Department of Computer Science

# Bayesian Multi-View Factor Models for Drug Response and Brain Imaging Studies

Eemeli Leppäaho

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
**DOCTORAL DISSERTATIONS** 178/2018

# Bayesian Multi-View Factor Models for Drug Response and Brain Imaging Studies

**Eemeli Leppäaho**

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall F239a of the school on 12th October 2018 at 12.

**Aalto University**
**School of Science**
**Department of Computer Science**

**Supervising professor**
Prof. Samuel Kaski, Aalto University, Finland

**Preliminary examiners**
Assoc. Prof. Jussi Tohka, University of Eastern Finland, Finland
Assoc. Prof. Morten Mørup, Technical University of Denmark, Denmark

**Opponent**
Dr. Irina Rish, T. J. Watson Research Center, USA

Aalto University publication series
**DOCTORAL DISSERTATIONS** 178/2018



ISBN 978-952-60-8184-7 (printed)
ISBN 978-952-60-8185-4 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-8185-4

Unigrafia Oy
Helsinki 2018

Finland

441 697
Printed matter

Aalto University

# Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

**Author**
Eemeli Leppäaho

**Name of the doctoral dissertation**
Bayesian Multi-View Factor Models for Drug Response and Brain Imaging Studies

**Publisher** School of Science

**Unit** Department of Computer Science

**Series** Aalto University publication series DOCTORAL DISSERTATIONS 178/2018

**Field of research** Machine learning

**Manuscript submitted** 6 March 2018 **Date of the defence** 12 October 2018

**Permission to publish granted (date)** 20 August 2018 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

**Abstract**

This thesis investigates knowledge inference from measurements of multiple data sources, motivated by technologies in a wide range of domains allowing effective measurement of several related, but heterogeneous data sources. In life sciences, examples of this kind of "multi-view" data are brain imaging data of multiple subjects along with description of the experimental stimuli, as well as drug response studies including measurements regarding the expression level, copy number variation and mutation of genes in cell lines. Data analyses have been typically related to analyzing the structure of a single data source, or the effect of one data source to another. The multi-view data inspected in this thesis results in a more complex problem: besides the structure of each of the data sources, the relations between the data sources are of high interest as well.

This thesis addresses modern multi-view data analysis problems using Bayesian latent variable models. They are a natural choice for developing models in order to gain knowledge about multiple data sources and their relations; they allow for missing values in the data, incorporating prior information to the modelling problem and estimating the uncertainty present in the inference. The key contributions of this thesis include formulating a low-rank data source relation model and presenting biclustering using sparse priors, as well as a relaxed formulation of tensor factorization. All the developed models have been published as open-source software, enabling wide-spread use and further development.

The presented machine learning tools are demonstrated using drug response and brain imaging studies, for both of which predictive performance above state-of-the-art level is achieved. In the drug response studies, the models were able to accurately relate similar drugs, as well as detect known cancer genes affecting the responsiveness of cells to certain drugs. In the brain response studies the benefits of the presented methods were shown via increased accuracy in predicting brain responses, whereas the relaxed tensor decomposition allowed for a novel way of utilizing measurements for multiple subjects. Finally, the advantage of using a low-dimensional latent space is illustrated in a genome-wide association study in an especially challenging domain: when there exist measurements for only two hundred subjects, yet there exist some thousands of features regarding the subjects, with the study discovering a relevant gene associated with components of brain activity.

**Keywords** Bayesian modelling, bioinformatics, brain imaging, factor analysis, multi-view modelling, tensor factorization

**ISBN (printed)** 978-952-60-8184-7 **ISBN (pdf)** 978-952-60-8185-4

**ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

**Location of publisher** Helsinki **Location of printing** Helsinki **Year** 2018

**Pages** 168 **urn** http://urn.fi/URN:ISBN:978-952-60-8185-4

**Tekijä**
Eemeli Leppäaho

**Väitöskirjan nimi**
Bayesiläisiä monilähdemalleja lääkevaste- ja aivokuvantamiskokeisiin

**Julkaisija** Perustieteiden korkeakoulu

**Yksikkö** Tietotekniikan laitos

**Sarja** Aalto University publication series DOCTORAL DISSERTATIONS 178/2018

**Tutkimusala** Koneoppiminen

**Käsikirjoituksen pvm** 06.03.2018 **Väitöspäivä** 12.10.2018

**Julkaisuluvan myöntämispäivä** 20.08.2018 **Kieli** Englanti

☐ **Monografia** ☒ **Artikkeliväitöskirja** ☐ **Esseeväitöskirja**

**Tiivistelmä**

Tässä työssä tutkitaan tiedon hankkimista monilähdeaineistoista. Nykyään monilla aloilla on mahdollista kerätä tehokkaasti mittauksia useista toisiinsa liittyvistä mutta heterogeenisistä datalähteistä. Biotieteissä esimerkkejä tällaisista monilähdeaineistoista ovat usean koehenkilön aivokuvantamismittaukset yhdistettynä kokeessa käytetyn ärsykkeen kuvaukseen sekä lääkevastekokeet, jotka sisältävät mittauksia solulinjojen geenien ilmentymisistä, kopioiden määrästä ja mutaatioista. Data-analyysiongelmissa tutkimuskohde on tyypillisesti ollut joko yksittäisen datalähteen rakenne tai yhden datalähteen vaikutus toiseen. Tässä työssä tarkasteltuihin monilähdeaineistoihin liittyy haastavampi ongelma, sillä jokaisen lähteen sisäisen rakenteen lisäksi halutaan tarkastella myös lähteiden välisiä suhteita.

Tässä työssä monilähdedata-analyysiongelmia ratkotaan bayesiläisillä piilomuuttujamalleilla. Ne soveltuvat hyvin mallien kehittämiseen useille datalähteille ja niiden välisille suhteille; ne sallivat puuttuvat arvot aineistossa sekä mahdollistavat prioritiedon huomioon ottamisen mallintamisessa ja epävarmuuden arvioinnin mallin päättelyssä. Tärkeimpinä kontribuutioina tässä työssä esitellään matalaulotteinen suhdemalli datalähteille, demonstroidaan biklusterointia harvoilla prioreilla sekä muotoillaan relaksoitu tensorihajotelma. Kaikki kehitetyt mallit on julkaistu avoimesti, jotta niitä voidaan edelleenkehittää ja käyttää laajasti.

Esiteltyjä koneoppimismalleja sovellettiin lääkevaste- ja aivokuvantamiskokeisiin. Molemmissa sovelluksissa ylitettiin aiempi huipputaso ennustustarkkuuksissa. Lääkevastekokeissa malleilla onnistuttiin assosioimaan samankaltaisia lääkkeitä ja havaittiin tunnettuja syöpägeenejä, jotka vaikuttivat solujen herkkyyteen tietyille lääkkeille. Aivokuvantamiskokeissa esitelty relaksoitu tensorihajotelma hyödynsi useiden koehenkilöiden mittauksia uudenlaisella tavalla. Lisäksi tässä työssä osoitettiin matalaulotteisen piiloavaruuden hyödyllisyys genominlaajuisessa assosiaatiotutkimuksessa erityisen haastavassa koeasetelmassa, jossa mittauksia on vain kahdestasadasta henkilöstä ja fenotyyppi koostuu tuhansista piirteistä. Sen avulla löydettiin merkityksellinen geeni, joka selittää aivoaktiivisuuden osatekijöitä.

**Avainsanat** aivokuvantaminen, bayesiläinen mallintaminen, bioinformatiikka, faktorianalyysi, monilähdemallintaminen, tensorihajotelmat

**ISBN (painettu)** 978-952-60-8184-7 **ISBN (pdf)** 978-952-60-8185-4

**ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

**Julkaisupaikka** Helsinki **Painopaikka** Helsinki **Vuosi** 2018

**Sivumäärä** 168 **urn** http://urn.fi/URN:ISBN:978-952-60-8185-4

# Preface

This work has been carried out in the Probabilistic Machine Learning (PML) group at the Department of Computer Science in Aalto University School of Science. I have been a part of the Finnish Center of Excellence in Computational Inference Research (COIN) and the Helsinki Institute for Information Technology HIIT. This research and thesis has been partially funded by projects ReKnow2 of Tekes and aivoAALTO of Aalto University.

I would like to thank my instructor and supervisor, Prof. Samuel Kaski, for introducing me to the ways of machine learning research, as well as constantly pushing me towards sharper thinking. I am grateful for the opportunities provided by working in a high-level research group with excellent collaboration opportunities.

I would like to express my sincerest gratitude to my excellent go-to brain imaging experts Prof. Riitta Salmelin and Dr. Hanna Renvall, who have helped me grasp some of the complexities of the human brain. Your contributions have helped me create machine learning that will hopefully prove valuable for brain imaging studies beyond this thesis.

I am very grateful for all my co-authors, without whom this dissertation would not exist, as well as for the whole PML group. Especially, I would like to thank Arto Klami and Seppo Virtanen for invaluable guidance in the beginning of my doctoral studies, and Suleiman Khan and Muhammad Ammad-ud-din for aiding me in numerous challenges regarding drug response studies. Additionally, I am thankful for Jussi Gillberg for our numerous scientific and non-scientific discussions, as well as to the people of PML group with whom I have played foosball or plancked, providing valuable refreshment during the mental work.

Finally, above all, I would like to thank Sini for providing continuous support and love during my studies, as well as Niilo for being the loveliest boy in the world.

Espoo, July 17, 2018,

Eemeli Leppäaho

# Contents

# List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

**I** Arto Klami, Seppo Virtanen, Eemeli Leppäaho, and Samuel Kaski. Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9):2136–2147, 2015.

**II** Suleiman A. Khan, Eemeli Leppäaho and Samuel Kaski. Bayesian multi-tensor factorization. *Machine Learning*, 105(2):233–253, 2016.

