Department of Information and Computer Science

Multivariate Multi-Way Modelling of Multiple High-Dimensional Data Sources

Ilkka Huopaniemi

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 117/2012

Multivariate Multi-Way Modelling of Multiple High-Dimensional Data Sources

Ilkka Huopaniemi

Doctoral dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the School of Science for public examination and debate in Auditorium T2 at the Aalto University School of Science (Espoo, Finland) on the 12th of October 2012 at noon (at 12 o'clock).

Aalto University
School of Science
Department of Information and Computer Science

Supervising professor
Prof. Samuel Kaski

Preliminary examiners
Prof. Antti Penttinen, University of Jyväskylä
Dr. Simon Rogers, University of Glasgow, United Kingdom

Opponent
Dr. Colin Campbell, University of Bristol, United Kingdom

Aalto University publication series
DOCTORAL DISSERTATIONS 117/2012

ISBN 978-952-60-4782-9 (printed)
ISBN 978-952-60-4783-6 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-4783-6

Unigrafia Oy
Helsinki 2012

Finland

Publication orders (printed book):
ihuopani@cc.hut.fi

441 697
Printed matter

Aalto University

Abstract

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

Author
Ilkka Huopaniemi

Name of the doctoral dissertation
Multivariate Multi-Way Modelling of Multiple High-Dimensional Data Sources

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 117/2012

Field of research Computer and Information Science

Manuscript submitted 12 June 2012 **Date of the defence** 12 October 2012

Permission to publish granted (date) 20 August 2012 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

A widely employed strategy in current biomedical research is to study samples from patients using high-throughput measurement techniques, such as transcriptomics, proteomics, and metabolomics. In contrast to the static information obtained from the DNA sequence, these techniques deliver a ``dynamic fingerprint'' describing the phenotypic status of the patient in the form of absolute or relative concentrations of hundreds, or even tens of thousands of molecules: mRNA, proteins, metabolites and lipids. The huge number of variables measured opens up new possibilities for biomedical research; harnessing the information contained in such 'omics' data requires advanced data analysis methods.

The standard setup in biomedical research is comparing case (diseased) and control (healthy) samples and determining differentially expressed molecules that are then considered potential bio-markers for disease. In modern biomedical experiments, more complicated research questions are common. For instance, diet or drug treatments, gender and age play central roles in many case-control experiments and the measurements are often in the form of a time-series. Due to these additional covariates, the experimental setting becomes a multi-way experimental design, but few tools for proper data-analysis of high-dimensional data with such a design exist. Moreover, the task of integrating multiple data sources with different variables is nowadays often encountered in two classes of biomedical experiments: (i) Multiple omics types or samples from several tissues are measured from each patient (paired samples), (ii) Translating biomarkers between human studies and model organisms (no paired samples). These data integration tasks usually additionally involve a multi-way experimental design.

In this dissertation, a novel Bayesian machine learning model for multi-way modelling of data from such multi-way, single-source or multi-source setups is presented, covering the majority of situations commonly encountered in statistical analysis of omics data coming from current biomedical research. The problem of high dimensionality is solved by assuming that the data can be described as highly correlated groups of variables. The Bayesian modelling approach involves training a single, unified, interpretable model to explain all the data. This approach can overcome the main difficulties in omics analysis: small sample-size and high dimensionality, multicollinearity of data, and the problem of multiple testing. This approach also enables rigorous uncertainty estimation, dimensionality reduction and easy interpretability of results from a complex setup involving multiple covariates and multiple data sources.

Keywords Bayesian methods, data integration, machine learning, multi-way ANOVA, small sample-size

ISBN (printed) 978-952-60-4782-9 **ISBN (pdf)** 978-952-60-4783-6

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

Location of publisher Espoo **Location of printing** Helsinki **Year** 2012

Pages 181 **urn** http://urn.fi/URN:ISBN:978-952-60-4783-6

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Ilkka Huopaniemi

Väitöskirjan nimi
Usean korkeaulotteisen datalähteen analyysi monisuuntaisessa koeasetelmassa

Julkaisija Perustiedeiden korkeakoulu

Yksikkö Tietojenkäsittelytieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 117/2012

Tutkimusala Informaatiotekniikka

Käsikirjoituksen pvm 12.06.2012 **Väitöspäivä** 12.10.2012

Julkaisuluvan myöntämispäivä 20.08.2012 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Yleinen modernissa biolääketieteellisessä tutkimuksessa käytetty menetelmä on tehdä mittauksia potilaista saaduista näytteistä transkriptomiikkaa, proteomiikkaa, metabolomiikkaa ja lipidomiikkaa käyttäen. Näillä 'omiikka'-tekniikoilla pystytään samanaikaisesti mittaamaan jopa kymmenien tuhansien molekyylien (lähetti-RNAn, proteiinien, metaboliittien, lipidien) konsentraatiot. Näiden potilaan tilaa kuvaavien muuttujien suuri määrä avaa uusia mahdollisuuksia lääketieteelle, mutta informaation löytäminen valtavasta havaintoaineistosta edellyttää edistyneitä data-analyysimenetelmiä.

Tässä väitöskirjassa on tutkittu omiikka-aineistojen tilastollista analyysia, kun näytteet (potilaat) on mitattu monisuuntaisessa koeasetelmassa. Yksisuuntainen koeasetelma tarkoittaa molekyylien konsentraatioiden suuruuden vertaamista esimerkiksi terveiden ja diabetesta sairastavien potilaiden välillä. Monisuuntaisessa koeasetelmassa potilasta kuvaa kaksi (tai useampi) kovariaattia, kuten taudin lisäksi sukupuoli, ikä tai annettu lääke, ja mittaukset voivat myös muodostaa aikasarjan. Biolääketieteellisistä kokeista tulevien tietoaineistojen analyysissa joudutaan usein myös yhdistämään useasta eri lähteestä tulevia aineistoja. Mittaukset tehdään monesti usealla eri omiikkamenetelmällä tai useasta eri kudoksesta, tai samaa tautia voidaan tutkia ihmispotilaissa ja malliorganismissa. Omiikka-aineistojen analyysin suurin ongelma on se, että näytteiden määrä on usein pieni, vaikka muuttujien määrä on suuri.

Tässä väitöskirjassa on kehitetty bayesilaiseen tilastotieteeseen perustuva koneoppimismalli, jolla pystytään analysoimaan yhdestä tai useasta lähteestä tulevia havaintoaineistoja, joissa näytteet on kerätty monisuuntaisessa koeasetelmassa. Menetelmä pystyy löytämään aineistosta usean kovariaatin vaikutukset sekä niiden yhteisvaikutukset ja toimii hyvin myös, kun näytteiden määrä on pieni ja muuttujien määrä suuri. Koska menetelmä on bayesilainen, tulosten epävarmuus pystytään arvioimaan luotettavasti. Menetelmän soveltuvuusalue kattaa merkittävän osan modernin biolääketieteen tutkimuksessa syntyvistä omiikka-aineistoista.

Avainsanat bayesilaiset menetelmät, datalähteiden yhdistäminen, koneoppiminen, monisuuntaiset koeasetelmat

ISBN (painettu) 978-952-60-4782-9 **ISBN (pdf)** 978-952-60-4783-6

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

Julkaisupaikka Espoo **Painopaikka** Helsinki **Vuosi** 2012

Sivumäärä 181 **urn** http://urn.fi/URN:ISBN:978-952-60-4783-6

Preface

This work was done in the Adaptive Informatics Research Centre of the Laboratory of Computer and Information Science (Department of Information and Computer Science since 2008), Helsinki University of Technology (TKK), i.e., as of 2010 the Aalto University School of Science. The work has been supported by the Graduate School of Computer Science and Engineering of TKK, as well as by project funding from TEKES through the MultiBio research consortium and the MASI program. Helsinki Doctoral Programme in Computer Science - Advanced Computing and Intelligent Systems (Hecse) supported my participation to scientific conferences and workshops abroad during the thesis work. I later switched to The Graduate School in Computational Biology, Bioinformatics, and Biometry (ComBi), i.e., as of 2010 the Finnish Doctoral Programme in Computational Sciences (FICS) who also supported my participation to scientific conferences and workshops and my research visit abroad. Tekniikan edistämissäätiö (TES) supported my research visit abroad. I also had a pleasure belonging to the Helsinki Institute for Information Technology (HIIT).

I want to thank my supervisor professor Samuel Kaski for guiding me on my path of becoming a machine learning researcher. I joined the MI group as a physics student with little background in machine learning or bioinformatics, I just knew that was what I was interested in. I am very grateful that he gave me a chance, even when the beginning was difficult. I got guidance and a chance to work in an interdisciplinary environment, solving interesting and important problems together with top people. He gave me an excellent example of how to identify research questions, solve modelling problems, publish, give presentations and build collaborations. These years have given me a set of skills that will surely be extremely useful during my future career.

Developing machine learning methods might be difficult without data, and I want to express my gratitude to our most important collaborator, professor Matej Orešič for providing us with many state-of-the-art lipidomics datasets. Those datasets had fascinating experimental designs and biomedical research questions, which laid the ground for identifying the key data analysis problems that needed to be solved in order to develop the novel models.

Special thanks go to Tommi Suvitaival. He arrived at the team at the time I most desperately needed help for tackling the problems that felt overwhelming. With his excellent skills and ability to learn quickly, he was an amazing work pair and we indeed succeeded in publishing good papers.

When a new guy joins in a lab and starts a research project, he needs an instructor. I could not have had a better person than docent Janne Nikkilä to instruct me. He helped me to get started with my research, and we hit our heads to the wall together trying to use existing methods to solve the almost impossible data analysis task at hand. Finally, we got on the right track on what kind of new modelling would be required. When the moment arrived, he took the role of a mentor, giving me the initiative, which allowed me to take the responsibility of planning the details of our methods. The model blueprint I one day plotted on the whiteboard of his office became the essence of this dissertation.

During the years of collaboration with VTT, I attended numerous formal and informal meetings with Matej Orešič and his collaborators. I want to thank all the people I met in those meetings: I always felt I was playing a small part in something big. In particular, the last paper of this thesis was the real test for the validity of the model I had developed: application to a completely new dataset not used in developing the model. I want to thank professor Marja-Riitta Taskinen, docent Matti Jauhiainen and doctor Laxman Yetukuri for having set up the interesting medical research question that we managed to answer, and for the conversations we had during our meetings.

Thanks for useful comments to the pre-examiners of the thesis: professor Antti Penttinen and doctor Simon Rogers.

During my studies, I was surrounded by a wonderful group of young researchers: the MI group. Many thanks to all its current and former members for help and friendship; especial thanks to Arto Klami for help, and for Gayle Leen and my true peer Leo Lahti for many inspiring con-

versations and time spent together. Thanks also to all the people at the Department of Information and Computer Science. Nicolau Gonçalves and Elina Karp were the main stays in the office where I stayed for several years. During those years, I got several offers to change the room to a smaller and more quiet one, but I always declined.

During the last year, I've continued my journey deeper into the world of science at the Charles R. Bronfman Institute for Personalized Medicine, Mount Sinai School of Medicine, New York. I've had many great scientific conversations with great people and received useful comments for my thesis.

The doctoral studies are a long and special period in one's life, and during these years I experienced many happy and successful moments, but also some sad and difficult ones. During the good and the bad moments, a bunch of people were always there for me. Very special thanks to my mom Maria, and my dad Olli who sadly passed away before he could see the completion of my doctoral studies. They greatly supported all the different phases of my education on my way to become a researcher, starting from the childhood adventures exploring the chemistry set, microscope, telescope and the books 'Keksijän käsikirja' and 'Suomen Luonto'. Thanks also to my siblings Markku and Anne, all my friends, and my relatives in the Huopaniemi family. Alongside my parents, my godfather Timppa always greatly encouraged me to do doctoral studies. I'm glad I did. Finally, the last phase of my studies, writing the thesis and simultaneously preparing for and getting started with a new period of life abroad, was much busier and occasionally more stressful than I had imagined. I am grateful to my girlfriend Karoliina for her love, standing on my side and helping me during this time.

New York City, September 10, 2012,

Ilkka Huopaniemi

Contents

Preface 1

Contents 5

List of Publications 9

Author's Contribution 11

1. Introduction 13

1.1 Biomarker discovery 14

1.2 Multi-way experimental designs 17

1.3 Measuring and integrating multiple data sources 19

1.4 Importance of model organisms and translation 21

1.5 Problem definition and contributions of this dissertation . . 22

1.5.1 Problem setting: ANOVA-type modelling 22

1.5.2 Contributions of the dissertation in multi-way, multi-source modelling 23

1.5.3 The model 23

1.5.4 Applicability area of the model 24

1.5.5 Multi-way learning as a novel branch of machine learning 25

1.6 Organization of the thesis 25

2. "Omics data" 27

2.1 Central dogma 27

2.2 Integrating omics data 28

2.3 Metabolomics and lipidomics 29

2.4 Analysis of omics data 31

3. Bayesian statistics 33

3.1 Generative models 33
3.2 Bayesian learning paradigm 34
3.3 Hierarchical models 35
3.4 Bringing prior knowledge into model structures 35
3.5 Plate diagram notation 36
3.6 Gibbs sampling 36

4. Multivariate modelling of omics data **39**
4.1 Data and statistical challenges 40
4.2 Univariate analysis 41
4.3 Clustering 41
4.4 Unsupervised component models: PCA 42
4.5 Supervised models: Classification 43
4.5.1 Using classifiers for biomarker discovery 44
4.5.2 Sparse approaches and regularization 45
4.6 Testing known groups 46
4.7 Summary: modelling correlated groups of variables 47

5. Existing multi-way ANOVA-type models **49**
5.1 ANOVA and MANOVA 49
5.2 General linear model 50
5.3 Linear mixed models and time-series modelling 50
5.4 Problems of ANOVA and MANOVA 52
5.5 Multivariate many-step approaches 53
5.6 Multivariate Bayesian approaches 53
5.7 Summary: Our modelling approach compared to the existing approaches 54
5.8 Multi-way learning compared to other advanced machine learning genres 55
5.8.1 Multi-class classification 56
5.8.2 Multi-task learning 56
5.8.3 Multi-label prediction 57

6. Integration of multiple data sources **59**
6.1 Unsupervised data integration 60
6.2 Supervised data integration 61
6.3 Cross-species analysis and translation 62
6.3.1 Known matching between the variables 63
6.3.2 Unknown matching between the variables 65

7. The unified multi-way, multi-source model **67**
7.1 Single-source multi-way modelling 68
7.1.1 Relationship to PCA-approaches 70
7.1.2 Relationship to LMMs 70
7.2 Multiple data sources 70
7.2.1 The 'data source' as a covariate 71
7.2.2 Paired samples 73
7.2.3 No paired samples 73
7.3 Multi-level covariate 74
7.4 Covariate having partly unknown structure 75
7.5 Using the Bayesian posterior distribution to perform a statistical test 76
7.6 Repeated measures 76
7.7 Imperfect multi-way design 77
7.8 Biological prior knowledge of existing clusters 78
7.9 Model complexity selection 78
7.10 Summary 78

8. Future improvements **79**
8.1 Multimodality of the posterior distribution 79
8.2 Multiple components in CCA 79
8.3 Finding the optimal number of clusters 80
8.4 Modelling non-linearity of biological data 81
8.5 (M)ANCOVA-type modelling 81
8.6 Single-variable clusters 81

9. Conclusions **83**
9.1 Contribution to single-source multi-way modelling 83
9.2 Contribution to multi-source, multi-way modelling 84
9.3 On the results obtained 85
9.4 Multi-way learning 87

10. Discussion **89**
10.1 Applicability of the model to other data types 89
10.2 Future use of unified Bayesian multi-way models 90

Bibliography **91**

Publications **105**

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 Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Matej Orešič, and Samuel Kaski. Two-way analysis of high-dimensional collinear data. *Data Mining and Knowledge Discovery*, 19(2):261-276, June 2009.

II Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Matej Orešič, and Samuel Kaski. Multivariate multi-way analysis of multi-source data. *Bioinformatics*, 26:i391-i398, July 2010.

III Ilkka Huopaniemi, Tommi Suvitaival, Matej Orešič, and Samuel Kaski. Graphical multi-way models. *In Jose Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag editors, Machine Learning and Knowledge Discovery in Databases - ECML PKDD 2010, volume 6321 of Lecture Notes in Computer Science*, pages 538-553. Springer-Verlag, Berlin / Heidelberg, September 2010.

IV Tommi Suvitaival, Ilkka Huopaniemi, Matej Orešič, and Samuel Kaski. Cross-species translation of multi-way biomarkers. *In Timo Honkela, Wlodzislaw Duch, Mark Girolami, and Samuel Kaski, editors, Artificial Neural Networks and Machine Learning - ICANN 2011, volume 6791 of Lecture Notes in Computer Science*, pages 209-216. Springer Berlin / Heidelberg, June 2011.

V Laxman Yetukuri, Ilkka Huopaniemi, Artturi Koivuniemi, Marianna Maranghi, Anne Hiukka, Heli Nygren, Samuel Kaski, Marja-Riitta Taskinen, Ilpo Vattulainen, Matti Jauhiainen, and Matej Orešič. High density lipoprotein structural changes and drug response in lipidomic profiles following the long-term fenofibrate therapy in the FIELD substudy. *PLoS ONE*, 6(8):e23589, July 2011.

Author's Contribution

Publication I: "Two-way analysis of high-dimensional collinear data"

The ideas in the publications have been a result of teamwork. The models in each Publication I-V have been designed together, the author having the initiative, being supervised by Prof. Kaski, and getting useful feedback from the other authors. The author is responsible for most modelling details in each publication, including deriving the formulas, deriving Gibbs sampling equations, designing the dimensionality reduction scheme, and designing the statistical testing scheme.

In Publication I, a novel Bayesian model for multi-way modelling of single-source omics datasets is presented. The author is also responsible for writing the first version of the manuscript and has a shared contribution in implementing the model.

Publication II: "Multivariate multi-way analysis of multi-source data"

A novel Bayesian model for multi-way modelling of multi-source omics datasets is presented. The author is responsible for most modelling details, writing the first version of the manuscript and has a shared contribution in implementing the model and in running the experiments.

Publication III: "Graphical multi-way models"

An overall unified model is presented, binding three multi-way modelling tasks: the multi-way, multi-source model from Publication II, and two novel multi-way models. One is for multi-way, single-source modelling,

one covariate having an unknown structure. The other is for translating biomarkers between multi-species experiments having a multi-way design. The author is responsible for most modelling details and writing the first version of the manuscript

Publication IV: "Cross-species translation of multi-way biomarkers"

A refined and extended algorithm for translating biomarkers between multi-species experiments having a multi-way design is presented, following the idea presented in Publication III. The author is responsible for most modelling details and participated in writing the manuscript.

Publication V: "High density lipoprotein structural changes and drug response in lipidomic profiles following the long-term fenofibrate therapy in the FIELD substudy"

Publication V is a shared biomedical work based on a drug intervention study called FIELD. The manuscript combines lipidomics analysis for FIELD patients, basic statistical analysis, advanced data analysis by multi-way modelling and molecular dynamics simulations of lipids in HDL particles. The lipidomic dataset has a multi-way experimental design, and the model presented in Publication I was applied to the dataset, additionally requiring further extensions. The author is responsible for the advanced multi-way statistical modelling part of the manuscript by planning and implementing the extensions, running the experiments, and writing the first version of the sections of the manuscript that were related to multi-way modelling.

1. Introduction

"Multi-scale data integration, including genomic, expression, metabolite, protein, and clinical information, will ultimately define the future of patient care" - (Eric Schadt, 2011)

Using machine learning to solve bioinformatics data analysis problems

Mankind has a challenge: In most scientific fields, the amount of measured information has exploded due to rapid technological advances and rapidly decreasing costs in measurement techniques. The question of how to convert data into knowledge has become a central issue, increasing remarkably the importance of novel scientific fields such as machine learning and data mining.

Modern biomedical research (alongside data coming from the Internet) is perhaps the clearest example of this development. Advances in molecular profiling and DNA sequencing techniques have enabled high-throughput measurements; simultaneous profiling of many or all of the genes or molecules of a given type. As a result, the amount of data has increased dramatically and measurement results can no longer be interpreted by visual inspection alone but require advanced computational data analysis methods. To tackle this challenge, the increasingly prominent field of bioinformatics has become a necessary component of biomedical research. Because of emerging biomedical research opportunities, new methodological research questions appear rapidly, maintaining bioinformatics in a state of constant change. In the analysis of biomedical data, one continually encounters situations where existing data-analysis methods are not applicable to the novel problem to be solved, and novel methods need to

be developed.

Machine learning is a scientific field that deals with developing algorithms that allow computers to learn from observed data. Machine learning overlaps with statistics and computer science but is its own scientific discipline with a large and active research community. Machine learning is useful in solving difficult, very high-dimensional data analysis problems that require novel modelling approaches in order to provide a meaningful dimensionality reduction and interpretation of the data. Machine learning algorithms take advantage of computers and computational power, which has fortunately greatly increased simultaneously with the amount of available data.

One of the widest and most important research questions common to biomedical research, bioinformatics and other machine learning application areas, is how to integrate information from multiple data sources instead of using a single data source. Data integration is a very active and interesting research question for methodological machine learning research as well.

This dissertation falls in the areas of machine learning and bioinformatics, more specifically in the areas of generative modelling and Bayesian methods. In a part of this dissertation, I also deal with integration of multiple data sources also known as multi-view learning. In this dissertation, Bayesian machine learning is used to tackle one of the most important questions in statistical data analysis, that is, multi-way modelling; finding the effects of multiple covariates and their interactions in the data. The focus is on high-dimensional, small sample-size single and multi-source data.

