \frac{1}{2\xi}\left(\sigma(\xi) - \frac{1}{2}\right)$. With this, $\log p(y|\beta) = \sum_{i=1}^n \log(\sigma((2y_i-1)x_i^T\beta))$ can be bounded by

$$\begin{aligned}\log p(y|\beta) \geq \frac{1}{2}\sum_i (2y_i-1)x_i^T\beta - \sum_i \eta(\xi_i)(x_i^T\beta)^2 \\ + \sum_i \left(\log(\sigma(\xi_i)) - \frac{1}{2}\xi_i + \eta(\xi_i)\xi_i^2\right).\end{aligned} \tag{5.20}$$

As this approximation restores a quadratic form in β, the remaining updates can be adopted from the case of a linear model above with the additional variational parameter ξ (see Appendix 5.A.2.2 for details).

5.3 Results

5.3.1 Results on simulated data

First, we evaluated the method on simulated data to test its ability to recover the model coefficients and hyper-parameters per group. For this, a random $\mathbf{X}$ matrix was generated from a multivariate normal distribution with mean zero and a Toeplitz covariance structure $\Sigma_{ij} = \rho^{|i-j|}$, and the response was simulated from a linear model with normal error. The

p predictors were split into $G = 6$ groups of equal size, and the coefficients were simulated from the model as described in Equations (5.6) and (5.7) with fixed π_k and γ_k for each group. In particular, we set $\gamma_k = 0.01$ for $k = 1, 2$, $\gamma_k = 1$ for $k = 3, 4$ and $\gamma_k = 100$ for $k = 5, 6$. For each pair of groups with same γ-value the sparsity level π_k was varied between ν and $\min(1, 1.5\nu)$ for a certain value of ν determining the sparsity level from 0 (sparse) to 1 (dense). We then varied the number of features p, the number of samples n, the correlation strength ρ, the noise precision τ and the sparsity level ν (Table 5.1) and generated for each setting ten independent data sets. We evaluated the recovery of the hyper-parameter γ and π for each group and compared the predictive performance and computational complexity to those of related methods including ridge regression [90], Lasso [184], elastic net [216], adaptive Lasso [215], sparse group Lasso, group Lasso [67], GRridge [203], varbvs [32] and IPF-Lasso [22]. Here, ridge regression, Lasso, elastic net and varbvs are covariate-agnostic methods, i. e. they ignore the group annotations of the features. The group Lasso methods account for the group structure but use a joint penalty parameter, and GRridge, IPF-Lasso and graper are covariate-aware and adapt the relative strength of the penalty to the groups.

Table 5.1: Simulation parameters

Here, p denotes the number of features, n the number of samples, ρ the correlation strength in $\mathbf{X}$, τ the noise precision and ν the sparsity level.

p	n	ρ	τ	ν
60, 120,...,1200	100	0	1	0.2
300	20, 40,...,500	0	1	0.2
300	100	0, 0.1,...,0.9	1	0.2
300	100	0	0.01, 0.1,...,100	0.2
300	100	0	1	0.001, 0.01, 0.05,...,1

5.3.1.1 Recovery of hyper-parameters

The algorithm accurately recovered the relative importance of different groups (encoded by γ_k) and the group-wise sparsity level (encoded by π_k) across a large range of settings as shown in Figure 5.1. The method failed to recover those parameters accurately only if the ratio between sample size and number of features was too small or the sparsity parameter ν was too close to 1. These settings were challenging for all methods as can be seen in Section 5.3.1.2, where we evaluated estimation and prediction performance in comparison to other methods. In addition, the groups had to contain sufficiently many predictors to reliably estimate group-wise parameters, as seen in Figure 5.1b. We also noted that a low signal-to-noise ratio could impede the estimation of hyperparameters as can be seen from the group with a very large γ value (meaning low coefficient amplitudes as in group 5 and 6) and low precision values (τ) of the noise term.

5.3.1.2 Prediction and estimation performance

Next, we compared the estimation of the true model coefficients and the prediction accuracy on an independent test set of $n = 1,000$. Overall, the method showed improved performance for a large range of sample sizes, correlations, numbers of features, noise variances and active features, both in terms of the root mean squared error on y as well as for estimation of β (Figure 5.2). Among the non-sparse methods graper with a non-factorized mean-field assumption clearly outperformed the factorized mean-field assumption as well

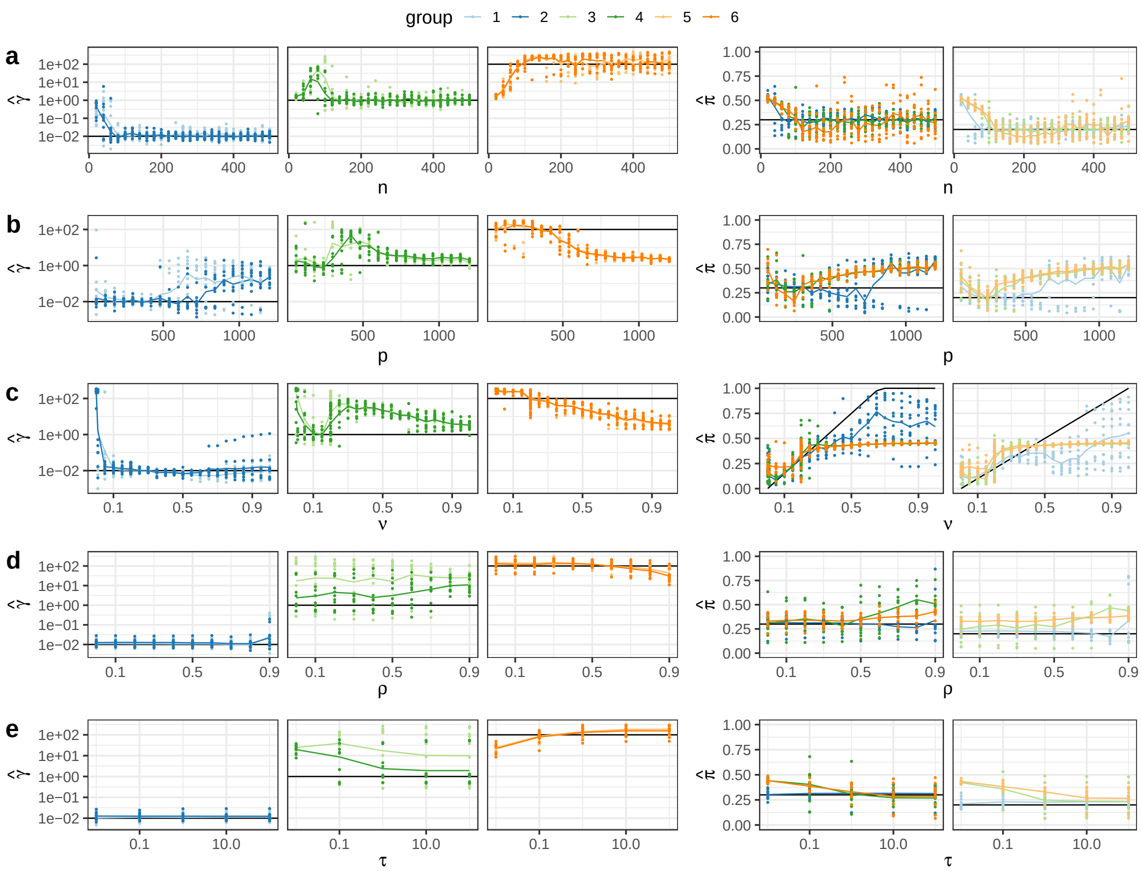


Figure 5.1: Recovery of hyperparameters on simulated data. Estimated values of the hyperparameter γ (left column) and π (right column) when varying each of the model parameters **(a-e)** while keeping the other four parameters fixed as described in Table 5.1. The line denotes the mean recovered hyperparameter across 10 random instances of simulated data, while points represent single instances. Colours denote the different groups ($k = 1, \ldots, 6$) and the black line indicates the true value of γ (left) and π (right) used in the simulation. Each panel displays groups with the same value of γ (left) and π (right).

as GRridge and group Lasso. The covariate-agnostic ridge regression performed worst in most cases. Sparse methods performed in general better in this simulation example, as the underlying model had a large fraction of zero coefficients. Here, we observed that graper was comparable to IPF-Lasso, which is the most closely related method. Only in settings with a very high number of active predictors or strong correlations between the predictors (ρ close to one) the method was outperformed by the IPF-Lasso.

5.3.1.3 Scalability

While the additional group-wise optimization comes at a computational cost, the variational approach runs inference in time complexity linear in the number of features p, samples n and groups G. Only in the case of a multivariate variational distribution, the complexity is quadratic in the larger of n and p and cubic in the smaller of the two. When varying the number of samples n, features p and groups G we observed comparable run times as for Lasso (Figure 5.3). Differences were mainly observed for p: For larger p, graper required slightly longer times than Lasso. This difference was more pronounced when using a sparsity promoting spike and slab prior, where additional parameters need to be inferred. As expected, the multivariate approach of graper became considerably slower for large p and showed comparable run times to the sparse group Lasso. The number of groups mainly influenced the computation times of IPF-Lasso, which scales exponentially in the number of groups. Here, graper provided a by far more scalable approach.

5.3.2 Application to data from high-throughput biology

5.3.2.1 Drug response prediction in leukaemia samples

Next, we exemplify the method's performance on real data by considering an application to biological data, where predictors were obtained from different assays. Using assay type as external covariates we used the method to integrate data from the different assays (also referred to as omics types) in a study on chronic lymphocytic leukaemia (CLL) [47]. This study combined drug response measurements with molecular profiling including gene expression and methylation. Briefly, we used normalized RNA-Seq expression values of the $5,000$ most variable genes, the DNA methylation M-values at the 1% most variable CpG sites as well as the ex-vivo cell viability after exposure to 61 drugs at 5 different concentrations as predictors for the response to a drug (ibrutinib) that was not included into the set of predictors. In total, this resulted in a model with $n = 121$ patient samples and $p = 9,553$ predictors.

We first applied the different regression methods to the data on their original scale. Since the features have different scales (e.g., the drug responses vary from around 1 (neutral) to 0 (completely toxic), the normalized expression values from 0 to 20 and the methylation M-values from -10 to 8), this ensures that the omics type information is an informative covariate: It results in larger effect sizes of the drug response data and smaller effect sizes of the methylation and expression data compared to scaled predictors. In this setting, incorporating knowledge on the assay type into the penalized regression showed clear advantages in terms of prediction performance: The covariate-aware methods (GRridge, IPF-Lasso and graper) all improved upon the covariate-agnostic Lasso, ridge regression or elastic net (Figure 5.4a). Also the group Lasso methods, which incorporate the group information but apply a single penalty parameter, could not adapt to the scale differences. The inferred hyper-parameters γ of graper highlighted the larger effect sizes of the drug response feature group, which was strongly favoured by the penalization (Figure 5.4b).

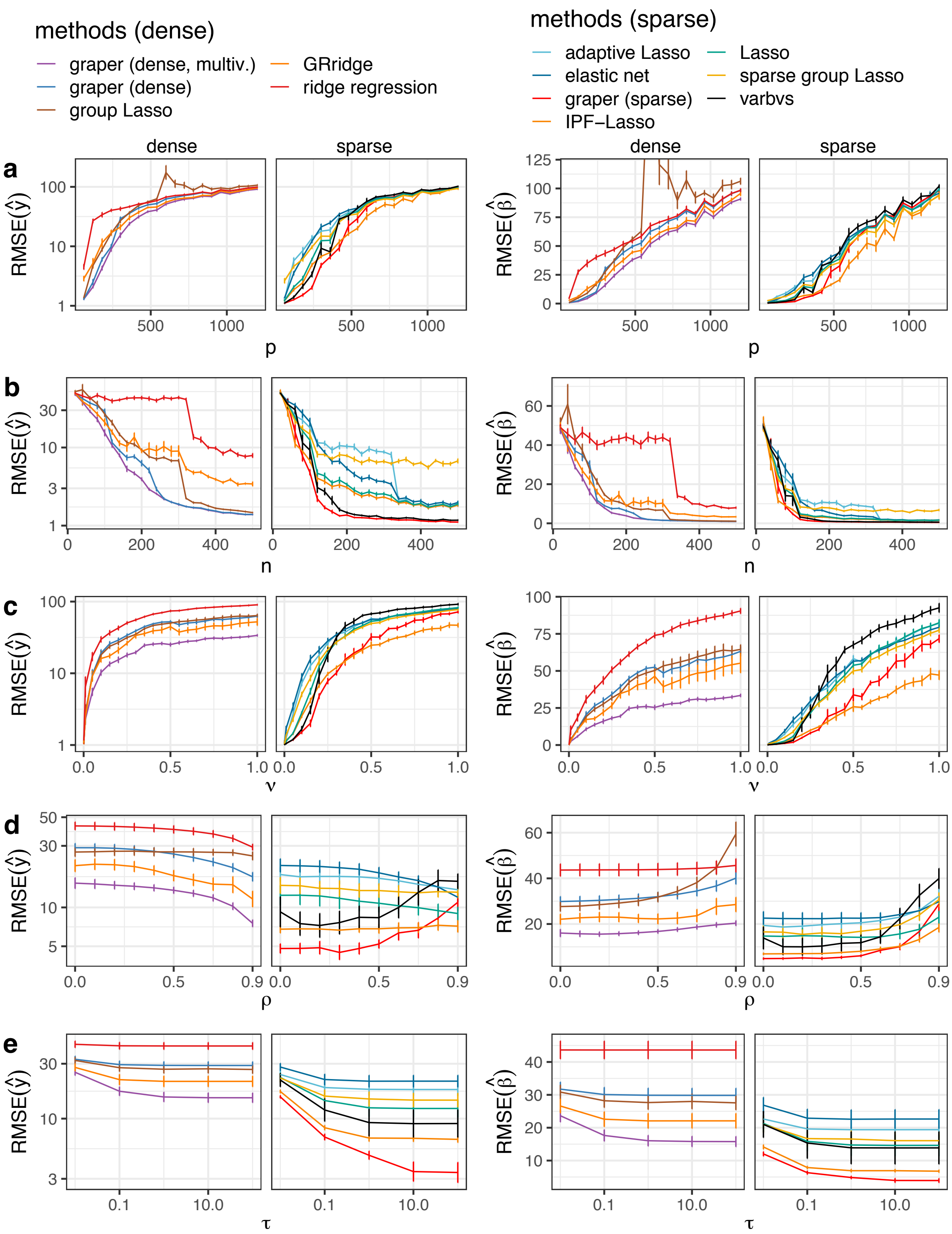


Figure 5.2: Prediction and estimation performance on simulated data. Shown are the root mean squared error (RMSE) of the predicted response $\hat{Y} = X^T\hat{\beta}$ (left) and the estimate $\hat{\beta}$ (right) for different methods when varying one of the simulation parameters **(a-e)** as described in Table 5.1. The prediction error is assessed on $n = 1,000$ test samples. The line denotes the mean RMSE across 10 random instances of simulated data with bars denoting standard errors. The two panels separate methods with sparse estimates of β (right) from non-sparse methods (left). (Group Lasso is counted as non-sparse method as it is not sparse within groups.)