**III** Kerstin Bunte, Eemeli Leppäaho, Inka Saarinen and Samuel Kaski. Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics*, 32(16):2457–2463, 2016.

**IV** Eemeli Leppäaho, Muhammad Ammad-ud-din, and Samuel Kaski. GFA: exploratory analysis of multiple data sources with group factor analysis. *Journal of Machine Learning Research*, 18(39):1–5, 2017.

**V** Xiangju Qin, Paul Blomstedt, Eemeli Leppäaho, Pekka Parviainen and Samuel Kaski. Distributed Bayesian matrix factorization with limited communication. *Submitted to a journal*, 27 pages, 2018.

**VI** Eemeli Leppäaho, Hanna Renvall, Elina Salmela, Juha Kere, Riitta Salmelin, and Samuel Kaski. Discovering heritable modes of MEG spectral power. *Submitted to a journal*, 30 pages, 2018.

# Author's Contribution

### Publication I: "Group factor analysis"

The ideas and experiments in this article were designed jointly. The author implemented the low-rank version of GFA with the aid of Dr. Virtanen. The author had a lead role in performing the simulation studies, and performed the drug response study. The writing of the article was a combined effort.

### Publication II: "Bayesian multi-tensor factorization"

The author designed and implemented the relaxed MTF model, as well as designed and performed both the simulation studies and the brain imaging study. The manuscript was written jointly.

### Publication III: "Sparse group factor analysis for biclustering of multiple data sources"

The author designed and implemented the GFA model utilizing data sources paired in two modes, as well as designed and implemented the experiments jointly with Ms. Saarinen. The analysis of the drug response study, as well as the writing of the manuscript were a joint effort.

### Publication IV: "GFA: exploratory analysis of multiple data sources with group factor analysis"

The author implemented the published software, and had the main responsibility in designing and writing the article.

### Publication V: "Distributed Bayesian matrix factorization with limited communication"

The author participated in the design and implementation of the models, and in performing and visualizing the experiments. The manuscript was written jointly.

### Publication VI: "Discovering heritable modes of MEG spectral power"

The author had the main responsibility in designing the experiments, and implemented them, providing the results and figures. The manuscript was written jointly, with main contributions from the author and Dr. Renvall.

# List of Abbreviations and Symbols

### Abbreviations

| | |
|---|---|
| ARD | automatic relevance determination |
| BRRR | Bayesian reduced-rank regression |
| CCA | canonical correlation analysis |
| CP | CANDECOMP/PARAFAC decomposition |
| DNA | deoxyribonucleic acid |
| FA | factor analysis |
| fMRI | functional magnetic resonance imaging |
| GFA | group factor analysis |
| GWAS | genome-wide association study |
| KL-divergence | Kullback-Leibler divergence |
| MBFA | multi-battery factor analysis |
| MCMC | Markov chain Monte Carlo |
| MEG | magnetoencephalography |
| MTF | multi-tensor factorization |
| PCA | principal component analysis |
| RMSE | root mean squared error |
| RNA | ribonucleic acid |
| SNP | single nucleotide polymorphism |
| VB | variational Bayesian |

## Symbols

| | |
|---|---|
| $\mathcal{X}, \mathcal{Y}$ | Tensors |
| $\mathbf{X}, \mathbf{Y}$ | Matrices |
| $\mathbf{X}^{(m)}$ | The $m$th matrix in list $\mathbf{X}$ |
| $\mathbf{X}_{N \times D}$ | Matrix with $N$ rows and $D$ columns |
| $\mathbf{X}^\top$ | Transpose of matrix $\mathbf{X}$ |
| $\mathbf{x}, \mathbf{y}$ | Vectors |
| $\mathbf{x}_i$ | The $i$th row of matrix $\mathbf{X}$ (column vector) |
| $\mathbf{x}_{:,j}$ | The $j$th column of matrix $\mathbf{X}$ |
| $\mathcal{N}(\mu, \sigma^2)$ | Gaussian distribution with mean $\mu$ and variance $\sigma^2$ |
| $\mathcal{G}(a, b)$ | Gamma distribution with shape $a$ and rate $b$ |
| $p(x)$ | Probability density function of $x$ |
| $p(x\|y)$ | Conditional probability density function of $x$ given $y$ |
| $\mathbb{E}(x)$ | Expected value of random variable $x$ |
| $\mathbf{I}_D$ | Identity matrix with $D$ rows and columns |

# 1. Introduction

Modern technologies allow for effective measurement of large amounts and various types of data, calling for suitable analysis methods for knowledge extraction. Great amounts of research has been conducted on how to model the ever larger data sources while providing generalizability to new measurements as well. Meanwhile, there has been lesser focus on ways to model the relations of heterogeneous data sources (also called data views).

Traditional data analysis methods, developed to study the relations of variables in a single data source, do not ideally generalize to multiple heterogeneous data sources, the relations of which can be even more important than the structure within them. This includes Bayesian latent variable models, which are a key tool in machine learning, as they form a principled framework for understanding the generation of data, incorporating prior information in the model, as well as evaluating the uncertainties of the model.

The analysis of drug response and brain imaging studies with machine learning tools is a key contribution of this thesis. The former can include measurements related to the chemical structure of the drugs, as well as the gene expression, copy number variation and methylation of the studied cell lines. Modern brain imaging studies, on the other hand, can contain natural stimuli with rich features (such as an auditory story), as well as measurements from multiple subjects, each different from one another. Studying the relations and structure of these data sources can bring insight into the problem at hand, be it the chemical action mechanism of drugs or the language processing of the brain, as well as result in accurate predictions of missing parts of the data.

This thesis studies Bayesian latent variable models that allow gaining insight on the relations of multiple data sources. The research conducted is based on a recent model group factor analysis (GFA) [1], which is pio-

neering work for modelling relations within multi-view data. Application of multi-view models, such as GFA, in life sciences requires carefully taking into account existing prior information regarding the measurements, and the similarities of observations, as well as whole data sources.

## 1.1 Contributions and organization of the thesis

This thesis presents new Bayesian latent variable models for joint modelling of multiple data sources. These models are designed to suit the needs of applications in life sciences, with specifically drug response and brain imaging studies in mind.

Publication I provides a literature survey of models related to group factor analysis and presents a novel low-rank formulation for modelling relations of data sources. The approach is demonstrated in an analysis of functional magnetic resonance imaging (fMRI) data with subjects listening to a piece of music, as well as in a drug response analysis, evaluating similarities of drugs.

Publication II presents a framework for joint decomposition of paired matrices and tensors, and further extends it by allowing a more flexible decomposition for the tensors, called relaxed multi-tensor factorization. The performance of the methods is demonstrated in a brain imaging study, where an auditory story is decomposed jointly with related brain responses, and in a toxicogenomics problem, where structural descriptors of drugs, gene expression of cells and drug toxicity are related.

Publication III examines GFA with sparse priors as a biclustering method for multiple data sources, and extends the formulation to allow for data sources paired in two modes. The presented methods are utilized in a drug response study, where the biclustering prior results in an interpretable and predictive posterior. The tools needed for this work are collected in a software package, implementing the full data analysis pipeline and a replication of the drug response study, presented in Publication IV.

Publication V presents a framework that allows posterior inference for challengingly large models (i.e., big data), by partitioning the data into parts, and performing a three-step inference procedure, parallelizing over the data partitioning. The performance of this approximate inference technique is demonstrated in movie recommendation and drug-protein interaction tasks.

Finally, Publication VI applies Bayesian reduced-rank regression (BRRR;

[2]) to explain heritable and environmentally affected modes of brain activity. The study demonstrates how modern machine learning tools are able to perform genome-wide association even for studies where there are more phenotypes (i.e., outcome variables) than subjects.

This thesis is organized as follows: Chapter 2 motivates the use of Bayesian latent variable models, discussing their properties and inference. Chapter 3 presents the multi-source models used and developed in the thesis. The application of these models in drug response studies is discussed in Chapter 4, and the application in brain imaging studies in Chapter 5. The thesis is concluded with a discussion in Chapter 6.

# 2. Bayesian Latent Variable Models

Statistical inference is focused on gaining knowledge and making predictions based on observations, and Bayesian latent variable models are no exception. Formally, given observations $\mathbf{Y}$, we assume a generative process

$$p\left(\mathbf{Y}|\mathbf{\Theta}\right), \tag{2.1}$$

defining the probability of the observations $\mathbf{Y}$ given latent (i.e., unobserved) variables $\mathbf{\Theta}$. Equation (2.1) is known as the likelihood of the observations, and it is generally motivated by the assumption that the observations can be described with a lower dimensional (more interpretable) set of latent variables. In Bayesian modelling the latent variables are also given a prior $p(\mathbf{\Theta})$, stating the prior beliefs about them before any data is observed. The prior can reflect various degrees of belief: a very even (flat) distribution over the possible space of $\mathbf{\Theta}$ is called an uninformative prior, as opposed to an informative prior. The former allows, given the likelihood, the observations $\mathbf{Y}$ alone to determine the latent variables $\mathbf{\Theta}$, whereas the latter can be used to affect them based on prior knowledge or assumptions.

In Bayesian latent variable models, defined by the prior and the likelihood, inferring the latent variables (specifically, their distribution) relies on the use of Bayes' theorem:

$$p(\mathbf{\Theta}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{\Theta})p(\mathbf{\Theta})}{p(\mathbf{Y})}, \tag{2.2}$$

where $p(\mathbf{\Theta}|\mathbf{Y})$ is called the posterior distribution of the latent variables, and $p(\mathbf{Y})$ is the probability of observing $\mathbf{Y}$ without regard to $\mathbf{\Theta}$. The posterior distribution is the final product of Bayesian latent variable models, allowing for gaining knowledge and making predictions based on the observations $\mathbf{Y}$. Having the distribution also allows for quantifying the uncertainty of the conclusions, as well as the predictions.