1.1 Biomarker discovery

The traditional avenue of biomedical research, stemming from the Mendelian era, has been the study of associations of genome, particularly gene alleles, to observable traits such as diseases. Whereas the genome represents a static blueprint of an individual, it has been long known that human phenotype, the overall physiological state of an individual at a given moment, results from both the genotype and environmental factors. The range of environmental factors contributing to an individual's phenotype is huge, consisting of the current state and the accumulation of environmental influence over a lifetime. Examples of major accumulating

factors in human phenotype are lifestyle, diet, smoking, drinking, aging, and physical activity [1, 2, 3, 4, 5, 6]. On the other hand, factors that characterize the current phenotypic state are: whether an individual is suffering from a disease, whether he is using medication, Body Mass Index (BMI), and age, for instance.

The overall aim of medicine is to improve the health and well-being of people and the purpose of biomedical research is to increase the possibilities and efficiency of medical care. One current direction of medical research is towards preventive health care. The phases of medical treatment can be divided into three categories: (i) treat the symptoms, (ii) cure the disease, and (iii) prevent the disease onset. One cornerstone of the concept of **personalized medicine** [1, 7], meaning customized medical care based on biological and other data, is to treat the disease risk [1, 8] before the disease develops.

Although the evaluation of the patient is possible for a physician from clinical examination and symptoms described by the patient, a standard practice of clinical medicine has long been to use laboratory tests: collect blood samples and measure the concentrations of certain molecules known to be associated with the onset or activity of a suspected disease. These known molecules, called **biomarkers**, can substantially help in diagnosing the disease and deciding on a treatment. Widely known examples of biomarkers are elevated concentration of glucose as a consequence of diabetes [9], C-reactive protein (CRP) [10] as a marker for inflammation and cancer markers [11].

There is an ever-growing continuous effort in the biomedical scientific community to find novel and more accurate biomarkers to characterize disease status or disease activity. Biomarker discovery has been traditionally limited to testing the validity of one (or a few) suspected candidate biomarker(s) at a time due to constraints in laboratory techniques. The advent of high-throughput profiling methods, such as transcriptomics (gene expression), proteomics and metabolomics, has revolutionized the potential for biomarker discovery. The word transcriptomics refers to the messenger RNA (mRNA) molecules, proteomics to proteins, and metabolomics to metabolites. High-throughput profiling has enabled researchers to move from targeted profiling of a few candidate molecules to simultaneous measurement of the concentrations of a very large number of molecules of a given type. The number of metabolites (such as lipids) detected can be hundreds, and the number of transcripts (mRNA) tens of

thousands, representing the whole profile of activity of the chosen 'omic' molecule type in the organism. In this dissertation, I will use the word *omics* to describe transcriptomics, proteomics and metabolomics (lipidomics); they all result in similar concentration-type, continuous-valued data. Any omics profile, such as transcriptomic [11, 12], metabolic [13], or integrated profile [2], is a fingerprint of the dynamical phenotypic status of an individual at a given time, and it contains valuable information of the health status of the individual. The large number of molecules and their concentrations, represented as activities of an omics profile, contain a potentially unlimited supply of candidate biomarkers for detecting disease-related information.

An even more ambitious goal, compared to standard disease diagnostics, is to search for **early biomarkers for disease** [8, 11, 14, 15]. Such biomarkers could alert physicians to the possibly disease-causing malfunction in the physiological state of a patient before the actual disease onset. The promise of early biomarkers may ultimately make it possible to design therapies that help prevent the onset of the disease.

A third field where biomarker discovery is active is pharmacogenomics [7], where biomarkers are used to predict how well a patient will respond to a drug treatment and whether the drug has adverse effects [7, 16, 17]. Although the main interest in pharmacogenomics has previously been in using genotypic biomarkers for prediction [7, 11, 18], dynamic transcriptomics [11, 19, 20, 21] and metabolomics [22] biomarkers are starting to prove useful as well.

To test whether a candidate biomarker is associated with a disease, one needs to design an experiment and do statistical testing on the collected data. The standard approach is a comparative study: to collect a population of diseased samples (case) and healthy samples (control) and study differential concentration or expression of the molecules by statistical tests. Elevated concentrations in the case population compared to the control population are called up-regulations, and lower concentrations in the case population are called down-regulations. Another measure used in biomarker discovery, not considered further in this dissertation, is the ratio of concentrations of two molecules [11]. Any discovered statistically significant difference in the concentration of a molecule is a potential biomarker for disease, although a plethora of problems exist [8, 11]. Careful validation steps, including repeated experiments in other laboratories and clinical validation, are usually required for a biomarker to be accepted

[23] for clinical use.

From a wider viewpoint beyond clinical biomarker discovery, the key interest of biomedical research is understanding how biological systems function. The same statistical principles of comparative experimental designs are commonly used in many biomedical studies. A common approach is perturbing a system by an intervention or by knocking out a particular gene [4, 24], and examining how the behavior of the system is altered compared to a normal, healthy system. The observed differences in the behavior can help to gain insight into the physiology of the system. These studies usually lead to experimental designs and need for data analysis similar to the ones in biomarker discovery.

The most serious fundamental difficulty in the data analysis of modern biomedical studies is that, whereas the number of variables (molecules) is large, the number of samples (patients) is often small. This may be due to economical or ethical reasons or simply because of a small number of patients with a given condition being available. Whereas the large number of profiled molecules offers a huge potential for discovering biomarkers and studying biological systems, the small number of samples represents fundamental problems [25, 26] for the statistical methodology used in simultaneously testing a large number of variables for disease associations. The scientific community is currently in the process of searching for feasible computational approaches to deal with the "small n, large p"-conditions: conditions where the assumptions, into which a century of work in traditional multivariate statistics has leaned, do not hold.

There are often two additional key problems that need to be addressed in the analysis of biomedical data: multi-way experimental designs, stemming from there usually being multiple covariates in the experiment in addition to the case-control comparison; and data integration, that is, analysis of measurements of multiple types. There exist no standard, widely accepted data analysis methodologies for multivariate analysis of high-dimensional data in the case of multiple covariates, especially when the data come from multiple data sources.

1.2 Multi-way experimental designs

The conceptualization of an experiment that looks for disease biomarkers by comparing case and control populations is, in many cases, an oversimplification. There are two reasons that may account for this: (i) in

biomedical experiments, there are usually confounding factors that may bias the results unless dealt with properly, and (ii), in many experiments, the effects of multiple covariates in the data, and the effects of their interactions, are often of particular interest to study.

In statistics, a covariate is a variable that is potentially predictive of the response variable(s). In this dissertation, variables that annotate individuals (disease status, gender, treatment) are covariates and the omics data are the response variables.

In some multi-way experimental designs, there is a clear distinction between the covariates of interest (such as disease status) and confounding factors. Confounding factors are covariates that are correlated both with the covariate of interest and the response variables. In human studies looking for disease associations, for instance, gender, age, BMI, drug treatment, or race can have large effects on the concentrations of molecules. In the analysis of omics data, it is important to try to take into account all such potential confounding factors in order to obtain unbiased disease associations. A basic approach is to stratify the data analysis problem into smaller parts, such as comparing healthy and diseased males and females in different age groups separately (as an example, see [9, 27]). The side effect of stratifying the analysis is that it leads to an even smaller number of samples available in each sub-population [28], worsening the "small n, large p"-problem. Additionally, interpreting the results from multiple separate analyses is more tedious. Instead, it would be advantageous to formulate the data analysis problem as a multi-way experimental design to be able to estimate the effects of all the covariates and their interactions in the data jointly.

In many experiments, the main research question is to study the effects of all the relevant covariates and their interactions. One of the most common biomedical experiment types is studying the effects of drug, diet, or other interventions [9, 29, 30, 31, 32]. Other examples are studying the effects of a gene knock-out [29] or the effects of gender and age as additional risk factors [29]. Treatment groups and other relevant descriptors are common covariates in experimental designs. For instance, in the designed diseased-healthy drug intervention multi-way experiment of Publication II, a drug effect may be interpreted as a direct drug side effect; the interaction of drug and disease is actually the effect of interest, indicating whether the drug cures the disease. As another example, patients of different genders have been shown to have differential omics profiles

[27, 33]. The interaction effect of disease and gender may be interesting for determining whether different genders have differential disease effects or differential responses to drugs.

One aim of personalized medicine is to move away from the broad definitions of a disease in a large population to defining disease subtypes on groups of similar patients. The aim is personalized treatment instead of the model that "one treatment fits all patients equally". Patient populations will therefore be increasingly looked at as subgroups defined by multiple covariates.

Experimental designs are often categorized into designed and observational studies. Designed study refers to a highly controlled study, for example a laboratory study, where patients are randomized into treatment groups. Observational study refers to studies where the assignment of patients into treatment groups is outside the control of the investigator; an example is human data accumulating from everyday clinical practice.

In summary, many clinical and biomedical settings have multiple covariates that are either confounding or interesting and coming from either a designed or an observational study. This type of data analysis problem can be formulated as having a multi-way experimental design, where observations (samples or individuals) have been divided into populations of measurements according to multiple covariates. The data analysis can then be done by multi-way modelling, which is the topic of this dissertation.

1.3 Measuring and integrating multiple data sources

There is a growing trend in biomedical research to measure multiple omics data sources from each individual, because different data sources are biologically complementary and the cost of the experiments are becoming increasingly tolerable. Such multi-source omics data come from two types of experiments: 1) measuring multiple different omics types and 2) collecting measurements from multiple tissues. Doing the multi-source measurements is relatively cheap and easy, whereas analyzing the data is more challenging. There are no widely accepted methods for integrating multiple data sources if the analysis is to be taken beyond a standard study of association between variables and a disease outcome.

It is widely believed [34, 35, 36] that the future of medicine is in the integration of genomic, transcriptomic, proteomic, and metabolic data with

clinical data. It is therefore increasingly common to measure multiple or all the omics data types: gene expression, proteomics and metabolomics from each individual [2]. It has been realized that more relevant information of biological phenomena can be gained from a combination of all the data types than from a single data type and that different omics types are biologically complementary. These types of datasets are often accompanied by high-throughput genetic sequence data and a varying amount of clinical information.

Another line of biomedical research leading to multiple data sources from each individual, is taking measurements from multiple tissues. There are two types of applications for multi-tissue experiments: 1) the practical application of predicting the state of a disease tissue from an easily collectible tissue, such as blood plasma, and 2) a more general physiological interest of studying disease-related relationships between multiple tissues.

In the center of most studies of diseases, there is a disease tissue of interest, such as pancreas in diabetes, kidney in chronic kidney disease, lung in lung cancer [29], or in general any cancer tissue [11, 36, 37]. The concentrations of molecules in target tissues and organs usually contain the most relevant information concerning the disease. However, taking tissue biopsies is difficult due to their invasive nature and it is doubtful that they could be brought into standard clinical practice [11, 31]. As discussed earlier, the aim of medical practice is to use the least invasive methods [2] such as blood, urine [9], fecal samples, even breath [38] for biomarker detection. Although the question of molecules of an easily collectible body fluid (blood) being associated with a disease is relevant for diagnostic purposes [39], the deeper underlying question is whether the blood molecules carry information of the target disease tissue [8, 40, 41]. The relevant research question is whether there are shared disease-related effects between blood and tissue molecules [41]. It would be of great diagnostic and scientific interest to determine reliably which blood plasma molecules are correlated with target tissue molecules and whether these associations carry disease-related information.

Another aim of multi-tissue experiments is to obtain information from multiple tissues in order to have a more holistic view [42] of the physiological or biological state of the individual. Such experiments can be used to find out which tissues [4, 24, 43] show disease-dependent changes and which do not, or whether there are disease-related dependencies between

the tissues. These types of experiments also include studying relationships between different omics and other data sources from multiple easily accessible sample types, such as between blood and gut microbiota measured from fecal samples [43, 44, 45, 46, 47] or between blood lipidomics and lipoprotein compositions [48].

From a statistical perspective, all the presented data integration tasks involving multiple omics data types, multiple tissues, multiple non-invasive sources, or any of their combination, are identical. They can all be formulated as data integration of multiple data sources with paired samples and different variables in different data sources. The pairing here means that multiple data sources have been measured from the same patient. Different omics types clearly have different variables (mRNA, proteins, metabolites) and we also assume that different tissues, in general, have different variables even if the omics type is the same. Some of the molecules in different tissues may be chemically identical to each other, but they may have different roles in different tissues. Because of the widespread use and great number of applications of this type of data integration, there is an endless demand for suitable computational methodologies. A practical point of view is that, as the number of data sources and relevant covariates grows, combined with the already overwhelming dimensionality of the data, there will be an increasing demand for more compact representations of the relevant findings in the data.

In this thesis, a novel computational approach for data integration is presented: integration of multiple data sources with paired samples in the context of an underlying multi-way experimental design. The underlying assumption to be studied is whether there is a dependence between the data sources that is associated with one or multiple covariates and their interactions.

1.4 Importance of model organisms and translation

Since improving the health of individuals is the ultimate goal of biomedical research, the information gained from clinical studies done on human patients is of primary importance. However, model organisms are often used in biomedical research as disease models [9, 49] and the effects of drugs are often tested on model organisms in pre-clinical phases [50] before proceeding to human studies. Furthermore, since biopsies are usually difficult or impossible to obtain from humans [31], multi-tissue studies are

mostly limited to model organisms. Most of the data used in this thesis are from human studies, but the multi-tissue dataset in Publications II and III is from a mouse study.

A particularly relevant research question is whether disease and treatment related findings found in a model organism actually have a correspondence in human clinical studies [9, 29, 49]. As different species do not, in general, fully share the same lipids and proteins, and chemically similar biomolecules may have different roles in different species [51], a new statistical modelling problem emerges: how to translate findings made in a model organism into human clinical studies. The task is to find whether there are molecules that behave similarly in response to disease and other covariates and their interactions in multiple species, for instance, human and mouse [49].

The possibilities of studying omics data are not limited to humans and mammals, but experiments are often done *in vitro* in cultured cells or cell lines [32], stem cells [52], plants [53] and yeast [54].

The methodology developed in this thesis covers the modelling problem of how to translate biomarker discovery results between model organisms and human studies. From a statistical perspective, the data analysis problem is to integrate multiple data sources when the samples have not been paired. This data integration task is also solved in the context of both datasets (species) having a similar multi-way experimental design. For instance, the datasets from both species can consist of healthy controls and individuals with the same disease [9], a similar drug treatment design or a time-series.

1.5 Problem definition and contributions of this dissertation

The contribution of this dissertation is to present a novel, computational, Bayesian model for multivariate multi-way ANOVA-type modelling of continuous-valued, high-dimensional, single-source and multi-source data. This model can be used even when the number of samples is small.

1.5.1 Problem setting: ANOVA-type modelling

The omics data are continuous-valued, high-dimensional data where variables represent absolute or relative concentrations of the molecules. Each sample is associated with multiple discrete covariates, such as disease

status and gender. Multi-way experimental design here means that the samples can be organized to populations according to the levels of the multiple covariates: for instance, diseased males are one population.

Multi-way ANOVA-type analysis is a well-established task in classical statistics for continuous-valued, univariate data, in order to model the effects of multiple covariates and their interactions in the data. The task has been traditionally solved by multi-way Analysis of Variance (ANOVA). ANOVA is a statistical test for determining whether the mean value of a variable is different in multiple populations of measurements. The multi-way modelling problem becomes much more complicated for high-dimensional, small sample-size data, and there are few previous approaches in these conditions. In particular, ANOVA -type modelling has not been previously studied in the context of multiple data sources.

1.5.2 Contributions of the dissertation in multi-way, multi-source modelling

A solution for multi-way modelling of high-dimensional, small sample-size data in the case of a single data source is provided in Publication I. This is followed by defining and solving how to do multi-way ANOVA-type analysis for multi-source data, when different data sources have different variables. The presented methodology covers both the case of multiple data sources with paired samples (that is, measurements of different data sources taken from the same individual) and the case without paired samples that arises from translating potential multi-way biomarkers between species. The case of paired samples is introduced in Publication II and revisited in Publication III. The case without paired samples is introduced in Publication III; a more advanced sampling algorithm is introduced in Publication IV. Another theoretical extension of multi-way modelling is one of the covariates having a previously unknown structure, introduced in publication III. In Publication V, the single-source method is applied to a new lipidomic study with an imperfect multi-way experimental design and a repeated measures setup.

1.5.3 The model

The achievement of this dissertation is that all the closely-related multi-way modelling settings can be handled with a unified Bayesian model. The different setups need a slightly different structuring, but in each case

a variant of the unified multi-way, multi-source model can be constructed. It is shown that, although the combination of multiple data sources and multiple covariates is very complicated, it is possible to define an overall generative model that can be assumed to have generated all the observed data. The model is interpretable and the desired statistical testing results are directly obtained from the parameters of the model. In contrast to joint multi-way, multi-source modelling, existing statistical methods can only solve simpler tasks. The advantage of an overall generative model is that model parameters of the whole model can be learned jointly, which improves uncertainty estimation of the model. A careful uncertainty estimation is crucial when the number of samples is small.

The unified model has an integrated dimensionality reduction that is based on the biologically relevant assumption that there are groups of correlated variables. The effects of covariates are modelled on these groups instead of single variables.

1.5.4 Applicability area of the model

The applicability area of the model covers most of the multi-way, single-source and multi-source experimental designs encountered in today's biomedical research that is focused on omics data. Relevant experiment types include clinical biomarker discovery, studying organisms in response to interventions or perturbations, studies on model organisms, and translational studies. Any relevant covariates present in these types of experiments, being confounders or covariates of interest, can be included as long as they are discrete or discretized. Also, any data integration task where different omics profiles are obtained from the same individual, being from different tissues or different omics types, can be analyzed with this model. The translation model is usable in multi-species cases with different data sets (species) having a similar multi-way experimental design and different variables in different species. Modelling similarities of multiple diseases is another possible application of the translation model.

The model can also deal with time-series datasets, if time-point can be seen as one of the covariates in the multi-way design. Three time-series modelling cases are possible: (i) time can be considered as a standard covariate, as in Publication I, (ii) irregular measurement times can be aligned into latent states [Publication III and IV], and (iii) repeated measures designs can be formulated by having an 'individual effect' as an additional covariate [Publication V].

The common factor to all the datasets studied in this dissertation is that they are lipidomics datasets with a multi-way experimental design. The datasets are either from single-source or multi-source settings; some datasets are time-series. The multi-source setups have either paired samples or no paired samples. The developed Bayesian model is applicable to all these multi-way modelling cases.

1.5.5 Multi-way learning as a novel branch of machine learning

As one theoretical contribution of this dissertation, multi-way learning is defined as a branch of machine learning. In the machine learning community, learning of the association of data to an external covariate has been mostly restricted to supervised learning: regression or classification. There are currently three popular advanced supervised approaches that are closely related to multi-way learning: multi-task learning, multi-label prediction, and multi-class classification. I will argue that multi-way learning is a different learning task that has remained mostly un-tackled in the machine learning literature, despite the importance of ANOVA and related methods in classical statistics.

1.6 Organization of the thesis

The remaining chapters are organized as follows. In Chapter 2, I present a biological review of different omics data types as parts of the biological information chain and discuss integration of these data types. In Chapter 3, I present the Bayesian learning paradigm and justify why it was chosen as the modelling approach. In Chapter 4, I present a review of the standard multivariate modelling approaches that are commonly used in modelling omics data, and discuss their inadequacies in modelling high-dimensional, small sample-size data having a multi-way experimental design. In Chapter 5, I present the existing multi-way modelling approaches to conclude that none of the existing methods perfectly fits to the multivariate case. In Chapter 6, I present relevant existing approaches that can be used for integrating multiple data sources and conclude that none of the existing approaches can fully address multi-way experimental designs. In Chapter 7, I present our Bayesian model for multivariate multi-way modelling of single-source and multi-source data, which is followed by a discussion of possible future improvements in Chapter 8. Finally, I

present the main conclusions of this dissertation in Chapter 9 and a discussion of the applicability of this work in Chapter 10.

2. “Omics data”

The topic of this dissertation is how to analyze and integrate data from omics experiments: transcriptomics, proteomics, metabolomics, and lipidomics. In this Chapter, a brief introduction to the biological background of these molecular types is presented. The Chapter introduces the central dogma of biology that illustrates the relationship of the different molecular types in order to explain the need of integrating the different omics data types. This is followed by a section describing lipidomics, a part of metabolomics, given that lipidomics datasets were used as case studies in this dissertation. The origin of correlations between metabolites or lipids is also discussed to justify why modelling correlated groups of variables was chosen as the main assumption of modelling omics data. Finally, the different phases of the data analysis pipeline from preprocessing to statistical analysis are presented.

2.1 Central dogma

The central dogma of biology states that genetic information contained within the DNA sequence of genes is transcribed into messenger RNA (mRNA) molecules, and this information is further translated into the amino acid chain of a certain protein. Proteins catalyze chemical reactions, where metabolites are converted from one to another, which maintains metabolism and homeostasis (stable cellular conditions).

The static genotypic information flows into the phenotype dynamically, influenced by the environment. A variety of signaling and regulatory mechanisms control the dynamic functioning of cells and in fact the whole organism in response to environmental factors, developmental factors, or simply for maintaining homeostasis. The genes need to be expressed to maintain a desired metabolism, which is the endpoint [11] of the biologi-

cal information flow.