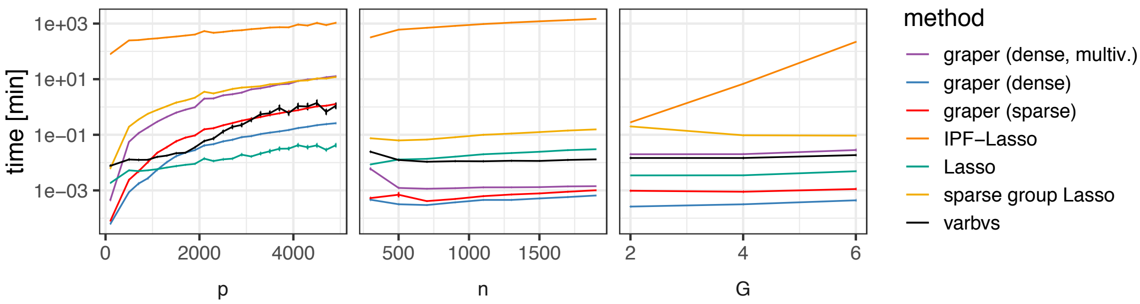


Figure 5.3: Comparison of run times. Shown are average run time (in minutes) for different methods when varying the number of samples n, features p and groups G. Each parameter is varied at a time while holding the others fixed to $n = 100$, $p = 300$ or $G = 6$. Shown are the average times across 50 random instances of simulated data with error bars denoting one standard error.

To address differences in feature scale, a common choice made by many implementations of penalized regression (e.g., glmnet [65]) is to scale all features to unit variance. Indeed, for the data at hand, this transformation was particularly beneficial for the covariate-agnostic methods, and their prediction performances became more similar to those of the covariate-aware methods. However, for dense methods such as ridge regression the covariate information on the omics type remained important (Figure 5.5a). Sparse methods in general resulted in very good predictions as the response to ibrutinib can be well explained by a very sparse model containing only few drugs with related mode of action. By learning weights for each omics type graper directly highlighted the importance of the drug data as predictors (Figure 5.5b).

In general, standardization of all features is unlikely to be an optimal choice, since in many applications there is a relation between information content and amplitude. Here, standardization would drown informative high-amplitude features and 'blow up' noisy low-amplitude features (see Appendix 5.A.3.1).

5.3.2.2 Age prediction from multi-tissue gene expression data

As a second example for a covariate in genomics we considered the tissue type. Using data from the GTEx consortium [126] we asked whether the tissue type is an informative covariate in the prediction of a person's age from gene expression. Briefly, we chose five tissues that were available for the largest number of donors and from each tissue considered the top 50 principal components on the RNA-Seq data after normalization and variance stabilization [127]. In total, this gave us $p = 250$ predictors from $G = 5$ tissues for $n = 251$ donors.

We observed a small advantage for methods that incorporate the tissue type as a covariate (Figure 5.6a): GRridge, IPF-Lasso and graper all had a smaller prediction error compared to covariate-agnostic methods. In particular, graper resulted in comparable prediction performance to IPF-Lasso whilst requiring less than a second for training compared to 40 minutes for IPF-Lasso. The learnt relative penalization strength and sparsity levels of graper can again provide insights into the relative importance of the different tissue types. In particular, we found lower penalization for blood vessel and muscle and higher penalization for blood and skin (Figure 5.6b). This is consistent with previous studies on a per-tissue basis, where gene expression in blood vessel has been found to be a good predictor for age, while blood was found to be less predictive [209].

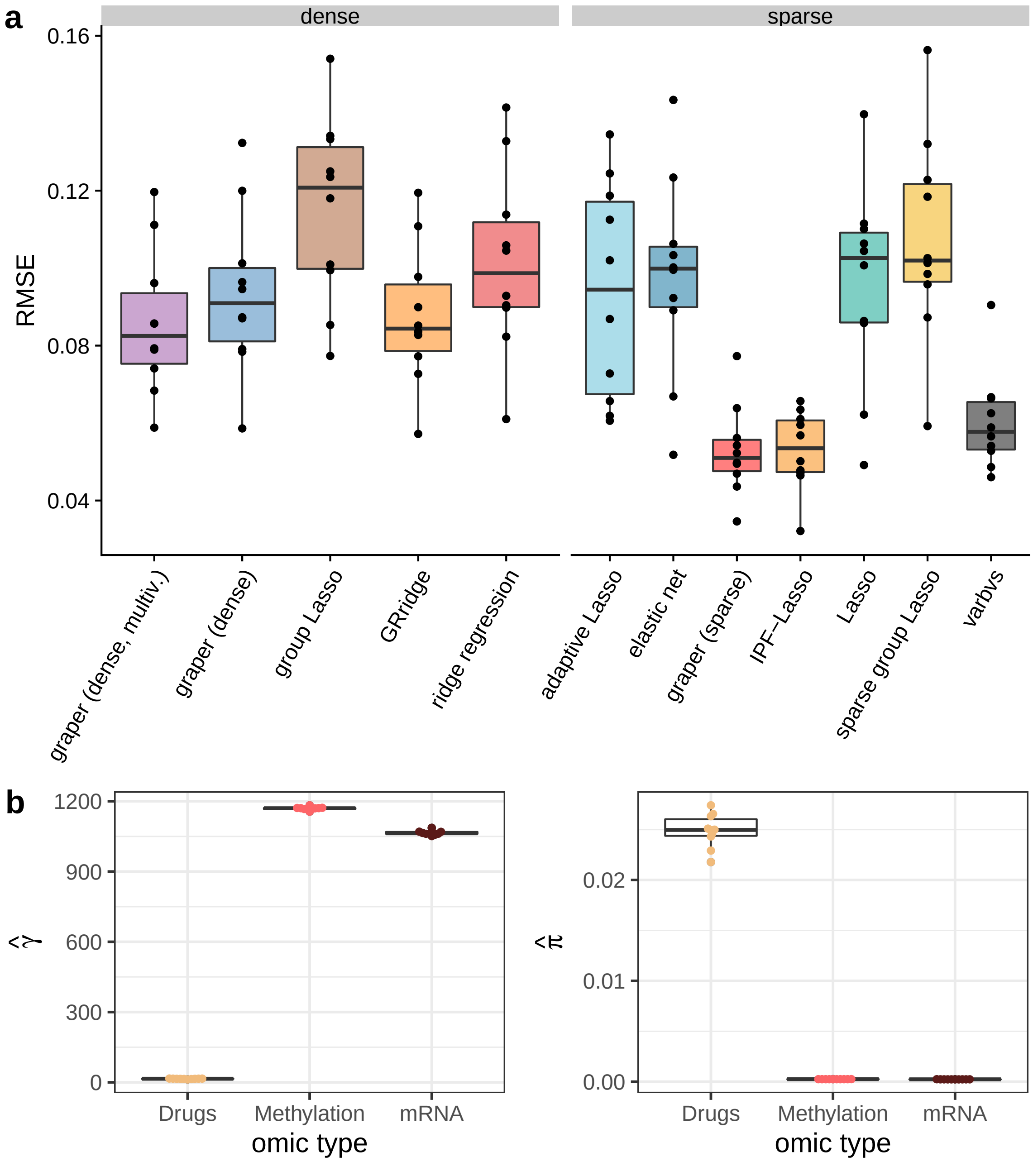


Figure 5.4: Application to the CLL data with scale differences between assays. **(a)** Comparison of the root mean-squared error (RMSE) for the prediction of samples' viability after treatment with ibrutinib. Performance was evaluated in a 10-fold cross-validation scheme, the points denote the individual RMSE for each fold. **(b)** Inferred hyperparameters by graper (sparse) in the different folds for the three different omics types (γ on the left and π on the right).

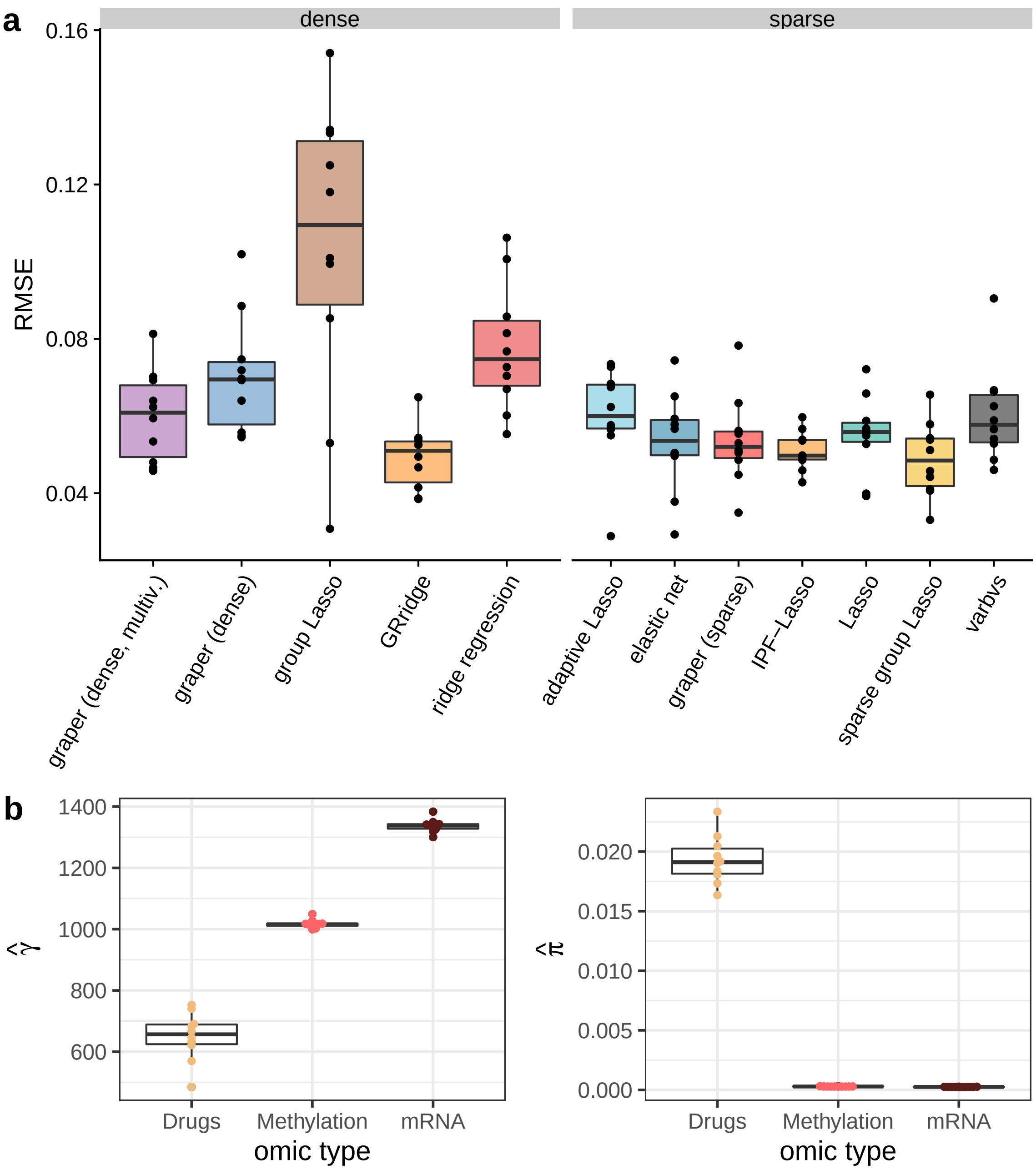


Figure 5.5: Application to the CLL data with standardized predictors. (a) Comparison of the root mean-squared error (RMSE) for the prediction of samples' viability after treatment with ibrutinib. Performance was evaluated in a 10-fold cross-validation scheme, the points denote the individual RMSE for each fold. **(b)** Inferred hyperparameters by graper (sparse) in the different folds for the three different omics types (γ on the left and π on the right).