Matrix factorization methods, used to present a data matrix $\mathbf{Y}$ in a lower dimensional space, are discussed in this chapter, along with tensor factorization methods, applied to higher-order data matrices. This is followed with a discussion of sparse models and model inference.

## 2.1 Matrix factorization

Let us assume observations presented in a matrix $\mathbf{Y}_{N\times D}$, such that $\mathbf{y}_{i,:}$ is an observation of $D$ features. In order to gain knowledge about $\mathbf{Y}$, if $D$ is very small, it is often meaningful to inspect a visualization of the observations and look at some simple statistics of the different features. However, as $D$ increases, this will no longer be effective, or even feasible. Assuming that the features are not independent, it is meaningful to present the observation matrix as

$$\mathbf{Y} \approx \mathbf{X}\mathbf{W}^\top, \tag{2.3}$$

where $\mathbf{X}_{N\times K}$ is a low-dimensional ($K < \min(N, D)$) description of the observations, and $\mathbf{W}_{D\times K}$ is a low-dimensional description of the features, that is, a mapping from the latent dimension (the $K$ components) to the observation space. Here the parameter $K$ defines the dimensionality of the latent space, i.e., the complexity of the factorization model. It is typical to assume zero-mean columns in $\mathbf{Y}$ (centering done as a preprocessing step), as this avoids having to set a feature mean parameter. The most well-known matrix factorization methods are factor analysis (FA; [3]) and principal component analysis (PCA; [4]).

Usually the likelihood for Equation (2.3) is defined as a normal distribution

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{N} \mathcal{N}\left(\mathbf{y}_i|\mathbf{W}\mathbf{x}_i, \mathbf{\Lambda}^{-1}\right), \tag{2.4}$$

where $\mathbf{\Lambda}$ is the precision matrix (i.e., the inverse covariance matrix) of the residuals $\mathbf{y}_i - \mathbf{W}\mathbf{x}_i$. Alternative choices include Laplace distribution, better suited for data containing outliers [5] and Gaussian scale mixture, allowing more robust parameter inference if both the rows and columns have varying noise scales [6]. The choice of $\mathbf{\Lambda}$ is crucial for the likelihood, as in principle the mean term $\mathbf{W}\mathbf{x}_i$ is used to model the latent signal of the observations, whereas $\mathbf{\Lambda}$ models the covariance structure of the residual noise. Fully parameterized precision, as used for example in Bayesian probabilistic matrix factorization [7], assumes that the residual contains structure not captured by the component model. On the other

hand, limiting the precision to be a diagonal matrix, or even of form $\tau \mathbf{I}_D$, such as in [8], assumes that the component model captures all the covariance present in $\mathbf{Y}$. It is also important to note that here we assume independent and identically distributed (i.i.d.) observations, which allows decomposing the likelihood in Equation (2.4) into a product of $N$ independent factors.

Furthermore, for Bayesian purposes, the most common prior distribution for elements of $\mathbf{X}$ and $\mathbf{W}$ are normal distributions with zero mean and unit variance, resulting in dense parameters. The variance of these distributions can be adjusted if informative priors regarding the scale of the components are desired. Sparse priors will be discussed in Section 2.3. For a fully parameterized $\mathbf{\Lambda}$, it is typical to assume a Wishart distribution, with e.g. scale matrix $\mathbf{I}$ and $D$ degrees of freedom [7], whereas the elements of a diagonal precision matrix are typically given gamma distributions with shape $a^\tau$ and rate $b^\tau$. Setting $a^\tau = b^\tau \leq 1$ leads to what is generally considered an uninformative prior. If prior knowledge exists about the signal-to-noise ratio, it can be used to define an informative prior for the precision matrix $\mathbf{\Lambda}$.

## 2.2 Tensor factorization

Tensor factorization extends matrix factorization principles to tensors, i.e., "matrices" with 3 or more modes. Here only tensors with 3 modes, i.e., observations $\mathcal{Y}_{N \times D \times L}$ are discussed, but the principles can be directly generalized to any number of modes, each associated with parameters describing the reconstruction from the latent space. A common likelihood for the observations is defined as

$$p(\mathcal{Y}|\mathbf{X}, \mathbf{W}, \mathbf{U}) = \prod_{i=1}^{N} \prod_{j=1}^{D} \prod_{l=1}^{L} \mathcal{N}\left(y_{i,d,l} | \sum_{k=1}^{K} x_{i,k} w_{j,k} u_{l,k}, \tau_{d,l}^{-1}\right), \tag{2.5}$$

again assuming i.i.d. observations, with $\mathbf{U}$ introduced as the parameter associated with the third mode of the tensor. The factorization in Equation (2.5) is equivalent to the CANDECOMP/PARAFAC (CP) factorization [9, 10], also knows as the tensor rank decomposition. Other types of tensor factorization schemes have been studied as well [11]. The prior distributions of the parameters can be chosen using the same principles as with matrix factorization.

## 2.3 Sparse models

In many applications of Bayesian latent variable models it is restrictive to assume that certain low-dimensional structure is present in all the observations and features. An alternative assumption can be made by setting sparse priors for the models parameters, generally resulting in observations (or features) being associated only with structure strongly present in them. In many tasks sparsity can also help in regularizing the model to avoid overfitting, as well as result in more easily accessible knowledge about the data. In the domain of matrix factorization, one alternative for imposing sparsity for latent weights $\mathbf{W}$ is to define them an automatic relevance determination (ARD) prior [12]:

$$\begin{aligned} w_{d,k} &\sim \mathcal{N}\left(0, \alpha_{d,k}^{-1}\right) \\ \alpha_{d,k} &\sim \mathcal{G}\left(a^{\alpha}, b^{\alpha}\right), \end{aligned} \tag{2.6}$$

where the Gamma-distributed $\alpha_{d,k}$ is used to learn the scale of each effect. By setting suitable values for shape $a^{\alpha}$ and rate $b^{\alpha}$, $w_{d,k}$ that contribute only little to the likelihood will be pushed towards zero by learning a high value for $\alpha_{d,k}$. For example $a^{\alpha} = b^{\alpha} = 10^{-14}$ results in an uninformative prior that is able to induce sparsity [1]. The ARD prior can also be used to perform model complexity selection [1, 8]. If the elements of $\mathbf{W}$ are given precision $\alpha_k$, opposed to $\alpha_{d,k}$ in Equation (2.6), components contributing only little to the likelihood will be pushed towards zero. Thus initializing the model with a high $K$, and using a threshold to omit weak components results in inferring the model complexity. Here the prior of the residual noise plays a crucial role: if a large proportion of noise is assumed, more components will be shut down.

The ARD prior is a conjugate prior, resulting in normal and gamma distributions for $\mathbf{W}$ and $\boldsymbol{\alpha}$, conditional on the other model parameters. It has the drawback that some threshold is needed to distinguish "approximately zero" and non-zero values. An alternative prior imposing binary sparsity is

$$\begin{aligned} w_{d,k} &\sim h_{d,k}\mathcal{N}\left(0,1\right) + \left(1 - h_{d,k}\right)\delta_0 \\ h_{d,k} &\sim Bernoulli\left(\pi_k\right) \\ \pi_k &\sim Beta\left(a^{\pi}, b^{\pi}\right), \end{aligned}$$

called the spike-and-slab prior [13]. Here $\pi_k$ controls how likely it is that $\mathbf{w}_{:,k}$ is used to explain data (with $h_{d,k} = 1$ corresponding to the slab and

0 to the spike). The spike-and-slab prior can be combined with ARD by changing the precision of the normal distribution to allow also learning differing scales for different features and/or components.

## 2.4 Model inference

At the simplest, the posterior of the model parameters $\boldsymbol{\Theta}$ can be inferred with the Bayes' theorem. However, the denominator in Equation (2.2) needs to be estimated as

$$p(\mathbf{Y}) = \int p(\mathbf{Y}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})d\boldsymbol{\Theta}\,;$$

an integral which usually is not tractable, that is, presentable in closed form. Not being able to estimate the denominator of the Bayes' theorem necessitates the use of approximation schemes. Different schemes can be divided broadly in two categories. In the first, the posterior is approximated as

$$p(\boldsymbol{\Theta}|\mathbf{Y}) \approx q(\boldsymbol{\Theta}),$$

where $q$ denotes the user chosen (simple) functional form. In the other, Markov chain Monte Carlo (MCMC), the goal is to obtain samples from the posterior distribution, and use a finite number of these samples to approximate it.

### 2.4.1 Variational Bayesian inference

In variational Bayesian (VB) inference the posterior is approximated by a variational distribution $q(\boldsymbol{\Theta})$. The log-marginal distribution can be formulated as

$$\begin{aligned}\ln p(\mathbf{Y}) &= \int q(\boldsymbol{\Theta}) \ln p(\mathbf{Y}) d\boldsymbol{\Theta} \\ &= \int q(\boldsymbol{\Theta}) \ln \frac{p(\mathbf{Y}, \boldsymbol{\Theta})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta} - \int q(\boldsymbol{\Theta}) \ln \frac{p(\boldsymbol{\Theta}|\mathbf{Y})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta} \\ &:= \mathcal{L}(q) + D_{KL}(q||p),\end{aligned}$$

where it is divided in two terms that are interpreted as the lower bound, $\mathcal{L}(q)$, and the KL-divergence between the posterior and the variational distribution, $D_{KL}(q||p)$. As $\ln p(\mathbf{Y})$ is a constant, it it possible to maximize the lower bound in order to minimize the divergence between the posterior $p(\boldsymbol{\Theta}|\mathbf{Y})$ and $q(\boldsymbol{\Theta})$.