Recent research has suggested, however, that the information contained in the DNA sequence may alone be insufficient in predicting human disease phenotype, rendering the classical central dogma inadequate. This suggests that environmental influences may actually be inherited through mechanisms such as epigenomics (reversible, possibly heritable modifications in DNA without altering DNA sequence) and gut microbiome (intestinal bacteria) compositions.

2.2 Integrating omics data

One of the most important focus areas of biomedical research has become how to integrate the different phases of the biological information chain, in other words the different omics types, to obtain a comprehensive view of the dynamic state of an organism [42]. It has been realized that studying any single molecular type alone is a reductionist approach and insufficient for understanding the functioning of the organism. There are enormous research opportunities as each step of the information chain, from DNA to RNA (gene expression), and further to proteins (proteomics) and metabolism (metabolomics), can nowadays be measured in a high-throughput manner to deliver a profile of a large number of its molecules.

The relationships of different omics data types and relationships of omics data with other relevant biological data types have been widely studied. The case that has attracted the most interest has been the study of functional relationships of the DNA sequence and gene expression data. One common approach is studying whether Single Nucleotide Polymorphisms (SNP), that is, variations of a single nucleotide between individuals, have a regulatory role on gene expression; such SNPs are called Expression Quantitative Trait Loci (EQTL) [55, 56, 57, 58]. The relationship of DNA and gene expression can also be determined by studying whether gene Copy Number Variations (CNV) (duplications of sequences of DNA in the genome) influence gene expression [59, 60]. Numerous studies have also been carried out on other data integration options. For example, a common approach is to study associations between the concentrations of two omics types, such as proteomics and metabolomics [61] or transcriptomics and proteomics [62]. Another approach is to study whether there are similar or negatively-correlated responses to drug treatments or other covariates in transcriptomics and metabolomics [24, 63], or metabolomics and

gut microbiota [43, 46, 47], for instance. The notion of EQTL has also been extended to studying whether SNPs affect the concentrations of metabolites [64] or lipids [65, 66]. An example of a more biologically focused approach is studying the relationships between proteins and metabolites in the context of proteins as catalysts in a metabolic network model to predict metabolic flux [67]. Some methods have been developed for integrating more than two data types [2, 42, 54, 68, 69] or studying the relationships of omics data to clinical data [70, 71, 72, 73].

In summary, the biological or mathematical definitions for what integration means, vary a lot depending on the research question and the experimental settings. The common approaches for continuous-valued omics data are searching for a similar response to a single covariate or associations between the molecular profiles of different omics types. The contribution of this dissertation to the biological data integration field is to present a formal model for integrating continuous-valued omics data types in the context of an underlying multi-way experimental design. In this work, I provide a novel definition to justify the data integration: if a similar response to multiple covariates and their interactions is found in multiple omics datasets, there is a connection between these omics types that is related to the disease or other covariates. The mathematical formulation will be presented in Section 7.2.

2.3 Metabolomics and lipidomics

Lipidomics is an important class of metabolomics. Lipidome refers to the complete set of lipids in a cell or a tissue, whereas metabolome refers to the complete set of metabolites.

The key physiological role of metabolism is to convert nutrients to energy in order to maintain cellular functions and, more generally, homeostasis. Despite the unifying term metabolomics, the range of chemical properties of the different metabolite families is large and consists of small amino acids, lipids, bile acids, and keto-acids [3, 74]. There is no currently existing experimental technique that can measure all of them simultaneously. It is common to concentrate on high-throughput measurements in one of the metabolite classes, such as lipids. The most common measurement techniques in metabolomics are Nuclear Magnetic Resonance (NMR), Gas Chromatography mass spectrometry (GC/MS) and Liquid Chromatography / Mass Spectrometry (LC/MS).

Lipids are a diverse class of metabolites with important roles as building blocks of cellular membranes, energy storage and cell signaling [3]. A broad definition of lipids is that they are hydrophobic or amphiphilic molecules that originate from ketoacyl or isoprene building blocks [75]. The LC/MS is the main experimental technique for measuring a lipidomic profile from the blood plasma or other tissues.

A particular characteristic of metabolomics and lipidomics is that concentrations of metabolites are highly correlated and there are correlated groups of metabolites [3, 5, 76]. Maintaining a certain metabolite and lipid composition, particularly their concentrations, is crucial for the correct functioning of an organism. There are biochemical networks composed of interconnected pathways where metabolites are converted to others, which results in concentrations between metabolites being correlated [76]. Another special characteristic to be taken into account in modelling is that abundances of metabolites vary by orders of magnitude [76, 77], however, groups of metabolites with very different abundances are nonetheless often highly correlated and respond similarly to external covariates.

A key concept in lipidomics is allostasis and allostatic responses [3]. In contrast to homeostasis, which refers to maintaining stable internal conditions in an organism, allostasis refers to maintaining homeostasis by making a physiological or behavioral change. When a pathological metabolic state develops, a biological stress disturbs the normal homeostatic mechanisms of the cell. This disturbance is compensated by an allostatic response, such as activation of alternative pathways. An allostatic response [3] can be a result of a malfunction in the biological system and it can eventually lead to disease onset. Allostatic load refers to the stress caused by the activation of the alternative pathways. When the organism can no longer tolerate the harmful allostatic load, the failure in the biological system leads to the onset of a disease. Because of the allostatic response, the original change in lipid metabolism does not necessarily show up in lipidomics data; one rather detects a secondary change (side effect), which is a result of an allostatic mechanism compensating for the original change.

In summary, entire pathways are usually affected by environmental changes and by the resulting allostatic responses leading to up- or down-regulations of entire pathways. Many publications [30, 43, 46, 78] report having found that groups of lipids were found up- or down-regulated to-

gether; lipids within a group from the same lipid family, such as fatty acids, triglycerides, sphingomyelins or phosphaticylcholins. To exploit the knowledge of the existence of correlation structures, I argue that especially for lipidomics, studying groups of similarly behaving, correlated clusters of variables (lipids) is the most biologically justified modelling approach. This approach enables the direct detection of up- or down-regulation of groups of variables as a response to covariates and therefore, this was taken as the main assumption of the dimensionality reduction part of our model.

The dimensionality reduction part of our model was indeed first developed [Publication I] to be feasible especially for metabolomics and lipidomics data that we used in our case studies. However, preliminary experiments on gene expression data [data not shown] have shown that correlated groups of variables is also a good assumption for gene expression. We assume this holds true also for proteomics.

2.4 Analysis of omics data

In this section, I discuss the different phases of the data analysis chain of omics data: preprocessing and modelling [5]. Within this data analysis chain, my work concentrates on exploratory statistical modelling of preprocessed omics data.

When the measurements from blood or tissue samples have been done using for example MS or NMR methods in metabolomics, or by RNA micro-arrays or RNA-sequencing in transcriptomics, the data have to be analyzed. During the preprocessing phase, raw data from the measurement device are processed into a data matrix where rows are samples (individual patients) and columns are variables (absolute or relative concentrations of the molecules of a chosen omics type). Preprocessing starts by converting raw signals from the measurement device into intensities of the molecules. In MS methods, this is done by detecting and identifying signal peaks from raw data. Other preprocessing procedures include normalization and removal of experimental artifacts and systematic biases. The lipidomic datasets used in this dissertation had been preprocessed using the MZmine software [79].

The modelling step attempts to answer the central research question of the study, for instance to determine the effect of a disease. The two main lines of modelling are [5] statistical modelling and more biologically-

focused modelling. In statistical modelling, the aim is to find associations between and within omics variables and clinical variables (covariates). Exploratory statistical analysis refers to describing and summarizing main characteristics of the data in an easily understandable form, having emphasis on novel interesting findings and hypothesis generation. Confirmatory statistical analysis refers to testing and confirming pre-defined hypothesis. Examples of more biologically-focused modelling tasks are pathway analysis [5], metabolic flux analysis [80], deciphering gene regulatory networks, and searching for transcription factor binding sites.

3. Bayesian statistics

The Bayesian formalism has been chosen as the modelling paradigm for the methods presented in this dissertation. The main characteristics of the data are small number of samples, high dimensionality and a need for radical dimensionality reduction. For the methods to work, the dimensionality reduction scheme has to be able to take into account the characteristics of the data. The problem setup, including multiple data sources and covariates, is complicated. The possibility of bringing prior knowledge flexibly into the hierarchical model structure is the key to solving these challenges and modelling the whole setup jointly as a single unified model with joint uncertainty estimation.

In this chapter, I introduce the Bayesian learning paradigm and explain how its advantages to our work stem from its basic principles. I first introduce the advantages of building generative models and describe how model parameters can be learned by Bayesian inference. Then I present the concept of hierarchical models to illustrate their usefulness in our modelling task, which is followed by an introduction to plate diagram notation that is used to visualize model structures. Finally, I discuss Gibbs sampling, an approximate inference method that was used in learning the model.

3.1 Generative models

An important class of statistical models are generative probabilistic models where the main idea is to assume a statistical model that has generated the observed data. The model parameters can then be learned from the observed data, which enables the interpretation of the data by identifying the generative process. The model parameters are often learned by maximum likelihood learning or by full Bayesian inference (details be-

low).

Factor Analysis (FA) is one of the most widely known statistical multivariate methods, based on an underlying generative model. The key idea is that a few underlying factors (and a noise model) generate the variation in the observed high-dimensional data. When learning the model, the model parameters of interest are the ones that model the structure of the factors because they indicate which variables function together. Factor Analysis is closely related to Principal Component Analysis (PCA) [81] but, as FA is based on a true generative statistical model, it can be used as a building block in a unified hierarchical model, as is done in Chapter 7.

Generative models differ from the discriminative models used in supervised learning (regression, classification, see Chapter 4.5) in the sense that they assume a model that has generated the observed data. Discriminative models, in contrast, attempt to learn only a model that explains the class label (classification) or continuous outcome variable (regression), given the observed data.

When a probabilistic generative model is assumed in modelling, its parameters can be learned in the classical (non-Bayesian) paradigm by various optimization algorithms that lead into a Maximum Likelihood (ML) estimate. Maximum likelihood means the most likely model parameters Θ to explain the observed data X. The probability distribution of X given a model M and the model parameters, called likelihood, is $p(X|\Theta, M)$.

3.2 Bayesian learning paradigm

The Bayesian modelling paradigm is an increasingly popular method in statistical learning. The main advantage of using full Bayesian inference is that a probability distribution over the model parameters is determined and it can be used directly as a rigorous uncertainty estimate of the results in the form of confidence intervals. The posterior distribution describes the full uncertainty of the model parameters given the observed data and the assumed model structure.

The central idea of the Bayesian paradigm is inverse statistics: given the observed data X and a model M, find the posterior probability distribution $p(\Theta|X, M)$ of the model parameters Θ. The Bayesian formula reads

as

$$p(\Theta|X, M) = \frac{p(X|\Theta, M)p(\Theta|M)}{p(X|M)}. \quad (3.1)$$

Bayesian modelling makes it possible to use prior information in modelling. The prior term $p(\Theta|M)$ describes the prior probability distribution of model parameters given the model, and any prior knowledge can be included into this term. In the absence of prior information, uninformative priors are commonly used. The term $p(X|M)$ is a normalization constant.

Using Bayesian statistics is especially advantageous when the number of available data samples is small, since it makes it possible to do a principled uncertainty estimation of the result even in that case. Full Bayesian inference for complex models is applicable also with small sample-sizes, whereas ML theory is asymptotic and requires a large number of samples. At the large sample-size limit where data dominates the priors, the results given by Bayesian inference converge to the results given by maximum likelihood estimation.

3.3 Hierarchical models

The possibility of building hierarchical models is a unique feature of Bayesian modelling. In a hierarchical model, a prior distribution $p(\theta|M)$ for a model parameter θ depends of another model parameter θ' as $p(\theta|M) \sim p(\theta|\theta', M)$ and θ' is also to be learned. In practice, all the assumptions of the process that has generated the observed data can be included in a hierarchical generative model. In our model, we use a hierarchy of dimensionality reduction, data integration and multi-way modelling, which will be explained in Chapter 7.

3.4 Bringing prior knowledge into model structures

The concept of prior information in Bayesian modelling is not restricted to prior distributions of parameters or to prior parameter values, such as *a priori* known disease prevalence rates. One important usage of prior knowledge is the design of model structures. Any expert knowledge or belief (such as biological or medical) can be included in the design of the likelihood distribution of the model M to reflect the best belief of the process that has generated the observed data. Hierarchical models are a flexible way to include this type of prior knowledge.

From the perspective of Bayesian modelling, the contribution of this dissertation is, with proper utilization of prior knowledge, to construct a hierarchical model structure M that solves the multi-way, multi-source modelling task in "small n, large p" - conditions. To solve this modelling task, the properties of Bayesian modelling are utilized in two ways. Firstly, it is possible to design a dimension reduction scheme that models the data in a suitable way (correlated groups of variables) so that the relevant information in the observed data is preserved and transmitted to the next levels in the hierarchy. Secondly, building hierarchical models makes it possible to include the tasks of learning the effects of multiple covariates and learning from multiple data sources in a single unified model and to couple them with the dimension reduction scheme. The advantage of using a unified model comes from the proper propagation of uncertainties between different model parts.

3.5 Plate diagram notation

Plate diagrams [82] are the standard notation of the machine learning community for representing Bayesian model structures, and therefore assumptions of conditional independence between random variables as a graph. Plate diagrams have also been used to visualize the model structures in the publications of this dissertation. As an example, see Figure 3.1. In this representation, nodes stand for different variables, shaded nodes being the observed variables and white nodes unobserved model parameters. The arrows depict conditional dependencies between the model parameters. Plates symbolize replications of variables, for instance multiple patients.

3.6 Gibbs sampling

Learning and inference in Bayesian models can in principle be done by exact analytical inference; in practice however, the analytical solution is often intractable due to complex model structures and approximate inference is needed. Alongside variational methods and Laplace approximation, Markov Chain Monte-Carlo (MCMC) sampling methods [83] are commonly used in learning Bayesian models. MCMC methods are used for drawing samples from the posterior distribution by constructing an

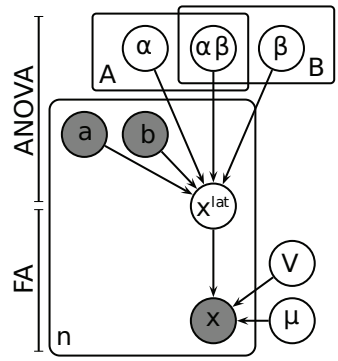

Figure 3.1. The plate diagram of the model in Publication I. Reprinted with kind permission from Springer Science and Business Media: Publication III, Figure 2 a).

MCMC chain where each new sample is drawn by conditioning on the previous sample.

Gibbs sampling is a popular MCMC method due to some appealing properties: it is relatively easy to derive Gibbs sampling equations for complex model structures and implement samplers. Conjugate prior distributions [83] are required in order to use Gibbs sampling. In the case of non-conjugate priors, Metropolis-Hasting steps can be included in the sampling formulas.

The main problems of Gibbs sampling and other MCMC methods are slow convergence of the sampler and the known fact that the sampler can get stuck in one posterior mode. The latter complicates finding the global optimum, since complex Bayesian posterior distributions are often highly multimodal. We used Gibbs sampling as the inference method in our model, although the posterior distribution is multimodal due to the complexity of the modelling problem. The implications and possible solutions of this problem will be discussed in Section 8.1.

4. Multivariate modelling of omics data

Most of the current omics studies have an underlying multi-way, even multi-source experimental design but the statistical analysis is usually done by standard simplistic tools. Proper data analysis, that takes into account the full multi-way nature of the data, is rarely done due to methodological difficulties and a lack of suitable methods.

Omics data are simply continuous-valued multivariate data and hence existing statistical tools are usually applicable and are often adopted. Statistical analysis can be done by supervised methods or unsupervised methods. Supervised methods can be used to learn a model of the associations between the omics data and a covariate, whereas the purpose of unsupervised methods is to find hidden structures from the data without the use of known training labels (covariates). Standard supervised approaches are limited to defining the statistical analyses as (one-way) case-control setups. In this chapter, I review the standard approaches commonly used for modelling omics data, including univariate statistical tests, unsupervised multivariate methods such as PCA and clustering, and supervised multivariate methods: classification and testing known groups. I also present relevant recent research to overcome the difficulties of multivariate analysis in the case of small sample-size and high dimensionality.

This thesis explores a novel approach for multi-way modelling; to date, few methods have been developed that use modern machine learning approaches. The aim of this thesis is to develop a single unified model that takes properly into account the underlying multi-way (even multi-source) experimental design. There is a need for such advanced methods since they allow a more meaningful interpretation of the data given the research question asked (see Chapter 1.2). The motivation of this chapter is to describe the standard statistical ideas since they will be used as build-

ing blocks in our overall model, but also to describe their inadequacies in order to justify the model structure chosen for our multi-way model. An example of applying the full multi-way model as an existing tool is provided in Publication V, where our model was applied to a new lipidomics dataset with a multi-way design.

4.1 Data and statistical challenges

The characteristics of the data in a standard data analysis of omics data are as follows.

- Data: Continuous-valued multivariate data with populations consisting of biological replicates. The data comes as an input data matrix with samples as rows and variables as columns.

- Statistical task: I concentrate on the most common tasks: modelling differential expression between populations, such as case and control, and modelling and discovering groups of variables based on correlations.

- The "small n, large p"-condition: High dimensionality p of the data is a challenge for data analysis. Firstly, it is difficult to comprehend a large amount of data with a lot of redundant information by visual inspection. Dimensionality reduction is usually required to find the interesting phenomena in the data. Secondly, the known problem of doing statistical analysis by univariate tests is multiple testing (Section 4.2). Thirdly, the combination of high dimensionality and small sample-size n is also problematic for most multivariate methods. Many classical multivariate methods are based on inverting a correlation (covariance) matrix and these cannot be used, since the covariance matrix becomes singular. Some multivariate methods can be used, but since the high-dimensional data space is populated only by a few data points (samples), the methods are prone to overfitting.

- Multicollinearity: The existence of correlated groups of variables, called multicollinearity, is a characteristic of omics data, known *a priori* from scientific biological knowledge. The similarly behaving correlated groups of variables often also have similar differential expression. Multicollinearity is a problem for some methods but, on the other hand, discovering

the correlated groups is biologically interesting.

4.2 Univariate analysis

The standard statistical analysis of omics data includes univariate statistical tests, such as Student t-test, Wilcoxon rank sum and one-way ANOVA, for studying the associations of a single variable at a time to a covariate. The statistical significance of the association is quantified by a p-value. Fold-change is another relevant measure used for studying differential expression [84], but not considered in this dissertation. The statistical significance levels obtained from these univariate statistical tests are a thoroughly known and a widely accepted method for deciding the significance of differential expression in biomarker discovery.

An example of stratifying a case/control, time-series multi-way experimental design, and using t-tests and fold change for comparing case and control groups at each time point, is provided in [14].

The main problem of univariate tests is that of multiple testing where, with a large number of parallel statistical tests, there is always a certain number of false positives. This problem can be alleviated to some extent by multiple testing correction, such as False Discovery Rate (FDR) [26]. Another problem is that discovering groups of genes or metabolites functioning together is biologically relevant, but univariate analysis of high-dimensional data omits correlations between the variables and only produces long lists of p-values.

The multicollinear nature of omics data and, therefore, the need for multivariate modelling and dimensionality reduction are widely acknowledged. Studying correlations, such as Pearson correlations, between the variables is the first step towards modelling similarities between the variables.

4.3 Clustering

Clustering is a basic unsupervised statistical tool for partitioning objects into sets of clusters. One can either cluster samples or variables depending on the data analysis task. Since omics data is multicollinear, clustering the variables (genes, metabolites) is a standard way of finding simi-

larly behaving clusters of biological molecules. The most commonly used clustering tools are hierarchical clustering [85] and K-means clustering. Pearson correlations or Euclidean distance are common choices for the distance measure.

4.4 Unsupervised component models: PCA

Multicollinearity of data, and resulting redundancy in the information, is often tackled by finding the principal components (or main latent factors) in the data. PCA is the most widely used multivariate method in the statistical analysis of omics data and other continuous-valued data types. The question of finding which variables are correlated can be accessed from the loading matrix of the learned PCA model. Since the loading matrix represents the relationship of the original variables to the latent factors, the latent factors can be used for visualization or further statistical analysis.

The goal of clustering and component models is to facilitate the interpretation of the data by studying the behavior of similarly behaving variables together. However, while clustering provides well-defined clusters of variables, the disadvantage of the standard component models (PCA, FA) is that each component has a non-zero loading from all the original variables. In practice, it is possible to interpret the latent components only if each component has a non-zero loading from a small number of original variables. The traditional solution for this problem is to use a cutoff for the loading strengths; a more recent solution is developing sparse approaches (Section 4.5.2).

Another disadvantage of standard unsupervised component models is that they do not model the association of data to covariates while forming the components. Although component models are a good dimensionality reduction approach for further analysis and the latent variables can be used in subsequent supervised tasks, the latent factors may not retain the biologically relevant association to covariates.

The model developed in this thesis uses a component model as a dimensionality reduction scheme to form latent factors for further analysis. We make a clusteredness assumption, where each variable belongs only to one factor.

4.5 Supervised models: Classification

Classifiers are a common method of choice for multivariate modelling of omics data when covariate information is available. Classifiers are supervised, usually discriminative models where the idea is to use known class labels to train a classifier that maximally separates two or more populations of data points (classes). The main purpose of classifiers is to predict the class label of a new data point, but the learned classifier can also be used as a model of the association between variables and a covariate. The performance of the learned classifier can be quantified by measuring the classification accuracy, in other words how well the classifier can predict the class label of new test samples with unknown class labels that where not used in training the classifier. Some classifiers, based on variable selection or latent factor models, are easily interpretable in the sense that it is straightforward to report which variables influence the discrimination between the classes the most.