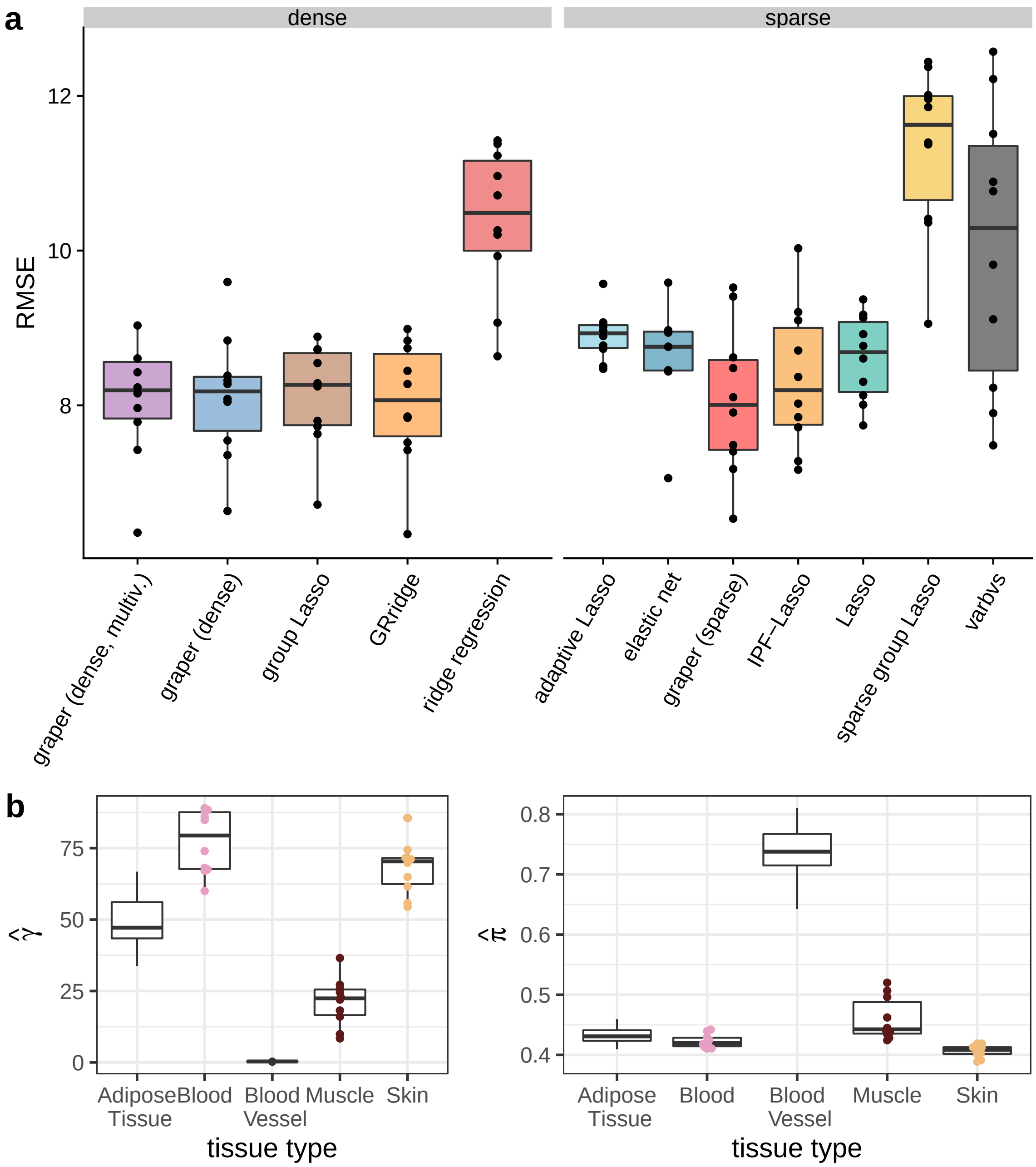


Figure 5.6: Application to the GTEx data. (a) Comparison of root mean-squared error (RMSE) for the prediction of donor age (in years). Performance is evaluated in a 10-fold cross-validation scheme, the points denote the individual RMSE for each fold. **(b)** Inferred hyperparameters by graper (sparse) in the different folds for the five tissues (γ on the left and π on the right).

5.4 Discussion

We propose a method that can use information from external covariates to guide penalization in regression tasks and that can provide a flexible and scalable alternative to approaches that were proposed recently [22, 203]. We illustrated in simulations and data from biological applications that if the covariate is informative of the effect sizes in the model, these approaches can improve upon commonly used penalized regression methods that are agnostic to such information. We investigated the use of important covariates in genomics such as omics type or tissue. The performance of our approach is in many cases comparable to the IPF-Lasso method [22], while scalability is highly improved in terms of the number of feature groups, thereby extending the range of possible applications.

The variational inference framework provides improved scalability compared to Bayesian methods that are based on sampling strategies. Variational Bayes methods have already been employed in the setting of Bayesian regression with spike and slab priors [31, 32]. However, these methods do not incorporate information from external covariates. A drawback of variational methods are too concentrated approximations to the posterior distribution. Nevertheless, they have been shown to provide reasonable point estimates in regression tasks [31], which we focused on here. Due to the mean-field assumption strong correlations between active predictors can lead to suboptimal results of graper. Here, a multivariate mean-field assumption in the variational Bayes approach can be of advantage, suggested as an alternative above. However, it comes at the price of higher computational costs. What is not addressed in our current implementation is the common problem of missing values in the data; if extant, they would need to be imputed beforehand.

While our approach is related to methods that adapt the penalty function in order to incorporate structural knowledge, such as the group Lasso [210], sparse group Lasso [67] or fused Lasso [185], these approaches apply the same penalty parameter to all the different groups and perform hard in- or exclusion of groups instead of the softer weighting proposed here. Alternatively, the loss function can be modified to incorporate prior knowledge based on a known set of 'high-confidence predictors' as in [102]. The existence and identity of such 'high-confidence predictors', however, is often not clear.

In contrast to frequentist regression methods, the Bayesian approach provides direct posterior-inclusion probabilities for each feature that can be useful for model selection. To obtain frequentist guarantees on the selected features it could be promising to combine the approach with recently developed methods for controlling the false discovery rate (FDR), such as the knockoffs [30]. For this, feature statistics can be constructed based on the estimated coefficients or inclusion probabilities from our model as long as the knockoffs obtain the same covariate information as their true counterpart.

An interesting question that we have not addressed is the quest for rigorous criteria when the inclusion of a covariate by differential penalization is of advantage. This question is not limited to the framework of penalised regression but affects the general setting of shrinkage estimation. While joint shrinkage of a set of estimates can be very powerful in producing more stable estimates with reduced variance, care needs to be taken on which measurements to combine in such a shrinkage approach. As in the case of coefficients in the linear model setting, external covariates could be helpful for this decision and facilitate a more informed shrinkage. However, allowing for differential shrinkage will re-introduce some degrees of freedom into the model and can only be advantageous if the covariate provides 'sufficient' information to balance this. For future work, it would be of interest to find general conditions for when this is the case, thereby enabling an informed choice of covariates in practice.

We provide an open-source implementation of our method in the R package graper. In addition, vignettes and scripts are made available that facilitate the comparison of graper with various related regression methods and can be used to reproduce all results contained in this work.

5.A Appendix

Here we provide details on the variational inference scheme (Section 5.A.1), the updates in our model (Section 5.A.2) and practical considerations for training (Section 5.A.3). As before, $\mathbf{X}$ denotes the $n \times p$ matrix of observed predictors and y the n-vector of observed response values. With $\mathbf{X}_{ij}$ we denote the $(i,j)^{th}$ element of the matrix $\mathbf{X}$ and with $\mathbf{X}_{\cdot,j}$ its j^{th} column. Furthermore, $y_i \in \mathbb{R}$ denotes the i^{th} response value and $x_i \in \mathbb{R}^p$ the i^{th} predictor vector, corresponding to the i^{th} row in $\mathbf{X}$. We will use $\mathbb{E} = \mathbb{E}_q$ to denote expectations with respect to the variational distribution q. As before, g denotes the function defined in Equation (5.5) that maps an index $j \in \{1,\dots,p\}$ to a group $g(j) \in \{1,\dots,G\}$ and $\mathcal{G}_k$ denotes the set of indices in a group k, i. e. $\mathcal{G}_k = \{j \in \{1,\dots,p\} | g(j) = k\}$.

5.A.1 Variational inference

To arrive at a simple iterative algorithm, we make use of the following lemma, which provides an update rule for each factor in the variational distribution [21].

Lemma 1 *Under the mean-field assumption and for a fixed j the evidence lower bound defined in Equation (5.14) is maximised by*

$$\log(q_j^*(\theta_j)) = \mathbb{E}_{-j}(\log(p(y,\theta))) - \text{const},$$

where the expectation is taken under the current variational distribution $\prod_{l \neq j} q(\theta_l)$.

This can be easily seen by writing

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_q\left(\log \frac{p(y,\theta)}{q(\theta)}\right) \\
&= \int q(\theta)(\log p(y,\theta) - \log q(\theta))d\theta \\
&= \int q_j(\theta_j) \int (\log p(y,\beta,\gamma,\tau) - \log q(\theta_j)) \prod_{i \neq j} q(\theta_i) d\theta_{-j} d\theta_j \\
&\quad - \int \sum_{i \neq j} \log q(\theta_i) \prod_{i \neq j} q(\theta_i) \int q(\theta_j) d\theta_j d\theta_{-j} \\
&= \int q_j(\theta_j)(\mathbb{E}_{-j}(\log p(y,\theta)) - \log q(\theta_j))d\theta_j - \text{const} \\
&= \int q_j(\theta_j) \log\left(\frac{\exp \mathbb{E}_{-j}(\log p(y,\theta))}{q(\theta_j)}\right) d\theta_j - \text{const} \\
&= -D_{\mathrm{KL}}(q_j \,||\, \exp \mathbb{E}_{-j}(\log p(y,\theta))).
\end{aligned}$$

Hence, after normalising, the distribution $q_j^*(\theta_j)$ is given by

$$q_j^*(\theta_j) = \frac{\exp(\mathbb{E}_{-j}[\log p(y,\theta)])}{\int \exp(\mathbb{E}_{-j}[\log p(y,\theta)])d\theta_j}.$$

5.A.2 Update equations for the variational inference

5.A.2.1 Linear regression model

In the linear model we assume that the likelihood is given by

$$y|\beta,\tau \sim \mathcal{N}\left(\mathbf{X}\beta, \frac{1}{\tau}\mathbb{1}\right).$$

With the priors as described in Section 5.2.3 and again denoting $\beta = sb$ the joint distribution is given by

$$p(y, b, s, \gamma, \pi, \tau) = p(y|b, s, \tau)p(b, s|\pi, \gamma)p(\gamma)p(\pi)p(\tau).$$

Hence,

$$\begin{aligned}
\log p(y, b, s, \gamma, \pi, \tau) = \text{const} &+ \frac{n}{2}\log(\tau) - \frac{\tau}{2}||y - \mathbf{X}(b \odot s)||_2^2 \\
&+ \sum_{j=1}^{p} \left\{\log(\pi_{g(j)})s_j + \log(1-\pi_{g(j)})(1-s_j)\right\} \\
&+ \sum_{j=1}^{p} \left\{\frac{1}{2}\log(\gamma_{g(j)}) - \frac{\gamma_{g(j)}}{2}b_j^2\right\} \\
&+ \sum_{k=1}^{G} \left\{(r_\gamma - 1)\log(\gamma_k) - d_\gamma \gamma_k\right\} \\
&+ \sum_{k=1}^{G} \left\{(d_\pi - 1)\log(\pi_k) + (r_\pi - 1)\log(1-\pi_k) - \log(B(d_\pi, r_\pi)\right\} \\
&+ (r_\tau - 1)\log(\tau) - d_\tau \tau,
\end{aligned}$$

where $\odot$ denotes the Hadamard-product. The dense model without the spike and slab component arises as a special case when dropping π and s from the model and setting $\beta = b$.

For the start we will make a full mean-field assumption, i.e.

$$q(b, s, \gamma, \pi, \tau) = \prod_{j=1}^{p} q(b_j, s_j)q(\gamma)q(\pi)q(\tau),$$

allowing only a joint distribution for (b_j, s_j) due to their strong dependencies [188].