A common assumption for the variational distribution follows mean field

theory [14], where subsets of the parameters, denoted with $\boldsymbol{\Theta}_i$, are assumed independent:

$$q(\boldsymbol{\Theta}) = \prod_i q_i(\boldsymbol{\Theta}_i).$$

Given the factorized variational distribution, the maximum of the lower bound w.r.t. $q_j(\boldsymbol{\Theta}_j)$ is achieved at

$$\mathbb{E}[\ln p(\mathbf{Y}, \boldsymbol{\Theta})|\{q_i\left(\boldsymbol{\Theta}_i\right)\}_{i \neq j}] + \text{const.} \tag{2.7}$$

This results in an iterative mean field variational Bayesian algorithm, where each $q_j(\boldsymbol{\Theta}_j)$ is updated at a time according to Equation (2.7). The iteration is needed since the optimum of $q_j(\boldsymbol{\Theta}_j)$ depends on the expectations of the other factors. The lower bound is convex with respect to each of the factors $q_i(\boldsymbol{\Theta}_i)$, so convergence to a local optimum is guaranteed [15]. For models inferred with VB, it is customary to use conjugate priors, resulting in closed-form distributions in Equation (2.7). In other cases the lower bound can be maximized utilizing for example numerical methods [16]. Detailed derivation of the VB procedure can be found in [17].

### 2.4.2 Gibbs sampling

Markov chain Monte Carlo methods are intended for sampling from a probability distribution of interest, such as the posterior distribution $p(\boldsymbol{\Theta}|\mathbf{Y})$. They operate iteratively by returning a sample $\boldsymbol{\Theta}^{(i+1)}$, given a state $\boldsymbol{\Theta}^{(i)}$. There are alternative ways of forming the sampling distribution $p(\boldsymbol{\Theta}^{(i+1)}|\boldsymbol{\Theta}^{(i)}, \mathbf{Y})$, such that it converges to the distribution of interest as $i \to \infty$.

Gibbs sampling is an MCMC method where the model parameters $\boldsymbol{\Theta}$ are split into subsets $\{\boldsymbol{\Theta}_j\}_{j=1}^J$ that are sampled sequentially from

$$p\left(\boldsymbol{\Theta}_j^{(i+1)}|\{\boldsymbol{\Theta}_k^{(*)}\}_{k \neq j}, \mathbf{Y}\right), \tag{2.8}$$

where the asterisk denotes the most current sample of the parameters. In practice, Gibbs sampling is initialized with some reasonable parameter values $\boldsymbol{\Theta}^0$, and the sampling is continued until it converges. The convergence can be estimated based on e.g., the Geweke diagnostic [18]; for a review of different methods, see [19]. Notably, as the sampling assumes the remaining parameters to be fixed, a single update does not fully capture dependencies between the parameters. To counter this, it is customary to analyze a collection of Gibbs samples instead of just a few. It is helpful to assume conjugate priors for the parameters, allowing the sampling in Equation (2.8) to be performed from a closed-form distribution.

### 2.4.3 Prediction and missing value handling

Many data sets contain features that are only partially observed, and ideally all of them should be included in the model, in order to get maximal statistical strength. Here we limit to analysis regarding data missing at random, i.e., assumption that whether an element is observed is independent of its value. Missing values are handled in the Bayesian context by defining the likelihood only for the observed parts of $\mathbf{Y}$. In matrix factorization, this corresponds to the likelihood

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{N} \prod_{j=1}^{D} \mathcal{N}\left(y_{i,j}|\mathbf{x}_i^\top \mathbf{w}_j, \tau_j\right)^{J_{i,j}}, \tag{2.9}$$

where $J_{i,j}$ equals 1 for observed elements of $\mathbf{Y}$ and 0 for the missing values. This formulation allows natural use of all the observed data, and implies that e.g. $\mathbf{w}_j$ is updated based on the observations of the $j$th feature only, and the rest can be predicted using the inferred $\mathbf{w}_j$ and $\mathbf{X}$. Here it is worth noting that parameters corresponding to features (or samples) with no observations will follow their prior distribution, generally resulting is predictions with zero mean. For simplicity, Equation 2.9 is limited to matrix factorization with a diagonal covariance matrix (with elements $\boldsymbol{\tau}$), but the missing value handling generalizes beyond this in a straightforward manner.

It is also possible to infer the model parameters sequentially, which can be useful e.g. in case of large data (resulting in high computational time) that is augmented later with smaller batches of data. In this case the parameters are first inferred for the larger data set $\mathbf{Y}$, and the new parameters later for the new data batch $\mathbf{Y}'$, given the shared parameters. For example in matrix factorization, the inference task for new observations of the data can be formulated as

$$p(\mathbf{X}'|\mathbf{Y}', q(\mathbf{W})), \tag{2.10}$$

where $q(\mathbf{W})$ has been learned earlier from $\mathbf{Y}$. The inference in Equation 2.10 is generally really fast, as it does not depend on the original (larger) data $\mathbf{Y}$, and only the latent variables $\mathbf{X}'$ need to be sampled, opposed to alternating between the dependent $\mathbf{X}$ and $\mathbf{W}$. Here $q(\mathbf{W})$ denotes either the variational distribution or the collection of posterior samples of $\mathbf{W}$. Further, $\mathbf{X}'$ and $\mathbf{W}$ can be used to predict any missing values in $\mathbf{Y}'$.

### 2.4.4 Scalable inference

Model inference in Bayesian latent variable models is often a computationally demanding task: updating the variational distributions or sampling new parameters will generally include matrix operations, such as products and inverses. When the data set $\mathbf{Y}$ is large, it can be prohibitively expensive to iterate the inference procedures sufficiently long for them to be practical. Furthermore, the data set $\mathbf{Y}$ can be too large to store in the memory of a standard computer, or even a supercomputer. In this kind of situations, it is necessary to divide the data into parts, and perform inference in parallel with a large number of CPUs for these manageable submodels, with the goal of getting the best possible approximation of the posterior distribution. Communication between supercomputer nodes handling the different parts of the data allows for accurate inference, but as the number of nodes increases, the communication time begins to dominate the computational complexity [20]. To counter this, it is possible to infer the submodels with minimal communication, using a framework called embarrassingly parallel MCMC [21, 22, 23, 24]. There, each of the submodels is inferred independently of each other, after which they are aggregated together to approximate the full posterior distribution. Various techniques for aggregating subset posteriors have been reviewed in [25].

The aggregation step of embarrassingly parallel MCMC is extremely challenging if the model parameters $\mathbf{\Theta}$ are unidentifiable, which is usually the case especially in matrix factorization, where the data reconstruction $\mathbf{X}\mathbf{W}^\top$ equals that of $\mathbf{X}\mathbf{R}\mathbf{R}^{-1}\mathbf{W}^\top$, if $\mathbf{R}$ is any rotation matrix. To alleviate this problem, Publication V presents posterior propagation framework, build on top of Bayesian matrix factorization [7], where the posteriors of submodels are passed onwards as priors of related submodels. Specifically, if the rows of $\mathbf{Y}$ are split into $I$ parts and the columns into $J$ parts, with one subset denoted as $\mathbf{Y}^{(i,j)}$, the procedure is started by inferring the posterior

$$p\left(\mathbf{X}^{(1)}, \mathbf{W}^{(1)} | \mathbf{Y}^{(1,1)}\right),$$

using Gibbs sampling. This is followed by the second stage, where Gaussian approximations of the samples of $\mathbf{X}^{(1,1)}$ and $\mathbf{W}^{(1,1)}$, denoted with function $q$, are set as the priors of the corresponding subsets of samples and features, respectively. The following submodel inferences can be run

in parallel using $I + J - 2$ cores:

$$p\left(\mathbf{X}^{(i)}, \mathbf{W}^{(1)} | \mathbf{Y}^{(i,1)}, q\left(\mathbf{W}^{(1)}\right)\right) \ \forall i = 2, ..., I$$
$$p\left(\mathbf{X}^{(1)}, \mathbf{W}^{(j)} | \mathbf{Y}^{(1,j)}, q\left(\mathbf{X}^{(1)}\right)\right) \ \forall j = 2, ..., J.$$

Finally, the rest of the submodels,

$$p\left(\mathbf{X}^{(i)}, \mathbf{W}^{(j)} | \mathbf{Y}^{(i,j)}, q\left(\mathbf{X}^{(i)}\right), q\left(\mathbf{W}^{(j)}\right)\right) \ \forall i = 2, ..., I \wedge j = 2, ..., J$$

can be inferred in parallel using $(I - 1)(J - 1)$ cores. The usage of the posteriors as priors of further stages is shown to aid in preserving identifiability, and in most cases, it is sensible to aggregate the parameters $\mathbf{X}^{(:)}$ and $\mathbf{W}^{(:)}$ of different submodels by simply averaging them (with the excessive weight of $q\left(\mathbf{X}^{(1)}\right)$ and $q\left(\mathbf{W}^{(1)}\right)$ subtracted, as they are propagated multiple times). In the experiments of Publication V, using the posteriors of the first stages subsequently as priors was shown to induce dependencies for the largely parallel inferences, resulting in accurate prediction of missing values in a short time. The procedure was shown to outperform three state-of-the-art posterior aggregation algorithms applied to naive embarrassingly parallel matrix factorization. Furthermore, the presented posterior propagation methodology is directly applicable to any Bayesian matrix factorization method.

# 3. Analysis of Multiple Data Sources

Many practical uses of machine learning methods, such as the Bayesian latent variable models, occur in settings where there exist data from multiple different sources, each somehow related to the other. Suitable models can be used to uncover the relations of the data sources, as well as to predict missing data from all the relevant observations. Matrix and tensor decomposition based approaches will be discussed here, as well as related regression methods, where one data source is set to be predicted by the others.