Some classifiers do dimensionality reduction by forming supervised latent factors consisting of groups of correlated variables when learning the classifier from the data. This latent factor-space is used for the classification task, but it can also be used for interpreting the components. Partial Least Squares - Discriminant Analysis (PLS-DA) [86, 87] and its advanced version Orthogonal Projections to Latent Structures OPLS [30, 61, 88] are among the most commonly used classifiers for omics data, since they can form latent factors in "small n, large p"- conditions and can deal with multicollinear data. Linear Discriminant Analysis is a similar classical method, but it is based on inverting a data covariance matrix, and hence requires a regularization scheme, such as in [89], in order to be applicable in "small n, large p"-conditions.

Variable selection is another approach for dimensionality reduction of classifiers, and it also enhances interpretability by highlighting the most significant variables. Variable selection is a good approach for searching for a small set of significant variables. However, if the separation of classes is due to correlated groups of variables, variable selection may not work well since the set of selected variables may not be stable. Each of these redundant variables is an almost equally good explanation, and different variables may get selected in different trials. Some of the most popular classifiers for omics data using variable selection are Decision Trees and its extension Random Forest [90], and optimization methods such as

LASSO [91], that induce sparsity in the discriminative model. Support Vector Machines (SVM) [92] and K-Nearest Neighbors (KNN) [93], on the other hand, are examples of well performing classifiers where interpretation may be harder.

4.5.1 Using classifiers for biomarker discovery

Interpretable classifiers have become a popular method for biomarker discovery [9, 20], since the standard tool — univariate statistical testing — is not a satisfactory approach for high-dimensional data and there are no widely used multivariate statistical tests for high-dimensional data. Classifiers enable multivariate modelling of the variables contributing to the differences between case and control classes. The justification for using them for biomarker discovery is that variables that have predictive value for a class label (case, control) may be biologically relevant. There are, however, several problems in this approach:

Firstly, classifiers are usually discriminative models, whereas statistical testing (including ANOVA-type modelling) is a generative modelling task. It is important to realize that in classification the task is to find a sufficient distinction between different classes, which can be achieved by finding only a few strongly discriminative variables. When classification is combined with variable selection or assumptions of sparsity, the goal is explicitly to find a minimal set of variables. The statistical question of ANOVA-type analysis, on the other hand, is to study which of all the possible variables have statistically significant differential means.

Secondly, overfitting is a serious problem for supervised multivariate methods, especially when the number of samples is small and dimensionality high. Overfitting means that a classifier searches for a maximal separation between the classes in the learning data but, when there is a large number of variables and a small number of samples, the classifier may fit to noise and classification accuracy may be poor. Serious care to guard against overfitting has to be taken, usually in the form of a Bayesian analysis or using resampling methods, such as cross-validation, bootstrapping [94] or permutation testing.

When using classifiers that form a supervised latent factor space, such as PLS-DA and OPLS, one has to very cautious [95] if considering using the latent factors for further analysis, such as plotting the factor scores, visualization or doing further statistical testing on them. The supervised latent factors have been constructed so that they find a maximal separa-

tion between classes. If the same data that were used for estimating the latent factor space are represented in the latent factor space, the class separability will seem over-optimistic, drastically so in "small n, large p" - conditions.

In summary, unsupervised component models are at a risk of not retaining biologically relevant covariate-related variation when forming latent factors, whereas classifiers find the most strongly separating direction in the data and are at risk of overfitting. Therefore, both approaches are problematic for the purpose of reducing dimensionality for further multiway modelling. Our aim has been to develop a modelling approach that is a compromise between these two extremes. We construct a generative model that is not supervised, to seek maximal separability. However, unlike unsupervised component models, it does model the effects of covariates as latent effects acting on the latent factor space.

Whereas the use of resampling methods is one possible tool against overfitting, Bayesian models estimate inherently a posterior distribution of the model parameters, which is another approach for uncertainty estimation. The possibility of performing uncertainty estimation of dimensionality reduction jointly with further analysis is the advantage of fully Bayesian models.

4.5.2 Sparse approaches and regularization

Solving the main difficulties of the multivariate analysis of high-dimensional and small sample-size data is an active research area in machine learning, statistics and bioinformatics since these conditions are prevalent in many application areas. The main problems are singularity of the covariance matrix, overfitting, and difficulties in interpreting latent components that are linear combinations of all the original variables.

A currently popular approach is to improve existing multivariate methods by either including a regularization scheme or using sparsity-enforcing priors or constraints. Regularization means adding a penalty term to the cost function of the optimization problem that is solved when learning the model. For methods using a covariance matrix, such as LDA, the intention is to make the covariance matrix non-singular [89, 96]. Also the loading matrix of a classifier can be regularized to enforce strong loadings for significant variables and weak loadings for non-significant variables, which helps in interpretation and against overfitting.

Sparsity-enforcing priors or constraints are an improvement over regu-

larization that (attempt to) actually enforce the non-significant loadings to zero. The additional promise of sparsity is a facilitated interpretation of the results as the loading matrix of a classifier or component model has a non-zero loading only for the significant variables.

The L1-regularization, or LASSO [91], is currently the most actively studied sparsity approach for supervised models. The aim of L1-regularization is to do variable selection by enforcing the loadings of the non-significant variables to zero, retaining only the most influential variables. LASSO is an example of a method that could be used for finding a small set of single-variable biomarkers, which is a good approach for data types where each individual input variable has strong effects on covariates. This approach is, however, problematic for multicollinear data where entire clusters of variables up- or down-regulate together.

Elastic Net [97] is an improved sparsity approach for multicollinear data that encourages joint sparsity patterns for groups of variables using a combined L1-L2-regularization. The need for properly finding and modelling correlated groups of variables in sparsity models has been noted recently [98, 99].

In Bayesian methods, the most common sparsity-enforcing prior for variable selection and statistical hypothesis testing in supervised methods is the point mass mixture prior [25, 100], also known as the spike and slab prior.

4.6 Testing known groups

The biological fact that variables act together as groups is widely known also in gene expression. Testing the joint differential expression of sets of variables that are known, *a priori*, to belong to the same pathway, is a popular approach in the analysis of gene expression data. The standard methods are Gene Set Enrichment Analysis (GSEA) [101] and Gene Set Analysis [102]; the concept has also been extended to metabolomics data [103]. In an advanced extension [104], multi-way experimental designs were actually taken into account. A recent machine learning approach for testing known groups is presented in [105].

The disadvantages of testing known, pre-defined sets are that a correlated cluster may also be only a subset from a gene set [39] or composed of subsets from various gene sets. An additional, exploratory drawback is that by testing known sets, no new and potentially interesting sets can be

found.

4.7 Summary: modelling correlated groups of variables

I have now presented biological, technical, and interpretational considerations to claim that generative modelling of correlated clusters of variables is the most justified approach for ANOVA-type modelling of omics data.

To recapitulate, it is known that there are correlations between groups of metabolites or similarly behaving genes and it is biologically relevant to find these groups. It has indeed been commonly reported in lipidomics, for instance, that clusters of correlated lipids from the same lipid family were found to be up- or down-regulated similarly [30, 78]. The dimensionality reduction part of the model has to be able to model joint up- and down-regulations of groups of variables efficiently.

Modelling omics data takes place in "small n, large p"-conditions. To facilitate interpretation of the data, the redundancy of information in multicollinear data should be decreased by dimensionality reduction. Sparse components, having a non-zero loading from only a subset of variables, should be used for easier interpretation. Unfortunately, most supervised multivariate methods run into problems with overfitting or singularity of the covariance matrix. Unsupervised component models, on the other hand, may not find the biologically relevant variation in the data. The main problem of univariate tests is multiple testing.

The optimal approach to reach the goal of this thesis - multi-way modelling of single-source and multi-source omics data using groups of correlated variables as the dimensionality reduction scheme - is a combination of all the presented approaches. We assume explicitly clusters of correlated variables that respond similarly to covariates to construct latent factors. Statistical testing takes place in a low-dimensional latent factor space representing the clusters, with the entire modelling being done within a unified Bayesian model.

We construct the dimensionality reduction scheme as a FA model, where the latent factors are constrained such that each variable belongs only to one factor. This approach is related to sparse component models in the sense that each factor consists of only a subset of variables. This retains the benefits of sparse approaches: easier interpretation of the components and overcoming the "small n, large p"-problem.

The reason for using the clustering approach instead of the common

sparsity approaches developed for supervised methods, is the difference between supervised learning and statistical testing. Whereas the supervised approaches and especially their sparse variants aim to find only the smallest possible subset of variables (or groups of variables [98, 99]) to separate two classes, ANOVA-type analysis aims to do statistical significance testing for all the variables. There are also highly correlated clusters of variables that do not respond to external covariates, and the model should discover this result as well.

Unsupervised component models with sparsity-priors have also been developed [100, 106] and could in principle be used as an alternative to the clustering approach. We chose the clustering model because the clusteredness assumption is relevant for the omics data. It would be of interest to study whether equally good results can be obtained with sparsity assumptions.

5. Existing multi-way ANOVA-type models

In this Chapter, I review existing approaches that can be used for multi-way modelling of continuous-valued data. First, I introduce the basic tools ANOVA and MANOVA and explain how they are special cases of the framework of General Linear Models. I then discuss the problems of ANOVA and MANOVA for high-dimensional data and present previous approaches to solve these problems by an additional dimensionality reduction step. Then, I present relevant Bayesian approaches for multi-way modelling and, finally, I review the connection of multi-way modelling to the related advanced machine learning approaches: multi-task learning, multi-label prediction and multi-class classification. The conclusion of this Chapter is that none of these methods is a rigorous approach for multivariate multi-way modelling of high-dimensional data; an even more important note is that none of these approaches enables an obvious extension to multi-way modelling of multi-source data.

5.1 ANOVA and MANOVA

The univariate ANOVA [107] was the first and is currently the most widely known method for the task of modelling the effects of multiple covariates and their interactions in populations of measurements. Multi-way ANOVA is an extension of the F-test used in one-way ANOVA and the Student's t-test. The central assumption of multi-way modelling is that a generative model explains the observed covariate-related variation in the data (in the two-way case) as

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab} + \epsilon. \tag{5.1}$$

Here a and b ($a = 0, \dots A$ and $b = 0, \dots B$) are the two independent covariates. The main effects $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_b$ and the interaction effect $(\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}$

model the variation from the baseline level (grand mean) μ. Despite the confusing notation, the interaction effect $(\alpha\beta)_{ab}$ is just another term: the interaction effect of disease and treatment, for example. The ϵ is a noise term. A Bayesian formulation of ANOVA was presented in [83].

Multivariate ANOVA (MANOVA) is a multivariate generalization of ANOVA. The MANOVA model is the same as the ANOVA model in Equation 5.1 except that all the terms are vectors, whereas in the univariate ANOVA-model they are scalars.

5.2 General linear model

ANOVA and MANOVA are special cases of the General Linear Model (GLM), which is a general term for the generative model

$$\mathbf{X} = \mathbf{DB} + \text{noise}, \tag{5.2}$$

where $\mathbf{X}$ is the data matrix, $\mathbf{D}$ is a design matrix of the levels of known covariates and $\mathbf{B}$ contains the regression coefficients. The statistical significance of the elements of $\mathbf{B}$ implies which covariates in $\mathbf{D}$ explain the data in $\mathbf{X}$. The GLM incorporates the cases of univariate and multivariate data for discrete and continuous-valued covariates. ANOVA is the special case of GLM in the case of discrete covariates and univariate data. In MANOVA, the data are multivariate. In the case of both discrete and continuous covariates, the methods are called Analysis of Covariance (ANCOVA) and Multivariate Analysis of Covariance (MANCOVA). Joint Bayesian modelling of discrete and continuous covariates, which is essentially a mixture of ANOVA and regression, has been presented for example in [29, 108]. (M)ANCOVA-type analysis is not, however, discussed in this thesis. The GLM incorporates also the Student t-test, F-test and linear regression. Generalized Linear Model is a generalization of GLM that additionally includes non-linear relationships between the data and the covariates, such as the logistic function in logistic regression. I do not discuss non-linear relationships in this thesis.

5.3 Linear mixed models and time-series modelling

Whereas the standard (M)ANOVA-type linear model consists of standard fixed effects, Linear Mixed Models (LMM) extend (M)ANOVA-type analysis such that they can also include random effects. Fixed effects mean ef-

fects of interest, or treatments, that have been chosen in the experimental design and are relevant regarding the chosen research question. Random effects model out real, but often uninteresting effects in the data, such as experimental artifacts from repetitions on an individual [109, 110], gene-specific effects in a clustering model [111, 112] or confounding factors due to a population structure [113]. The equation for a linear mixed model is often written as

$$\mathbf{X} = \mathbf{DB} + \mathbf{VU} + \text{noise}, \tag{5.3}$$

which extends the linear model having fixed effects with the inclusion of random effects $\mathbf{V}$. The $\mathbf{U}$ are the regression coefficients from random effects to the data. The $\mathbf{D}$ indeed consists of fixed zeros and ones that relate known covariates to the samples, whereas the random effects $\mathbf{V}$ are assumed to have been sampled from a Gaussian distribution and they have to be learned as well.

LMMs have been a popular choice for modelling time-series omics data. In ANOVA-type or LMM-modelling of time-series data, time-point is usually the fixed effect; the main covariate to be modelled or one of the covariates in the case of a multi-way design. Time-series modelling of omics data is a big field of research of its own [114], where the central research questions are how to optimally take the time-series nature of data into account and how to predict future time-points. However, other than the possibilities within ANOVA-type modelling, I do not discuss modelling the time-series nature of data in this thesis.

A particularly interesting application of LMMs is clustering genes in time-series omics data [111, 112]. In this approach, cluster-specific time-course is the interesting fixed effect and the gene-specific deviation from the cluster-specific time-course is a random effect. This is also a model-based clustering approach where the cluster-specific time-course becomes the “model”. The random effects are here used to model out uninteresting gene-specific time-effects, which helps in forming the clusters.

The relationship of our modelling approach to LMMs is two-fold. We will concentrate on fixed effects only and do not consider mixed models. However, we will use the same idea of model-based clustering where the “model” for each cluster can be used to describe the effects of covariates.

5.4 Problems of ANOVA and MANOVA

The general challenges of univariate statistical tests for high-dimensional data (Section 4.2) also hold for multi-way ANOVA, although the problem of multiple testing can be alleviated to some extent with multiple testing corrections. There are also more advanced test statistics developed for multi-way experimental designs [84, 115]. Access to the information of correlated clusters can be sought by clustering the variables before the analysis, or by grouping the variables [116] according to the p-values obtained from the univariate statistical tests.

MANOVA is the multivariate generalization of ANOVA and defines a formal multivariate statistical test for testing the effects of covariates on populations of measurements, taking correlations between variables into account. MANOVA was originally designed for low-dimensional data in $n > p$-conditions. Unfortunately, in "small n, large p" - conditions that are ubiquitous in omics data, the covariance matrix becomes singular. Another problem of MANOVA is that it tests the difference between two populations in terms of all the variables simultaneously and, therefore, only gives the statistical significance of the overall effect. This result is not sufficiently informative for high-dimensional data and the highly relevant information of which variables were up- or down-regulated has to be deduced by other methods, such as univariate tests. Furthermore, as ANOVA and MANOVA only give the statistical significance of the effect, the direction, up or down, has to be deduced by other means.

The univariate ANOVA and multivariate MANOVA are two extremes for how to solve the multi-way modelling problem, both of them facing both technical and interpretational problems due to the "small n, large p"-conditions. Few methods have been presented for multivariate multi-way modelling, essentially solving the general linear model, in "small n, large p"-conditions. It is obvious that a compromise between testing a single variable (ANOVA) and all the variables simultaneously (MANOVA) is sought; the common modelling choice is a dimensionality reduction into a low-dimensional latent factor space where the statistical testing is done. A few such classical and Bayesian approaches have been developed.

5.5 Multivariate many-step approaches

To my knowledge, no methods for multivariate multi-way modelling of high-dimensional data with a single unified model exist in classical statistical literature. In contrast to a unified model, the concept of many-step approach is here used to refer to a pipeline of separate methods, such as prior dimensionality reduction followed by statistical testing. All the presented classical methods for multi-way modelling of high-dimensional data are many-step approaches.

The working solutions presented so far combine a prior dimensionality reduction by PCA, followed by modelling the effects of covariates by statistical tests in the reduced-dimensional latent factor space. Examples are ANOVA on the PCA scores [117], 50-50 MANOVA [118] where a MANOVA follows a PCA dimensionality reduction and ANOVA-Simultaneous Component Analysis (ASCA) [119]. These approaches are prone to the standard problems of PCA (Section 4.4).

The possibility of estimating a whole model jointly and propagating uncertainties between different model parts properly makes unified Bayesian models a more elegant approach compared to many-step approaches. The Bayesian model of this dissertation has an integrated dimensionality reduction into latent factors where multi-way and multi-source modelling are done. Another probabilistic method, very similar to our method in Publication I, was presented later in [74]. In that model, probabilistic PCA was used in combination with generative modelling of the effects of multiple covariates.

5.6 Multivariate Bayesian approaches

One Bayesian approach for multi-way ANOVA-type modelling of high-dimensional data is modelling the dataset by univariate linear models that are coupled together by joint sparsity priors [29, 32, 108]. This approach defines a regression from all the covariates to each variable. A joint sparsity prior, point-mass mixture priors in this work, gives a non-zero loading only for the regression coefficients most strongly modelling the association between the variables and covariates. This is a good approach against the multiple testing problem and the induced sparsity helps in interpreting the dataset.

Another well known approach that uses the point-mass mixture priors

is sparse factor regression [25]. In this supervised multivariate approach, sparse latent factors are used to predict an external covariate. Using the external covariate to supervise the latent factors can overcome the problem of standard PCA where covariates are not taken into account when learning the principal components. The point-mass mixture priors make it possible to form sparse factors where each factor is associated only with a subset of variables and vice versa. This supervised approach can be used to find the groups of variables that best predict the covariate. A similar non-Bayesian approach for (one-way) supervised PCA has been proposed in [120].

A further extension called sparse latent factor regression model [100] was developed by combining the univariate ANOVA models, coupled by a joint sparsity priors, with sparse latent factor models into a joint linear model. The model is used so that known covariate-related variation is modelled by the univariate ANOVA models. This variation is often assumed to be uninteresting experimental bias and is explained away. The remaining variation not explained by the covariates is assumed to be the interesting information and is modelled by the sparse latent factors. The latent factors can be either unsupervised or supervised for a prediction purpose. In the supervised case, the response variables can be Gaussian, right-censored, categorical or binary class labels, where the latter are dealt with logistic or other link functions.

5.7 Summary: Our modelling approach compared to the existing approaches

The ANOVA-type modelling is a generative modelling task where the covariates are assumed to explain the variation in the data that is high dimensional in the case of omics data. The Bayesian sparse latent factor regression model [100] is closely related to our model in Publication I (Chapter 7) in the sense that sparse latent factors, each consisting of only a subset of the variables, are formed in a data-driven manner. The difference in the models is that in [100] (as well as in any other models [25, 29, 32] in the same framework), multi-way modelling is not done in the multivariate sense. This model can either do multi-way modelling using the univariate models, or the multivariate latent factors can be supervised for the purpose of multiple classification tasks, or both. As discussed in Chapter 4.5, classification is a different task compared to statis-

tical testing. The multivariate component model part of the sparse latent factor regression model therefore can not rigorously separate the effects of multiple covariates from the effects of their interactions.

In contrast, our model (Publication I, Section 7.1) does multivariate multi-way modelling by estimating the statistical significance of the effects of covariates and their interactions in the latent factor space, where each latent factor represents a cluster of correlated variables. The other model that does multivariate multi-way modelling is 50-50 MANOVA, presented in Section 5.5, since it does the MANOVA test on the PCA scores. However, this approach is a many-step approach.

The most important aspect of our modelling approach is that it enables the extension of the multi-way model naturally into the multi-source cases as a single unified model. Compared to the sparse latent factor regression model [100], our model has an additional latent factor space where the effects of the covariates are modelled. This additional hierarchical structuring in our generative model makes it possible to straightforwardly define yet another layer of latent factors that are used for integrating multiple data sources. In contrast, the extensions of the sparse latent factor regression model used for multi-species translation [29, 32] are many-step approaches. In that model, a list of genes responding to a covariate in one dataset, is used as prior information for further modelling in the other dataset.

5.8 Multi-way learning compared to other advanced machine learning genres

In the machine learning literature, we have introduced the name "multi-way learning" to define the multi-way modelling task: finding and evaluating the statistical significance of the effects of multiple independent covariates and their interactions. Multi-way learning is a generative modelling task and clearly different from standard discriminative models, such as classifiers. Binary classifiers are, however, commonly used in analyzing data having a multi-way experimental design. This is done naively by either stratifying the case-control comparison according to the additional covariates or by predicting each covariate at time.

There are currently three popular advanced machine learning approaches that are closely-related to multi-way learning: multi-task learning, multi-label prediction, and multi-class classification. In this section, I review

the connection of these approaches to multi-way learning and justify why multi-way learning is a different learning task.

5.8.1 Multi-class classification

When considering the relationship of multi-way learning to other machine learning genres, it is important to notice that each sample is associated with multiple covariates and there is a structure between them. This means that in a setting consisting of "diseased treated" and "diseased untreated" patients, for instance, data from both groups is used to estimate the disease effect.