Denoting with θ all individual parameter components in the mean-field assumption, the updates are given by

$$\log(q_j(\theta_j)) = \mathbb{E}_{-j}\log(p(y, \theta)),$$

as shown above (Lemma 1). Thanks to conjugacy between the chosen priors and the likelihood these updates maintain the distributional family of θ_j reducing the inference to updates of their parameters in each step. Explicitly, this leads to the following updates in step l:

Updates for β (b and s) For β one notes

$$\begin{aligned}
&\log(q(b_j, s_j)) \\
&\quad = -\mathbb{E}\frac{\tau}{2}\mathbb{E}_{-j}||y - \mathbf{X}(b \odot s)||_2^2 + \mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}}s_j - \frac{\mathbb{E}\gamma_{g(j)}}{2}b_j^2 + \text{const} \\
&\quad = -\mathbb{E}\frac{\tau}{2}\left(b_j s_j \sum_k \left(-2y_k \mathbf{X}_{kj} + 2\sum_{l \neq j}\mathbf{X}_{kl}\mathbf{X}_{kj}\mathbb{E}(s_l b_l)\right) + s_j b_j^2 \sum_k \mathbf{X}_{kj}^2\right) \\
&\qquad + \mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}}s_j - \frac{\mathbb{E}\gamma_{g(j)}}{2}b_j^2 + \text{const}.
\end{aligned}$$

This can be written as

$$q(b_j, s_j) = q(s_j = 0)q(b_j|s_j = 0) + q(s_j = 1)q(b_j|s_j = 1),$$

where

$$\begin{aligned} b_j|s_j = 0 &\sim \mathcal{N}(0, (\mathbb{E}\gamma_{g(j)})^{-1}), \\ b_j|s_j = 1 &\sim \mathcal{N}(\mu_j^{(l)}, \sigma_j^{(l)2}), \end{aligned}$$

with

$$\begin{aligned} \sigma_j^{(l)2} &= (\mathbb{E}\tau||\mathbf{X}_{\cdot,j}||_2^2 + \mathbb{E}\gamma_{g(j)})^{-1}, \\ \mu_j^{(l)} &= \sigma_j^{(l)2}\mathbb{E}\tau\left(-\sum_{k=1}^{n}\sum_{l\neq j}^{p}\mathbf{X}_{kj}\mathbf{X}_{kl}\mathbb{E}(\beta_l) + \mathbf{X}_{\cdot,j}^T y\right). \end{aligned}$$

To make this scale linearly in p in the inner loop we follow [31] and keep track of $v = \mathbf{X}\mu$ and update this only in the new component $v \leftarrow v + (\mu_j^{(\text{new})} - \mu_j)\mathbf{X}_{\cdot,j}$.

The marginal distribution of s_j is given by $s_j \sim \text{Ber}(\psi_j^{(l)})$ with $\psi_j^{(l)}$ obtained from

$$\begin{aligned} \text{logit}(\psi_j^{(l)}) &= \mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}} - \frac{1}{2}\log(\mathbb{E}\tau||\mathbf{X}_{\cdot,j}||_2^2 + \mathbb{E}\gamma_{g(j)}) + \frac{1}{2}\log(\mathbb{E}\gamma_{g(j)}) \\ &\quad + \frac{(\mathbb{E}\tau)^2\left(\mathbf{X}_{\cdot,j}^T y - \sum_{k=1}^n\sum_{l\neq j}^p \mathbf{X}_{kj}\mathbf{X}_{kl}\mathbb{E}(b_l s_l)\right)^2}{2(\mathbb{E}\tau||\mathbf{X}_{\cdot,j}||_2^2 + \mathbb{E}\gamma_{g(j)})^{-1}} \\ &= \mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}} + \frac{1}{2}\log(\mathbb{E}\gamma_{g(j)}) + \frac{1}{2}\log(\sigma_j^2) + \frac{1}{2}\frac{\mu_j^2}{\sigma_j^2}. \end{aligned}$$

This is derived by integrating the joint density of $q(b_j, s_j)$ to obtain the marginal density of s_j. Denoting the normal density with $\varphi(\cdot;\mu,\sigma^2)$ we have

$$\begin{aligned} q(s_j) &= \int q(b_j, s_j)db_j \\ &\propto \exp\left(\mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}}s_j\right) \\ &\qquad \int \exp\left(-\mathbb{E}\frac{\tau}{2}\mathbb{E}_{-j}||y - \mathbf{X}(b\odot s)||_2^2 - \frac{\mathbb{E}\gamma_{g(j)}}{2}b_j^2\right)db_j \\ &\propto \exp\left(\mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}}s_j\right) \\ &\qquad \int \varphi(b_j;\mu_j(s_j),\sigma_j^2(s_j))\sqrt{2\pi\sigma_j^2(s_j)}\exp\left(\frac{\mu_j(s_j)^2}{2\sigma_j^2(s_j)}\right)db_j \\ &\propto \exp\left(\mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}}s_j\right)\sqrt{\sigma_j^2(s_j)}\exp\left(\frac{\mu_j(s_j)^2}{2\sigma_j^2(s_j)}\right)\cdot 1 \\ &= \exp\left(\mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}}s_j + \frac{1}{2}\log\sigma_j^2(s_j) + \frac{\mu_j(s_j)^2}{2\sigma_j^2(s_j)}\right). \end{aligned}$$

Hence,

$$
\begin{aligned}
\log q(s_j) &= \text{const} + \mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}}s_j + \frac{1}{2}\log\sigma_j^2(s_j) + \frac{\mu_j(s_j)^2}{2\sigma_j^2(s_j)} \\
&= \text{const} + s_j\mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}} - \frac{1}{2}\log(s_j\mathbb{E}\tau||\mathbf{X}_{\cdot,j}||_2^2 + \mathbb{E}\gamma_{g(j)}) \\
&\quad + \frac{s_j^2(\mathbb{E}\tau)^2\left(\mathbf{X}_{\cdot,j}^T y - \sum_{k=1}^n\sum_{l\neq j}^p \mathbf{X}_{kj}\mathbf{X}_{kl}\mathbb{E}(b_l s_l)\right)^2}{2(s_j\mathbb{E}\tau||\mathbf{X}_{\cdot,j}||_2^2 + \mathbb{E}\gamma_{g(j)})^{-1}} \\
&= \text{const} + s_j\left\{\mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}} - \frac{1}{2}\log(\mathbb{E}\tau||\mathbf{X}_{\cdot,j}||_2^2 + \mathbb{E}\gamma_{g(j)})\right. \\
&\quad \left. + \frac{1}{2}\log(\mathbb{E}\gamma_{g(j)}) + \frac{(\mathbb{E}\tau)^2\left(\mathbf{X}_{\cdot,j}^T y - \sum_{k=1}^n\sum_{l\neq j}^p \mathbf{X}_{kj}\mathbf{X}_{kl}\mathbb{E}(b_l s_l)\right)^2}{2(\mathbb{E}\tau||\mathbf{X}_{\cdot,j}||_2^2 + \mathbb{E}\gamma_{g(j)})^{-1}}\right\}.
\end{aligned}
$$

In the last steps note $s \in \{0,1\}$. Comparing this to $s \sim \text{Ber}(\psi)$ where $\log(q(s)) = \text{const} + s\,\text{logit}(\psi)$ we get the above formula for $\psi^{(l)}$.

Taken together, $\beta_j = s_j b_j \sim \delta_0(1-\psi_j^{(l)}) + \psi_j^{(l)}\mathcal{N}(\mu_j^{(l)}, \sigma_j^{(l)2})$.

Updates for $\gamma = (\gamma_1, \ldots, \gamma_G)$

$$
\begin{aligned}
\log q(\gamma) &= \text{const} + \sum_{j=1}^p\left\{\frac{1}{2}\log(\gamma_{g(j)}) - \frac{\gamma_{g(j)}}{2}\mathbb{E}b_j^2\right\} \\
&\quad + \sum_{k=1}^G\{(r_\gamma - 1)\log(\gamma_k) - d_\gamma\gamma_k\} \\
&= \text{const} + \sum_{k=1}^G\left\{\log(\gamma_k)(r_\gamma - 1 + \frac{1}{2}|\mathcal{G}_k|) - \gamma_k(d_\gamma + \frac{1}{2}\sum_{j\in\mathcal{G}_k}\mathbb{E}b_j^2)\right\}
\end{aligned}
$$

Thus, $\gamma_k \sim \text{Gamma}(\alpha_k^{\gamma,(l)}, \beta_k^{\gamma,(l)})$ are independent gamma distributions with parameters in step l given by

$$
\begin{aligned}
\alpha_k^{\gamma,(l)} &= r_\gamma + \frac{1}{2}|\mathcal{G}_k|, \\
\beta_k^{\gamma,(l)} &= d_\gamma + \frac{1}{2}\sum_{j\in\mathcal{G}_k}\mathbb{E}b_j^2.
\end{aligned}
$$

Updates for τ

$$
\log q(\tau) = \text{const} + \frac{n}{2}\log(\tau) - \frac{\tau}{2}\mathbb{E}||y - \mathbf{X}\beta||_2^2 + (r_\tau - 1)\log(\tau) - d_\tau\tau
$$

Thus, $\tau \sim \text{Gamma}(\alpha^{\tau,(l)}, \beta^{\tau,(l)})$ is a gamma distribution with parameters in step l given by

$$
\begin{aligned}
\alpha^{\tau,(l)} &= r_\tau + \frac{n}{2}, \\
\beta^{\tau,(l)} &= d_\tau + \frac{1}{2}\mathbb{E}\beta||y - \mathbf{X}\beta||_2^2.
\end{aligned}
$$

Updates for $\pi = (\pi_1, \ldots, \pi_G)$

$$\begin{aligned}
\log q(\pi) = \text{const} &+ \sum_{j=1}^{p}\{\log(\pi_{g(j)})\mathbb{E}s_j + \log(1-\pi_{g(j)})(1-\mathbb{E}s_j)\}\\
&+ \sum_{k=1}^{G}\{(d_\pi - 1)\log(\pi_k) + (r_\pi - 1)\log(1-\pi_k) - \log(B(d_\pi, r_\pi)\}\\
= \sum_{k=1}^{G}&\{\log(\pi_k)(d_\pi - 1 + \sum_{j\in\mathcal{G}_k}\mathbb{E}s_j) + \log(1-\pi_k)(r_\pi - 1 + \sum_{j\in\mathcal{G}_k}(1-\mathbb{E}s_j))\}
\end{aligned}$$

Thus, $\pi_k \sim \text{Beta}(\alpha_k^{\pi,(l)}, \beta_k^{\pi,(l)})$ are independent beta distributions with parameters in step l given by

$$\begin{aligned}
\alpha_k^{\pi,(l)} &= d_\pi + \sum_{j\in\mathcal{G}_k}\mathbb{E}s_j,\\
\beta_k^{\pi,(l)} &= r_\pi + \sum_{j\in\mathcal{G}_k}(1-\mathbb{E}s_j).
\end{aligned}$$

Expected values required The updates above involve the calculation of expected values under the current variational distribution q. These are given by

$$\begin{aligned}
\mathbb{E}\tau &= \frac{\alpha^\tau}{\beta^\tau},\\
\mathbb{E}\gamma_k &= \frac{\alpha_k^\gamma}{\beta_k^\gamma},\\
\mathbb{E}\log\frac{\pi_k}{1-\pi_k} &= \psi(\alpha_k^\pi) - \psi(\beta_k^\pi),\\
\mathbb{E}s_j &= \psi_j,\\
\mathbb{E}||y - \mathbf{X}\beta||_2^2 &= y^T y - 2y^T\mathbf{X}\mu_\beta + \sum_{i,j}(\mathbf{X}^T\mathbf{X})_{ij}(\Sigma_{ij}^\beta + \mu_i^\beta\mu_j^\beta),\\
\mathbb{E}b_j &= \psi_j\mu_j,\\
\mathbb{E}b_j^2 &= (1-\psi_j)\left(\mathbb{E}\gamma_{g(j)}^{-1}\right) + \psi_j\left(\mu_j^2 + \sigma_j^2\right),\\
\mathbb{E}\beta_j = \mathbb{E}b_j s_j &= \mu_j\psi_j,\\
\mathbb{E}\beta_j^2 = \mathbb{E}b_j^2 s_j &= (\mu_j^2 + \sigma_j^2)\psi_j.
\end{aligned}$$

Here, ψ denotes the digamma function $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ and

$$\begin{aligned}
\mu^\beta &= (\mathbb{E}\beta_j)_{j=1}^p = (\mathbb{E}b_j s_j)_{j=1}^p,\\
\Sigma^\beta &= \text{diag}(\text{Var}(\beta_j)_{j=1}^p) = \text{diag}((\mathbb{E}\beta_j^2 - (\mathbb{E}\beta_j)^2)_{j=1}^p).
\end{aligned}$$

Note that here and in the following we dropped the step index (l) of all parameters from the notation for simplicity.