## 3.1 Group factor analysis

Besides the internal stucture within matrices, also the relations between different data matrices are generally of interest when analyzing a group of matrices $\mathbf{Y}^{(1)}$, $\mathbf{Y}^{(2)}$, ..., $\mathbf{Y}^{(M)}$. Besides variation that is shared across all the data sources, more specific relations can occur. Matrix factorization could be applied directly to $\mathbf{Y}^{(:)}$ as well, if these are collected together in a larger matrix $\mathbf{Y}'$. This kind of a treatment makes an implicit assumption that the different data sources are homogeneous, and can be directly thought of as just one larger data matrix. Often this kind of an assumption is unrealistic, and even if the inferred model could be able to learn differences between parts of the matrix $\mathbf{Y}'$, it will likely degrade the model's predictive and interpretative performance. A solution for explicit handling of multiple data sources in factor analysis has been coined as multi-battery factor analysis (MBFA) [26, 27]. MBFA assumes that there are $K$ components shared by all the data sources, as well as a specific number of component specific to each data source, explaining away variance that is not related to any of the other data sources. The MBFA problem has gained attention recently, with some solutions includ-

ing [28, 29, 30, 31, 32]. A related problem is that of canonical correlation analysis (CCA), which aims to find the dependencies of two matrices [33, 34].

Extending beyond MBFA and CCA, ideally any subset of the data sources should be able to contain structure not present in the other data sources. This would allow for maximal flexibility in identifying the latent structure, and would prevent any irrelevant data sources from affecting the components. A straightforward solution for this would be specifying $K$ components to affect each subset of the data sources [35], but this results in an exponential model complexity with respect to the number of data sources $M$. A solution avoiding this problem was presented in [8] for two data sources, and extended to multiple data sources in [1] and Publication I, as well as in [36, 37]. Specifically, each data source $\mathbf{Y}^{(m)}_{N\times D_m}$ is given the likelihood

$$\mathbf{Y}^{(m)} = \prod_{i=1}^{N} \mathcal{N}\left(\mathbf{y}_i^{(m)} | \mathbf{W}^{(m)}\mathbf{x}_i, \mathbf{\Lambda}^{(m)-1}\right),$$

where $\mathbf{W}^{(m)}_{D_m\times K}$ is given a group sparse prior determining the scale of its columns $\mathbf{w}^{(m)}_{:,k}$. Multiple formulations exist for this group sparsity prior, but in practice they serve the same purpose: if only little evidence of component $k$ is found in data source $m$, then $\mathbf{w}^{(m)}_{:,k}$ is pushed (close) to zero. This in turn means that the model is able to divide each of the total of $K$ components to explain structure in any subset of the data sources. This includes components that are shared across all the data sources, as well as those specific to a single data source. The model is intended to explain all (or most) of the covariance structure using the flexible component model, and thus it uses a diagonal precision matrix $\mathbf{\Lambda}^{(m)}$ , or possibly even $\mathbf{\Lambda}^{(m)} = \tau_m \mathbf{I}_{D_m}$. Either choice allows learning different scales of residual variance for the different data sources, which can be crucial when analyzing heterogeneous data collections, as it allows giving more weight to the most reliable observations. Non-Bayesian models addressing the group factor analysis problem include [38] and [39].

### 3.1.1 Model sparsity

Two ways of imposing group sparsity are discussed here. To begin with, the ARD prior is used in Publication I as follows:

$$\mathbf{w}^{(m)}_{:,k} \sim \mathcal{N}\left(0, \alpha^{-1}_{m,k}\mathbf{I}_{D_m}\right)$$
$$\alpha_{m,k} \sim \mathcal{G}\left(10^{-14}, 10^{-14}\right).$$

Here $\alpha_{m,k}$ determines the scale of component $k$ in the $m$th data view: high precision implies that the component has little effect in it. The prior of each element of $\boldsymbol{\alpha}$ has an expected value of 1, and variance of $10^{14}$, making the prior very uninformative. The GFA model defined with this ARD prior is inferred with variational Bayesian techniques.

Publication IV uses a more flexible spike-and-slab prior defined as:

$$w_{d,k}^{(m)} \sim h_{m,k}\mathcal{N}\left(0, \alpha_{m,k}^{-1}\right) + (1 - h_{m,k})\,\delta_0 \tag{3.1}$$

$$\begin{aligned} \alpha_{m,k} &\sim \mathcal{G}\,(1,1) \\ h_{m,k} &\sim Bernoulli\,(\pi_{m,k}) \\ \pi_{m,k} &\sim Beta\,(1,1)\,, \end{aligned} \tag{3.2}$$

where $h_{m,k}$ is a binary variable indicating whether component $k$ is present if the $m$th data source. Alternatively, if feature-wise sparsity is sought, it is possible to replace $h_{m,k}$ with $h_{d,k}^{(m)}$ in Equations (3.1) and (3.2), as described in Publication III and [40]. In this case variable $\pi_{m,k}$ (the probability of associating component $k$ with data view $m$), ties the data sources together.

As both the group sparsity priors contain variable $\boldsymbol{\alpha}$ determining the scale of the components, it is not necessary to include this information in the prior of $\mathbf{X}$. Hence, if components that are present in all the samples of the data are sought, $\mathcal{N}(0,1)$ is a reasonable prior for the elements of $\mathbf{X}$, as used in Publication I. However, if it is more meaningful to assume the components to be present in a subset of the observations only, it is possible to give $\mathbf{X}$ a feature-wise sparse spike-and-slab prior, as demonstrated in Publication III, where the model is also extended to account for data sets paired in two modes (extending for arbitrarily paired data source would be possible, as demonstrated in [41]). Feature-wise sparsity in both the modes results in a novel method of biclustering multiple data sources. Some relevant work for biclustering a single data source are presented in [42, 43, 44].

Software package GFA[1] implements a Gibbs sampler for all the options discussed above. The software is introduced in Publication IV. Alternative sparsity priors for GFA include the three-parameter beta prior [45], that utilizes separate variables to induce global (related to the whole data collection), component-specific and local (feature-wise) sparsity [46].

[1] Available in `https://cran.r-project.org/package=GFA`.

### 3.1.2 Low-rank approximation

If the number of data sources $M$ grows very high, the group sparsity assumption of GFA may be too flexible, not fully benefitting on the statistical strength of numerous related data sources. To counteract this, Publication I uses the ARD prior to enable group sparsity, but sets:

$$\log \boldsymbol{\alpha} = \mathbf{U}\mathbf{V}^\top + \mathbf{a}\mathbf{1}^\top + \mathbf{1}\mathbf{b}^\top$$
$$u_{m,r},\, v_{k,r},\, a_m,\, b_k \sim \mathcal{N}(0,1),$$

where $\mathbf{U}_{M\times R}$ is a low-dimensional description of component activities in data sources, and $\mathbf{V}_{K\times R}$ similarly a low-dimensional representation of the component structure. The length $M$ vector $\mathbf{a}$ models the mean profiles of the data sources (some with more/less activity that the others), and the length $K$ vector $\mathbf{b}$ the mean profiles of the components. Here $R$ determining the rank of $\boldsymbol{\alpha}$ is a crucial parameter determining the complexity of the group sparsity. The experiments in Publication I are performed for a variety of ranks, and the results describe the complexity of the investigated data sets; also methods for choosing $R$ based on predictive RMSE, lower bound and computational time are discussed. Further, the applicability of using $\mathbf{U}$ to visualize data sources, along with their relations, is shown. For $R = 2$ this can be done by simply making a scatter plot of the matrix $\mathbf{U}$, resulting in data sources with similar component structure shown near each other.

## 3.2 Multi-tensor factorization

Generalizing GFA for a collection of tensors $\{\mathcal{Y}^{(m)}_{N\times D_m\times L_m}\}_{m=1}^M$ (including matrices with $L_m = 1$) is achieved using CP decomposition by defining the following likelihood:

$$p(\mathcal{Y}|\boldsymbol{\Theta}) \sim \prod_{m=1}^{M}\prod_{i=1}^{N}\prod_{d=1}^{D_m}\prod_{l=1}^{L_m}\mathcal{N}\left(y^{(m)}_{i,d,l}|\sum_{k=1}^{K} x_{i,k}w^{(m)}_{d,k}u^{(m)}_{l,k}), \tau_m^{-1}\right). \tag{3.3}$$

The model is completed by giving $\mathcal{N}(0,1)$ priors to the elements of $\mathbf{X}$, and spike-and-slab prior defined in Equation (3.1) for $\mathbf{W}^{(m)}$. Further, $\mathbf{U}^{(m)}$ is set as $\mathbf{1}_{1\times K}$ for matrices ($L_m = 1$), and given the $\mathcal{N}(0,1)$ prior for tensors. This model has been presented in [47] and Publication II, coined as multi-tensor factorization (MTF), enabling modelling the relations of data sources of different shapes, bound together by the shared latent components $\mathbf{X}$. The posterior distribution is inferred using Gibbs sampling.

Earlier research regarding coupled matrix-tensor factorization (generally limited to one tensor only) has been performed using non-probabilistic models [48, 49, 50, 51], assuming an underlying CP decomposition for tensors and using a gradient-based least squares optimization approach.