A multi-way experimental design can in principle be naively considered as a multi-class classification problem where each combination of covariates, such as "diseased treated", is considered an individual class. Even standard classifiers, such as PLS-DA and LDA, can be used in this sense. An example of converting a data analysis problem with a multi-way, multi-source design into a series of multi-class PLS-DA classifiers, each source at a time, was presented in [121]. In this approach, one naturally loses the information that samples with "treatment 1 in early time-point" and samples with "treatment 1 in late time-point" are related due to both having "treatment 1". Multi-class classification cannot be used easily to estimate the statistical significance of the effects of multiple covariates and their interactions.

To my knowledge, the concept of multi-way classification taking into account the multi-way experimental design and being able to estimate interaction effects has not been introduced, although defining such a concept could result in a useful methodology.

5.8.2 Multi-task learning

Multi-task learning [122] is a popular machine learning approach that can be used for learning the association between the data and multiple covariates. The leading principle of multi-task learning [58, 123, 124] is that there are multiple related (usually supervised) learning tasks and learning these tasks jointly makes it possible to borrow statistical strength from one another. This improves learning results in contrast to learning each task separately, which is usually evaluated in the form of classification accuracy. Searching for task relatedness [125] is another goal of multi-task learning.

Some multi-task learning methods deal with multiple sets of samples with the same input space, each set having its own learning task [123, 126, 127]. Another common setting is learning multiple related tasks (covariates) from the same dataset [128, 58], the latter task being closely related to multi-way learning. However, the leading principles of the two genres "learning multiple related tasks jointly" and "modelling the effects of multiple independent covariates and their interactions" are clearly contradictory. The common multi-task learning assumption of finding a common set of discriminative variables predictive of all the relevant tasks is not a relevant assumption in multi-way learning. In multi-way learning, covariates are independent and each is *a priori* assumed to up- or down-regulate different variables, although overlap is naturally possible. Therefore, multi-way learning should be considered as a different genre.

There is, however, a connection between multi-task learning and multi-way learning; the dimensionality reduction of our multi-way model is done by representing each cluster of variables by a latent factor and the effects of multiple covariates and their interactions are learned in this low-dimensional latent factor space. The latent factor space is the same for all of the samples (and all the covariates) and, therefore, the effects of each covariate are learned in the same space. In multi-task learning, there is a similar aim to learn all the tasks in a shared variable, or latent variable space, which supposedly provides increased statistical strength compared to learning each task in a different space.

5.8.3 Multi-label prediction

Multi-label prediction is a recent machine learning genre developed to address modern application problems of molecular biology and Internet data, such as gene function prediction or text and image annotation. In multi-label classification, for instance, the task is to predict a large number of class labels associated with the samples.

The setting is similar to multi-way learning in the sense that each sample is associated with multiple covariates. The main interest of multi-label prediction is, however, in learning or utilizing label correlations and label structure. A good example is Hierarchical Multi-Label Classification [129, 130], which deals with labels having a hierarchy such that the sample belonging to a class automatically belongs to all ancestor classes in the hierarchy.

Again, the central difference is that multi-way learning is a generative

learning approach, aiming to estimate the statistical significance of the effects and interaction effects of a few central, independent covariates. Multi-label prediction is a discriminative classification approach and does not have proper means for doing this evaluation.

6. Integration of multiple data sources

As the integration of different biomedical data types is believed to be increasingly important in the future medicine, data integration itself is an active research field in the methodological machine learning research, often called multi-view learning. Its applications and modelling problems span a wide range of statistical questions and combinations of heterogeneous data types, such as continuous-valued data, categorical data, relational (network) data, rank data, text data, "Internet clicking"- data and image data.

Also in this dissertation, I present methodological contributions to integrating multiple continuous-valued data sources, but now in the context of multi-way modelling where no previous multi-source methods have been presented. When integrating multiple data sources, a central question is what kind of pairing information is available to connect the data sources. We will do multi-way, multi-view learning in two cases: **paired samples**, which appear when measurements are taken from multiple tissues or by multiple different omics techniques from each patient; and in the case of **no paired samples**, which appear from cross-species translation between multiple species. The latter problem is much more difficult in the statistical sense. Another question is whether variables are matched between the data sources. In our models, we focus on the general case and assume no *a priori* matched variables since in our applications, different data sources have in general different molecules. Having matching information of variables available would make it possible to build more powerful statistical models, however, the applicability area of such models is restricted.

When integrating information from multiple data sources, an assumption for connecting the multiple datasets is required. In this Chapter, I first review two relevant existing approaches that can be used for in-

tegrating multiple continuous-valued data sources with paired samples: unsupervised generative modelling that can be used to find what is shared between the datasets, and supervised approaches that define similar classification results in multiple datasets as the justification for integrating the data sources. Other proposed means of connecting different data sources are projecting different data spaces into a latent factor space, such that neighborhood relationships are preserved [131] and shared causality [34]. I then review the existing approaches for cross-species translation-type data integration, where samples are not paired, and show that most existing approaches are based on having a matching of variables available, whereas our model is not restricted to such an assumption.

The model presented in this dissertation complements the existing multi-source approaches by a novel assumption for data integration: a shared response to multiple covariates and their interactions. According to this assumption, a similar response to multiple covariates and their interactions can be discovered in both datasets and it can be modelled by an underlying shared latent variable that is assumed to have generated the data in both data sources.

6.1 Unsupervised data integration

Canonical Correlation Analysis (CCA) [132] is a widely used method for finding dependencies between two or more datasets with paired samples and different variable-spaces. CCA can be used to answer the question of what is shared between two data sources, being a generalization of correlation to multivariate data. CCA has recently attracted considerable attention as the importance of integrating multiple data sources has been noticed.

The underlying assumption of the generative model of CCA [133] is that there is a shared latent process that has generated a part of the variation in both of the observed data sources and, additionally, there is data source-specific variation. This formulation makes it evident that CCA is closely related to other, single data-source factor models, such as FA and PCA. These models can, therefore, be used conveniently together as building blocks of a hierarchical unified model, as was done in Publication II. CCA extends FA and PCA such that in the generative model, there is a shared latent factor common to both data sources (together with source-specific effects). A Bayesian formulation of CCA was presented

in [134, 135] and several other formulations have been presented lately [136, 137].

A central theme in the current research on CCA, as well as on other factor models, is finding sparse components both to facilitate interpretation of high-dimensional data and to deal with the "small n, large p"-conditions. The main approaches of the Bayesian genre for sparsity-inducing priors of CCA are an advanced version of the Automatic Relevance Determination (ARD) prior [136] and an Indian Buffet Process prior [137]. Non-Bayesian approaches are usually based on L1 and/or L2 -regularized optimization methods [57, 138, 139] or kernel methods [140]. In our model, the CCA-component integrates multiple low-dimensional latent factor spaces that result from a dimensionality reduction at a lower level in the hierarchy. Further sparsity approaches are therefore unnecessary for the CCA-component in our model.

In applications having both paired samples and matching information between the variables available, the matching information can be used to reduce the number of parameters of the CCA-model, which makes the model more powerful [59].

Other canonical correlation-type approaches for unsupervised data integration are Co-Inertia Analysis (CIA) [141, 142, 143] and O2-PLS [61, 144], which is a generalization of OPLS. Methods that assume shared latent variables between multiple data sources can be also used for clustering the samples (patients) [145]. The goal of finding unsupervised dependencies between datasets without modelling response to covariate(s) can also be achieved by regressing from one dataset to another [48, 143, 146], for instance by PLS.

Overall, methods that search dependencies between data sources are one modelling option for connecting multiple data sources, however, since the unsupervised data integration approaches do not take the covariate information into account, they need to be extended for the purpose of multi-way modelling. We will do that in Section 7.2.

6.2 Supervised data integration

The standard supervised approach for data integration is to learn a separate classifier to predict a binary class label (covariate) in each data source and combine the classification results. The usual goal is to determine whether additional data sources, often of different data type, can improve

classification accuracy, as was done in the original multi-view learning paper [147] and in several later approaches [70, 71, 72, 73, 121]. These studies have indeed shown that integrating classification results from multiple data sources can improve the classification accuracy over using a single data source. The combined multi-source, supervised one-way learning is a setting closely related to multi-source, multi-way modelling, but the ultimate goal of multi-source, multi-way modelling is again different: learning the association and statistical significance of (multi-source) data to multiple covariates and their interactions, and in particular, explaining the covariate-related dependencies between the data sources.

Corresponding to our interest in combining multi-way learning and multi-view learning, the multi-task learning community has lately presented various methods for combining multi-task learning and multi-view learning [148]. This indicates a growing interest in learning joint models in setups involving multiple data sources and multiple covariates (or tasks).

6.3 Cross-species analysis and translation

Cross-species analysis of omics data is a research genre with increasing importance since model organisms are used as models of a disease or treatment response and the findings have to be translated into human clinical studies. The use of omics techniques enables a biomarker discovery-type approach to cross-species analysis: studying how the concentrations of different molecules respond (up- or down-regulate) to covariate(s). The genomes of different species have a partially shared, conserved component due to joint genetic ancestry [12]. Due to evolutionary changes, however, different species have additionally a species-specific genetic component. As a result, partially similar gene expression [12] and other omics phenotypes can be expected between different species as well, including partially same molecules (mRNA/proteins/metabolites) and partially similar physiological functions.

Modelling cross-species omics data is a difficult data integration task. It differs from the data integration tasks with paired samples (Subsections 6.1 and 6.2) in the sense that there are no paired samples, as naturally there is no pairing between an individual human and mouse, for instance. Despite this difficulty, the goal of these two data integration settings is the same: to find what is shared between the data sources.

An important issue is whether a known correspondence, or matching

between different molecules in the two species exists. Two different types of models are usable depending on whether the variables can be *a priori* matched (at least partially) between the species, or whether no such pairing information is available. Having *a priori* matching information makes the modelling task much easier by considerably more statistical strength being available. Most importantly, the matching offer a well-defined setting for finding shared patterns between the datasets. For mRNA data, this matching information is often known since the traditional research question of cross-species analysis has been to find orthologous genes [12] based on the similarity of the DNA sequence coding the gene. If an orthologous gene has a similar role in a cellular physiological process across species, it is called a core gene [12, 50] in the case of gene expression data. If an orthologous gene has a different role in the different species, it is called divergent gene [50].

In some cross-species research questions, such as in translational lipidomics studies between human patients and model organisms [49], there is not necessarily any matching information available. Although the experimental techniques identify similar chemical molecules in multiple species [31], these molecules can be divergent genes or metabolites or different molecules in different species may have taken the same role in some physiological systems. If the correspondence of the roles of a chemically similar molecule in different species is not well known, it is better not to assume any known matchings. Therefore, a research question of great importance is how to actually map the, say, lipidomics disease phenotypes and responses to drugs between clinical human studies and model organisms [31, 49]. By using omics data, this problem can be approached in a data-driven way. Also, the validity of a specific animal model as a model for a certain disease is often debated by the biomedical community [9] and computational cross-species analysis can help to validate [31] these models. In lipidomics, for instance, the mappings of phenotypes between species are to date mostly unknown. The scientific community is only beginning to search for means for finding this information, [49] being one of the first attempts. Computational models based on *a priori* known matching between variables are not usable here.

6.3.1 Known matching between the variables

When orthology information of genes is available, the central research question of cross-species analysis of mRNA data is finding the core genes,

conserved cross-species gene modules [12, 68] that give evidence of evolutionary mechanisms. Expression meta-analysis [12] refers to comparing expression of genes in different species under similar (usually one-way) experimental conditions. When matching information of orthologous genes is available between all the variables, expression meta-analysis is conceptually straightforward: comparing lists of p-values, at the simplest, although such results have not always been encouraging [12].

I give two examples of using the known matching of variables. Firstly, when translating findings between observational human studies and *in vitro* cultured human cells or cancer cell lines, matching between practically all the variables is present [32], which provides a straightforward starting-point for translation. Secondly, in cross-species analysis of gene expression data, there is usually a list of orthologous genes available [12, 29, 149, 150]; this list is a subset of all the genes in each species. For instance, the authors in [29] report that out of 12000 human gene probesets and 12000 mouse probesets, there are 7000 known matches. The analysis can, therefore, be restricted to this common subset. The translation is done in [29] and [32] using as prior information the list of model organism genes that respond to a covariate, the model organism being cell lines [32] or mouse [29]. The subsequent analysis on human samples is done by restricting the latent factor modelling to the prior list of genes. In these experiments, both human samples and model organisms have the same or related disease, although the full experimental designs are not equal. Biclustering both genes and conditions is another approach for multi-species integration [68] when matching of genes are available.

It has also been shown that knowing some matchings *a priori* can be used to find more matchings from the data [50, 150]. On the other hand, [150] questioned the exactness of known orthologs between genes in multiple species that are solely based on sequence similarities. Instead, they used the degree of sequence similarity as a probability of match, transforming exact prior affiliations to soft prior probabilistic matchings to be confirmed by additional omics data.

Another worthwhile question is whether cross-species translation should be done assuming one-to-one matching of individual variables or by assuming matching of clusters of variables. Although there indeed exists, in some cases, prior information for one-to-one correspondences of orthologous genes [12, 50, 149, 150], the redundant information resulting from multicollinearity of omics data makes the question of finding one-to-one

matching difficult. Most computational approaches are indeed clustering-based methods [12, 32, 68, 150, 151], and we also assume matching for clusters of correlated variables.

6.3.2 Unknown matching between the variables

A novel, considerably more difficult computational problem of "how to actually find matching from the data" appears when no *a priori* matching information is available or it is not reasonable to use it. This task has been tackled by a CCA-type approach [49, 51, 152] and by a Co-Inertia Analysis-based method [151]. In the lipidomics application of [51], however, the search for matching was facilitated by restricting the search space to candidate sets according to prior information on lipid functional classes and chemical properties. These methods do not take the covariate-information into account, which makes it difficult to interpret what the similarity in matched variables (molecular profiles) actually is.

The model presented in this dissertation tackles the problem of finding matching without any *a priori* matching information between the variables in different species and, additionally, translating the results between the datasets coming from the two species. As a summary: no existing method can tackle this task. The methods that assume an *a priori* known matching between the variables naturally can not be used to find the matching. The existing methods that do not assume an *a priori* matching [51, 151] do not model the response to covariates, which would enable to translate these responses.

The novel contribution of our model (Section 7.2.3, Publications III and IV) in the cross-species translation line of research is to search for translation by searching for a similar response to multiple covariates and their interactions in the context of a similar underlying multi-way experimental design in the two species. The method therefore finds what is shared in the two datasets, which can be modelled by a shared underlying latent effect. Using this approach, the model can both find the variables that respond similarly determining that such variables are matched, and model the shared (core) response to the covariates, which translates the findings (response to covariates) between the species. The method is therefore directly usable to the task of mapping disease and treatment-related findings between model organisms and clinical studies.

7. The unified multi-way, multi-source model

The contribution of this dissertation is to extend multivariate multi-way ANOVA-type modelling to multiple data sources with different, unmatched variables in each data source and to make multi-way modelling possible in "small n, large p"-conditions. Two data integration cases are considered: (i) multiple data sources with paired samples (multiple tissues or multiple omics types measured from each individual), and (ii) multiple data sources without paired samples (translating results between human patients and model organisms). Multi-way modelling is additionally extended to cases where the structure of one of the covariates is partly unknown. These modelling problems are closely related as illustrated in Figure 7.1. We have developed a unified Bayesian model that can be structured for all the multi-way modelling cases with slight modifications to the model structure according to the experimental design.

In this chapter, I first present how the problem of high dimensionality and small sample-size can be overcome by a dimensionality reduction approach, where clusters of similarly behaving variables are modelled as latent factors. The effects of the covariates and their interactions are modelled in the low-dimensional latent factor space. I then present how multiple data sources can be combined by integrating the low-dimensional latent factor spaces they are represented by. This is followed by the extension of multi-way modelling to the case where one of the covariates has a partly unknown structure; the idea is to learn the unknown structure jointly with multi-way modelling. Finally, I will discuss the technical details of the model and how we handled the imperfect experimental design of the data in Publication V.

In all the multi-way modelling cases, the goal is to train a unified generative model that is assumed to have generated all the observed data. After learning the parameters of the model, the posterior distribution provides

a solution to the multi-way modelling task. The statistical significance of the effects, the decomposition to shared and data source-specific effects, and the cluster assignments of the variables can all be directly obtained from the posterior distribution.

7.1 Single-source multi-way modelling

Our approach for solving the multi-way modelling task of a single data source in "small n, large p"-conditions [Publication I] is the following. We reduce the dimensionality of the data by modelling clusters of correlated groups of variables by a factor analyzer and the effects of covariates are modelled in the latent factor space. The factor analysis model is

$$\begin{aligned} \mathbf{x}_j^{lat} &\sim \mathcal{N}(0, \mathbf{\Psi}^x), \\ \mathbf{x}_j &\sim \mathcal{N}(\boldsymbol{\mu}^x + \mathbf{V}^x \mathbf{x}_j^{lat}, \mathbf{\Lambda}^x) \ . \end{aligned} \tag{7.1}$$

Here $\mathbf{x}_j$ is a p-dimensional observation vector for sample $j = 1, \ldots, n$ and $\mathbf{V}^x$ is the projection matrix ($p \times K$) that is assumed to generate the data vector $\mathbf{x}_j$ from the K-dimensional latent variable $\mathbf{x}_j^{lat}$. The K is the number of components (clusters). The $\mathbf{V}^x \mathbf{x}_j^{lat}$ models common variation of the data around the p-dimensional mean vector $\boldsymbol{\mu}^x$.

The regularizing assumption required to overcome the "small n, large p"-problem is that latent factors are composed of clusters of (correlated) variables such that each variable belongs exactly to one component (details in Publication I). The elements of $\mathbf{V}^x$ have been restricted to being positive only; therefore, each cluster consists of only positively correlated variables. These components can be seen as sparse factors, since each factor involves only a subset of the original variables, unlike in the standard FA and PCA where components are a linear combination of all the variables.

The effects of multiple covariates act directly on the latent factors as

$$\mathbf{x}_j^{lat} \sim \mathcal{N}(\boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}, \mathbf{\Psi}^x), \tag{7.2}$$

where $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_b$ are the K-dimensional main effects and $(\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}$ are the K-dimensional interaction effects for covariates $a = 0, \ldots A$ and $b = 0, \ldots B$. This equation holds also to the case of one of the covariates having partly unknown structure, presented in Section 7.4.

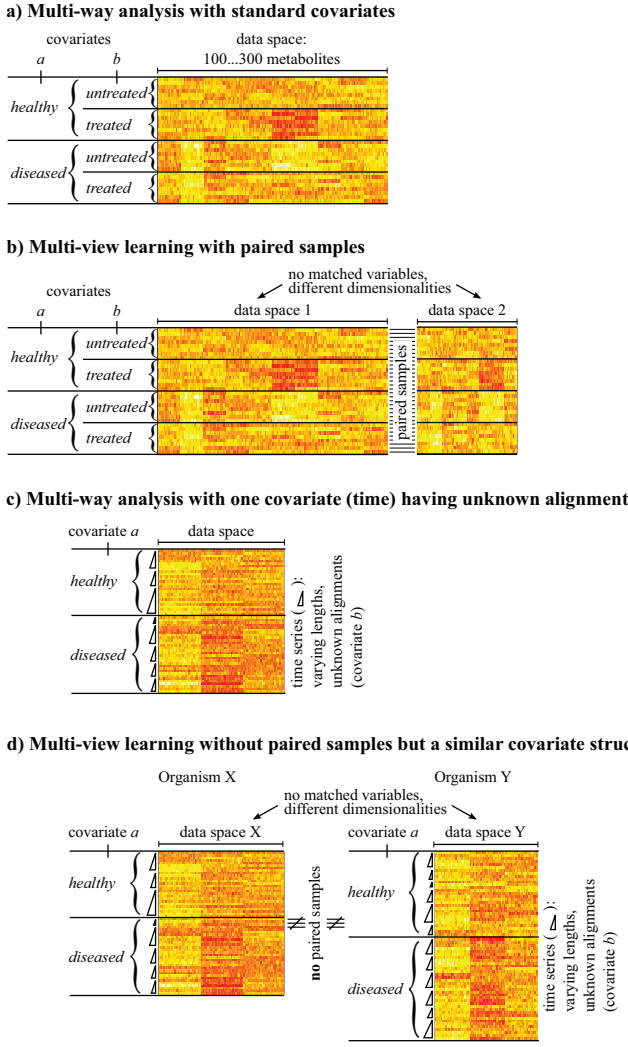

Figure 7.1. Illustration of the four data analysis tasks in this paper. (a) Standard ANOVA setup, but with high dimensionality (variables) compared to the number of samples (rows). (b) Extension to multiple data sources with paired samples. (c) Extension to time with unknown alignment. (d) Extension to multiple data sources without paired samples. The images represent data matrices, where rows are samples and columns are variables. The illustration represents the experimental design of each task, composed of a combination of standard covariates (disease, treatment), time-series information, and integration of multiple data sources. Reprinted with kind permission from Springer Science and Business Media: Publication III, Figure 1.

7.1.1 Relationship to PCA-approaches

The intuitive idea of doing multi-way modelling in a low-dimensional latent factor space is similar to the many-step approaches where PCA is followed by multi-way analysis on the PCA scores [74, 118, 119]. For the multi-source case with paired samples, the intuitive idea is doing a sparse CCA followed by multi-way modelling on the scores in a shared CCA-space. The added benefit of our unified model is that the uncertainty estimation of dimensionality reduction and multi-way modelling is done jointly.