Calculation of the evidence lower bound The evidence lower bound bounds the log model evidence from below and can be calculated in each step to monitor convergence. Recall

$$\log(y) = \mathcal{L}(q) + D_{\mathrm{KL}}(q\,||\,p)$$

with

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q \left(\log \frac{p(y, b, s, \gamma, \pi, \tau)}{q(b, s, \gamma, \pi, \tau)} \right) \\ &= \mathbb{E}_q \left(\log p(y, b, s, \gamma, \pi, \tau) \right) + H(q(b, s, \gamma, \pi, \tau)) \\ &= \mathbb{E}_q \left(\log p(y, b, s, \gamma, \pi, \tau) \right) + \sum_{j=1}^{p} H(q(b_j, s_j)) \\ &\quad + H(q(\gamma)) + H(q(\pi)) + H(q(\tau)),\end{aligned}$$

where $H(q) = \int -q(\theta) \log q(\theta) d\theta$ denotes the differential entropy. The terms from the joint model density are given by

$$\begin{aligned}\mathbb{E}_q \log p(y, b, s, \gamma, \tau) = {} & \mathbb{E}_q \log p(y|b, s, \tau) + \mathbb{E}_q \log p(b|\gamma) + \mathbb{E}_q \log p(s|\pi) \\ & + \mathbb{E}_q \log p(\gamma) + \mathbb{E}_q \log p(\pi) + \mathbb{E}_q \log p(\tau)\end{aligned}$$

with

$$\begin{aligned}\mathbb{E}_q \log p(y|\beta, \tau) &= \frac{n}{2} \mathbb{E} \log(\tau) - \frac{1}{2} \mathbb{E}\tau ||y - \mathbf{X}(b \odot s)||_2^2 - \frac{n}{2} \log(2\pi), \\ \mathbb{E}_q \log p(b|\gamma) &= \sum_j \left(\frac{1}{2} \mathbb{E} \log(\gamma_{g(j)}) - \frac{1}{2} \mathbb{E}\gamma_{g(j)} b_j^2 - \frac{1}{2} \log(2\pi) \right), \\ \mathbb{E}_q \log p(s|\pi) &= \sum_j \left(\mathbb{E} s_j \log(\pi_{g(j)}) + \mathbb{E}(1 - s_j) \log(1 - \pi_{g(j)}) \right), \\ \mathbb{E}_q \log p(\gamma) &= \sum_k \left((r_\gamma - 1) \mathbb{E} \log(\gamma_k) - d_\gamma \mathbb{E}\gamma_k - \log(\Gamma(r_\gamma)) + r_\gamma \log(d_\gamma) \right), \\ \mathbb{E}_q \log p(\pi) &= \sum_k \left((d_\pi - 1) \mathbb{E} \log(\pi_k) + (r_\pi - 1) \mathbb{E} \log(1 - \pi_k) - \log \mathrm{B}(d_\pi, r_\pi) \right), \\ \mathbb{E}_q \log p(\tau) &= (r_\tau - 1) \mathbb{E} \log(\tau) - d_\tau \mathbb{E}\tau - \log(\Gamma(r_\tau)) + r_\gamma \log(d_\tau).\end{aligned}$$

Here, $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ denoted the beta function. The required expectations in addition to those used in the updates are easily obtained using the known distributions and parameters of the variational density in each iteration and the fact that q factorizes, i.e.

$$\begin{aligned}\mathbb{E} \log(\tau) &= \psi(\alpha_\tau) - \log(\beta_\tau), \\ \mathbb{E}\tau ||y - \mathbf{X}(b \odot s)||_2^2 &= \mathbb{E}\tau \mathbb{E} ||y - \mathbf{X}(b \odot s)||_2^2, \\ \mathbb{E} \log(\gamma_k) &= \psi(\alpha_k^\gamma) - \log(\beta_k^\gamma) \\ \mathbb{E}\gamma_{g(j)} \beta_j^2 &= \mathbb{E}\gamma_{g(j)} \mathbb{E}\beta_j^2, \\ \mathbb{E} \log(\pi_k) &= \psi(\alpha_k^\pi) - \psi(\alpha_k^\pi + \beta_k^\pi), \\ \mathbb{E}(1 - \log(\pi_k)) &= \psi(\beta_k^\pi) - \psi(\alpha_k^\pi + \beta_k^\pi).\end{aligned}$$

The entropies are derived from the known expression for the entropy of the gamma, beta, Bernoulli and normal distribution, i.e.

$$\begin{aligned}
H(q(b_j, s_j)) &= H(q(b_j|s_j)) + H(q(s_j))\\
H(q(b_j|s_j)) &= \frac{1}{2}(\log(2\pi) + 1) - \frac{1}{2}\log(s_j \mathbb{E}\tau ||\mathbf{X}_{\cdot,j}||_2^2 + \gamma_{g(j)})\\
H(q(s_j)) &= -(1-\psi_j)\log(1-\psi_j) - \psi_j \log(\psi_j)\\
H(q(\gamma)) &= \sum_k \left(\alpha_k^\gamma - \log(\beta_k^\gamma) + \log(\Gamma(\alpha_k^\gamma)) + (1-\alpha_k^\gamma)\psi(\alpha_k^\gamma)\right)\\
H(q(\pi)) &= \sum_k (\log \mathrm{B}(\alpha_k^\pi, \beta_k^\pi) - (\alpha_k^\pi - 1)\psi(\alpha_k^\pi) - (\beta_k^\pi - 1)\psi(\beta_k^\pi)\\
&\qquad + (\alpha_k^\pi + \beta_k^\pi - 2)\psi(\alpha_k^\pi + \beta_k^\pi))\\
H(q(\tau)) &= \alpha^\tau - \log(\beta^\tau) + \log(\Gamma(\alpha^\tau)) + (1-\alpha^\tau)\psi(\alpha^\tau).
\end{aligned}$$

Multivariate mean-field approximation for β The assumption that the variational distribution $q(\beta)$ factorizes across all predictors can be very strong. Therefore, a more accurate approximation of the true posterior can be obtained by allowing for a p-variate distribution for β.

For $s = 1$, i.e. no spike term in the model, and hence $\beta = b$ the joint distribution in the updates is then given by

$$\log q(\beta) = \text{const} - \frac{\mathbb{E}(\tau)}{2}||y - \mathbf{X}\beta||_2^2 + \sum_{j=1}^p \left\{ -\frac{\mathbb{E}(\gamma_{g(j)})}{2}\beta_j^2 \right\}$$

Thus, $\beta \sim \mathcal{N}(\mu^{(l)}, \Sigma^{(l)})$ is a normal distribution with parameters

$$\begin{aligned}
\mu^{(l)} &= \mathbb{E}(\tau)\Sigma^{(l)}\mathbf{X}^T y,\\
\Sigma^{(l)} &= (\mathbb{E}(\tau)\mathbf{X}^T\mathbf{X} + D)^{-1} \quad \text{with } D = \mathrm{diag}((\mathbb{E}(\gamma_{g(j)}))_{j=1}^p).
\end{aligned}$$

The other updates stay the same, where the covariance matrix Σ is now no longer diagonal as previously. As this update requires the inversion of a $p \times p$ matrix, a limiting factor for applying the multivariate mean-field approximation is its computational complexity. When n is small compared to p a better solution is to employ the Woodbury-Matrix identity [206], i.e.

$$\Sigma^{(l)} = D - D\mathbf{X}^T((\mathbb{E}(\tau))^{-1}\mathbb{1}_n + \mathbf{X}D\mathbf{X}^T)^{-1}\mathbf{X}D,$$

which requires the inversion of a $n \times n$ matrix only. This multivariate assumption can be useful in the presence of strong correlations between the predictors. In the case where $\mathbf{X}^T\mathbf{X}$ is diagonal we obtain a similar form than for a fully factorized variational distribution.

The evidence lower bound is obtained analogous to the fully factorized case with a multivariate normal distribution and dropping the terms involving s and π. In particular

$$\mathcal{L}(q) = \mathbb{E}_q(\log p(y, \beta, \gamma, \tau)) + H(q(\beta)) + H(q(\gamma)) + H(q(\tau)),$$

with

$$H(q(\beta)) = \frac{p}{2}(\log(2\pi) + 1) + \frac{1}{2}\log(|\Sigma|).$$

5.A.2.2 Logistic regression model

In order to adapt the model to binary data, we change the likelihood of Y to a Bernoulli distribution and consider a generalized linear model with logistic link function, i.e.

$$y_i|\beta \sim \mathrm{Ber}(\sigma(x_i^T\beta)) \quad \text{with} \quad \sigma(z) = \frac{1}{1+\exp(-z)}.$$

The priors on the model coefficients β remain the same as in the linear model, the noise variance τ is dropped from the model. While the model is strongly related to the case of the normal response variable, the challenge here lies in the fact that with the Bernoulli distribution for Y we loose the conjugacy of the prior from the linear model. To solve this and still obtain a fast and explicit inference scheme, we use an approximation of the sigmoid function by an exponential of a quadratic term, thus restoring conjugacy.

As $\sigma(-a) = 1 - \sigma(a)$ we can write

$$\begin{aligned}\mathbb{P}(y_i = 1|\beta) &= \sigma(x_i^T\beta),\\ \mathbb{P}(y_i = 0|\beta) &= \sigma(-x_i^T\beta),\end{aligned}$$

and hence the likelihood is given by

$$p(y_i|\beta) = \sigma((2y_i - 1)x_i^T\beta).$$

Following [101] we use the following lower bound on the sigmoid

$$\sigma(z) \geq \sigma(\xi)\exp\left(\frac{1}{2}(z-\xi) - \eta(\xi)(z^2-\xi^2)\right), \qquad \eta(\xi) = \frac{1}{2\xi}\left(\sigma(\xi) - \frac{1}{2}\right).$$

This introduces an additional variational parameter ξ, which we update alongside the other updates to improve this approximation in each iteration.

Using this approximation we can bound the joint density of the model by

$$\begin{aligned}p(y,\beta,\gamma) &= p(y|\beta)p(\beta|\gamma,\pi)p(\gamma)p(\pi)\\ &\geq h(\beta,\xi)p(\beta|\gamma,\pi)p(\gamma)p(\pi),\end{aligned}$$

with

$$\begin{aligned}\log h(\beta,\xi) = &\frac{1}{2}\sum_i(2y_i-1)x_i^T\beta - \sum_i \eta(\xi_i)(x_i^T\beta)^2\\ &+ \sum_i\left(\log(\sigma(\xi_i)) - \frac{1}{2}\xi_i + \eta(\xi_i)\xi_i^2\right).\end{aligned} \tag{5.21}$$

With the fully factorised mean-field assumption we get the following updates:

$$\begin{aligned}\log(q(b_j,s_j)) &= \text{const} + \log h(\beta,\xi) - \frac{\mathbb{E}(\gamma_{g(j)})}{2}b_j^2 + \mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}}s_j\\ &= \text{const} + \frac{1}{2}\sum_i(2y_i-1)x_i^T\beta - \sum_i\eta(\xi_i)(x_i^T\beta)^2\\ &\quad - \frac{\mathbb{E}(\gamma_{g(j)})}{2}b_j^2 + \mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}}s_j\\ &= \text{const} + \left(\sum_i(y_i - \frac{1}{2})\mathbf{X}_{ij}\right)b_js_j - s_jb_j^2\sum_{i=1}^n\eta(\xi_i)\mathbf{X}_{ij}^2\\ &\quad - 2b_js_j\sum_{i=1}^n\eta(\xi_i)\sum_{l\neq j}\mathbf{X}_{il}\mathbf{X}_{ij}\mathbb{E}\beta_l - \frac{\mathbb{E}(\gamma_{g(j)})}{2}b_j^2 + \mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}}s_j.\end{aligned}$$

Analogous to the linear model we can derive the following updates for the coefficients: $b_j|s_j = 0 \sim \mathcal{N}(0, \mathbb{E}\gamma_{g(j)}^{-1})$ and $b_j|s_j = 1 \sim \mathcal{N}(\mu_j, \sigma_j^2)$ with

$$\sigma_j^2 = \left(2\sum_{i=1}^n \eta(\xi_i)\mathbf{X}_{ij}^2 + \mathbb{E}\gamma_{g(j)}\right)^{-1},$$

$$\mu_j = \sigma_j^2 \left(-2\sum_{i=1}^n \eta(\xi_i) \sum_{l\neq j}^p \mathbf{X}_{ij}\mathbf{X}_{il}\mathbb{E}\beta_l + \mathbf{X}_{\cdot,j}^T(y - \frac{1}{2})\right).$$

The probability for $s_j = 1$ is given by

$$\operatorname{logit}(\psi_j^{(l)}) = \mathbb{E}\log\frac{\pi_{g(j)}}{1-\pi_{g(j)}} - \frac{1}{2}\log(\mathbb{E}\gamma_{g(j)}^{-1}) + \frac{1}{2}\log(\sigma_j^2) + \frac{1}{2}\frac{\mu_j^2}{\sigma_j^2},$$

as in the linear model.

In the case of a multivariate mean-field assumption on β we obtain

$$\begin{aligned}
\log q(\beta) &= \text{const} + \log h(\beta,\xi) + \sum_{j=1}^p\left\{-\frac{\mathbb{E}(\gamma_{g(j)})}{2}\beta_j^2\right\} \\
&= \text{const} + \frac{1}{2}\sum_{i=1}^n(2y_i - 1)x_i^T\beta - \sum_{i=1}^n \eta(\xi_i)(x_i^T\beta)^2 + \sum_{j=1}^p\left\{-\frac{\mathbb{E}(\gamma_{g(j)})}{2}\beta_j^2\right\} \\
&= \text{const} + \left(\sum_{i=1}^n (y_i - \frac{1}{2})x_i^T\right)\beta - \beta^T\left(\sum_{i=1}^n \eta(\xi_i)x_i x_i^T\right)\beta \\
&\quad + \sum_{j=1}^p\left\{-\frac{\mathbb{E}(\gamma_{g(j)})}{2}\beta_j^2\right\}.
\end{aligned}$$

Thus, $\beta \sim \mathcal{N}(\mu, \Sigma)$ with parameters

$$\mu = \Sigma\sum_{i=1}^n\left(y_i - \frac{1}{2}\right)x_i,$$

$$\Sigma = \left(2\sum_{i=1}^n\{\eta(\xi_i)x_i x_i^T\} + D\right)^{-1} \quad \text{with } D = \operatorname{diag}((\mathbb{E}\gamma_{g(j)})_{j=1}^p).$$

Relationship to the linear model Note that the analogy to the linear update becomes explicit, when interpreting Equation (5.21) as a normal density on pseudo-data [168] defined by

$$\tilde{y}_i = \frac{2y_i - 1}{4\eta(\xi_i)}.$$

Then it can be easily seen that

$$\log h(\beta,\xi) = \log p(\tilde{y}|\beta) + c(\xi),$$

where

$$\tilde{y}_i|\beta \sim \mathcal{N}(x_i^T\beta, (2\eta(\xi_i))^{-1}).$$

Replacing the precision parameter τ in the linear case with the precision of the pseudo-data (which is now sample-specific) can give us above updates directly from the linear model.