### 3.2.1 Relaxed decomposition

The key assumption of CP factorization, namely that each $\mathbf{Y}^{(m)}_{:,:,l}$ (called a tensor slab) shares the same $\mathbf{X}$ and $\mathbf{W}^{(m)}$, with merely a deviation of scale allowed by $\mathbf{U}^{(m)}$, can often be limiting. In the context of brain imaging, this corresponds to assuming that all the subjects share exactly the same (spatial) response structure, if the data dimensions are ordered as temporal, spatial and subject. On the other hand, treating the tensor slabs independent a priori limits the statistical strength gained by the repeated measurements. To provide a continuum between these choices, Publication II proposes the following model:

$$
\begin{aligned}
y^{(m)}_{n,d,l} &\sim \mathcal{N}\left(\mathbf{x}_n^\top \mathbf{w}^{(m)}_{l,d}, \tau^{-1}_{m,l}\right) \\
w^{(m)}_{l,d,k} &\sim h^{(m)}_{l,k}\, \mathcal{N}\left(u^{(m)}_{l,k} v^{(m)}_{d,k}, (\alpha^{(m)}_{d,k})^{-1}\right) + \left(1 - h^{(m)}_{l,k}\right) \delta_0 \qquad (3.4) \\
x_{n,k}, u^{(m)}_{l,k} &\sim \mathcal{N}(0,1) \\
v^{(m)}_{d,k} &\sim \begin{cases} 0, & \text{if } L_m = 1, \\ \mathcal{N}\left(0, (\beta^{(m)}_{d,k})^{-1}\right), & \text{otherwise.} \end{cases} \\
\alpha^{(m)}_{d,k} &\sim \begin{cases} Gamma(1,1), & \text{if } L_m = 1, \\ \lambda, & \text{otherwise.} \end{cases} \\
h^{(m)}_{l,k} &\sim Bernoulli(\pi_k) \\
\pi_k &\sim Beta(1,1) \\
\beta^{(m)}_{d,k},\, \tau_{m,l},\, \lambda &\sim Gamma(1,1).
\end{aligned}
$$

For data matrices ($L_m = 1$) this corresponds to a standard GFA model with the spike-and-slab prior. Each tensor, however, is constructed from a relaxed CP decomposition, where Equation (3.4) allows for deviation from the mean profile $\mathbf{U}\mathbf{V}^\top$, governed by the precision parameter $\lambda$. Infinite precision would correspond to using a strict CP decomposition, as in MTF, whereas $\lambda \to 0$ would eliminate the dependencies between the tensor slabs. The uninformative Gamma prior, however, allows learning the degree of similarity that the tensor slabs have, in terms of the CP decomposition.

### 3.3 Reduced-rank regression

In many applications of machine learning, it is sufficient to be able to make predictions for one data source given the other(s), and the generative process of the rest is of negligible interest. For this kind of tasks, there exists a wide range of different regression models from regularized linear regression [52] to highly non-linear random forests [53]. In the scope of this dissertation, the most interesting are (linear) reduced-rank regression models, introduced in [54], that infer a latent space through which the regression takes place. Denoting $\mathbf{Y}$ as the target variables and $\mathbf{X}_{N\times P}$ as the covariates, the difference to the factorization models discussed is that, in general, only the latent space relevant to $p(\mathbf{Y}|\mathbf{X})$ is sought, while $p(\mathbf{X})$ is considered uninteresting. Bayesian reduced-rank regression[2] [2] is defined as

$$\mathbf{Y} = (\mathbf{X}\mathbf{\Psi} + \mathbf{\Omega})\mathbf{\Gamma} + \mathbf{E},$$

where $\mathbf{\Psi}_{P\times K}$ maps the covariates into the $K$-dimensional latent space, and $\mathbf{\Gamma}_{K\times D}$ projects the latent space into the observation space. The $\mathbf{\Omega}_{N\times K}$ contains factors representing latent noise, and $\mathbf{E}_{N\times D}$ describes residual noise in the observation space. The priors for elements of $\mathbf{\Psi}$ and $\mathbf{\Gamma}$ are set as normal distributions with two precision parameters, one implementing element-wise sparsity, and the other providing overall scale of the components, allowing automatic model complexity detection [2].

[2] http://research.cs.aalto.fi/pml/software/latentNoise/

# 4. Multi-view Models for Drug Response Studies

A fundamental question in life sciences is how chemical entities affect living cells. The responses are often diverse, and genome-wide measurements, along with suitable modelling tools for the complex data, are intended to provide insight to the action mechanisms [55]. This is especially of interest in drug response studies, where the effects of drugs on cells, either in living biological entities (*in vivo*, e.g. in humans) or outside of their biological context (*in vitro*), are inspected. Modern tools allow efficiently surveying the expression of the whole genome for a large number of cells [56], that is, inspecting how different genes are expressed as proteins, which are the functional units of the cell. The gene expression of cells before and after administrating a drug are naturally key to understanding the effects of the drugs, but fully uncovering their action mechanisms requires other data sources as well. Relevant sources include DNA (and RNA) methylation, that is, the process of adding methyl groups affecting the activity of a DNA segment, without changing the actual sequence [57]. Similarly, copy number variation, i.e., the number of times a DNA sequence is repeated in the genome, as well as mutations in the genes, inspected using exome sequencing are key to functioning of the cell [58, 59].

Publication I studies a data collection measured for studying connections among diseases, as well as drug action and the related chemical processes [60]. The collection contains the gene expressions of three cancer cell lines (*HL60, MCF7* and *PC3*; *in vitro*), after administrating $N = 684$ different drugs. Along with chemical descriptors (VolSurf) of the drugs, the data are organized as $M = 430$ matrices with $N$ samples each: 13 data sources of different types of chemical descriptors, as well as 139 functional pathways containing a total of 1032 genes for each of the cancer cell lines. GFA with low-rank ARD prior separates the distinct cell lines from each other without using any labels, showing that their expression levels

have significantly differing patterns. The quality of the factorization, i.e., modelling the relations between the functional pathways and the chemical descriptors, was evaluated in a drug retrieval task similar to [61]. That is, the latent variables $\mathbf{X}$ were used to evaluate the similarities of the drugs, and the quality of these results was assessed using an external database of drug function. Starting with the raw data, the VolSurf descriptors resulted in a much greater precision than the gene expression. GFA with a rank above 2 (and below full rank) managed to improve the VolSurf precisions, whereas the compared methods (FA, sparse FA, MBFA and similarity component analysis [38]) resulted in poor precisions, close to the level of gene expression only. Similar results were obtained when comparing the predictive RMSE for a left-out data set, given the others (model trained with $N = 616$). These results imply that GFA is able to learn meaningful relations of the chemical descriptors and the related physiological responses, and can be useful in drug retrieval tasks.

If a data collection of interest contains very heterogeneous measurements, it can be useful to utilize sparse priors for both $\mathbf{X}$ and $\mathbf{W}$, resulting in bicluster components, each of which affect a subset of the samples and features. This approach is adopted in Publication III, studying a data collection provided in a public NCI-DREAM drug sensitivity prediction challenge [62]. The collection contains measurements for $N = 53$ cancer cell lines, with the goal of predicting the drug responses for a left-out set of 18 cell lines. GFA was inferred to describe the relations of gene expression, DNA methylation, protein abundance and exome sequence, as well as the drug sensitivity measurements. As the number of observations is very low, the residual noise was given a mildly informative prior, assuming signal-to-noise ratio 0.5. Furthermore, the model parameters were inferred with 50 independent Gibbs sampling chains, each starting from a different initial set of parameters; the predictions were averaged over all the acquired posterior samples. This application of GFA resulted in a prediction score that would have won the DREAM challenge, demonstrating that GFA is able to infer action mechanisms of drugs. Furthermore, including functional connectivity fingerprints (calculated with PaDEL-Descriptor [63]) in the analysis, paired with the second mode of the drug sensitivity data, improved the score by a small amount. Analysis of the acquired components (biclusters) showed e.g. structure differentiating basal and luminal cell lines and associations with multiple cancer genes.

# 5. Multi-view Models for Brain Imaging Studies

The human brain is an incredibly complex organ that allows acquiring and processing information, problem solving and creativeness, among other things. The brain contains tens of billions of neurons [64], each connected to several thousand other neurons by synapses. This neural network is responsible for the advanced information processing of humans and, as its functioning is not fully understood yet, a challenging subject of study for many. Traditionally the functioning of the brain has been assessed through its host in behavioral studies, but modern technologies allow more direct measures: Magnetoencephalography (MEG; [65]) measures changes in magnetic fields around the brain, caused by neurons firing, resulting in temporally high-resolution recordings. Functional magnetic resonance imaging [66], on the other hand, measures the oxygenation of blood in the brain, serving as a proxy for neuronal activity, that is fueled with oxygenated blood. In contrast to MEG, this results in high spatial, but low temporal resolution, with one image recorded throughout the brain typically every 2 seconds. Both the imaging modalities result in large amounts of measurements with complex structure and large amount of variance irrelevant to the experiment (as it is difficult to isolate the brain), creating the need for suitable analysis methods.

Brain imaging studies have traditionally focused on studying the effects of relatively simple stimuli, or certain aspects of more complex stimuli, in order to obtain a reasonable signal-to-noise ratio [67]. However, studies show that responses to rich natural stimuli elicit more diverse brain responses than the simplified setups; see e.g. [68, 69, 70]. This raises the need for intelligent analysis methods that are able to account for complex relations between the stimuli and the brain responses. Furthermore, individual subjects can deviate significantly from the average in terms of their brain activity, and means for accounting this are needed as well.

The ability of relaxed multi-tensor factorization to account for inter-subject differences was studied in Publication II with measurements of 9 subjects listening to an auditory book for an hour in an MEG scanner [71]. The MEG recordings were stored in a tensor $\mathcal{Y}_{N\times D\times L}$ with $N = 28547$ samples, $D = 70$ features (standard mapping of the 204 MEG channels to a lower dimensional space) and $L = 9$. The tensor was factorized along with a 13-dimensional description of the power spectrum of the auditory stimulus. The quality of the factorization was estimated using the predictive RMSE of the model from the auditory features to the brain responses: relaxed MTF resulted in the best predictions with all the tested training set sizes (ranging from 5 to 35 minutes of recordings). For very short recordings, the strict CP assumption of MTF worked as a decent regularizer, yielding accuracy close to that of relaxed MTF. With more observations, however, the assumption limited the predictive performance, whereas GFA (each subject treated independently of each other) achieved predictions comparable to relaxed MTF. This shows that with enough data, also overly flexible assumptions can give good results. With longer time courses, a non-probabilistic coupled matrix-tensor factorization method [48] performed similarly to MTF, whereas its asymmetric version [49] did not relate the auditory signal and the brain responses at all. Relaxed MTF resulted in 1 to 4 shared component between the audio and the brain over different sample sizes: one component observed in all the posterior samples described the energy of the speech signal (with two peaks, corresponding to words and pauses between words), which was associated with auditory areas of the brain. As all the story descriptors were related to the power spectrum of the audio, only limited commonalities between the audio and the brain could be found.

Genome-wide association studies (GWAS) have been performed to estimate the genetic determinant of neurophysiological disorders, see e.g. [72, 73], as well as to analyze the brain activity of healthy subjects [74, 75, 76]. The studies are statistically challenging, as they aim to infer which genes or single nucleotide polymorphisms (SNPs), with dimensionality between tens of thousands and tens of millions, affect certain high-dimensional phenotypes [77]. In this kind of tasks, in order to maintain maximal statistical strength, it is beneficial to perform the data analysis using regression methods, where only latent variables relevant to the phenotype are inferred. We studied the experiment reported in [78] using Bayesian reduced-rank regression, where the latent space is utilized only to explain

brain activity given the genotype data. Publication VI reports a regression analysis for measurements of 150 thousand SNPs regarding 201 subjects, used to explain a total of 9 minutes of MEG recordings for each subject, with three conditions: eyes open, eyes closed, and making voluntary hand movements. In order to make the free-form recordings co-occurring, we inspected their power spectra (204 MEG channels $\times$ 21 frequency bands). BRRR allowed performing GWAS in this challenging domain in a statistically powerful way [79] by searching for low-dimensional descriptions of the MEG power spectrum. The approach resulted in statistically significant associations between the genome and the MEG power spectrum, providing insight into the genetic determination of brain activity.

We also applied BRRR to identify familial components from the MEG recordings of siblings in Publication VI, resulting in modes of brain responses (the latent space) similar within families. These modes could be identified already with a couple of seconds of data, and they served as fingerprints identifying later time courses of subjects' themselves accurately.

# 6. Discussion and Conclusions

This thesis has shown that latent variable models for multiple data sources are able to identify low-dimensional structure describing data collections, resulting in accurate predictions of missing values. These properties are demonstrated in the included publications with simulated data, as well as drug response and brain imaging studies, including a genome-wide association study.

This thesis has introduced several novel multi-view methods, each suitable for somewhat different application needs. In Publication I, group factor analysis was extended to allow implicit modelling of the data source relations, especially suitable for collections of a large number of data sources. This approach extended previous research on multi-battery factor analysis (e.g. [26, 32]) and provided a novel tool for visualizing the relations of data sources. GFA was also adapted for data sources paired in two modes with sparse priors in Publication III, resulting in interpretable biclusters describing the relations of different measurements on cancer cell lines and drugs, and providing superior predictions for the drug sensitivities of the cell lines. The work was published as an open-source software package in Publication IV, allowing easy application by researchers and practitioners in a wide range of problems.

Furthermore, joint factorization of multiple tensors was extended in Publication II by proposing a framework forming a continuum between strict tensor factorization and in some cases overly flexible matrix factorization. This approach enables accurate joint modelling of similar data sources (e.g., recurrent experiments, such as brain imaging measurements of several subjects), while allowing for differences between them. In contrast, traditional tensor factorization methods, such as [47, 48], resort to stricter assumptions.

The applications discussed have been for data collections of moderate

size, allowing model inference using either variational Bayesian techniques or Gibbs sampling for the full data collection. However, for significantly larger data collections this may no longer be feasible in a reasonable time. This issue was addressed in Publication V by proposing a framework for partitioning a data matrix, and inferring smaller submodels in parallel. This approach, coined as posterior propagation, was shown to produce reasonably good predictions in a fraction of the wall-clock time of the inference with full data. The framework is generalizable for multi-view data as well, but this is left as future work.

Finally, Bayesian reduced-rank regression was shown to be able to perform a genome-wide association study with as few as 200 subjects in Publication VI, with the phenotype (MEG spectral power) consisting of thousands of features and the genotype of 150 thousand SNPs. Joint modelling of the two data sources (genotype and the phenotype) through a low-dimensional latent space was shown to enable this kind of an application, resulting in interesting findings regarding both the heritable features of brain activity, and their genetic determination.

In summary, this thesis has presented new Bayesian latent variable models that are suitable for gaining knowledge from various types of data collections and the relations of the data sources. The models were shown to exceed state-of-the-art level predictive power, and result in interpretable findings regarding drug response and brain imaging studies. In this thesis, only linear relations between the data sources were sought, and non-linear extensions of the presented latent variable models could provide even further influence for this type of analysis. Finally, there lies great potential for gaining insights and encountering novel modelling problems in applying the presented methods wider within life sciences, as well as beyond them.

# References


[1] S. Virtanen, A. Klami, S. A. Khan, and S. Kaski, "Bayesian group factor analysis," in *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pp. 1269–1277, 2012.

[2] J. Gillberg, P. Marttinen, M. Pirinen, A. J. Kangas, P. Soininen, M. Ali, A. S. Havulinna, M.-R. Järvelin, M. Ala-Korpela, and S. Kaski, "Multiple output regression with latent noise," *Journal of Machine Learning Research*, vol. 17, no. 122, pp. 1–35, 2016.

[3] L. L. Thurstone, "Multiple factor analysis," *Psychological Review*, vol. 38, no. 5, pp. 406–427, 1931.

[4] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of Educational Psychology*, vol. 24, no. 6, p. 417, 1933.

[5] N. Wang, T. Yao, J. Wang, and D.-Y. Yeung, "A probabilistic approach to robust matrix factorization," in *Proceedings of the European Conference on Computer Vision*, pp. 126–139, 2012.

[6] B. Lakshminarayanan, G. Bouchard, and C. Archambeau, "Robust Bayesian matrix factorisation," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 425–433, 2011.

[7] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 880–887, 2008.

[8] S. Virtanen, A. Klami, and S. Kaski, "Bayesian CCA via group sparsity," in *Proceedings of the 28th International Conference on Machine Learning*, pp. 457–464, 2011.

[9] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[10] R. A. Harshman, "Foundations of the PARAFAC procedure: models and conditions for an explanatory multimodal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[11] T. Kolda and B. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[12] D. MacKay, "Bayesian methods for backpropagation networks," *Models of Neural Networks III*, vol. 6, pp. 211–254, 1996.

[13] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[14] G. Parisi and R. Shankar, *Statistical field theory*. AIP, 1988.

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

[16] D. Blei and J. Lafferty, "Correlated topic models," in *Advances in Neural Information Processing Systems 18*, pp. 147–154, 2006.

[17] C. Bishop, *Pattern Recognition and Machine Learning*, vol. 4. Springer, 2006.

[18] J. Geweke, "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," in *Bayesian Statistics*, vol. 4, pp. 169–193, Oxford University Press, New York, 1992.

[19] M. K. Cowles and B. P. Carlin, "Markov chain Monte Carlo convergence diagnostics: a comparative review," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 883–904, 1996.

[20] T. Vander Aa, I. Chakroun, and T. Haber, "Distributed Bayesian probabilistic matrix factorization," in *Proceedings of the International Conference on Cluster Computing*, pp. 346–349, IEEE, 2016.

[21] S. Minsker, S. Srivastava, L. Lin, and D. Dunson, "Robust and scalable Bayes via a median of subset posterior measures," *arXiv preprint arXiv:1403.2660*, 2014.

[22] W. Neiswanger, C. Wang, and E. Xing, "Asymptotically exact, embarrassingly parallel MCMC," in *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pp. 623–632, AUAI Press, 2014.

[23] S. L. Scott, A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch, "Bayes and big data: the consensus Monte Carlo algorithm," *International Journal of Management Science and Engineering Management*, vol. 11, no. 2, pp. 78–88, 2016.

[24] X. Wang and D. Dunson, "Parallelizing MCMC via Weierstrass sampler," *arXiv preprint arXiv:1312.4605*, 2013.

[25] E. Angelino, M. J. Johnson, and R. P. Adams, "Patterns of scalable Bayesian inference," *arXiv preprint arXiv:1602.05221*, 2016.

[26] M. W. Browne, "The maximum-likelihood solution in inter-battery factor analysis," *British Journal of Mathematical and Statistical Psychology*, vol. 32, pp. 75–86, 1979.

[27] L. R. Tucker, "An inter-battery method of factor analysis," *Psychometrika*, vol. 23, pp. 111–136, 1958.

[28] C. Archambeau and F. Bach, "Sparse probabilistic projections," in *Advances in Neural Information Processing Systems 21*, pp. 73–80, 2009.

[29] F. Deleus and M. V. Hulle, "Functional connectivity analysis of fMRI data based on regularized multiset canonical correlation analysis," *Journal of Neuroscience Methods*, vol. 197, no. 1, pp. 143–157, 2011.

[30] X. Qu and X. Chen, "Sparse strucured probabilistic projections for factorized latent spaces," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pp. 1389–1394, 2011.

[31] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel, "Joint and individual variation explained (JIVE) for integrated analysis of multiple data types," *Annals of Applied Statistics*, vol. 7, no. 1, pp. 523–542, 2013.

[32] P. Ray, L. Zheng, Y. Wang, J. Lucas, D. Dunson, and L. Carin, "Bayesian joint analysis of heterogeneous data," tech. rep., Duke University, 2013.

[33] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Tech. Rep. 688, Department of Statistics, University of California, Berkeley, 2005.

[34] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3, pp. 321–377, 1936.

[35] S. K. Gupta, D. Phung, B. Adams, and S. Venkatesh, "A Bayesian framework for learning shared and individual subspaces from multiple data sources," in *Advances in Knowledge Discovery and Data Mining, 15th Pacific-Asia Conference, PAKDD*, pp. 136–147, 2011.

[36] A. Damianou, C. Ek, M. Titsias, and N. Lawrence, "Manifold relevance determination," in *Proceedings of the 29th International Conference on Machine Learning*, pp. 145–152, 2012.

[37] S. K. Gupta, D. Phung, and S. Venkatesh, "A Bayesian nonparametric joint factor model for learning shared and individual subspaces from multiple data sources," in *Proceedings of the 12th SIAM International Conference on Data Mining*, pp. 200–211, 2012.

[38] K. V. Deun, T. F. Wilderjans, R. A. v. Berg, A. Antoniadis, and I. V. Mechelen, "A flexible framework for sparse simultaneous component based data integration," *BMC Bioinformatics*, vol. 12, no. 448, 2011.

[39] Y. Jia, M. Salzmann, and T. Darrell, "Factorized latent spaces with structured sparsity," in *Advances in Neural Information Processing Systems 23*, pp. 982–990, 2010.

[40] T. Suvitaival, J. A. Parkkinen, S. Virtanen, and S. Kaski, "Cross-organism toxicogenomics with group factor analysis," *Systems Biomedicine*, vol. 2, no. 4, pp. 71–80, 2014.

[41] A. Klami, G. Bouchard, and A. Tripathi, "Group-sparse embeddings in collective matrix factorization," in *International Conference on Learning Representations*, 2014.

[42] J. A. Hartigan, "Direct clustering of a data matrix," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123–129, 1972.

[43] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[44] J. N. Morgan and J. A. Sonquist, "Problems in the analysis of survey data, and a proposal," *Journal of the American Statistical Association*, vol. 58, no. 302, pp. 415–434, 1963.

[45] A. Armagan, M. Clyde, and D. B. Dunson, "Generalized beta mixtures of gaussians," in *Advances in Neural Information Processing Systems 24*, pp. 523–531, 2011.

[46] S. Zhao, C. Gao, S. Mukherjee, and B. E. Engelhardt, "Bayesian group factor analysis with structured sparsity," *Journal of Machine Learning Research*, vol. 17, no. 196, pp. 1–47, 2016.

[47] S. A. Khan and S. Kaski, "Bayesian multi-view tensor factorization," in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 656–671, 2014.

[48] E. Acar, T. G. Kolda, and D. M. Dunlavy, "All-at-once optimization for coupled matrix and tensor factorizations," *arXiv preprint arXiv:1105.3422*, 2011.

[49] E. Acar, A. J. Lawaetz, M. A. Rasmussen, and R. Bro, "Structure-revealing data fusion model with applications in metabolomics," in *Proceeding of the 35th Annual International Conference on Engineering in Medicine and Biology Society*, pp. 6023–6026, IEEE, 2013.

[50] E. Acar, M. A. Rasmussen, F. Savorani, T. Naes, and R. Bro, "Understanding data fusion within the framework of coupled matrix and tensor factorizations," *Chemometrics and Intelligent Laboratory Systems*, vol. 129, pp. 53–63, 2013.

[51] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Towards mobile intelligence: Learning from GPS history data for collaborative recommendation," *Artificial Intelligence*, vol. 184, pp. 17–37, 2012.

[52] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[53] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[54] A. J. Izenman, "Reduced-rank regression for the multivariate linear model," *Journal of Multivariate Analysis*, vol. 5, no. 2, pp. 248–264, 1975.

[55] E. C. Butcher, E. L. Berg, and E. J. Kunkel, "Systems biology in drug discovery," *Nature Biotechnology*, vol. 22, no. 10, pp. 1253–1259, 2004.

[56] J. Kononen, L. Bubendorf, A. Kallionimeni, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter, and O.-P. Kallionimeni, "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nature Medicine*, vol. 4, no. 7, pp. 844–847, 1998.

[57] A. Razin and A. D. Riggs, "DNA methylation and gene function," *Science*, vol. 210, no. 4470, pp. 604–610, 1980.

[58] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.-L. Kuo, C. Chen, Y. Zhai, *et al.*, "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays," *Nature Genetics*, vol. 20, no. 2, pp. 207–211, 1998.

[59] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure, "Exome sequencing as a tool for Mendelian disease gene discovery," *Nature Reviews Genetics*, vol. 12, no. 11, pp. 745–755, 2011.

[60] J. Lamb, E. Crawford, D. Peck, J. Modell, I. Blat, M. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. Ross, M. Reich, H. Hieronymus, G. Wei, S. Armstrong, S. Haggarty, P. Clemons, R. Wei, S. Carr, E. Lander, and T. Golub, "The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[61] S. A. Khan, A. Faisal, J. P. Mpindi, J. A. Parkkinen, T. Kalliokoski, A. Poso, O. P. Kallioniemi, K. Wennerberg, and S. Kaski, "Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs," *BMC Bioinformatics*, vol. 13, no. 112, 2012.

[62] J. C. Costello, L. M. Heiser, *et al.*, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnology*, vol. 32, no. 12, pp. 1202–1212, 2014.

[63] C. W. Yap, "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints," *Journal of Computational Chemistry*, vol. 32, no. 7, pp. 1466–1474, 2011.

[64] D. Pelvig, H. Pakkenberg, A. Stark, and B. Pakkenberg, "Neocortical glial cell numbers in human brains," *Neurobiology of Aging*, vol. 29, no. 11, pp. 1754–1762, 2008.

[65] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalography — theory, instrumentation, and applications to noninvasive studies of the working human brain," *Reviews of Modern Physics*, vol. 65, no. 2, pp. 413–505, 1993.

[66] S. Ogawa, T.-M. Lee, A. R. Kay, and D. W. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proceedings of the National Academy of Sciences*, vol. 87, no. 24, pp. 9868–9872, 1990.

[67] J. Ylipaavalniemi, E. Savia, S. Malinen, R. Hari, R. Vigário, and S. Kaski, "Dependencies between stimuli and spatially independent fMRI sources: Towards brain correlates of natural stimuli," *NeuroImage*, vol. 48, no. 1, pp. 176–185, 2009.

[68] A. J. Anderson, J. R. Binder, L. Fernandino, C. J. Humphries, L. L. Conant, M. Aguilar, X. Wang, D. Doko, and R. D. S. Raizada, "Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation," *Cerebral Cortex*, pp. 1–17, 2016.

[69] P. Hagoort and P. Indefrey, "The neurobiology of language beyond single words.," *Annual Review of Neuroscience*, vol. 37, pp. 347–362, 2014.

[70] A. G. Huth, W. A. De Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, pp. 453–458, 2016.

[71] M. Koskinen and M. Seppä, "Uncovering cortical MEG responses to listened audiobook stories," *NeuroImage*, vol. 100, pp. 263–270, 2014.

[72] S. G. Potkin, J. A. Turner, G. Guffanti, A. Lakatos, J. H. Fallon, D. D. Nguyen, D. Mathalon, J. Ford, J. Lauriello, F. Macciardi, *et al.*, "A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype," *Schizophrenia Bulletin*, vol. 35, no. 1, pp. 96–108, 2009.

[73] L. Shen, S. Kim, S. L. Risacher, K. Nho, S. Swaminathan, J. D. West, T. Foroud, N. Pankratz, J. H. Moore, C. D. Sloan, *et al.*, "Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort," *NeuroImage*, vol. 53, no. 3, pp. 1051–1063, 2010.

[74] E. Salmela, H. Renvall, J. Kujala, O. Hakosalo, M. Illman, M. Vihla, E. Leinonen, R. Salmelin, and J. Kere, "Evidence for genetic regulation of the human parieto-occipital 10-Hz rhythmic activity," *European Journal of Neuroscience*, vol. 44, no. 3, pp. 1963–1971, 2016.

[75] C. M. Smit, M. J. Wright, N. K. Hansell, G. M. Geffen, and N. G. Martin, "Genetic variation of individual alpha frequency (IAF) and alpha power in a large adolescent twin sample," *International Journal of Psychophysiology*, vol. 61, no. 2, pp. 235–243, 2006.

[76] M. N. Smolka, G. Schumann, J. Wrase, S. M. Grüsser, H. Flor, K. Mann, D. F. Braus, D. Goldman, C. Büchel, and A. Heinz, "Catechol-O-methyltransferase Val158Met genotype affects processing of emotional stimuli in the amygdala and prefrontal cortex," *Journal of Neuroscience*, vol. 25, no. 4, pp. 836–842, 2005.

[77] D. P. Hibar, O. Kohannim, J. L. Stein, M.-C. Chiang, and P. M. Thompson, "Multilocus genetic analysis of brain images," *Frontiers in Genetics*, vol. 2, pp. 1–11, 2011.

[78] H. Renvall, E. Salmela, M. Vihla, M. Illman, E. Leinonen, J. Kere, and R. Salmelin, "Genome-wide linkage analysis of human auditory cortical activation suggests distinct loci on chromosomes 2, 3, and 8," *The Journal of Neuroscience*, vol. 32, no. 42, pp. 14511–14518, 2012.

[79] M. Vounou, T. E. Nichols, and G. Montana, "Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach," *NeuroImage*, vol. 53, no. 3, pp. 1147–1159, 2010.

Aalto-DD 178/2018

ISBN 978-952-60-8184-7 (printed)
ISBN 978-952-60-8185-4 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

**Aalto University**
**School of Science**
**Department of Computer Science**
**www.aalto.fi**

**BUSINESS +**
**ECONOMY**

**ART +**
**DESIGN +**
**ARCHITECTURE**

**SCIENCE +**
**TECHNOLOGY**

**CROSSOVER**

**DOCTORAL**
**DISSERTATIONS**