7.1.2 Relationship to LMMs

From the perspective of clustering, the Bayesian model does model-based clustering of variables where the latent factor scores (probabilistic in our model) are the "model" for each cluster. Multi-way modelling on these probabilistic factor scores on the next level up in the hierarchy provides directly the statistical significance of the effects of the covariates and their interactions for each cluster of variables. As noted in Section 5.3, this model-based clustering approach is related to using LMMs for clustering time-series data. Another connection to LMMs is that in the higher level of hierarchy of our model (Equation 7.2), the variation of the latent factors is separated into fixed multi-way (and multi-source) effects and latent space noise. The latent space noise is closely related to random effects in the sense that it models individual-specific variation that deviates from covariate effects for each cluster. Modelling the latent space noise enables clusters of correlated variables to be formed even when no significant fixed effects are found from the data.

7.2 Multiple data sources

I now proceed by extending the single-source model into multi-source cases; the different model variants are illustrated in Figure 7.2. In all cases, the effects of covariates are modelled with terms $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$ and $(\boldsymbol{\alpha\beta})_{ab}$. In the single-source cases, these effects act directly on the latent space representing the groups of correlated variables. In the multi-source case with paired samples, these effects act on another level of hierarchy; the shared latent factor space modelling shared variation between the sources.

In each case, the posterior distribution of the effects is estimated and can be used to directly estimate the statistical significance of the effects of the covariates and their interactions.

7.2.1 The 'data source' as a covariate

The key theoretical idea of extending multi-way modelling to multiple data sources is to consider the 'data source' as an additional covariate. In a standard multi-way model, there are covariates such as disease, treatment, time, and gender. In this subsection, I call these **standard covariates** [Publication III] to distinguish them from the 'data source' as a covariate. An ANOVA-model for two standard covariates is

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab} + \text{noise}. \tag{7.3}$$

Consider now a data analysis task with two standard covariates and two data sources with different variables. When the source is an additional covariate d, the model becomes

$$\begin{aligned}\mathbf{x}_d = \boldsymbol{\mu} + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab} + \boldsymbol{\gamma}_d + \\ (\boldsymbol{\alpha}\boldsymbol{\gamma})_{ad} + (\boldsymbol{\beta}\boldsymbol{\gamma})_{bd} + (\boldsymbol{\alpha}\boldsymbol{\beta}\boldsymbol{\gamma})_{abd} + \text{noise},\end{aligned} \tag{7.4}$$

where $\boldsymbol{\gamma}_d$ would denote the effect of source. However, since different data sources have different (and a varying number of) variables, this model cannot be applied as such. The main and interaction effects cannot act on two different data spaces with different dimensionalities.

It is possible, however, to build a hierarchical model where the latent effects are projected to the actual data spaces x and y with unknown projections f^x and f^y that can be estimated from the data. The equations are

$$\begin{aligned}\mathbf{x} &= \boldsymbol{\mu}^x + f^x(\boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}) + f^x(\boldsymbol{\alpha}^x_a + \boldsymbol{\beta}^x_b + (\boldsymbol{\alpha}\boldsymbol{\beta})^x_{ab}) + \epsilon, \\ \mathbf{y} &= \boldsymbol{\mu}^y + f^y(\boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}) + f^y(\boldsymbol{\alpha}^y_a + \boldsymbol{\beta}^y_b + (\boldsymbol{\alpha}\boldsymbol{\beta})^y_{ab}) + \epsilon\,.\end{aligned}$$

From now on I denote the two data sources as x and y, respectively. In this definition, the effect of the data source cannot be included as a main effect since different data sources have different variables. However, it is now possible to separate the main effects $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$, and $(\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}$ from the interaction effects of a standard covariate and the source $\boldsymbol{\alpha}^x_a$, $\boldsymbol{\beta}^x_b$, and $(\boldsymbol{\alpha}\boldsymbol{\beta})^x_{ab}$ (similarly for y). The latter effects have the interpretation as "source-specific" covariate effects, whereas the main effects are shared covariate effects

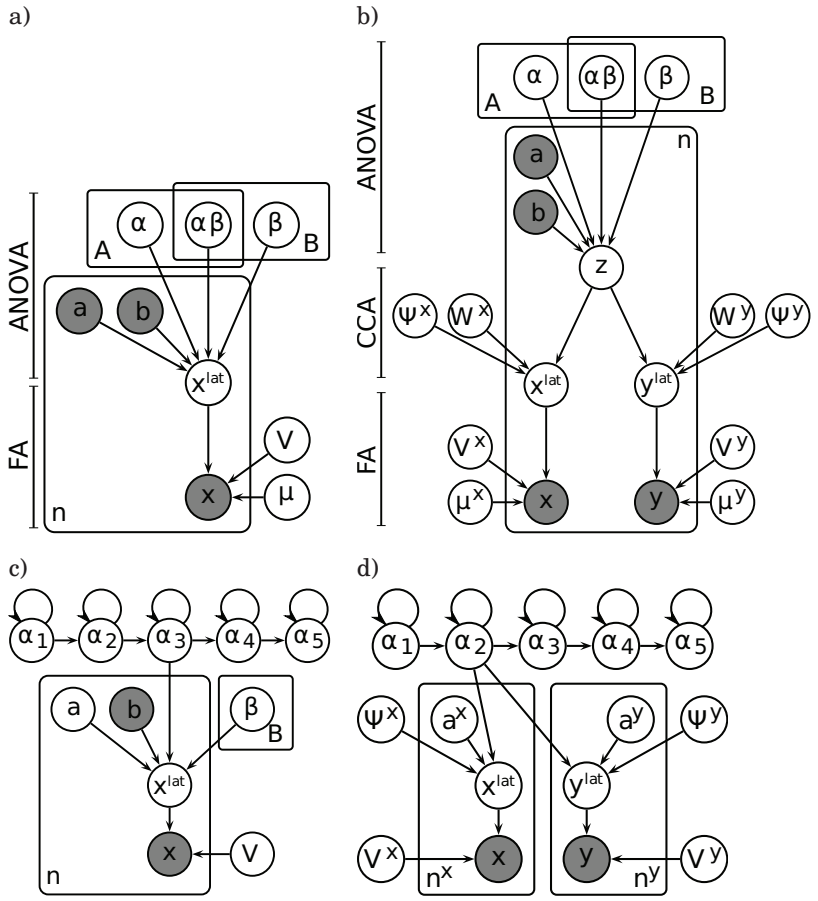

Figure 7.2. The introduced model variants. (a) The hierarchical latent-variable model for standard multi-way modelling of a single data source with standard covariates, under "large p, small n" conditions, (b) model for the multi-source case with paired samples, (c) time with unknown alignment, (d) multi-source case without paired samples, coupled only by shared latent effects (time-course with unknown alignment in this case). Reprinted with kind permission from Springer Science and Business Media: Publication III, Figure 2.

that act on both data sources. We here define that modelling shared effects of multiple covariates and their interactions is the underlying assumption of the integration of multiple data sources in the context of a multi-way design.

The idea of considering the source as a covariate and solving the multi-way, multi-source problem by a hierarchical machine learning model is valid for both paired and unpaired samples. In the case of unpaired samples, both data sources need to have a similar experimental design for the above definition to work.

7.2.2 Paired samples

The known pairing of the samples can be used for integrating the data sources [Publication II] by including the generative model of CCA [133, 134] to the unified model. This is illustrated in Figure 7.2 and in Figure 3 in Publication II. The equations of the hierarchical generative model are:

$$\begin{aligned}
\boldsymbol{\alpha}_0 = 0, \boldsymbol{\beta}_0 = 0, (\boldsymbol{\alpha}\boldsymbol{\beta})_{a0} = 0, (\boldsymbol{\alpha}\boldsymbol{\beta})_{0b} = 0 \\
\boldsymbol{\alpha}_a, \boldsymbol{\beta}_b, (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}, \boldsymbol{\alpha}^x_a, \boldsymbol{\beta}^x_b, (\boldsymbol{\alpha}\boldsymbol{\beta})^x_{ab} \sim \mathcal{N}(0, \mathbf{I}) \\
\mathbf{z}_j|_{j\in a,b} \sim \mathcal{N}(\boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}, \mathbf{I}) \\
\mathbf{z}^x_j|_{j\in a,b} \sim \mathcal{N}(\boldsymbol{\alpha}^x_a + \boldsymbol{\beta}^x_b + (\boldsymbol{\alpha}\boldsymbol{\beta})^x_{ab}, \mathbf{I}) \\
\mathbf{x}^{lat}_j \sim \mathcal{N}(\mathbf{W}^x_{\text{shared}}\mathbf{z}_j + \mathbf{W}^x_{\text{specific}}\mathbf{z}^x_j, \boldsymbol{\Psi}^x) \\
\mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu}^x + \mathbf{V}^x\mathbf{x}^{lat}_j, \boldsymbol{\Lambda}^x),
\end{aligned} \tag{7.5}$$

and similarly for $\mathbf{y}$. If shared variation in the two data sources can be modelled by a shared latent space $\mathbf{z}$ such that statistically significant shared effects $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$ and $(\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}$ can be found, then shared covariate-related variation was present in the two data sources.

The generative model has an intuitive interpretation: diseases $\boldsymbol{\alpha}_a$, treatments $\boldsymbol{\beta}_b$, and their interactions $\boldsymbol{\alpha}\boldsymbol{\beta}_{ab}$ have a shared effect on the entire organism $\mathbf{z}$ of individual j. This shared effect affects multiple tissues $\mathbf{x}^{lat}$ and $\mathbf{y}^{lat}$ and activates multiple pathways in each tissue; each pathway is represented by one dimension of $\mathbf{x}^{lat}$ or $\mathbf{y}^{lat}$. Each pathway then affects the concentrations of a cluster of correlated lipids in the observed data $\mathbf{x}$ and $\mathbf{y}$.

7.2.3 No paired samples

The case of no paired samples, motivated by the translation of biomarkers between two data sources with a similar covariate structure, is sta-

tistically much more difficult since without paired samples there does not exist a shared latent space $\mathbf{z}$. The learning task becomes a matching problem; finding groups of variables that respond similarly to multiple standard covariates in the two data spaces (species). The definition of the data integration task remains the same: search for shared covariate-related effects $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$, and $(\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}$. There is a dependence between the data sources if statistically significant effects can be found.

Two algorithms have been presented in this dissertation to solve the matching problem. In Publication III, a simple matching algorithm was presented to find only the shared response in the two sources. In Publication IV, a more advanced algorithm was presented, attempting to estimate also the source-specific effects. In Publication IV, two covariates where included; one of them being a covariate with partly unknown structure, namely, unaligned time.

7.3 Multi-level covariate

I have so far dealt with binary covariates, such as healthy/diseased or male/female, however, a discrete covariate can in general have multiple levels. Typical examples of a multi-level covariate are time, having multiple measurement times or age groups, and treatment, as there are often multiple drug alternatives or other treatments.

The model developed in this dissertation is applicable to experimental designs where one or several of the covariates in the multi-way design have multiple levels. There is, indeed, a time-series setting in several applications of this dissertation. In the second application of Publication I, age is one of the relevant covariates and we binned it to age groups. This is a working formulation, but from the point of view of time-series analysis it is a naive solution. In this application, disease is the other covariate and the interaction of time and disease has the interpretation of an age-specific disease effect.

In publications III and IV, the same dataset [14] was analyzed in a more advanced manner. There were two multi-level covariates: disease state and development state. The dataset was accompanied with antibody information, which allowed us to divide the disease covariate into multiple (fixed) disease development states. As for the development state covariate, it had an *a priori* unknown structure that we wanted to learn from the data.

As another example, we formulated the treatment covariate in the multi-way design of Publication V as a multi-level covariate that has levels 'treatment', 'placebo', and 'untreated', as will be explained in Section 7.7.

7.4 Covariate having partly unknown structure

We have also extended our model to multi-way designs where one of the covariates has a partly unknown structure. As a case study, we have used time with unknown alignment that is usually modelled by Hidden Markov Models (HMM).

The HMMs are a common time-series model in the machine learning literature; the idea is to align time-points into latent HMM-states. We have taken the point of view that a HMM model can be seen as a one-way multi-level ANOVA-model where the HMM-state of each sample is a covariate and the mean parameter of each HMM-state is the covariate effect. The difference to a standard one-way ANOVA-model is that the covariates associated with the samples are *a priori* unknown, but they can be learned from the data simultaneously with learning the mean parameters of HMM-states.

In this dissertation, I present a multi-way model where the unaligned time is one of the covariates together with others, such as disease. The previously unknown structure of time was learned jointly while learning the multi-way model. However, this concept is not limited to time and HMMs; there can be other similar applications where the previously unknown structure of a covariate has to be learned. Factorial HMM [153] is a related model where the observed data is generated by the combined effects of multiple HMM chains.

Another note in the relationship of this work to HMMs is that HMM is a generative model and the HMM-state is assumed to emit the observed data sample from the emission distribution defined by the mean parameter of each HMM-state. In our work, the HMM chain and the other covariate effects emit a latent variable together; this latent variable in turn generates the actual data point. This is consistent with our standard multi-way model where the covariate effects generate a latent variable, which in turn generates the data sample.

We have used standard formulations of the Bayesian HMM [154, 155]. The model structure is a restricted linear HMM-structure where only self-transitions and transitions to the next state are allowed.

7.5 Using the Bayesian posterior distribution to perform a statistical test

It is worth noticing that in this dissertation, the posterior distribution of the effects of covariates is used in order to perform a statistical test. When the model parameters have been learned from data, the posterior distribution of the covariate effects gives directly the statistical significance of the effects; a posterior distribution consistently above (below) zero implies a consistent positive (negative) effect with some level of significance. We have used a widely adopted threshold of 95%. Since the prior of the effects is zero-mean Gaussian, the posterior distribution models simultaneously two things: whether the effect is non-zero (significant) and whether the effect is positive or negative, implying an up-regulation or down-regulation, respectively.

A similar idea of variable selection and statistical hypothesis testing has been used in the Sparse factor regression models [25, 29, 32, 108, 100] and other related models using the sparsity-inducing point-mass mixture priors (Spike and Slab priors). In these works, the Spike and Slab prior models the regression coefficient between the observed variable and the latent factors or covariates. Whereas our prior for a covariate-effect is Gaussian, the Spike and Slab is a mixture of point mass at zero and a Gaussian distribution. This construction induces sparsity since the point-mass component forces most regression coefficients to zero. In our model, the statistical testing is done one level higher up in the hierarchy between the covariates and latent factors and the clustering of the variables at a lower level makes the components sparse. Since the dimensionality of the latent factor space is low, the effects can be estimated without assuming sparsity.

7.6 Repeated measures

Repeated measures ANOVA, also known as pairwise comparisons, is an important and well-founded concept in statistical testing and in LMMs. The idea is that, if the measurement has been done from each individual before and after a treatment, it is advantageous to take into account that the two measurements come from the same individual. Including the individual-specific effects in the ANOVA-model, in other words doing pairwise comparisons, makes the statistical test stronger since the

individual-specific variation is taken into account.

The experimental design of Publication V included such pairwise comparisons: measurements had been done before and after drug treatments. We found out that modelling pairwise comparisons can be included in our model by including an individual-specific effect. The assumption we make is that individual-specific variation affects clusters of variables.

7.7 Imperfect multi-way design

In Publication V, we presented an application of the model developed in this thesis on a novel lipidomic drug study where the dataset had an interesting experimental design: an imperfect design. Whereas a standard two-way design with binary covariates would include four populations of measurements, in this study there were three populations: two groups of human patients had been given fenofibrate drug treatment and another group had been given a placebo treatment. Fenofibrate raises Homocysteine (Hcy) levels of patients after treatment in varying amounts. A population from the highest quartile of elevated Hcy levels (called high Hcy group) and a population from the lowest quartile (called low Hcy group) had been chosen for the study. Placebo does not raise Hcy levels and, therefore, a placebo group with high Hcy concentrations is biologically not possible. Lipidomics measurements had been taken before and after treatment.

The research question was to study the effect of the fenofibrate treatment (compared to the placebo group) in lipid profiles and, additionally, to study the difference between high Hcy and low Hcy groups. The experimental design is clearly of multi-way nature, however, since there does not exist a low Hcy - placebo group, there is not a unique way to formulate this design for multi-way data analysis. In addition, there was a repeated measures design.

The aim was to formulate a single multi-way model that can answer both research questions: (i) effect of fenofibrate, and (ii) difference between low Hcy and high Hcy groups. This joint modelling task was solved in Publication V in the following way. All the samples taken before fenofibrate or placebo treatment belong to the control group ('before treatment'). The treatment covariate is a multi-level covariate having three levels: fenofibrate treatment, placebo and 'before treatment'. Hcy level (high/low) is the second covariate, however, it applies only to the patients treated

with fenofibrate and hence the multi-way design is imperfect. By using this formulation, we were able to estimate the effects of the fenofibrate treatment, placebo and the interaction of fenofibrate and Hcy level. Evaluating the statistical significance of these effects answered the two research questions. We also found the anticipated result that, for most lipid clusters, there is no placebo effect.

7.8 Biological prior knowledge of existing clusters

If prior information on lipid or gene clusters is available and it is desirable to use, it can be included directly as prior probabilities of clustering, see Publication I.

7.9 Model complexity selection

An important issue in Bayesian modelling is to compare different models and choose the one that best explains the data. In this dissertation, model comparison is a relevant issue in choosing the number of clusters and it was solved using predictive likelihood, as explained in Publication I. In the multi-source cases, the selection of the number of clusters is done for each data source separately. Another issue is choosing the number of HMM states. However, since solving both model complexity selection tasks simultaneously is a difficult problem, we have so far simply chosen the number of HMM states *a priori* according to an earlier HMM analysis on the same dataset [33].

7.10 Summary

In summary, the applicability area of the model presented in this dissertation covers most single-source or data integration (paired or non-paired samples) multi-way modelling tasks, possibly with one covariate with a partly unknown structure. The main message is that, although these data analysis tasks are very complicated, they can be formulated as a multi-way ANOVA-type problem where the design can consist of standard covariates and possibly multiple data sources.

8. Future improvements

In this chapter, I present potential technical improvements that would make our model even more usable in multi-way modelling of omics data.

8.1 Multimodality of the posterior distribution

Multiple modes of the posterior distribution of complicated Bayesian models is a known problem of MCMC methods, including Gibbs sampling used in this dissertation. When sampling multimodal posterior distributions, an MCMC chain can get stuck in one of the modes. The complicated multi-way (multi-source) model presented in this dissertation also encounters this general problem. In practical data analysis, the problem is that a unique globally optimal solution cannot be guaranteed and different MCMC chains give slightly different results, although our results have been relatively consistent. When modelling real-world datasets, there can be multiple modes and if clusterings obtained from parallel chains are different, combining these results can be somewhat difficult.

Two approaches could help to solve this problem. One is to replace MCMC with another approximate inference method, for instance variational approximation. Unfortunately, no currently existing approximate inference method is guaranteed to find a global optimum for complicated models. Another solution is to develop a proper approach for combining information from multiple chains.

8.2 Multiple components in CCA

The current multi-source method with paired samples [Publication II] estimates only one shared component between the data sources and one source-specific component for each data source. However, in a complex

high-dimensional real-world dataset, there can be more than one shared effect that have different responses to multiple covariates. Therefore, more than one component might be required to uncover all the relevant information from the data. In the formulation of our model, the shared effect is represented by a CCA-component having contribution from one or more clusters in each data source. It was demonstrated in Publication II that the method can find the shared effect and one source-specific effect in each data source simultaneously.

Since our modelling approach is an extension of CCA which is normally used for estimating multiple shared components, it is conceptually straightforward to include multiple shared components in our model. In practice, however, there is a challenge of unidentifiability of multiple components when using the Bayesian CCA model. The existing solution for the unidentifiability of probabilistic CCA [156] is not applicable here because our model is a modification of CCA where the effects of the covariates modelled as hyperpriors. Furthermore, it is difficult to simultaneously choose the number of shared components and the number of source-specific components in each data source. Solving these issues is an active topic in CCA research, and one recent promising solution has been found [157].

8.3 Finding the optimal number of clusters

Model complexity selection is one of the fundamental research problems in machine learning. We have constructed our Gibbs sampler such that the number of clusters is fixed *a priori* in order to have a parameter space with constant dimensionality. The clusters can also become empty. The task of finding the optimal number of clusters has been solved in this thesis by a predictive likelihood approach, as explained in Publication I. Although this is a well justified approach, in practical data analysis it is somewhat inconvenient and computationally extensive due to the use of cross-validation. The practicality of the method would improve by a better approach for choosing the number of clusters.

Non-parametric Bayesian methods, such as Dirichlet processes [158] and Indian Buffet process [159], are one promising approach to deal with the uncertainty in the number of clusters or components. There are, however, unresolved practical difficulties in using these methods, and they do not yet offer a working solution for our problem.

8.4 Modelling non-linearity of biological data

The question of linearity vs. non-linearity of biological data is a common issue, many experts claiming biological data is non-linear. Our model is a linear model as it follows the traditional concept of ANOVA. Since we are, however, using a hierarchical Bayesian model, we can replace Gaussian distributional assumptions by non-linear distributions.

However, learning non-linear relationships from the data requires considerably more data points than learning linear relationships, and non-linear models have an even more serious risk of overfitting. When the modelling takes place in "small n, large p"-conditions, it is questionable whether more complex relationships than linearity can be learned. Allowing non-linear relationships might also require redefining the statistical significance concepts for non-linear ANOVA, although the topic has already been studied to some extent in the formalism of Generalized Linear Models.