Update for the variational parameter ξ The update of the variational parameter ξ is given following [101] by

$$\xi_i^2 = x_i^T(\Sigma + \mu_l \mu_l^T)x_i,$$

which can be restricted to non-negative values of ξ due to the symmetry.

Evidence lower bound As before

$$\mathcal{L}(q) = \mathbb{E}_q\left(\log p(y, b, s, \gamma, \pi)\right) + \sum_{j=1}^{p} H(q(b_j, s_j)) + H(q(\gamma)) + H(q(\pi)).$$

The entropies can be calculated as in the linear model with the respective parameters of the variational distributions. The terms from the joint model density only differ in the first term

$$\begin{aligned}\mathbb{E}_q \log p(y, b, s, \gamma, \pi) = {} & \mathbb{E}_q \log p(y|b, s) + \mathbb{E}_q \log p(b|\gamma) \\ & + \mathbb{E}_q \log p(s|\pi) + \mathbb{E}_q \log p(\gamma) + \mathbb{E}_q \log p(\pi),\end{aligned}$$

which here is given by

$$\begin{aligned}\mathbb{E}_q \log p(y|\beta) &= \mathbb{E}_q \log \sigma((2y-1)\mathbf{X}\beta) \\ &\geq \mathbb{E}_q \Bigg(\frac{1}{2}\sum_i (2y_i - 1)x_i^T \mu - \sum_i \eta(\xi_i)(x_i^T\beta)^2 \\ &\qquad + \sum_i \left(\log(\sigma(\xi_i)) - \frac{1}{2}\xi_i + \eta(\xi_i)\xi_i^2 \right) \Bigg) \\ &= \frac{1}{2}\sum_i \log(2\eta(\xi_i)) - \frac{1}{2}\sum_i 2\eta(\xi_i)(\tilde{y}_i - x_i^T\mu)^2 + \text{const.}\end{aligned}$$

This provides a lower bound on the evidence lower bound in analogy to the linear model that is used to monitor convergence.

5.A.3 Practical considerations

5.A.3.1 Standardization of the predictors

In penalised regression a common preprocessing step is the standardization of all predictors to unit variance to ensure a presumably 'fair' penalty. This scaling is in 1:1 correspondence to differential penalty factors. Without standardization features on a larger scale would be preferred as they need a smaller coefficient relative to a feature with the same effect but measured on a smaller scale. However, standardization can be suboptimal as it does not distinguish between meaningful differences in variance (e.g. features that differ between two disease groups) and differences in variance due to the scale. While removal of the latter would be desirable, meaningful differences should be retained. For example, in many applications we measure high-amplitude signals that are informative jointly with low-amplitude features that originate mainly from technical noise. Here, standardization can be harmful (Figure 5.A.1). Hence, the question of whether to scale the predictors or not, is related to the question of whether the variance of a feature is an informative covariate.

By default, our method standardizes all features. However, if we want to maintain the difference of variances within each assay, our method can adaptively learn scale differences between assays by γ, thereby removing the need to standardize for adjustment between

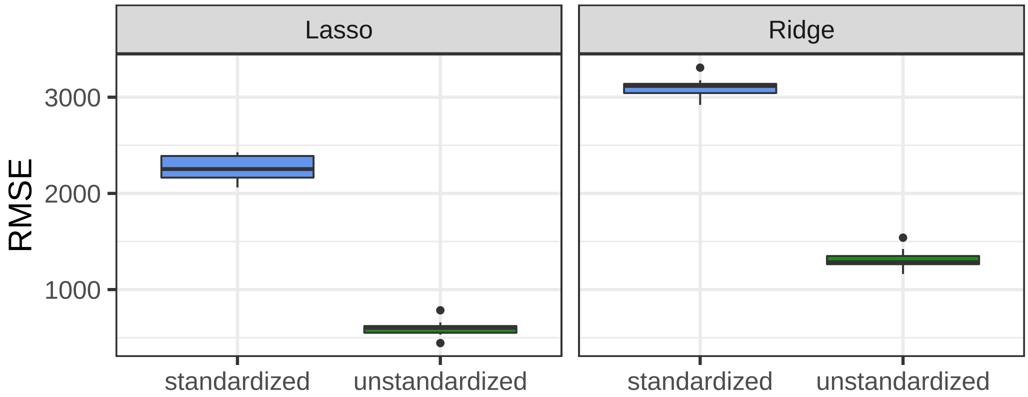


Figure 5.A.1: Simulation example illustrating the effect of standardization in settings with informative high-amplitude features and uninformative low-amplitude features. A number of $p = 600$ features was simulated from a standard normal distribution and multiplied by 10 (high-amplitude features, $p = 300$) or 1 (low-amplitude features, $p = 300$). The response was simulated from a normal model with coefficients given by 1 for the high-amplitude features and 0 otherwise. Lasso and ridge regression were fitted on a training set of $n = 500$ samples using either standardized predictors (blue) or predictors on the original scale (green). The resulting fits were evaluated in terms of the root mean squared error (RMSE) on an independent test set of $n = 500$ samples. The boxplots were obtained from ten independent instances of simulated data.

assays as seen in the CLL application. This could help to retain meaningful differences in the features' variance within one assay. Alternatively, it is also possible to standardize features but re-include information on their variance via the covariate, e.g., binning features based on their variance. A recent study on RNA-Seq data found no strong effect of standardization compared to no standardization [218]. Depending on the data set at hand it might, however, make sense to retain the original scale. For example with binary mutation data, where features are all on the same scale, standardization would favour mutations with lower frequencies.

5.A.3.2 Modelling an intercept

To include an intercept into the model, we apply centering of $\mathbf{X}$ and y before model fitting in the case of a linear model. For the logistic model this is not as straightforward and we follow [31] in the implementation, i.e. the intercept β_0 is assumed to have a normal prior $\mathcal{N}(0, \sigma_0^2)$ but considering the limiting case for σ_0 to infinity yielding an improper prior (essentially not penalizing the intercept).

CHAPTER 6

Conclusions and perspectives

With the growing number of studies involving multiple platforms, conditions, tissues, organisms, time points or locations, integrative approaches and joint analyses of heterogeneous data sets will remain essential. Potential promises to be gained from such studies are far reaching, both for basic research as well as for society as a whole. Improved health care based on molecular information of individuals is only one example. Now it is up to the scientific community at the interface of statistics, computer science, biology and medicine to turn this data into useful information and eventually (actionable) knowledge. In this thesis we have contributed methods for an integrative multivariate analysis of data that is comprised of heterogeneous modalities, taking multi-omics data as a motivating example. In particular, we have addressed the questions, how to use heterogeneous data jointly to uncover main structures in an unbiased manner, and how to relate heterogeneous features to a response of interest. Both questions are essential along the way of realizing the potential of the available data, and in Chapters 4 and 5, respectively, we have proposed possible answers. Importantly, the methods that are contained in this thesis are available as open-source software packages (i. e. MOFAtools, mofa and graper) that are accompanied by detailed documentation and vignettes containing example workflows. We hope that thereby we could improve the accessibility of multivariate integrative methods to researchers in genome biology as well as ensure reproducibility of our results. For the unsupervised method, MOFA, we have already encountered broad interest from the research community, highlighting the need for exploratory tools that combine information from different omics data sets.

Looking back at the statistical tool set that we have employed, we can first highlight the importance of structured regularization, which enabled us to account for the heterogeneity and high-dimensionality of the data in multivariate approaches. We focused here on feature groups with distinct properties, such as different data modalities. However, other types of structure can likewise be exploited using similar approaches, such as temporal, spatial or functional relationships between features. Second, we note, that at several points in this thesis, we have made transitions between frequentist and Bayesian perspectives on a statistical model. While historically these two paradigms were strongly opposed, a more pragmatic thinking has emerged by now and the two paradigms start to cross-fertilize one another, especially in the field of data science. While the concepts of inference and uncertainty quantification are still distinct, Bayesian estimates can be investigated for their frequentist properties, or frequentist estimates can be given Bayesian interpretations, and intermediate frameworks emerge. This can foster the development of powerful, adaptive and flexible methods.

While this thesis has enlarged the tool set available to integrate data from diverse sources, still many promising findings lie hidden in the data, and as the technological developments continuously progress and new data types emerge, we will also steadily encounter novel statistical challenges and opportunities. Currently, this is clearly visible in the field of single cell biology. Here, new characteristics of the data emerge such as a strong heterogeneity of the samples, a vast number of missing values and much larger sample sizes compared to bulk studies, which open up novel statistical approaches of modelling and inference, that hopefully in parts can draw upon ideas and methods presented in this thesis.

References

1. Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. A. & Pe'er, D. An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005–1017 (2010).
2. Åkerfelt, M., Morimoto, R. I. & Sistonen, L. Heat shock factors: integrators of cell stress, development and lifespan. *Nature Reviews Molecular Cell Biology* **11**, 545 (2010).
3. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415 (2013).
4. Alyass, A., Turcotte, M. & Meyre, D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics* **8**, 33 (2015).
5. Andersen, M. R., Vehtari, A., Winther, O. & Hansen, L. K. Bayesian inference for spatio-temporal spike and slab priors. *arXiv preprint arXiv:1509.04752* (2015).
6. Andersson, E., Pützer, S., Yadav, B., Dufva, O., Khan, S., He, L., Sellner, L., Schrader, A., Crispatzu, G., Oleś, M., *et al.* Discovery of novel drug sensitivities in T-PLL by high-throughput ex vivo drug testing and mutation profiling. *Leukemia* **32**, 774 (2018).
7. Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T., *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods* **13**, 229 (2016).
8. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W. & Stegle, O. Multi-Omics Factor Analysis - a framework for unsupervised integration of multi-omic data sets. *Molecular Systems Biology* (2018).
9. Auclair, G., Guibert, S., Bender, A. & Weber, M. Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. *Genome Biology* **15**, 545 (2014).
10. Bach, F. R. & Jordan, M. I. A probabilistic interpretation of canonical correlation analysis. *Technical Report 688, Department of Statistics, University of California, Berkeley* (2005).
11. Bandura, D. R., Baranov, V. I., Ornatsky, O. I., Antonov, A., Kinach, R., Lou, X., Pavlov, S., Vorobiev, S., Dick, J. E. & Tanner, S. D. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical Chemistry* **81**, 6813–6822 (2009).

12. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
13. Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics* **40**, 955 (2008).
14. Basilevsky, A. T. *Statistical factor analysis and related methods: theory and applications* (John Wiley & Sons, 2009).
15. Beal, J. *Variational algorithms for approximate Bayesian inference* (University College London, 2003).
16. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300 (1995).
17. Bergersen, L. C., Glad, I. K. & Lyng, H. Weighted Lasso with data integration. *Statistical Applications in Genetics and Molecular Biology* **10** (2011).
18. Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G. & Milanesi, L. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* **17**, 15 (2016).
19. Bishop, C. M. *Bayesian PCA* in *Advances in Neural Information Processing Systems* (1999), 382–388.
20. Bishop, C. M. Pattern recognition. *Machine Learning* **128**, 1–58 (2006).
21. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112**, 859–877 (2017).
22. Boulesteix, A.-L., De Bin, R., Jiang, X. & Fuchs, M. IPF-LASSO: Integrative L_1-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and Mathematical Methods in Medicine* **2017** (2017).
23. Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y. & Greenleaf, W. J. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486 (2015).
24. Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C. & Stegle, O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biology* **18**, 212 (2017).
25. Buettner, F. & Theis, F. J. A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics* **28**, i626–i632 (2012).
26. Bühlmann, P. & Van De Geer, S. *Statistics for high-dimensional data: methods, theory and applications* (Springer Science & Business Media, 2011).
27. Bunte, K., Leppäaho, E., Saarinen, I. & Kaski, S. Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics* **32**, 2457–2463 (2016).
28. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411 (2018).

29. Cancer Genome Atlas Research Network *et al.* Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169**, 1327–1341 (2017).

30. Candes, E., Fan, Y., Janson, L. & Lv, J. Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351* (2016).

31. Carbonetto, P. & Stephens, M. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis* **7**, 73–108 (2012).

32. Carbonetto, P., Zhou, X. & Stephens, M. varbvs: Fast variable selection for large-scale regression. *arXiv preprint arXiv:1709.06597* (2017).

33. Carvalho, C. M., Polson, N. G. & Scott, J. G. *Handling sparsity via the horseshoe* in *Artificial Intelligence and Statistics* (2009), 73–80.

34. Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., *et al.* Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414 (2016).

35. Chen, R. & Snyder, M. Promise of personalized omics to precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **5**, 73–82 (2013).

36. Cheow, L. F., Courtois, E. T., Tan, Y., Viswanathan, R., Xing, Q., Tan, R. Z., Tan, D. S., Robson, P., Loh, Y.-H., Quake, S. R., *et al.* Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature Methods* **13**, 833 (2016).

37. Chessel, D. & Hanafi, M. Analyses de la co-inertie de K nuages de points. *Revue de Statistique Appliquée* **44**, 35–60 (1996).

38. Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications* **9**, 781 (2018).

39. Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F. & Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences* **102**, 7426–7431 (2005).

40. Colomé-Tatché, M. & Theis, F. Statistical single cell multi-omics integration. *Current Opinion in Systems Biology* **7**, 54–59 (2018).

41. Crick, F. H. *On protein synthesis* in *Symp Soc Exp Biol* **12** (1958), 8.

42. Culhane, A. C., Perrière, G. & Higgins, D. G. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics* **4**, 59 (2003).

43. Damianou, A., Lawrence, N. D. & Ek, C. H. Multi-view learning as a nonparametric nonlinear inter-battery factor analysis. *arXiv preprint arXiv:1604.04939* (2016).

44. De la Cruz, O. & Holmes, S. The duality diagram in data analysis: examples of modern applications. *The Annals of Applied Statistics* **5**, 2266 (2011).

45. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38 (1977).

46. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & Van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nature Biotechnology* **33**, 285 (2015).

47. Dietrich, S., Oleś, M., Lu, J., Sellner, L., Anders, S., Velten, B., Wu, B., Hüllein, J., da Silva Liberio, M., Walther, T., *et al.* Drug-perturbation-based stratification of blood cancer. *The Journal of Clinical Investigation* **128**, 427–445 (2018).

48. Dobriban, E., Fortney, K., Kim, S. K. & Owen, A. B. Optimal multiple testing under a Gaussian prior on the effect sizes. *Biometrika* **102**, 753–766 (2015).

49. Donoho, D. L. & Grimes, C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences* **100**, 5591–5596 (2003).

50. Dray, S., Chessel, D. & Thioulouse, J. Co-inertia analysis and the linking of ecological data tables. *Ecology* **84**, 3078–3089 (2003).

51. Duzkale, H., Schweighofer, C. D., Coombes, K. R., Barron, L. L., Ferrajoli, A., O'Brien, S., Wierda, W. G., Pfeifer, J., Majewski, T., Czerniak, B. A., *et al.* LDOC1 mRNA is differentially expressed in chronic lymphocytic leukemia and predicts overall survival in untreated patients. *Blood* (2011).

52. Eckart, C. & Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* **1**, 211–218 (1936).

53. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**, 207–210 (2002).

54. Efron, B. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction* (Cambridge University Press, 2012).

55. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., *et al.* Least angle regression. *The Annals of Statistics* **32**, 407–499 (2004).

56. Engelhardt, B. E. & Adams, R. P. Bayesian structured sparsity from Gaussian fields. *arXiv preprint arXiv:1407.2235* (2014).

57. Engelhardt, B. E. & Stephens, M. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genetics* **6**, e1001117 (2010).

58. Fabbri, G. & Dalla-Favera, R. The molecular pathogenesis of chronic lymphocytic leukaemia. *Nature Reviews Cancer* **16**, 145 (2016).

59. Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., *et al.* The reactome pathway knowledgebase. *Nucleic Acids Research* **44**, D481–D487 (2015).

60. Falconer, E., Hills, M., Naumann, U., Poon, S. S., Chavez, E. A., Sanders, A. D., Zhao, Y., Hirst, M. & Lansdorp, P. M. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nature Methods* **9**, 1107 (2012).

61. Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., Kaper, F., Fan, J.-B., Zhang, K., Chun, J., *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods* **13**, 241 (2016).

62. Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G. & Kong, A. Unsupervised empirical Bayesian multiple testing with external covariates. *The Annals of Applied Statistics*, 714–735 (2008).
63. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16**, 278 (2015).
64. Fluhr, S., Boerries, M., Busch, H., Symeonidi, A., Witte, T., Lipka, D. B., Mücke, O., Nöllke, P., Krombholz, C. F., Niemeyer, C. M., *et al.* CREBBP is a target of epigenetic, but not genetic, modification in juvenile myelomonocytic leukemia. *Clinical Epigenetics* **8**, 50 (2016).
65. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1 (2010).
66. Friedman, J., Hastie, T. & Tibshirani, R. *The elements of statistical learning* **10** (Springer series in statistics New York, NY, USA: 2001).
67. Friedman, J., Hastie, T. & Tibshirani, R. A note on the group Lasso and a sparse group Lasso. *arXiv preprint arXiv:1001.0736* (2010).
68. Frismantas, V., Dobay, M. P., Rinaldi, A., Tchinda, J., Dunn, S. H., Kunz, J., Richter-Pechanska, P., Marovca, B., Pail, O., Jenni, S., *et al.* Ex vivo drug response profiling detects recurrent sensitivity patterns in drug resistant ALL. *Blood*, e26–e37 (2017).
69. Frost, H. R., Li, Z. & Moore, J. H. Principal component gene set enrichment (PCGSE). *BioData Mining* **8**, 25 (2015).
70. Fuchs, E. Keratins as biochemical markers of epithelial differentiation. *Trends in Genetics* **4**, 277–281 (1988).
71. Fukuyama, J. Adaptive gPCA: A method for structured dimensionality reduction. *arXiv preprint arXiv:1702.00501* (2017).
72. Gagnon-Bartsch, J. A., Jacob, L. & Speed, T. P. Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, 1–112 (2013).
73. Garg, R., Benedetti, L. G., Abera, M. B., Wang, H., Abba, M. & Kazanietz, M. G. Protein kinase C and cancer: what we know and what we do not. *Oncogene* **33**, 5225 (2014).
74. Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
75. Geeleher, P., Cox, N. J. & Huang, R. S. Cancer biomarker discovery is improved by accounting for variability in general levels of drug sensitivity in pre-clinical models. *Genome Biology* **17**, 190 (2016).
76. Geman, S. & Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 721–741 (1984).

77. Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Della Porta, M. G., Jädersten, M., Dolatshad, H., Verma, A., Cross, N. C., Vyas, P., *et al.* Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nature Communications* **6**, 5901 (2015).
78. Gross, S. M. & Tibshirani, R. Collaborative regression. *Biostatistics* **16**, 326–338 (2014).
79. GTEx Consortium *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: multi-tissue gene regulation in humans. *Science* **348**, 648–660 (2015).
80. Guan, Y. & Dy, J. *Sparse probabilistic principal component analysis* in *Artificial Intelligence and Statistics* (2009), 185–192.
81. Guo, F., Li, L., Li, J., Wu, X., Hu, B., Zhu, P., Wen, L. & Tang, F. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell research* **27**, 967 (2017).
82. Hamburg, M. A. & Collins, F. S. The path to personalized medicine. *New England Journal of Medicine* **2010**, 301–304 (2010).
83. Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., *et al.* Mapping the mouse cell atlas by Microwell-seq. *Cell* **172**, 1091–1107 (2018).
84. Harman, H. H. *Modern factor analysis* (University of Chicago press, 1976).
85. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biology* **18**, 83 (2017).
86. Hastie, T. & Stuetzle, W. Principal curves. *Journal of the American Statistical Association* **84**, 502–516 (1989).
87. Hastie, T. & Tibshirani, R. Efficient quadratic regularization for expression arrays. *Biostatistics* **5**, 329–340 (2004).
88. Hernández-Lobato, D., Hernández-Lobato, J. M. & Dupont, P. Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. *The Journal of Machine Learning Research* **14**, 1891–1945 (2013).
89. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
90. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
91. Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K. & Marchini, J. Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics* **48**, 1094 (2016).
92. Hotelling, H. Canonical correlation analysis (CCA). *Journal of Educational Psychology* (1935).
93. Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* **24**, 417 (1933).
94. Hothorn, T. & Lausen, B. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis* **43**, 121–137 (2003).
95. Hsiang, T. A Bayesian view on ridge regression. *Journal of the Royal Statistical Society. Series D (The Statistician)* **24**, 267–268 (1975).

96. Huang, Y. & Sanguinetti, G. BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biology* **18**, 123 (2017).

97. Ignatiadis, N., Klaus, B., Zaugg, J. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods* **13**, 577 (2016).

98. International Human Genome Sequencing Consortium *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860 (2001).

99. Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).

100. Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. EMPIAR: a public archive for raw electron microscopy image data. *Nature Methods* **13**, 387 (2016).

101. Jaakkola, T. S. & Jordan, M. I. Bayesian parameter estimation via variational methods. *Statistics and Computing* **10**, 25–37 (2000).

102. Jiang, Y., He, Y. & Zhang, H. Variable selection with prior information for generalized linear models via the prior Lasso method. *Journal of the American Statistical Association* **111**, 355–376 (2016).

103. Johnstone, I. M. & Lu, A. Y. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104**, 682–693 (2009).

104. Jolliffe, I. T., Trendafilov, N. T. & Uddin, M. A modified principal component technique based on the Lasso. *Journal of Computational and Graphical Statistics* **12**, 531–547 (2003).

105. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning* **37**, 183–233 (1999).

106. Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research* **33**, D428–D432 (2005).

107. Joyce, A. R. & Palsson, B. Ø. The model organism as a system: integrating'omics' data sets. *Nature Reviews Molecular Cell Biology* **7**, 198 (2006).

108. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nature Reviews Genetics* **19**, 299 (2018).

109. Khan, S. A., Virtanen, S., Kallioniemi, O. P., Wennerberg, K., Poso, A. & Kaski, S. Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. *Bioinformatics* **30**, i497–i504 (2014).

110. Kim, M., Rai, N., Zorraquino, V. & Tagkopoulos, I. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. *Nature Communications* **7**, 13090 (2016).

111. Klami, A. & Kaski, S. *Local dependent components* in *Proceedings of the 24th International Conference on Machine Learning* (2007), 425–432.

112. Klami, A., Virtanen, S. & Kaski, S. Bayesian canonical correlation analysis. *Journal of Machine Learning Research* **14**, 965–1003 (2013).

113. Klami, A., Virtanen, S., Leppäaho, E. & Kaski, S. Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems* **26**, 2136–2147 (2015).

114. Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A. & Kirschner, M. W. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

115. Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T., *et al.* ArrayExpress update - simplifying data submissions. *Nucleic Acids Research* **43**, D1113–D1116 (2014).

116. Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I. & Noble, W. S. A statistical framework for genomic data fusion. *Bioinformatics* **20**, 2626–2635 (2004).

117. Lawrence, N. D. *Gaussian process latent variable models for visualisation of high dimensional data* in *Advances in Neural Information Processing Systems* (2004), 329–336.

118. Lê Cao, K.-A., González, I. & Déjean, S. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* **25**, 2855–2856 (2009).

119. Lê Cao, K.-A., Martin, P. G., Robert-Granié, C. & Besse, P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* **10**, 34 (2009).

120. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3**, e161 (2007).

121. Lei, L. & Fithian, W. AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 649–679 (2018).

122. Leppäaho, E., Ammad-ud-din, M. & Kaski, S. GFA: exploratory analysis of multiple data sources with group factor analysis. *The Journal of Machine Learning Research* **18**, 1294–1298 (2017).

123. Li, A. & Barber, R. F. Multiple testing with the structure adaptive Benjamini-Hochberg algorithm. *arXiv preprint arXiv:1606.07926* (2016).

124. Li, C. & Li, H. Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The Annals of Applied Statistics* **4**, 1498 (2010).

125. Lock, E. F., Hoadley, K. A., Marron, J. S. & Nobel, A. B. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics* **7**, 523 (2013).

126. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.* The genotype-tissue expression (GTEx) project. *Nature Genetics* **45**, 580 (2013).

127. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

128. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology* **17**, 75 (2016).

129. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

130. Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods* **12**, 519 (2015).

131. MacKay, D. J. *Bayesian methods for backpropagation networks* in *Models of Neural Networks III* (Springer, 1996), 211–254.

132. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).

133. Maloum, K., Settegrana, C., Chapiro, E., Cazin, B., Leprêtre, S., Delmer, A., Leporrier, M., Dreyfus, B., Tournilhac, O., Mahe, B., *et al.* IGHV gene mutational status and LPL/ADAM29 gene expression as clinical outcome predictors in CLL patients in remission following treatment with oral fludarabine plus cyclophosphamide. *Annals of Hematology* **88**, 1215–1221 (2009).

134. Mazumder, R., Hastie, T. & Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research* **11**, 2287–2322 (2010).

135. McInnes, L. & Healy, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

136. Meinshausen, N., Bühlmann, P., *et al.* High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34**, 1436–1462 (2006).

137. Meng, C., Kuster, B., Culhane, A. C. & Gholami, A. M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **15**, 162 (2014).

138. Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M. & Culhane, A. C. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics* **17**, 628–641 (2016).

139. Mertins, P., Mani, D., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55 (2016).

140. Methé, B. A., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., Gevers, D., Petrosino, J. F., Abubucker, S., Badger, J. H., *et al.* A framework for human microbiome research. *Nature* **486**, 215 (2012).

141. Min, E. J., Safo, S. E. & Long, Q. Penalized Co-Inertia Analysis with Applications to -Omics Data. *Bioinformatics* (2018).

142. Minka, T. P. *Expectation propagation for approximate Bayesian inference* in *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence* (2001), 362–369.

143. Mitchell, T. J. & Beauchamp, J. J. Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**, 1023–1032 (1988).

144. Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M. & Shen, R. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 201208949 (2013).

145. Mohammed, H., Hernando-Herraez, I., Savino, A., Scialdone, A., Macaulay, I., Mulas, C., Chandra, T., Voet, T., Dean, W., Nichols, J., *et al.* Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Reports* **20**, 1215–1228 (2017).

146. Morabito, F., Cutrona, G., Mosca, L., D'Anca, M., Matis, S., Gentile, M., Vigna, E., Colombo, M., Recchia, A. G., Bossio, S., *et al.* Surrogate molecular markers for IGHV mutational status in chronic lymphocytic leukemia for predicting time to first treatment. *Leukemia Research* **39**, 840–845 (2015).

147. Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90 (2011).

148. Oakes, C. C., Seifert, M., Assenov, Y., Gu, L., Przekopowitz, M., Ruppert, A. S., Wang, Q., Imbusch, C. D., Serva, A., Koser, S. D., *et al.* DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nature Genetics* **48**, 253 (2016).

149. Paatero, P. & Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **5**, 111–126 (1994).

150. Park, T. & Casella, G. The Bayesian Lasso. *Journal of the American Statistical Association* **103**, 681–686 (2008).

151. Parkhomenko, E., Tritchler, D. & Beyene, J. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology* **8**, 1–34 (2009).

152. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genetics* **2**, e190 (2006).

153. Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G. & Sandberg, R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods* **10**, 1096 (2013).

154. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology* **16**, 241 (2015).

155. Plesingerova, H., Librova, Z., Plevova, K., Libra, A., Tichy, B., Skuhrova Francova, H., Vrbacky, F., Smolej, L., Mayer, J., Bryja, V., *et al.* COBLL1, LPL and ZAP70 expression defines prognostic subgroups of chronic lymphocytic leukemia patients with high accuracy and correlates with IGHV mutational status. *Leukemia & Lymphoma* **58**, 70–79 (2017).

156. Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904 (2006).

157. Quake, S. R., Wyss-Coray, T., Darmanis, S., Tabula Muris Consortium, *et al.* Single-cell transcriptomic characterization of 20 organs and tissues from individual mice creates a Tabula Muris. *bioRxiv*, 237446 (2018).

158. Queirós, A. C., Villamor, N., Clot, G., Martinez-Trillos, A., Kulis, M., Navarro, A., Penas, E. M. M., Jayne, S., Majid, A., Richter, J., *et al.* A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. *Leukemia* **29**, 598 (2015).

159. Raman, S., Fuchs, T. J., Wild, P. J., Dahl, E. & Roth, V. *The Bayesian group-Lasso for analyzing contingency tables* in *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), 881–888.

160. Remes, S., Mononen, T. & Kaski, S. Classification of weak multi-view signals by sharing factors in a mixture of Bayesian group factor analyzers. *arXiv preprint arXiv:1512.05610* (2015).

161. Rencher, A. C. *Methods of multivariate analysis* (John Wiley & Sons, 2003).

162. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* **16**, 85 (2015).

163. Rockova, V., Lesaffre, E., *et al.* Incorporating grouping information in Bayesian variable selection with applications in genomics. *Bayesian Analysis* **9**, 221–258 (2014).

164. Roweis, S. T. & Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326 (2000).

165. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 319–392 (2009).

166. Schelker, M., Feau, S., Du, J., Ranu, N., Klipp, E., MacBeath, G., Schoeberl, B. & Raue, A. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nature Communications* **8**, 2032 (2017).

167. Schölkopf, B., Smola, A. & Müller, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* **10**, 1299–1319 (1998).

168. Seeger, M. & Bouchard, G. *Fast variational Bayesian inference for non-conjugate matrix factorization models* in *Artificial Intelligence and Statistics* (2012), 1012–1018.

169. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* **14**, 618 (2013).

170. Shen, R., Olshen, A. B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).

171. Singh, A., Gautier, B., Shannon, C. P., Rohart, F., Vacher, M., Tebutt, S. J. & Le Cao, K.-A. DIABLO: from multi-omics assays to biomarker discovery, an integrative approach. *bioRxiv*, 067611 (2018).

172. Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W. & Kelsey, G. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods* **11**, 817 (2014).

173. Söderholm, S., Fu, Y., Gaelings, L., Belanov, S., Yetukuri, L., Berlinkov, M., Cheltsov, A. V., Anders, S., Aittokallio, T., Nyman, T. A., *et al.* Multi-omics studies towards novel modulators of influenza A virus–host interaction. *Viruses* **8**, 269 (2016).

174. Spearman, C. General Intelligence, objectively determined and measured. *The American Journal of Psychology* **15**, 201–292 (1904).

175. Srivastava, P. Roles of heat-shock proteins in innate and adaptive immunity. *Nature Reviews Immunology* **2**, 185 (2002).

176. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology* **6**, e1000770 (2010).

177. Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S. & Robinson, G. E. Big data: astronomical or genomical? *PLoS Biology* **13**, e1002195 (2015).

178. Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R. & Smibert, P. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* **14**, 865 (2017).

179. Tay, J. K. & Tibshirani, R. A latent factor approach for prediction from multiple assays. *arXiv preprint arXiv:1807.05675* (2018).

180. Tenenbaum, J. B., De Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).

181. Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J. & Frouin, V. Variable selection for generalized canonical correlation analysis. *Biostatistics* **15**, 569–583 (2014).

182. Tenenhaus, A. & Tenenhaus, M. Regularized generalized canonical correlation analysis. *Psychometrika* **76**, 257 (2011).

183. Thioulouse, J. *et al.* Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods. *The Annals of Applied Statistics* **5**, 2300–2325 (2011).

184. Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288 (1996).

185. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 91–108 (2005).

186. Tibshirani, R. & Wang, P. Spatial smoothing and hot spot detection for CGH data using the fused Lasso. *Biostatistics* **9**, 18–29 (2007).

187. Tipping, M. E. & Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**, 611–622 (1999).

188. Titsias, M. K. & Lázaro-Gredilla, M. *Spike and slab variational inference for multi-task and multiple kernel learning* in *Advances in Neural Information Processing Systems* (2011), 2339–2347.

189. Trachootham, D., Alexandre, J. & Huang, P. Targeting cancer cells by ROS-mediated mechanisms: a radical therapeutic approach? *Nature Reviews Drug Discovery* **8**, 579 (2009).

190. Trojani, A., Di Camillo, B., Tedeschi, A., Lodola, M., Montesano, S., Ricci, F., Vismara, E., Greco, A., Veronese, S., Orlacchio, A., *et al.* Gene expression profiling identifies ARSD as a new marker of disease progression and the sphingolipid metabolism as a potential novel metabolism in chronic lymphocytic leukemia. *Cancer Biomarkers* **11**, 15–28 (2012).

191. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R. B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).

192. Tucker, L. R. An inter-battery method of factor analysis. *Psychometrika* **23**, 111–136 (1958).

193. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research* **43**, D204–D212 (2014).

194. Van Der Maaten, L., Postma, E. & Van den Herik, J. Dimensionality reduction: a comparative. *Journal of Machine Learning Research* **10**, 66–71 (2009).

195. Vasconcelos, Y., De Vos, J., Vallat, L., Reme, T., Lalanne, A., Wanherdrick, K., Michel, A., Nguyen-Khac, F., Oppezzo, P., Magnac, C., *et al.* Gene expression profiling of chronic lymphocytic leukemia can discriminate cases with stable disease and mutated Ig genes from those with progressive disease and unmutated Ig genes. *Leukemia* **19**, 2002 (2005).

196. Verìssimo, A., Oliveira, A. L., Sagot, M.-F. & Vinga, S. DegreeCox – a network-based regularization method for survival analysis. *BMC Bioinformatics* **17**, 449 (2016).

197. Virtanen, S., Klami, A., Khan, S. & Kaski, S. *Bayesian group factor analysis* in *Artificial Intelligence and Statistics* (2012), 1269–1277.

198. Waldron, L., Pintilie, M., Tsao, M.-S., Shepherd, F. A., Huttenhower, C. & Jurisica, I. Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics* **27**, 3399–3406 (2011).

199. Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B. & Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* **11**, 333 (2014).

200. Wang, C. Variational Bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks* **18**, 905–910 (2007).

201. Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Cancer Genome Atlas Research Network, *et al.* The cancer genome atlas pan-cancer analysis project. *Nature Genetics* **45**, 1113 (2013).

202. Westra, H.-J., Jansen, R. C., Fehrmann, R. S., te Meerman, G. J., Van Heel, D., Wijmenga, C. & Franke, L. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27**, 2104–2111 (2011).

203. Wiel, M. A., Lien, T. G., Verlaat, W., Wieringen, W. N. & Wilting, S. M. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in Medicine* **35**, 368–381 (2016).

204. Witten, D. M. & Tibshirani, R. J. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology* **8**, 1–27 (2009).

205. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534 (2009).

206. Woodbury, M. A. Inverting modified matrices, Memorandum Rept. 42. *Statistical Research Group, Princeton University, Princeton, NJ* (1950).

207. Wu, A., Park, M., Koyejo, O. O. & Pillow, J. W. *Sparse Bayesian structure learning with dependent relevance determination priors* in *Advances in Neural Information Processing Systems* (2014), 1628–1636.

208. Xu, X., Ghosh, M., *et al.* Bayesian variable selection and estimation for group Lasso. *Bayesian Analysis* **10**, 909–936 (2015).

209. Yang, J., Huang, T., Petralia, F., Long, Q., Zhang, B., Argmann, C., Zhao, Y., Mobbs, C. V., Schadt, E. E., Zhu, J., *et al.* Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Scientific Reports* **5** (2015).

210. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49–67 (2006).

211. Zenz, T., Mertens, D., Küppers, R., Döhner, H. & Stilgenbauer, S. From pathogenesis to treatment of chronic lymphocytic leukaemia. *Nature Reviews Cancer* **10**, 37 (2010).

212. Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., *et al.* Ensembl 2018. *Nucleic Acids Research* **46**, D754–D761 (2017).

213. Zhao, Q., Shi, X., Xie, Y., Huang, J., Shia, B. & Ma, S. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Briefings in Bioinformatics* **16**, 291–303 (2014).

214. Zhao, S., Gao, C., Mukherjee, S. & Engelhardt, B. E. Bayesian group factor analysis with structured sparsity. *The Journal of Machine Learning Research* **17**, 6868–6914 (2016).

215. Zou, H. The adaptive Lasso and its oracle properties. *Journal of the American statistical association* **101**, 1418–1429 (2006).

216. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320 (2005).

217. Zou, H., Hastie, T. & Tibshirani, R. Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15**, 265–286 (2006).

218. Zwiener, I., Frisch, B. & Binder, H. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS One* **9**, e85150 (2014).

Acronyms

PCA principal component analysis

pPCA probabilistic principal component analysis

MOFA multi-omics factor analysis

FDR false discovery rate

ARD automatic relevance determination

ELBO evidence lower bound

SNF similarity network fusion

kNN k-nearest neighbour

BIC Bayesian information criterion

EM expectation-maximization

VB variational Bayes

SVD singular-value decomposition

MLE maximum likelihood estimator

GFA group factor analysis

CCA canonical correlation analysis

IBFA inter-battery factor analysis

CIA co-inertia analysis

MCMC Markov chain Monte Carlo

iid independent and identically distributed

KL Kullback-Leibler

CLL chronic lymphocytic leukaemia

IGHV immunoglobulin heavy chain variable region gene

HSP heat shock protein

ROS reactive oxygen species

mESC mouse embryonic stem cell

GWAS genome-wide association studies

QTL quantitative trait loci

Notation

Most mathematical notation is introduced within each chapter of this thesis, as the chapters are self-contained and in part contain published articles. Here, we provide an overview of some general notation as a reference:

Symbol	Description
$\mathbb{1}_r$	identity matrix of rank r; in some cases we simply use $\mathbb{1}$ if the rank is clear from the context
$\mathcal{N}(\mu, \sigma^2)$	normal (or Gaussian) distribution with mean μ and variance σ^2
$\mathrm{Gamma}(a, b)$	Gamma distribution with shape parameter a and rate parameter b
$\mathrm{Laplace}(\mu, b)$	Laplace distribution with mean μ and scale b
$\mathrm{Ber}(p)$	Bernoulli distribution with success probability p
$\mathrm{Beta}(a, b)$	Beta distribution with shape parameters a, b
$\mathrm{Poi}(\lambda)$	Poisson distribution with rate parameter λ
$B(a, b)$	Beta function in a, b
$\Gamma(a)$	Gamma function in a
$\mathrm{diag}(x)$	diagonal matrix containing the vector x on the diagonal
$\mathrm{diag}(x_1, \ldots, x_r)$	diagonal matrix containing $x_1, \ldots, x_r$ on the diagonal
$\mathbb{E}_q(x)$	expected value of x under the distribution q; in some cases we simply use $\mathbb{E}(x)$ or $\langle x \rangle$ if the distribution is clear from the context
$\mathcal{F}(x\|\theta)$	probability density function of a distribution $\mathcal{F}$ in x with parameters θ. For example, $\mathcal{N}\left(x \mid \mu, \sigma^2\right)$ denotes the density of a univariate normal distribution with mean μ and variance σ^2, and $\mathrm{Gamma}\left(x \mid a, b\right)$ the density of a gamma distribution with parameters a and b.
$\|\mathcal{S}\|$	cardinality of a set $\mathcal{S}$

Furthermore, in general we use bold capital letters (e. g. $\mathbf{X}$) to refer to a matrix. Elements of the matrix $\mathbf{X}$ are represented as $\mathbf{X}_{ij}$ or x_{ij}. The j^{th} column of a matrix is referred to using $\mathbf{X}_{\cdot,j}$ or $\mathbf{x}_{:,j}$ and the i^{th} row using $\mathbf{X}_{i,\cdot}$ or $\mathbf{x}_{i,:}$. Random variable are represented by capital non-bold letters (e. g. X) when describing their general properties, while finite samples are denoted by non-capital letters (e. g. $x_1, \ldots, x_n$).

List of Figures

List of Tables