8.5 (M)ANCOVA-type modelling

The current version of our model cannot be readily applied to (M)ANCOVA-type modelling, mixture of discrete and continuous-valued covariates. However, continuous-valued clinical variables, such as BMI and age, are common covariates in multi-way designs of omics datasets. There are two possible approaches for taking continuous-valued covariates into account. Firstly, the multi-way, multi-source model [Publication II] can be applied such that the omics data are one data source and the continuous-valued clinical variables are another data source. Secondly, the model can be modified by changing the distributional assumptions of the covariates to allow also continuous-valued covariates.

8.6 Single-variable clusters

As a result of the clusteredness assumption, we have implicitly defined multivariate multi-way modelling as modelling the effects of covariates and their interactions on clusters of variables, which complements the traditional approach of searching for single-molecule biomarkers. Instead of two separate analyses, it would be desirable to search for single-variable markers and up- and down-regulations of clusters of similarly-behaving

molecules jointly. The current model can find clusters consisting of only a single variable; in practice, few are found. To improve the practical usefulness of the method, the flexibility of Bayesian modelling could be utilized to further encourage the method to find also single-variable clusters. In this way, the improved method could do joint univariate and multivariate modelling.

9. Conclusions

"Factorial designs will become *de rigeur* within molecular and genome biology in the way they were in the early 20th century in agricultural research, and the need for relevant statistical analysis tools will be evermore central" - (Seo, Goldschmidt-Clermont and West, 2007)

In this dissertation, I have provided a formal theoretical and feasible practical approach for multivariate multi-way modelling of high-dimensional, small sample-size, single-source and multi-source datasets. The multi-way modelling task, modelling the effects of multiple covariates and their interactions in the data, is increasingly common in the analysis of data coming from current biomedical research. Omics datasets also come increasingly often from multiple sources (different omics types, tissues, species). Multi-way modelling has a long tradition in classical statistics where the problem has been traditionally defined and solved by the Analysis of Variance (ANOVA). However, ANOVA is a univariate method and cannot be adequately used for high-dimensional data, which results in a need of multivariate methods. We have developed such a method and an additional theoretical contribution of this dissertation is that we have generalized multivariate multi-way modelling to multiple data sources and to cases where one of the covariates has a partly unknown structure.

9.1 Contribution to single-source multi-way modelling

Since biomedical omics experiments are increasingly common and they usually have an underlying multi-way experimental design, there is a great need for methods capable of multivariate multi-way modelling of high-dimensional data with a small number of samples. As a result of methodological difficulties and lack of suitable multivariate multi-way

methods, data analysis of biomedical data is usually done by simpler approaches, such as (one-way) binary classifiers, univariate statistical tests or unsupervised methods (PCA and clustering). When using simpler methods, one has to either neglect the multi-way setup, which introduces confounding factors, or to stratify the analysis into multiple case-control comparisons. In the latter case, multiple separate models have to be learned, each with even a smaller number of samples available. Using such simpler approaches cannot lead to as good an interpretation of the data as full multi-way modelling, and there are technical issues involved.

The first goal of this dissertation was to develop a formal unified Bayesian model where the statistical significance testing is done in a reduced-dimensional latent factor space. In this way, the uncertainties between the model parts propagate properly, which is crucial when the number of samples is small.

Few other approaches have been presented for multivariate multi-way modelling of high-dimensional data. The Bayesian sparse latent factor regression model [25, 100] is closely related to our model. However, despite the various applications where the authors have applied their model, to my knowledge, they have not developed a model variant that does the statistical testing of the effects of multiple covariates and their interactions in the multivariate sense, that is, for groups of variables.

9.2 Contribution to multi-source, multi-way modelling

To my knowledge, no previous methods have been presented for multivariate multi-way modelling of multi-source data, and this dissertation is the first approach of the kind. The analysis of experimental omics data with a multi-way, multi-source setup is usually done using the same standard tools as in the analysis of single-source data. The analysis is done separately for each data source or the data sources are concatenated.

In this dissertation we have defined that the underlying assumption for integrating the data sources is the similar response in multiple datasets to multiple covariates and their interactions, which can be modelled by a latent shared effect. This holds true for both common cases: paired samples and no paired samples.

In the case of paired samples, integrating multiple different omics data types and integrating data from multiple tissues are the two important application areas. There are a few existing data integration approaches

for multi-source data; the problem of unsupervised approaches is that covariate information is not taken into account and supervised approaches are limited to one-way cases. As for the case of no paired samples (cross-species translation), most existing data integration approaches assume *a priori* matched variables. Our model has been developed for the more general case where such matching information is not available, but the actual task is to find the matchings based on the data. The existing methods that do not assume *a priori* known matching between the variables, do not take the covariate information into account. Our translation model thus contributes to the cross-species modelling line of research as a feasible approach for finding unknown matching of molecules between two species, additionally making it possible to translate biomarker-type findings between the species.

We were successful in developing a formal multi-way, multi-source model. Using a unified modelling approach makes it possible to define a proper statistical analysis question and obtain a holistic view of the biological system even when the dimensionality of the data, the number of data sources or the number of covariates grows.

9.3 On the results obtained

Our results showed that clusteredness of variables is a very good assumption for lipidomics data and we were successful in utilizing the assumption for dimensionality reduction to overcome the "small n, large p"- difficulty. Having clusters of variables also helps to interpret the data. Modelling clusters of correlated variables is also a good approach against the problem of multiple testing since the number of statistical tests decreases essentially from the dimensionality of the data to the number of clusters. Strongly correlated lipid clusters were found in all the lipidomic data sets studied in the applications. A few preliminary feasibility studies have shown that the assumption of clusteredness works well for gene expression and gut microbiota data as well [data not shown]. In particular, the results showed that all the lipids belonging to a cluster up- or down-regulate similarly as a response to one or multiple external covariates and their interactions, and statistically significant multivariate effects were found despite the extremely small number of samples available. Also, a desired result was that in each study, most clusters are unaffected by external covariates; this suggests that the method truly finds only the vari-

ables affected by the covariates and is not prone to overfitting, as many classifiers are. In this sense, the clusteredness assumption also helps to guard against single-variable false positives.

We also succeeded in applying the model as a ready tool to a novel 600-dimensional dataset in Publication V and the model was able to find multiple main effects and interaction effects. Furthermore, the extensibility of our Bayesian model was demonstrated here when the dataset had both an imperfect design and a repeated measures design. We were able to formulate the task as a multi-way modelling problem and extend the model to take this design into account properly within the unified model, which would not have been fully possible by standard statistical tools.

In the data integration cases with paired samples and without paired samples, the task was to find shared effects between the datasets. The results showed that the clustering also works as the necessary dimensionality reduction component for the complicated data integration cases, bringing the information in high-dimensional data to the shared latent variable space. The results showed that the method was capable of finding statistically significant shared effects between multiple data sources and source-specific effects.

We realize that the problems in sampling multimodal posterior distributions by MCMC methods concern our work as well, but our model seems to work well and fulfills the standard requirement that results from parallel MCMC chains are consistent. In this work, the models have been validated with experiments on simulated data. More extensive Bayesian model criticism has been left for future studies. Nevertheless, results obtained from our model from real lipidomics data were consistent with results obtained by simpler statistical methods: 50-50 MANOVA [Publication II], basic statistical analysis done in Publication V, and unpublished preliminary feasibility studies using the basic approaches: univariate tests, PLS-DA, and PCA [data not shown].

One challenge of traditional ANOVA-analysis are unbalanced designs, where different groups have different numbers of samples. The performance of our model in the case of unbalanced designs has not been systematically studied.

9.4 Multi-way learning

As a result of this dissertation, I claim that multi-way learning should be seen as its own machine learning genre. There are three leading trends in the machine learning community that are closely related to multi-way learning: multi-task learning, multi-label prediction, and multi-class classification. The existing methods in these genres solve different, usually discriminative tasks and these methods are not designed to do multi-way ANOVA-type modelling, which is a generative modelling task.

10. Discussion

10.1 Applicability of the model to other data types

The model I have presented is applicable to single-source or multi-source omics datasets: gene expression, proteomics, metabolomics (lipidomics) and their combination. This model can handle most modelling problems that can be formulated as a multi-way experimental design including any multi-tissue or any data integration setting having multiple types of omics data with paired measurements from each individual. In cross-species translation type of studies (no paired samples), two datasets from different species, with a similar experimental design, can be integrated with this model in order to match the variables and find a shared response to the multiple covariates. As an example of a such biological study, see [49]. Both data integration model variants have been developed to be used under the less restrictive assumption of no *a priori* known matching between the variables from different data sources. If matching between the variables is available, more powerful statistical approaches can be applied by either extending this model to take the pairing information into account or by using other models.

Although I have focused on discussing biomedical omics data in this dissertation, the model is readily applicable to other continuous-valued data types where the clusteredness assumption of variables can be made and where the research question is multi-way modelling of single-source or multi-source data. One such potential data type is gut microbiota [43, 44, 45, 46], which is assumed to consist of relative amounts of bacteria in bacterial populations (clusters). Gut microbiota datasets often have similar experimental designs and research questions and are often integrated with omics data. Another possible data type is Functional Mag-

netic Resonance Imaging (fMRI) data, which consists of intensity valued voxels with heavy correlations between neighboring voxels. The fMRI data are high-dimensional and require dimensionality reduction. Data integration settings for fMRI data are attracting increasing interest, although the approaches presented so far have been basic many-step approaches [160, 161]. Integrating MRI data with omics data [35, 162] is an example of an even more exciting scientific perspective that can be sought by data integration.

10.2 Future use of unified Bayesian multi-way models

As a result of this dissertation, I claim that unified Bayesian generative modelling is a good approach for holistic analysis of complex datasets that consist of an increasingly growing mixture of multi-way, multi-source and even multi-species settings. The results of this dissertation suggest that building unified models for analyzing data from complicated experimental setups is possible; the flexibility of the Bayesian formalism made it possible to include all the necessary prior information of the experimental design as necessary building blocks for a unified model: integration of multiple data sources, multi-way modelling and dimensionality reduction in the form of clusteredness assumption.

As large biobanks [1, 163, 164, 165, 166, 167] become available, there will be a growing need to integrate omics data with clinical data and genetic sequence data. Although the current version of our method is not readily applicable to discrete data types, the unified modelling approach of this dissertation is a good starting point for integrating new data types as additional data sources. As sub-components of Bayesian hierarchical models can be changed flexibly, new data types can be incorporated easily.

A few technical details remain unsolved so far, namely how to tackle the multimodality of complex posterior distribution and how to find multiple components in the case of multiple data sources with paired samples. At the current stage, the model can be used as an exploratory tool for finding interesting multi-way, multi-source structures in the data. As soon as means for finding a unique solution will be found, the model presented in this dissertation has potential to become the leading approach in multivariate multi-way modelling of high-dimensional data.

Bibliography

[1] J. Kaiser. Population databases boom, from Iceland to the U.S. *Science*, 298(5596):1158–1161, 2002.

[2] M. Kussmann, F. Raymond, and M. Affolter. Omics-driven biomarker discovery in nutrition and health. *Journal of Biotechnology*, 124(4):758 – 787, 2006.

[3] M. Orešič, V. A. Hänninen, and A. Vidal-Puig. Lipidomics: a new window to biomedical frontiers. *Trends in Biotechnology*, 26(12):647 – 652, 2008.

[4] H. Atherton, M. K. Gulston, N. Bailey, K.-K. Cheng, W. Zhang, K. Clarke, and J. Griffin. Metabolomics of the interaction between PPAR-alpha and age in the PPAR-alpha-null mouse. *Molecular Systems Biology*, 5(259), 2009.

[5] M. Orešič. Informatics and computational strategies for the study of lipids. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1811(11):991 –999, 2011.

[6] M. Orešič, J. Tang, T. Seppänen-Laakso, I. Mattila, S. Saarni, S. Saarni, J. Lönnqvist, M. Sysi-Aho, T. Hyotyläinen, J. Perälä, and J. Suvisaari. Metabolome in schizophrenia and other psychotic disorders: a general population-based study. *Genome Medicine*, 3(3):19, 2011.

[7] G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman. Bioinformatics challenges for personalized medicine. *Bioinformatics*, 27(13):1741–1748, 2011.

[8] L. Hood. Lee Hood outlines his vision of personalized medicine for the next 10 years. *Nature Biotechnology*, 29(3):191, 2011.

[9] R. M. Salek, M. L. Maguire, E. Bentley, D. V. Rubtsov, T. Hough, M. Cheeseman, D. Nunez, B. C. Sweatman, J. N. Haselden, R. D. Cox, S. C. Connor, and J. L. Griffin. A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. *Physiological Genomics*, 29(2):99–108, 2007.

[10] W. Tillett and J. T. Francis. Serological reactions in pneumonia with a nonprotein somatic fraction of pneumococcus. *Journal of Experimental Medicine*, 52:561–585, 1930.

[11] S. Mehta, A. Shelling, A. Muthukaruppan, A. Lasham, C. Blenkiron, G. Laking, and C. Print. Predictive and prognostic molecular markers for

cancer medicine. *Therapeutic Advances in Medical Oncology*, 2(2):125–148, 2010.

[12] Y. Lu, P. Huggins, and Z. Bar-Joseph. Cross species analysis of microarray expression data. *Bioinformatics*, 25(12):1476–1483, 2009.

[13] M. Assfalg, I. Bertini, D. Colangiuli, C. Luchinat, H. Schäfer, B. Schütz, and M. Spraul. Evidence of different metabolic phenotypes in humans. *Proceedings of the National Academy of Sciences*, 105(5):1420–1424, 2008.

[14] M. Orešič, S. Simell, M. Sysi-Aho, K. Nanto-Salonen, T. Seppänen-Laakso, V. Parikka, M. Katajamaa, A. Hekkala, I. Mattila, P. Keskinen, L. Yetukuri, A. Reinikainen, J. Lähde, T. Suortti, J. Hakalax, T. Simell, H. Hyöty, R. Veijola, J. Ilonen, R. Lahesmaa, M. Knip, and O. Simell. Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. *Journal of Experimental Medicine*, 205(13):2975–2984, 2008.

[15] F. van der Kloet, F. Tempels, N. Ismail, R. van der Heijden, P. Kasper, M. Rojas-Cherto, R. van Doorn, G. Spijksma, M. Koek, J. van der Greef, V. Mäkinen, C. Forsblom, H. Holthöfer, P. Groop, T. Reijmers, and T. Hankemeier. Discovery of early-stage biomarkers for diabetic kidney disease using MS-based metabolomics (FinnDiane study). *Metabolomics*, 8:109–119, 2012.

[16] J. Davis, E. Lantz, D. Page, J. Struyf, P. Peissig, H. Vidaillet, and M. Caldwell. Machine learning for personalized medicine: Will this drug give me a heart attack? In *Proceedings of Machine Learning in Health Care Applications Workshop. In conjunction with ICML 2008*, 2008.

[17] T. I. W. P. Consortium. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764, 2009.

[18] M. Welsh, L. Mangravite, M. W. Medina, K. Tantisira, W. Zhang, R. S. Huang, H. McLeod, and M. E. Dolan. Pharmacogenomic discovery using cell-based models. *Pharmacological Reviews*, 61(4):413–429, 2009.

[19] I. G. Costa, A. Schonhuth, C. Hafemeister, and A. Schliep. Constrained mixture estimation for analysis and robust classification of clinical time series. *Bioinformatics*, 25(12):i6–14, 2009.

[20] A. Julià, A. Erra, C. Palacio, C. Tomas, X. Sans, P. Barceló, and S. Marsal. An eight-gene blood expression profile predicts the response to infliximab in rheumatoid arthritis. *PLoS ONE*, 4(10):e7556, 2009.

[21] H. K. Dressman, A. Berchuck, G. Chan, J. Zhai, A. Bild, R. Sayer, J. Cragun, J. Clarke, R. S. Whitaker, L. Li, J. Gray, J. Marks, G. S. Ginsburg, A. Potti, M. West, J. R. Nevins, and J. M. Lancaster. An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *Journal of Clinical Oncology*, 25(5):517–525, 2007.

[22] R. Kaddurah-Daouk, R. A. Baillie, H. Zhu, Z.-B. Zeng, M. M. Wiest, U. T. Nguyen, K. Wojnoonski, S. M. Watkins, M. Trupp, and R. M. Krauss. Enteric microbiome metabolites correlate with response to simvastatin treatment. *PLoS ONE*, 6(10):e25482, 2011.

[23] E. P. Diamandis. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool. *Molecular & Cellular Proteomics*, 3(4):367–378, 2004.

[24] G. Medina-Gomez, S. L. Gray, L. Yetukuri, K. Shimomura, S. Virtue, M. Campbell, R. K. Curtis, M. Jimenez-Linan, M. Blount, G. S. H. Yeo, M. Lopez, T. Seppänen-Laakso, F. M. Ashcroft, M. Orešič, and A. Vidal-Puig. PPAR gamma 2 prevents lipotoxicity by controlling adipose tissue expandability and peripheral lipid metabolism. *PLoS Genetics*, 3(4):e64, 2007.

[25] M. West. Bayesian factor regression models in the large p, small n paradigm. *Bayesian Statistics*, 7:723–732, 2003.

[26] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[27] W. Mostertz, M. Stevenson, C. Acharya, I. Chan, K. Walters, W. Lamlert-thon, W. Barry, J. Crawford, J. Nevins, and A. Potti. Age- and sex-specific genomic profiles in non–small cell lung cancer. *JAMA: The Journal of the American Medical Association*, 303(6):535–543, 2010.

[28] D. Broadhurst and D. Kell. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics*, 2:171–196, 2006.

[29] D. M. Seo, P. J. Goldschmidt-Clermont, and M. West. Of mice and men: Sparse statistical modelling in cardiovascular genomics. *Annals of Applied Statistics*, 1(1):152–178, 2007.

[30] S. Hao, J. Xin, J. Lian, Q. Xie, D. Chen, Y. Guo, Y. Lu, G. Sheng, W. Xu, J. Huang, and L. Li. Establishing a metabolomic model for the prognosis of hepatitis b virus-induced acute-on-chronic liver failure treated with different liver support systems. *Metabolomics*, 7:400–412, 2011.

[31] D. Damian, M. Orešič, E. Verheij, J. Meulman, J. Friedman, A. Adourian, N. Morel, A. Smilde, and J. van der Greef. Applications of a new subspace clustering algorithm (COSA) in medical systems biology. *Metabolomics*, 3:69–77, 2007.

[32] J. E. Lucas, C. M. Carvalho, J. L.-Y. Chen, J.-T. Chi, and M. West. Cross-study projections of genomic biomarkers: An evaluation in cancer genomics. *PLoS ONE*, 4(2):e4523, 2009.

[33] J. Nikkilä, M. Sysi-Aho, A. Ermolov, T. Seppänen-Laakso, O. Simell, S. Kaski, and M. Orešič. Gender dependent progression of systemic metabolic states in early childhood. *Molecular Systems Biology*, 4:197, 2008.

[34] J. Tang, C. Tan, M. Orešič, and A. Vidal-Puig. Integrating post-genomic approaches as a strategy to advance our understanding of health and disease. *Genome Medicine*, 1(3):35, 2009.

[35] M. Orešič, J. Lötjönen, and H. Soininen. Systems medicine and the integration of bioinformatic tools for the diagnosis of Alzheimer's disease. *Genome Medicine*, 2(11):83, 2010.

[36] G. Blekherman, R. Laubenbacher, D. Cortes, P. Mendes, F. Torti, S. Akman, S. Torti, and V. Shulaev. Bioinformatics tools for cancer metabolomics. *Metabolomics*, 7:329–343, 2011.

[37] J. E. Lucas, C. M. Carvalho, D. Merl, and M. West. In-vitro to in-vivo factor profiling in expression genomics. In *Bayesian Modelling in Bioinformatics*, pages 293–316. Taylor-Francis, 2010.

[38] T. Risby and S. Solga. Current status of clinical breath analysis. *Applied Physics B: Lasers and Optics*, 85:421–426, 2006.

[39] M. C. Wu, L. Zhang, Z. Wang, D. C. Christiani, and X. Lin. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9):1145–1151, 2009.

[40] H. Zhang, A. Y. Liu, P. Loriaux, B. Wollscheid, Y. Zhou, J. D. Watts, and R. Aebersold. Mass spectrometric detection of tissue proteins in plasma. *Molecular & Cellular Proteomics*, 6(1):64–71, 2007.

[41] C. Hu, H. Wei, A. M. van den Hoek, M. Wang, R. van der Heijden, G. Spijksma, T. H. Reijmers, J. Bouwman, S. Wopereis, L. M. Havekes, E. Verheij, T. Hankemeier, G. Xu, and J. van der Greef. Plasma and liver lipidomics response to an intervention of rimonabant in apoe*3leiden.cetp transgenic mice. *PLoS ONE*, 6(5):e19423, 05 2011.

[42] A. Joyce and B. Palsson. The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology*, 7:198–210, 2006.

[43] V. R. Velagapudi, R. Hezaveh, C. S. Reigstad, P. Gopalacharyulu, L. Yetukuri, S. Islam, J. Felin, R. Perkins, J. Borén, M. Orešič, and F. Bäckhed. The gut microbiota modulates host energy and lipid metabolism in mice. *Journal of Lipid Research*, 51(5):1101–1112, 2010.

[44] P. Turnbaugh, R. Ley, M. Hamady, C. Fraser-Liggett, R. Knight, and J. Gordon. The human microbiome project. *Nature*, 449:804–810, 2007.

[45] P. Kovatcheva-Datchary, E. G. Zoetendal, K. Venema, W. M. de Vos, and H. Smidt. Review: Tools for the tract: understanding the functionality of the gastrointestinal tract. *Therapeutic Advances in Gastroenterology*, 2(4 suppl):9–22, 2009.

[46] M. Orešič, T. Seppänen-Laakso, L. Yetukuri, F. Bäckhed, and V. Hänninen. Gut microbiota affects lens and retinal lipid composition. *Experimental Eye Research*, 89(5):604 – 607, 2009.

[47] W. R. Wikoff, A. T. Anfora, J. Liu, P. G. Schultz, S. A. Lesley, E. C. Peters, and G. Siuzdak. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proceedings of the National Academy of Sciences*, 106(10):3698–3703, 2009.

[48] M. Sysi-Aho, A. Vehtari, V. Velagapudi, J. Westerbacka, L. Yetukuri, R. Bergholm, M.-R. Taskinen, H. Yki-Järvinen, and M. Orešič. Exploring the lipoprotein composition using Bayesian regression on serum lipidomic profiles. *Bioinformatics*, 23(13):i519–528, 2007.

[49] M. Sysi-Aho, A. Ermolov, P. V. Gopalacharyulu, A. Tripathi, T. Seppänen-Laakso, J. Maukonen, I. Mattila, S. T. Ruohonen, L. Vähätalo, L. Yetukuri, T. Härkönen, E. Lindfors, J. Nikkilä, J. Ilonen, O. Simell, M. Saarela, M. Knip, S. Kaski, E. Savontaus, and M. Orešič. Metabolic regulation in progression to autoimmune diabetes. *PLoS Computational Biology*, 7:e1002257, 2011.

[50] H.-S. Le, Z. N. Oltvai, and Z. Bar-Joseph. Cross-species queries of large gene expression databases. *Bioinformatics*, 26(19):2416–2423, 2010.

[51] A. Tripathi, A. Klami, M. Orešič, and S. Kaski. Matching samples of multiple views. *Data Mining and Knowledge Discovery*, 23:300–321, 2011.

[52] M. Scheideler, C. Elabd, L.-E. Zaragosi, C. Chiellini, H. Hackl, F. Sanchez-Cabo, S. Yadav, K. Duszka, G. Friedl, C. Papak, A. Prokesch, R. Windhager, G. Ailhaud, C. Dani, E.-Z. Amri, and Z. Trajanoski. Comparative transcriptomics of human multipotent stem cells during adipogenesis and osteoblastogenesis. *BMC Genomics*, 9(1):340, 2008.

[53] R. Hall, M. Beale, O. Fiehn, N. Hardy, L. Sumner, and R. Bino. Plant metabolomics: The missing link in functional genomics strategies. *The Plant Cell Online*, 14(7):1437–1440, 2002.

[54] E. A. Dennis. Lipidomics joins the omics evolution. *Proceedings of the National Academy of Sciences*, 106(7):2089–2090, 2009.

[55] E. E. Schad, S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, P. S. Linsley, M. Mao, R. B. Stoughton, and S. H. Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422:297–301, 2003.

[56] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, 2002.

[57] S. Waaijenborg, P. C. Verselewel de Witt Hamer, and A. H. Zwinderman. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7(3), 2008.

[58] S. Lee, J. Zhu, and E. Xing. Adaptive multi-task Lasso: with application to eQTL detection. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1306–1314. MIT Press, Cambridge,MA, 2010.

[59] L. Lahti, S. Myllykangas, S. Knuutila, and S. Kaski. Dependency detection with similarity constraints. In *Proceedings of MLSP 2009, IEEE International Workshop on Machine Learning for Signal Processing*, pages 89–94. IEEE, 2009.

[60] H. M. Horlings, C. Lai, D. S. Nuyten, H. Halfwerk, P. Kristel, E. van Beers, S. A. Joosse, C. Klijn, P. M. Nederlof, M. J. Reinders, L. F. Wessels, and M. J. van de Vijver. Integration of DNA copy number alterations and prognostic gene expression signatures in breast cancer patients. *Clinical Cancer Research*, 16(2):651–663, 2010.

[61] M. Rantalainen, O. Cloarec, O. Beckonert, I. D. Wilson, D. Jackson, R. Tonge, R. Rowlinson, S. Rayner, J. Nickson, R. W. Wilkinson, J. D. Mills, J. Trygg, J. K. Nicholson, and E. Holmes. Statistically integrated metabonomic–proteomic studies on a human prostate cancer xenograft model in mice. *Journal of Proteome Research*, 5(10):2642–2655, 2006.

[62] S. Rogers, M. Girolami, W. Kolch, K. M. Waters, T. Liu, B. Thrall, and H. S. Wiley. Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics*, 24(24):2894–2900, 2008.

[63] C. J. Lelliott, M. López, R. K. Curtis, N. Parker, M. Laudes, G. Yeo, M. Jimenez-Liñan, J. Grosse, A. K. Saha, D. Wiggins, D. Hauton, M. D. Brand, S. ORahilly, J. L. Griffin, G. F. Gibbons, and A. Vidal-Puig. Transcript and metabolite analysis of the effects of tamoxifen in rat liver reveals inhibition of fatty acid synthesis in the presence of hepatic steatosis. *The FASEB Journal*, 19(9):1108–1119, 2005.

[64] K. Suhre, S.-Y. Shin, A.-K. Petersen, R. P. Mohney, and et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, 477:54–60, 2011.

[65] T. M. Teslovich, K. Musunuru, A. V. Smith, A. C. Edmondson, and et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–713, 2010.

[66] A. A. Hicks, P. P. Pramstaller, Åsa Johansson, V. Vitart, and et al. Genetic determinants of circulating sphingolipid concentrations in European populations. *PLoS Genetics*, 5(10):e1000672, 2009.

[67] K. Yizhak, T. Benyamini, W. Liebermeister, E. Ruppin, and T. Shlomi. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics*, 26(12):i255–i260, 2010.

[68] P. Waltman, T. Kacmarczyk, A. Bate, D. Kearns, D. Reiss, P. Eichenberger, and R. Bonneau. Multi-species integrative biclustering. *Genome Biology*, 11(9):R96, 2010.

[69] G. R. Lanckriet, T. D. Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[70] O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau, and B. D. Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, 22(14):e184–e190, 2006.

[71] Y. Sun, S. Goodison, J. Li, L. Liu, and W. Farmerie. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics*, 23(1):30–37, 2007.

[72] A.-L. Boulesteix, C. Porzelius, and M. Daumer. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, 24(15):1698–1706, 2008.

[73] K.-A. Le Cao, E. Meugnier, and G. J. McLachlan. Integrative mixture of experts to combine clinical factors and gene markers. *Bioinformatics*, 26(9):1192–1198, 2010.

[74] G. Nyamundanda, L. Brennan, and I. Gormley. Probabilistic principal component analysis for metabolomic data. *BMC Bioinformatics*, 11(1):571, 2010.

[75] E. Fahy, S. Subramaniam, H. A. Brown, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell, Y. Seyama, W. Shaw, T. Shimizu, F. Spener, G. van Meer, M. S. VanNieuwenhze, S. H. White, J. L. Witztum, and E. A. Dennis. A comprehensive classification system for lipids. *Journal of Lipid Research*, 46(5):839–862, 2005.

[76] R. Steuer. Review: On the analysis and interpretation of correlations in metabolomic data. *Briefings in Bioinformatics*, 7(2):151–158, 2006.

[77] A. Danielsson, T. Moritz, H. Mulder, and P. Spégel. Development of a gas chromatography/mass spectrometry based metabolomics protocol by means of statistical experimental design. *Metabolomics*, 8:50–63, 2012.

[78] P. Puri, M. M. Wiest, O. Cheung, F. Mirshahi, C. Sargeant, H.-K. Min, M. J. Contos, R. K. Sterling, M. Fuchs, H. Zhou, S. M. Watkins, and A. J. Sanyal. The plasma lipidomic signature of nonalcoholic steatohepatitis. *Hepatology*, 50(6):1827–1838, 2009.

[79] M. Katajamaa, J. Miettinen, and M. Oresic. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, 22(5):634–636, 2006.

[80] U. Sauer. High-throughput phenomics: experimental methods for mapping fluxomes. *Current opinion in biotechnology*, 15(1):58–63, 2004.

[81] I. T. Jolliffe. *Principal Component Analysis*. Springer, 1986.

[82] M. I. Jordan. Graphical models. *Statistical Science*, 19(1):140–155, 2004.

[83] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis (2nd edition)*. Chapman & Hall/CRC, Boca Raton, FL, 2003.

[84] D. J. McCarthy and G. K. Smyth. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, 25(6):765–771, 2009.

[85] J. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[86] S. Wold, A. Ruhe, H. Wold, and W. Dunn. The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5:735–743, 1984.

[87] A.-L. Boulesteix and K. Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1):32–44, 2007.

[88] J. Trygg and S. Wold. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3):119–128, 2002.

[89] F. Tai and W. Pan. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics*, 23(23):3170–3177, 2007.

[90] L. Breiman. Random forests. *Machine Learning*, 1:5–32, 2001.

[91] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

[92] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[93] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to data mining*. Pearson Addison-Wesley, 2006.

[94] B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[95] J. Westerhuis, H. Hoefsloot, S. Smit, D. Vis, A. Smilde, E. van Velzen, J. van Duijnhoven, and F. van Dorsten. Assessment of PLSDA cross validation. *Metabolomics*, 4(1):81–89, 2008.

[96] G. Cao and C. Bouman. Covariance estimation for high dimensional data vectors using the sparse matrix transform. In *Advances in Neural Information Processing Systems 21*, pages 225–232, Cambridge,MA, 2009. MIT Press.

[97] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[98] A. Qi and F. Yan. Eigennet: A Bayesian hybrid of generative and conditional models for sparse learning. In *Advances in Neural Information Processing Systems 24*, Cambridge,MA, 2011. MIT Press.

[99] E. Grave, G. R. Obozinski, and F. Bach. Trace Lasso: a trace norm regularization for correlated designs. In *Advances in Neural Information Processing Systems 24*, Cambridge,MA, 2011. MIT Press.

[100] C. Carvalho, J. Chang, J. Lucas, J. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

[101] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005.

[102] B. Efron and R. Tibshirani. On testing the significance of sets of genes. *Ann. Appl. Stat.*, 1(1):107–129, 2007.

[103] M. Kankainen, P. Gopalacharyulu, L. Holm, and M. Orešič. MPEA-metabolite pathway enrichment analysis. *Bioinformatics*, 27(13):1878–1879, 2011.

[104] L. Wang, B. Zhang, R. D. Wolfinger, and X. Chen. An integrated approach for the analysis of biological pathways using mixed models. *PLoS Genet*, 4(7):e1000115, 2008.

[105] M. Lopes, L. Jacob, and M. J. Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems 24*, Cambridge,MA, 2011. MIT Press.

[106] Y. Guan and J. Dy. Sparse probabilistic principal component analysis. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *JMLR W&CP*, pages 185–192. JMLR, 2009.

[107] R. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Royal Society of Edinburgh from Transactions of the Society*, 52:399–433, 1918.

[108] J. E. Lucas, C. M. Carvalho, Q. Wang, A. H. Bild, J. R. Nevins, and M. West. Sparse statistical modelling in gene expression genomics. In *Bayesian Inference for Gene Expression and Proteomics*, pages 155–176. Cambridge University Press, 2006.

[109] M. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7:819–837, 2000.

[110] R. Wolfinger, G. Gibson, E. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. Paules. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8(6):625–637, 2001.

[111] S. K. Ng, G. J. McLachlan, K. Wang, L. Ben-Tovim Jones, and S.-W. Ng. A Mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, 22(14):1745–1752, 2006.

[112] G. Celeux, O. Martin, and C. Lavergne. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modeling*, 5(3):243–267, 2005.

[113] J. Listgarten, C. Kadie, E. Schadt, and D. Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 2010.

[114] Z. Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493–2503, 2004.

[115] G. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(3), 2004.

[116] Y. F. Leung, P. Ma, B. A. Link, and J. E. Dowling. Factorial microarray analysis of zebrafish retinal development. *Proceedings of the National Academy of Sciences*, 105(35):12909–12914, 2008.

[117] N. Bratchell. Multivariate response surface modeling by principal component analysis. *Journal of Chemometrics*, 3:579–588, 1989.

[118] O. Langsrud. 50-50 multivariate analysis of variance for collinear responses. *Journal of the Royal Statistical Society Series D-the Statistician*, 51:305–317, 2002.

[119] A. K. Smilde, J. J. Jansen, H. C. J. Hoefsloot, R.-J. A. N. Lamers, J. van der Greef, and M. E. Timmerman. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, 21(13):3043–3048, 2005.

[120] Y. Guo. Supervised exponential family principal component analysis via convex optimization. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 569–576, Cambridge, MA, 2009. MIT Press.

[121] B.-J. M. Webb-Robertson, L. A. Mccue, N. Beagley, J. E. Mcdermott, D. S. Wunschel, S. M. Varnum, J. Z. Hu, N. G. Isern, G. W. Buchko, K. Mcateer, J. G. Pounds, S. J. Skerrett, D. Liggitt, and C. W. Frevert. A Bayesian integration model of high-throughput proteomics and metabolomics data for improved early detection of microbial infections. In *Pacific Symposium on Biocomputing*, volume 14, pages 451–463, 2009.

[122] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[123] J. Zhang, Z. Ghahramani, and Y. Yang. Flexible latent variable models for multi-task learning. *Machine Learning*, 73:221–242, 2008.

[124] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11:2901–2934, 2010.

[125] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research*, 8:35–63, 2007.

[126] G. Obozinski, B. Taskar, and M. Jordan. Multi-task feature selection. Technical report, Department of Statistics University of California, Berkeley, June 2006.

[127] K. Puniyani, S. Kim, and E. P. Xing. Multi-population GWA mapping via multi-task regularized regression. *Bioinformatics*, 26(12):i208–i216, 2010.

[128] X. Yang, S. Kim, and E. P. Xing. Heterogeneous multitask learning with joint sparsity constraints. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2151–2159. MIT Press, Cambridge,MA, 2009.

[129] Z. Barutcuoglu, R. E. Schapire, and O. G. Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

[130] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73:185–214, 2008.

[131] N. Quadrianto and C. Lampert. Learning multi-view neighborhood preserving projections. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 425–432, New York, NY, USA, 2011. ACM.

[132] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[133] F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

[134] A. Klami and S. Kaski. Local dependent components. In Z. Ghahramani, editor, *Proceedings of ICML 2007, the 24th International Conference on Machine Learning*, pages 425–432. Omnipress, 2007.

[135] C. Wang. Variational Bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, 18:905–910, 2007.

[136] C. Archambeau and F. Bach. Sparse probabilistic projections. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 73–80. MIT Press, 2009.

[137] P. Rai and H. Daume. Multi-label prediction via sparse infinite CCA. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1518–1526. MIT Press, Cambridge, MA, 2009.

[138] D. Witten and R. Tibshirani. Extensions of sparse canonical correlation analysis, with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):Article 28, 2009.

[139] E. Parkhomenko, D. Tritchler, and J. Beyene. Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proceedings*, 1(Suppl 1):S119, 2007.

[140] D. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83:331–353, 2011.

[141] S. Dolédec and D. Chessel. Co-inertia analysis: an alternative method for studying species–environment relationships. *Freshwater Biology*, 31(3):277–294, 1994.

[142] A. Culhane, G. Perriere, and D. Higgins. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, 4(1):59, 2003.

[143] K.-A. Le Cao, P. Martin, C. Robert-Granie, and P. Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10(1):34, 2009.

[144] J. Trygg and S. Wold. O2-pls, a two-block (x–y) latent variable regression (lvr) method with an integral osc filter. *Journal of Chemometrics*, 17(1):53–64, 2003.

[145] P. Agius, Y. Ying, and C. Campbell. Bayesian unsupervised learning with multiple data types. *Statistical Applications in Genetics and Molecular Biology*, 8(27), 2009.

[146] S. Monni and M. Tadesse. A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Analysis*, 4(3):413 – 436, 2009.

[147] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100. ACM, New York, NY, USA, 1998.

[148] J. He and R. Lawrence. A graph-based framework for multi-task multi-view learning. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 25–32, New York, NY, USA, June 2011. ACM.

[149] Y. Lu, R. Rosenfeld, G. J. Nau, and Z. Bar-Joseph. Cross species expression analysis of innate immune response. *Journal of Computational Biology*, 17(3):253–268, 2010.

[150] H.-S. Le and Z. Bar-Joseph. Cross species expression analysis using a Dirichlet process mixture model with latent matchings. In J. Lafferty and *et al.*, editors, *Advances in Neural Information Processing Systems 23*, pages 1270–1278. MIT Press, Cambridge,MA, 2010.

[151] A. M. Gholami and K. Fellenberg. Cross-species common regulatory network inference without requirement for prior gene affiliation. *Bioinformatics*, 26(8):1082–1090, 2010.

[152] A. Tripathi, A. Klami, and S. Kaski. Using dependencies to pair samples for multi-view learning. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, pages 1561–1564. IEEE, 2009.

[153] Z. Ghahramani and M. I. Jordan. Factorial hidden markov models. In *Machine Learning*, volume 29, pages 245–273. MIT Press, 1997.

[154] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden markov model. In *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

[155] J. V. Gael, Y. W. Teh, and Z. Ghahramani. The infinite factorial hidden markov model. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1697–1704. MIT Press, Cambridge,MA, 2009.

[156] C. Archambeau, N. Delannay, and M. Verleysen. Robust probabilistic projections. In W. Cohen and A. Moore, editors, *Proceedings of the 23rd International Conference on Machine Learning*, pages 33–40. ACM, 2006.

[157] S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 457–464, New York, NY, 2011. ACM.

[158] C. Rasmussen. The infinite Gaussian mixture model. In S. A. Solla, T. K. Leen, and K.-R. Muller, editors, *Proceedings of Neural Information Processing Systems*, pages 554–560, Cambridge, MA, 2000. MIT Press.

[159] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 475–482, Cambridge, MA, 2006. MIT Press.

[160] 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*, 48:176–185, 2009.

[161] J. Sui, T. Adali, G. Pearlson, H. Yang, S. R. Sponheim, T. White, and V. D. Calhoun. A CCA+ICA based model for multi-task brain imaging data fusion and its application to schizophrenia. *NeuroImage*, 51(1):123 – 134, 2010.

[162] M. Sysi-Aho, J. Koikkalainen, T. Seppänen-Laakso, M. Kaartinen, J. Kuusisto, K. Peuhkurinen, S. Kärkkäinen, M. Antila, K. Lauerma, E. Reissell, R. Jurkko, J. Lötjönen, T. Heliö, and M. Orešič. Serum lipidomics meets cardiac magnetic resonance imaging: Profiling of subjects at risk of dilated cardiomyopathy. *PLoS ONE*, 6(1):e15744, 01 2011.

[163] C. A. McCarty, R. A. Wilke, P. F. Giampietro, S. D. Wesbrook, and M. D. Caldwell. Marshfield clinic personalized medicine research project (PMRP): design, methods and recruitment for a large population-based biobank. *Personalized Medicine*, 2(1):49–79, 2005.

[164] D. Roden, J. Pulley, M. Basford, G. Bernard, E. Clayton, J. Balser, and D. Masys. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clinical Pharmacology & Therapeutics*, 84:362–369, 2008.

[165] J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics*, 26(9):1205–1210, 2010.

[166] S. Streicher, S. Sanderson, E. Jabs, M. Diefenbach, M. Smirnoff, I. Peter, C. Horowitz, B. Brenner, and L. Richardson. Reasons for participating and genetic information needs among racially and ethnically diverse biobank participants: a focus group study. *Journal of Community Genetics*, 2:153–163, 2011.

[167] B. O. Tayo, M. Teil, L. Tong, H. Qin, G. Khitrov, W. Zhang, Q. Song, O. Gottesman, X. Zhu, A. C. Pereira, R. S. Cooper, and E. P. Bottinger. Genetic background of patients from a university medical center in Manhattan: Implications for personalized medicine. *PLoS ONE*, 6(5):e19166, 05 2011.

DISSERTATIONS IN INFORMATION AND COMPUTER SCIENCE

Aalto-DD111/2011 Dubrovin, Jori
Efficient Symbolic Model Checking of Concurrent Systems. 2011.

Aalto-DD118/2011 Hyvärinen, Antti
Grid Based Propositional Satisfiability Solving. 2011.

Aalto-DD136/2011 Brumley, Billy Bob
Covert Timing Channels, Caching, and Cryptography. 2011.

Aalto-DD11/2012 Vuokko, Niko
Testing the Significance of Patterns with Complex Null Hypotheses. 2012.

Aalto-DD19/2012 Reunanen, Juha
Overfitting in Feature Selection: Pitfalls and Solutions. 2012.

Aalto-DD33/2012 Caldas, José
Graphical Models for Biclustering and Information Retrieval in Gene Expression Data. 2012.

Aalto-DD45/2012 Viitaniemi, Ville
Visual Category Detection: an Experimental Perspective. 2012.

Aalto-DD51/2012 Hanhijärvi, Sami
Multiple Hypothesis Testing in Data Mining. 2012.

Aalto-DD56/2012 Ramkumar, Pavan
Advances in Modeling and Characterization of Human Neuromagnetic Oscillations. 2012

Aalto-DD97/2012 Turunen, Ville T.
Morph-Based Speech Retrieval: Indexing Methods and Evaluations of Unsupervised Morphological Analysis. 2012

Aalto-DD 117/2012

ISBN 978-952-60-4782-9
ISBN 978-952-60-4783-6 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Information and Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS