HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Electronics, Communications and Automation

Tommi Suvitaival

BAYESIAN TWO-WAY ANALYSIS OF HIGH-DIMENSIONAL COLLINEAR METABOLOMICS DATA

Thesis submitted for examination for the degree of Master of Science in Technology

Espoo 2.10.2009

Thesis supervisor:

Prof. Samuel Kaski

Thesis instructor:

M.Sc.(Tech.) Ilkka Huopaniemi

TEKNILLINEN KORKEAKOULU

DIPLOMITYÖN TIIVISTELMÄ

Tekijä: Tommi Suvitaival

Työn nimi: Korkeaulotteisen ja kollineaarisen metabolomiikkadatan bayesilainen kaksisuuntainen analyysi

Päivämäärä: 2.10.2009 Kieli: Englanti Sivumäärä: 7+59

Tiedekunta: Elektroniikan, tietoliikenteen ja automaation tiedekunta

Professuuri: Informaatiotekniikka Koodi: T-61

Valvoja: Prof. Samuel Kaski

Ohjaaja: DI Ilkka Huopaniemi

Kaksisuuntainen tehtävänasettelu on yleinen bioinformatiikan alalla. Tässä diplomityössä esitellään uusi bayesilaisen mallinnuksen menetelmä kaksisuuntaisen havaintoaineiston analysointiin. Menetelmä toimii myös vähän näytteitä sisältävillä korkeaulotteisilla havaintoaineistoilla.

Havaintoaineiston oletetaan jakautuvan populaatioihin kovariaattien mukaan, jotka tyypillisessä biologisessa kokeessa ovat yksilön terveydentila, sukupuoli, lääketieteellinen hoito sekä yksilön ikä. Esiteltävä menetelmä on suunniteltu arvioimaan näiden kovariaattien vaikutus havaintoaineiston kontrolliryhmän perustasoon verrattuna.

Menetelmä perustuu olettamukseen siitä, että havaintoaineiston piirteet muodostavat ryhmiä, joiden sisällä piirteet ovat voimakkaasti kollineaarisia. Tämä olettamus mahdollistaa piilomuuttajamalliin perustuvan dimensionaalisuuden pudotuksen, jonka ansiosta menetelmä on toimiva myös pienen näytemäärän havaintoaineistoille.

Menetelmä käsittelee havaintoaineistoa täysin bayesilaisittain, Gibbsin otannan avulla. Bayesilainen lähestymistapa tuottaa arvion sekä mallin ja havaintoaineiston yhteisjakaumalle että mallin jokaisen parametrin marginaalijakaumalle. Tämä mahdollistaa tulosten epävarmuuden arvioinnin sekä vertailun toisiin malleihin.

Uuden menetelmän toimivuutta esitellään metabolomiikan alalta olevan havaintoaineiston avulla. Aineisto sisältää lipidiprofiileja, jotka on mitattu terveistä lapsista ja lapsista, jotka myöhemmin sairastuvat tyypin 1 diabetekseen. Kahdessa erillisessä analyysissä tutkitaan sairauden ja sukupuolen sekä sairauden ja iän vaikutusta lipidiprofiileihin.

Avainsanat: varianssianalyysi. bayesilainen mallitus, faktorianalyysi, hierarkinen malli, metabolomiikka, pieni näytemäärä

Author: Tommi Suvitaival

Title: Bayesian Two-Way Analysis of High-Dimensional Collinear Metabolomics Data

Date: 2.10.2009 Language: English Number of pages: 7+59

Faculty: Faculty of Electronics, Communications and Automation

Professorship: Computer and Information Science Code: T-61

Supervisor: Prof. Samuel Kaski

Instructor: M.Sc.(Tech.) Ilkka Huopaniemi

Two-way experimental designs are common in bioinformatics. In this thesis, a new Bayesian model is proposed for the analysis of two-way data. The method also works for small sample-size data with a high number of features.

The data set is assumed to be divided into populations according to covariates, which in the case of a typical biological experiment are the health status, the gender, the medical treatment and the age of the individual. The proposed method is designed to estimate the effect of these covariates compared to the ground level of a control group of the data.

The method is based on the assumption that features of the data form groups that are highly collinear. This allows the use of a latent variable-based dimensionality reduction, which makes inference possible also for small sample-size data sets.

The method treats the data in a completely Bayesian way, which produces an estimate for the joint distribution of the model and the data, and marginal posterior distributions of all model parameters. This allows one to evaluate the significance and uncertainty of the results and to compare it to other models. Inference is carried out with Gibbs sampling.

The performance of the new method is demonstrated with a metabolomic data set by comparing lipidomic profiles from children who remain healthy to those who will later develop type 1 diabetes. In two separate studies, the effect of the disease and gender, and the effect of the disease and time, are estimated.

Keywords: ANOVA, Bayesian modelling, factor analysis, hierarchical model, metabolomics, small sample-size

Preface

This thesis was written during summer 2009. I joined the Statistical Machine Learning and Bioinformatics Group of Prof. Samuel Kaski in spring 2008 and have since been working on Bayesian modelling under instruction of M.Sc. Ilkka Huopaniemi.

The research group is a part of Adaptive Informatics Research Centre of Helsinki University of Technology, Helsinki Institute for Information Technology, and Department of Information and Computer Science at Helsinki University of Technology. My work has been funded by Tekes Transcendo project.

I would like to thank both Prof. Kaski and M.Sc. Huopaniemi for the professional guidance that I have received. The work presented in this thesis was published in the journal *Data Mining and Knowledge Discovery* and the model development will continue in the future. My part in the work mainly was to implement the new method, designed jointly with Prof. Kaski and M.Sc. Huopaniemi, and to perform the experiments using the method.

Finally, I would like to thank my family and friends for their support during the years of my studies.

Otaniemi, 2.10.2009

Tommi Suvitaival

Contents

List of abbreviations

ANOVA	analysis of variance
ARD	automatic relevance determination
ASCA	anova-simultaneous component analysis
CCA	canonical correlation analysis
CDF	cumulative density function
Da	daltons
DNA	deoxyribonucleic acid
EM	expectation maximization
FA	factor analysis
FDR	false discovery rate
GC	gas chromatography
LC	liquid chromatography
LDA	linear discriminant analysis
MANOVA	multivariate analysis of variance
MAP	maximum *a posteriori*
MCMC	Markov chain Monte Carlo
mRNA	messenger ribonucleic acid
MS	mass spectrometry
m/z or mZ	mass-to-charge ratio
NMR	nuclear magnetic resonance
ODE	ordinary differential equation
PCA	principal component analysis
PDF	probability density function
PLS	partial least squares
PSRF	potential scale reduction factor
RT	retention time

1 Introduction

This thesis presents a method for multi-way analysis of high-dimensional data with small sample-size.

Metabolomics is a field of research where tools of bioinformatics are used in the analysis of measured concentrations of a high number of small molecules in cells. Metabolites are molecules that are substrates and end products of biological processes in cells. These processes are catalyzed by proteins.

Bioinformatics is a field of research that concentrates on analyzing observations measured from biological organisms. Bioinformatics includes modern statistical methods that are able to extract information from high-throughput experiments, where large data sets with high number of features are produced. In addition to metabolite concentrations, typical sources of data in bioinformatics are the measured transcriptional activities of genes in transcriptomics, and the concentrations of proteins in proteomics. Metabolomics is studying the last part in the chain of information from the genome to cell function.

A main motivation of bioinformatics is to provide new knowledge related to causes of diseases. Another important task of bioinformatic methods is to predict the disease risk of new patients.

A powerful approach to disease-related questions is to compare patients who are diseased to those who are healthy. This comparison reveals changes in the biological state that the patient undergoes during the development of the disease. Such findings, considered as potential biomarkers, may lead the way to a biological breakthrough and, eventually, to development and commercialization of new medical therapies.

In addition to the healthy-diseased grouping, in many experiments samples can be further divided to populations according to additional covariates. For instance, samples may be divided into medically treated and untreated populations, male and female populations, or populations according to the time of the measurement. In such a multi-way setup, not only the healthy-diseased difference is of interest but also effects of other covariates and interactions of several covariates may provide useful information.

The modelling problem presented in this thesis is to separate the effects of the covariates and their interactions from other variation of the data. This is called multi-way analysis. The thesis has a focus on two-way modelling, where the effects of two covariates and their interaction are studied. Two- and multi-way analyses give information on how the covariates affect the observations. For metabolomic data, the result of a multi-way analysis can be, for instance, a list of metabolites that have differential concentration in healthy and diseased patients and the numerical quantity of the difference.

A common problem in modern biological experiments is the small number of available measurements from patients which, in the field of statistics, are called samples.

The main reason for the small number of samples is the high financial cost of a single measurement, which leads to a need of keeping the number of samples as small as possible. In many medical experiments, animals are grown in captivity and included in the experiment that includes medication and taking samples of live tissue. In such experiments, the sample-size is small not only because of the high financial costs of growing animals in laboratory conditions but also due to ethical questions.

The problem of small sample-size, or the $n < p$ problem, leads to constraints in the analysis. Many traditional statistical methods do not work for a data set with less samples than features.

The new method presented in this thesis has been rigorously developed to analyze small sample-size multi-way metabolomic data sets. A particular application is the Type 1 Diabetes prediction and prevention (DIPP) follow-up study [1]. In the study, lipidomic profile was measured from the blood serum of children with a genetic risk for type 1 diabetes. Some patients developed into type 1 diabetes, whereas some of the individuals under risk remained healthy.

Lipids are a sub-group of metabolites that form the cell membrane and also take part in many important biological processes such as cell signaling. Lipids are of particular interest because changes in their concentrations are related to many metabolic diseases, such as diabetes.

In the DIPP study, the lipidomic profile was measured from each child on several occasions during the first years of life. Thus, the data set consists of a time series of profiles for each patient. Due to the nature of the experiment, the data includes three covariates, which are: healthy-diseased, male-female, and the age of the patient at the time of the measurement.

The objective of the developed method is to take the additional covariates and special properties of metabolomic data into account and to provide information on how healthy and diseased individuals differ in their lipidomic profiles. It is of particular interest to find out which lipids are the ones having differential concentration due to the covariates. The new method is able to identify effects resulting from the covariates and effects resulting from the interaction of multiple covariates. In many studies, the covariate interaction effects, such as the the interaction of the disease and age, are of most interest.

None of the current methods, univariate or multivariate, is perfectly suited for the current problem setting of a multi-way high-throughput experiment with small number of samples. Univariate methods, which analyze only one feature at a time, are still usable in the $n < p$ case but do not provide reliable results. Unlike multivariate methods, univariate methods do not take correlations of features into account, thus missing one of the critical properties of metabolomics. Multivariate methods are generally more useful for analyzing multivariate data but many of them cannot be used in the small sample-size case.

The lack of available methods is the main motivation in developing a Bayesian model that is particularly suited to this application. The Bayesian approach is justified,

as it handles properly the problems following the small sample-size.

This thesis introduces the research field of metabolomics, common methods in analyzing metabolomic data and presents the new Bayesian method with demonstrations of its functioning. Section 2 consists of an overview of metabolomics discussing the characteristic properties of metabolomic data, technologies used in performing the measurements, pre-processing steps applied to the raw measurement data, and data analysis methods widely used in metabolomic analysis.

Widely used univariate statistical methods are presented in Section 3. Of the methods presented, especially analysis of variance (ANOVA) is important from the perspective of the this thesis, as it closely resembles the new method presented in this thesis.

Clustering, factor analysis and multivariate analysis of variance are multivariate methods that have similarities to the proposed new method. These methods are presented in Section 4. Many of these more traditional multivariate methods are not applicable for $n < p$ data.

Less established multivariate methods proposed for bioinformatics applications are presented in Section 5. They are usable for data with small number of samples but the methods are in many ways not perfectly suitable for analysis of multi-way metabolomics data.

The idea of Bayesian modelling is introduced in Section 6. The Bayesian approach provides a clear framework for estimating the uncertainty of the model and for incorporating prior knowledge into the statistical model. Assessment of the uncertainty is important especially at small sample sizes. Complex dependencies of variables can be modeled with hierarchical models, which consist of several layers of unobserved variables. The concepts of a hierarchical model and sampling are also explained in Section 6.

The new method presented in this thesis is a hierarchical Bayesian model. It performs dimensionality reduction and estimates the ANOVA-type effects arising from experimental treatments. The dimensionality reduction is critical in enabling the model for the small-sample size case, whereas the estimation of treatment effects is the main objective of the model. The building blocks of the method are explained in detail in Section 7.

Functioning of the new method is demonstrated in Section 8 with simulated data, especially with small sample-size data. To extract novel biological information, the model is applied to the DIPP lipidomic data set and the covariate effects in the experiment are estimated. The model is learned both with data from a single time point and with time series data.

2 Metabolomics data

Metabolomics is a newly emerging field of 'omics' research concerned with the high-throughput identification and quantification of small molecule metabolites in the metabolome. Metabolome is defined as the complete complement of all metabolites smaller than 1500 Da that are found in a specific cell, organ or organism. It includes natural reagents and final products of chemical processes taking place in the cell, such as lipids, but also chemicals that are not naturally part of the cell metabolism, such as drugs. Traditionally metabolomics, sometimes called metabonomics, has been seen as a diagnostic tool for metabolic classification of individuals [2].

Lipidomics is a sub-category of metabolomics focusing on quantitative study of lipids. The group of lipids is defined [3] as hydrophobic or amphiphilic small molecules that originate either entirely or in part from two distinct types of building blocks: carbanion-based condensations of thioester and carbocation-based condensations of isoprene. The first group includes fatty acids and the second includes sterols.

Metabolites and lipids are interesting because they reflect the effects of gene and protein regulation and thus provide vital information regarding the biological state of the biological system [4]. Lipids are especially interesting, as they play an important role in cell signaling, membrane architecture, transcriptional and translational modulation, cell-cell and cell-protein interactions and response to environmental changes over time.

Changes in lipid concentrations are a signal of changes in the metabolic state in a cell. Such changes are related to many metabolic diseases. Possible connection to a disease is also the motivation of the experiments presented in this thesis.

Lipids are classified into eight major categories: fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids, and polyketides [3]. The major difference between lipids and other major components of living tissue is their solubility to organic solvents [4]. Another unique characteristic of many lipids is their spontaneous arrangement into micelles or bilayer vesicles in an aqueous environment. The polar head groups of the molecules tend to remain associated with water, whereas the hydrocarbon tails form hydrophobic interactions.

Other traditional fields of 'omics' are genomics, transcriptomics and proteomics. Genome is the inherited information in the cell. Gene is the part of the genome that is associated in production of a protein. The first step in production of the protein is transcription of the gene into messenger ribonucleic acid (mRNA). In the second step, mRNA is translated into the protein such that each codon in the mRNA corresponds to one amino acid, which is a basic building block of proteins. Proteins are the functional molecules built by the cell. Proteins are critical in catalyzing biochemical reactions, where metabolites act as substrates and products.

Genomics is a field that studies the DNA sequence of an organism. Transcriptomics is the field of study that focuses on investigating the transcription. Also known as

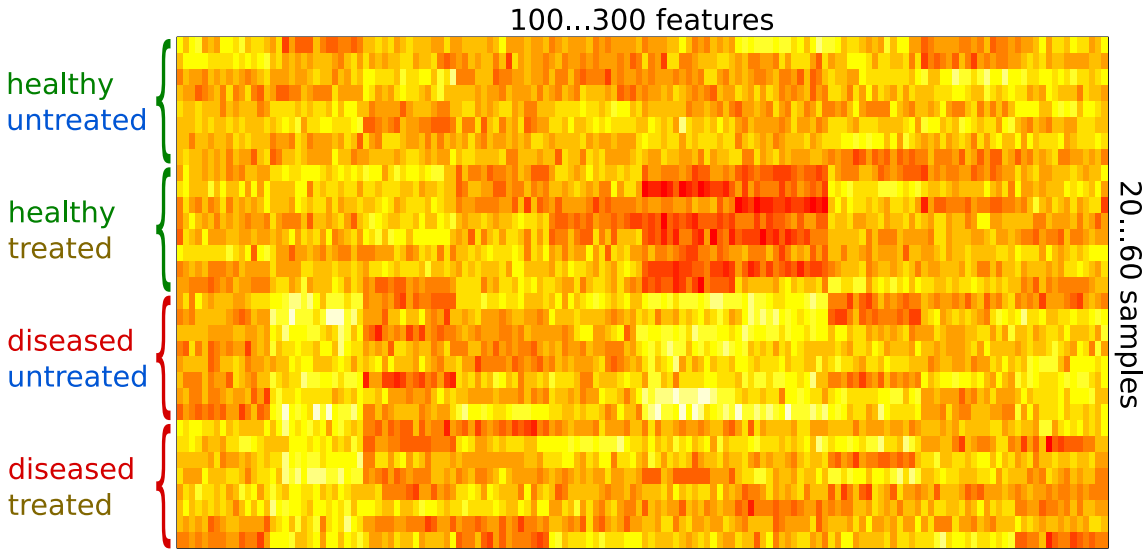


Figure 1: Small sample-size data matrix containing simulated normally distributed data with population-specific up- and down-regulations. Features are ordered according to similarity, and samples according to treatment covariates, which in this case are healthy-diseased and treated-untreated. A feature is a measured property of the studied object. In metabolomics, one feature is equivalent to one metabolite. Data of a biological origin is considerably more noisy. Also, the covariate effects are not as visible as in this simulated example.

expression profiling, it examines the expression level of mRNAs in the cell, often using high-throughput techniques based on DNA microarray technology. Proteomics is a field studying the structures and functions of proteins. Even though the protein production and function are regulated by many ways in the cell, the protein concentration is highly dependent on the transcriptional activity of the gene encoding the corresponding protein. The same connection to transcriptional activity holds for metabolites because their production is governed by proteins. [2]

2.1 Characteristics of the data

A metabolomic data set consists of metabolic profiles measured from different individuals under various treatments. Metabolic profile is a vector of measurements of metabolite concentrations in a biological cell. The profile usually consists of measurements of 20 to 200 metabolites. Such profile vectors from different individuals or treatments are united to form a data matrix, which consists of p rows and n columns, corresponding to p different metabolites and n individual profiles, respectively. One individual profile is called a sample. An example data set is shown in Figure 1.

Metabolomic data, as data in bioinformatics in general, is usually high-dimensional compared to the number of samples available. The number of samples is limited mainly by external constraints. Growing multitude of animals, such as mice, in

laboratory environment is costly. Also, such experiments, where genetically or otherwise manipulated animals are grown and might suffer because of the treatment, are ethically questionable.

Besides being high-dimensional, metabolomic data typically has high variation between individuals from which the measurements have been made. This feature does not only result from noise in the measuring process but mainly from biological variation between the individuals.

Repeats of measurements from one individual are called technical repeats and measurements from different individuals in similar conditions are called biological repeats. When animal cells are under study, the variation of technical repeats is usually small compared to variation of biological repeats. In plant cells, however, the variation of technical repeats is high because of stronger regulation in biological processes [5].

Metabolomic data differs strikingly from other fields of bioinformatics in the way the correlations between data features arise. In transcriptomics and proteomics, concentrations are mainly governed by a network of regulatory interactions, whereas metabolites are synthesized from other metabolites via a network of biochemical reactions [6]. Concentrations of members of such a reaction chain are naturally highly correlated.

If one wishes to build a model on the reaction pathways of a set of metabolites, measuring equipment with high temporal resolution is required. Such equipment are not widely available currently. Traditionally, metabolic pathway modelling is done by writing down and solving systems of time-dependent ordinary differential equations (ODEs) that describe the chemical reaction coefficients and rates of the metabolic system of interest. For finding underlying biological causes, this kind of modelling approach is extremely useful. However, the data analyzed in this work does not have high enough temporal resolution for this type of analysis and, thus, the topic is not discussed further in this thesis.

In this thesis, the focus is in analysis of experiments with temporally less frequent measurements. In contrast to pathway modelling, the time scale of the experiment is comparable to the life span of the species under study. In such a scale, metabolic changes do not arise from instant fluctuations in chemical processes and their regulation but from changes in individual's state, which is caused by aging, biological conditions or by the applied medical treatments. A metabolite or other substance that can be used as an indicator of a biological state is called a biomarker.

Metabolic profiling aims at identifying and quantifying the components in the sample, which in practice means estimating concentrations of metabolites. The concentration profiles can then be used in estimating the effect of a treatment on various metabolites and thereby in finding potential biomarkers. Such an approach, where data sets of even hundreds of metabolites are analyzed, is called non-targeted, global or shotgun metabolomics [7].

The complications in non-targeted metabolomics are discussed in later sections. An-

other approach to studying identified metabolites is targeted metabolomics, where the metabolites are pre-selected according to prior information. In the targeted approach, the problem of high number of metabolites compared to the number of samples is avoided but there is also the drawback that behavior of only a limited number of metabolites can be studied.

The chemometric approach differs from targeted and non-targeted profiling such that individual metabolites are not identified from the spectral data but the spectrum is analyzed as such. The approach is used for finding statistically significant differences between large groups of spectra collected on different samples or under different conditions. It is obvious that this approach will not provide information on which metabolites are affected by the applied treatment, but the spectra can still be used in finding statistically significant differences under different conditions or in classification of samples.

The clear advantage of the chemometric approach is that it is not required to identify unknown components in complex spectra, which in any case is a hard task, and that the differences can be studied without taking opinion on which the chemical components behind the phenomenon are. Because the approach does not provide information on the behavior of identified metabolites, it is fundamentally different from the approach of the new method presented in this thesis. Due to the limited scope of this thesis, the chemometric approach is not discussed in more detail.

A typical task in a metabolomic study is to find metabolites that have differential concentration in treatments of an experiment. Naturally this is possible only when metabolites have been identified from the raw data. There is interest in performing experiments with more complicated experimental designs than the simple one-way healthy-sick design. Such designs include multiple ways of treatment and multiple levels of treatment. For high-dimensional data with small number of samples, proper analysis of such experiments has not been possible due to lack of suitable statistical methods.

2.2 Measurement technologies

Spectral analysis forms the basis for all metabolomic studies. Samples are measured using non-chemical and non-colorimetric methods, such as mass spectrometric methods or nuclear magnetic resonance spectroscopy. These methods measure the spectrum of the sample. Spectrum is the relative intensity of the measured sample representing the chemical analysis. For instance, in mass spectrometry, the intensity of the sample is presented as function of the particle mass.

To study the measurements efficiently, the spectrum is divided into bins. Each bin contains the spectral information of a small region of the spectrum - and usually the close-by peaks. This binned information can then be compared between samples. Binning is an established approach in metabolomics, as each peak of the spectrum is considered to correspond to one type of molecule. Thus, the peak height is directly linked to concentration of the molecule. In the data analysis, each bin is treated as

one feature of the raw data.

An alternative approach to the chemometric approach is targeted metabolic profiling, where individual components of the spectrum are identified. The spectrum is assumed to be the sum of the spectra of pure metabolites that the sample includes. The identification requires use of spectral databases of known molecules to which the measured spectrum is compared. This approach is practiced in the methods that are presented in this thesis.

After collecting a sample from a biological tissue, the sample is stored frozen to avoid molecular degradation. Homogenization of the frozen sample usually takes place mechanically by glass beads and aims at leveling off differences of concentrations in different parts of the sample [8]. Proteins are precipitated chemically from the sample, usually by methanol.

After preparation of the sample, different types of metabolite molecules are separated to make it possible to measure their concentrations. Some of the most widely used separation technologies within the metabolomics field are gas and liquid chromatography (GC and LC, respectively).

Chromatographic methods separate different types of molecules, or compounds, based on their characteristics, such as polarity or hydrophobicity. The sample, or analyte, is introduced into a column. The column contains a mobile and a stationary phase. The analyte's motion through the column along the mobile phase is slowed by specific chemical or physical interactions with the stationary phase. Each compound traverses through the column in a specific time, which is dependent on properties of the analyte. Retention time (RT) is the time it takes for the compound to appear from the column after injection of the analyte. The appearance of the compound produces a peak in the measured intensity spectrum. The retention time is later used in identifying the compound.

Gas, or gas-liquid, chromatography differs from liquid chromatography such that the moving phase is gas instead of liquid, and the stationary phase is liquid instead of solid. GC is good for separating compounds that can be vaporized without decomposition.

All compounds cannot be resolved by chromatography alone, because several different compounds may have equal retention times. For identification of the compounds, it is required to measure also other properties. Typically the weight of the compound is measured with mass spectrometry (MS). In mass spectrometry, each compound is ionized, and the mass-to-charge ratio (m/z) of the ion is measured by simultaneously introducing it to electric and magnetic fields. The ionization is usually carried out with electrospray ionization technique.

The mass spectrometer is directly linked to the output of the chromatographic device to enable the measurement of the retention time and the mass-to-charge ratio consecutively. This setting leads to data, where each compound has a position on the mZ/RT plane. To identify the corresponding compounds, the mZ/RT coordinates are then compared to literature values of known compounds.

Nuclear magnetic resonance (NMR) spectroscopy is a detection method that does not rely on prior separation of the analytes. NMR is usually based on radiation absorption and emission of hydrogen atoms in a strong magnetic field (^{1}H NMR). Carbon-13 NMR (^{13}C NMR), a widely used NMR technique, is not commonly used in the metabolomics field [9]. The resonance frequency of a hydrogen atom is dependent on the molecule that it is bound to, which is why intensity peaks in the NMR spectrum correspond to molecules and their concentrations.

The combination of chromatography and mass spectrometry is usually preferred because of the huge advantage in terms of sensitivity over the NMR spectroscopy. One strength of the NMR is that the technique is potentially non-invasive, leading to many medical applications of the NMR effect to detect molecules *in vivo* [10]. This feature, however, is not of use when measuring internal tissues.

Differential profiling is an approach, where measurements from different populations of samples are compared to find differences in the observed profile. The approach requires a broad coverage of measured compounds, which can be fulfilled with LC-MS or GC-MS techniques. The use of these techniques is currently predominant in the metabolomics field. LC-MS was also used for producing the DIPP data that is analyzed in this thesis. Still, to be able to compare compound concentrations between individuals, it is necessary to perform pre-processing on the raw data.

2.3 Data pre-processing methods

The data obtained from the LC-MS or GC-MS equipment is noisy. It is convenient to prepare the data with pre-processing before the final statistical analysis. The conventional pre-processing steps are:

1. spectral filtering, which aims at reducing the complexity of spectra and removing the noise
2. peak detection, which finds the peaks corresponding to the compounds or fragments thereof
3. alignment, which aims at matching the corresponding peaks across multiple sample runs
4. normalization, which is the final step to reduce systematic error by adjusting the intensities within each sample run

The measurement data of the experiments in this thesis were prepared for further analysis using MZmine [11], which is a stand-alone all-in-one software platform for performing the above listed pre-processing steps of raw metabolic profile data.

MZmine performs filtering of spectral peaks in both the RT and m/z directions to remove peaks with the weakest intensities. The mZ/RT plane is sliced into bins that are considered as one-dimensional spectra. Peaks in neighboring slices are then

joined if they have similar m/z values. These peaks then form one peak in the RT direction.

Peaks from the same compound in different samples usually match closely in mass-to-charge values, but there can be variation in retention times between the measurement runs. Alignment step is used for matching peaks between multiple sample runs. Alignment usually leads to missing values if a certain peak is not found in every sample.

The usual causes for not finding the peak are the low concentration of the corresponding molecule and noise in the measuring process. Noise not only affects the intensity of the peak but also its position in the mZ/RT plane, which causes difficulties for the peak detection algorithm. MZmine provides a way of searching missing peaks from the raw data.

The last step, normalization, is used to reduce the systematic error in the data. Normalization can be performed using a linear method or by calculating a normalization based on standard compound intensities. Standard compounds are molecules that are added to the sample with known concentrations prior to the measurements. Each peak can then be normalized either relative to the standard compound closest to the peak, or by using a weighted contribution of all standard compounds.

Software used in pre-processing may have different algorithms but the principal operations are common for all programs. Parameters for peak width, shape, intensity, resolution, etc., have to be set for the peak-picking algorithm to find the metabolites using predefined tolerances for mass and retention time shifts. [12]

2.4 Common data-analysis methods

This sub-section provides a brief overview of the methods currently used in metabolomics. Technical details of the methods are discussed in Sections 3, 4 and 5.

Statistical data analysis methods are applied to the pre-processed data set, which is a full matrix of measurement values. The rows correspond to samples and columns correspond to metabolites, or features, as called in machine learning literature.

A common approach in present-day metabolomics is one-way analysis, where the effect of a single covariate is studied. One of the most simple, and still, most widely used method in bioinformatics is the Student's t-test [13], which decides whether a difference between two groups of samples is statistically significant. The t-test has numerous modifications to multivariate case [14], where the data consists of multiple features, treated separately and combined. The data in bioinformatics is practically always multivariate.

Analysis of variance (ANOVA) [13] is the most traditional method for analyzing multi-way experiments. Originally, ANOVA is a univariate method but there is a generalization of the method to the multivariate case, called multivariate ANOVA or MANOVA [15]. ANOVA is widely used in bioinformatics but it does not provide reliable results when applied to multivariate data such as in metabolomics. MANOVA

is not at all applicable for small sample-size data.

A successful approach to high-dimensional data analysis is dimensionality reduction, where multiple variables are combined as few hybrid variables. The relevant variation of the original variables is hoped to be transferred to the hybrid variables while the noise is averaged out. Further analysis is carried out on the lower-dimensional hybrid variables.

Principal component analysis (PCA) [16] is a common method used in dimensionality reduction. PCA performs a linear projection on the data, such that the variance of the projection is maximal. In several approaches, PCA is used as a dimensionality reduction method to feed lower-dimensional data to another method, such as MANOVA [17]. The PCA projection is not guaranteed to be at all relevant on the experimental setting and, thus, it is not likely that relevant biomarkers are found.

Anova simultaneous component analysis (ASCA) [18] is a recently proposed method for analyzing small sample-size multi-way data. ASCA performs a PCA-type projection on the data. From the projection, ANOVA-type treatment effects are estimated.

Partial least squares (PLS) regression is a widely used method in classification and prediction. It is related to PCA but, unlike PCA, it is supervised, meaning that also target values are used in finding the optimal projection. It is particularly suited when there is high multicollinearity among features, which is common in metabolomics.

Partial least squares discriminant analysis (PLSDA) [19] is a variant of PLS regression specialized in classification of samples. In the method, known class labels are used as target values in the learning. The problem of PLS-based methods is that they are prone to overfitting, which means that they give unreliable results.

Clustering is a usual approach in any multivariate analysis, where the task is to find groups of similar samples or features. There are numerous clustering methods and one of the most widely-used methods is hierarchical clustering. It is a traditional clustering method also in bioinformatics. The result of the method is nothing more than a tree-like grouping of features. Even though information on similarity of the features is obtained, any further analysis needs to be done by using other methods.

There is still great need for interpretable and reliable statistical methods in the field of metabolomics. The progress of statistical machine learning methods leads to a possibility of extracting more knowledge from the noisy biological experiments.

2.5 The biological experiment of this thesis

In this thesis, data from targeted profiling experiment is analyzed. The spectral peaks are identified and a group of relevant lipids are selected for further analysis. After peak detection and other pre-processing, the data is in the form of a full matrix, where there are tens to hundreds of features (or metabolites, i.e. rows) and some tens of samples (or patients, i.e. columns).

The originates from a screening performed for over 100,000 newborn infants in Fin-

land [1]. The infants were screened for a known genetic risk of acquiring type 1 diabetes. Over 8,000 children with genetic risk continued in the follow-up surveillance, where concentrations of four antibodies related to diabetes were measured during a long time interval. During the first two years of their life, the children were monitored roughly once in every three months, and after that, once or twice a year. 117 of the 8,000 children were picked at random into a more detailed study, where their metabolic profiles were measured from blood serum. In this profile study, there were both individuals who acquired diabetes and individuals who stayed healthy.

Diabetes is a condition in which the body does not produce enough, or properly respond to, insulin. Insulin is a hormone produced in the pancreas. It enables cells to absorb glucose in order to turn it into energy. Type 1 diabetes is an autoimmune disease that results in destruction of insulin-producing β-cells of the pancreas.

The presence of anti-islet antibodies, in particular GADA, are useful in identifying the disease and also in distinguishing type 1 diabetes from type 2 diabetes [20]. Nowadays type 1 diabetes is sub-categorized into type 1A and type 1B diabetes, of which the first one can be identified based on antibody levels. About 90 % of white children will have at least one of the anti-islet antibodies at disease diagnosis.

The samples were analyzed with ultra performance liquid chromatography coupled to mass spectrometry [1]. Software platform MZmine was used for pre-processing the raw data. A total of 117 individuals were included in the study, of which 50 acquired the disease and 67 remained healthy. From each individual, on average 10 measurement time points were obtained. In total, 53 lipids were identified in all these 1,196 samples. Later, the number of individuals in the data set were reduced to roughly 50 by removing individuals whose time series was not complete [21].

In an earlier analysis performed on the same data [1], metabolic changes were found already before the emergence of islet autoantibodies. Also intra-individual development has been studied using the data [21]. A hidden Markov model was used to align the developmental trajectories of individuals of one gender, which led to lower within-group variance than the age-based alignment.

Because of the number of samples being smaller than the number of features, traditional statistical methods break down due to the non-invertible covariance matrix. To identify the key components in the feature space, dimensionality reduction is required. A Bayesian two-way model was developed for the data because the traditional methods presented in the previous sub-section do not provide reliable results for collinear data with small number of samples compared to the number of features.

3 Univariate statistical methods

Relevant traditional methods used in analyzing univariate data are presented in this section. Being univariate means that there is a only single feature in the data set. Then, one sample is not a vector of values but a single scalar value. In metabolomics, this means that the data are analyzed one metabolite at a time, leaving all inter-metabolite correlations without attention.

Student's t-test is the most traditional way of comparing statistical difference of two values or means of two populations. Analysis of variance (ANOVA) extends the test to several populations of values. The presented univariate methods can also be applied to multivariate data sets but then the results are more likely to be incorrect because the methods are not designed for such use. Multivariate methods will be discussed in the next section.

3.1 *p*-value

Statistical hypothesis testing is an approach for determining whether an aspect of experimental data is statistically significant. The decisions are made using null hypothesis tests, which determine the probability of the observations assuming the null hypothesis is true. The null hypothesis formally describes some aspect of the statistical behaviour of a set of data. This description is treated as valid unless the actual behaviour of the data contradicts this assumption. The testing is performed by calculating a specific test statistic from the observed values and by comparing it to the cumulative density function (CDF) of the distribution the test statistic is assumed to follow.

p-value of the test is the value of the cumulative density function at the point of the test statistic value. The decision over the validity of the null hypothesis is then based on the p-value of the test statistic which tells how probable it is to obtain a test statistic value more extreme than the one that was calculated. If the value is too extreme, the decision is that the test statistic does not follow the assumed distribution and, thus, the null hypothesis is rejected.

In statistical significance testing, p-value of either 0.05 or 0.01 is often considered as a critical limit. If the p-value of observations is lower than this limit, the null hypothesis is rejected.

3.2 Student's *t*-test

The traditional way of comparing the mean of a normally distributed population either to a value specified in a null hypothesis, or to the mean of another normally distributed population, is Student's t-test. The former is based on the fact that the test statistic T calculated from n-sample mean $\overline{x}_n$, standard deviation s_n, and null

hypothesis mean μ_0,

$$T = \frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} \sim t(n-1), \tag{1}$$

follows the Student's t-distribution with $n-1$ degrees of freedom, if the null hypothesis holds. This quantity T is called the test statistic. In the equation, both sample mean and sample standard deviation are estimated from the data.

The Student's t-distribution is lower-peaked and longer-tailed than the normal distribution. Thus, a variable that is t-distributed has more uncertainty than a normally distributed variable. This arises from the fact that in addition to the mean, also the standard deviation is estimated from the data.

3.3 Analysis of variance

Analysis of variance (ANOVA) extends the test of equal means to the case of more than two populations. When the division is based on one factor, a one-way analysis is applied, whereas when there are more factors, a multi-way analysis is applied.

The new method presented in this thesis includes a module similar to the traditional multi-way ANOVA. However, the new method is more informative, as it not only provides one p-value for each population but also provides a distribution estimate for the effect. The distribution estimate is useful in assessing the uncertainty of the estimate.

3.3.1 One-way ANOVA

In the one-way analysis of variance, one factor divides the sample into k populations. One-way ANOVA then separates the grand mean and population-specific effects. The statistical model of one-way ANOVA [13] is parametrized as

$$y_{ki} = \mu_i + \varepsilon_{ki}, \tag{2}$$

where y_{ki} is the observation k belonging to population i, μ_i is the population-specific mean, and ε_{ki} the residual term of the observation ki, that is, the part of the observation that is not explained by the linear model. The same can be expressed more intuitively as

$$y_{ki} = \mu + \tau_i + \varepsilon_{ki}, \tag{3}$$

where μ the grand mean of the data, τ_i the population i-specific effect, or deviation from the grand mean, and ε_{ki} the same residual term as previously. The effects are restricted to sum over all samples to zero:

$$\sum_{i=1}^{I} n_i \tau_i = 0, \tag{4}$$

where n_i is the number of samples in population i and I is the total number of populations. The residual term is assumed to be independent and identically normally distributed between samples:

$$\varepsilon_{ki} \sim \mathcal{N}\left(0, \sigma^2\right), \tag{5}$$

where σ^2 is the residual variance.

The null hypothesis H_0 in ANOVA is that all population-specific means μ_i are equal:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_I = \mu. \tag{6}$$

The statistical testing of populations is performed by decomposing the variance of the sample into variances between and within populations. Traditionally, the calculations have been done by using square sums because it requires less computation. The decomposition is simply

$$SST = SSG + SSE, \tag{7}$$

where SST, SSG and SSE correspond to square sums of total, population-specific and sample-specific variation, respectively.

The residual square sum that corresponds to the sample-specific variance is defined as

$$SSE = \sum_{i=1}^{I} \sum_{k=1}^{n_i} \left(y_{ki} - \overline{y}_i\right)^2, \tag{8}$$

where $\overline{y}_i$ is the mean of the population i. The total square sum, corresponding to the total variance, is defined as

$$SST = \sum_{i=1}^{I} \sum_{k=1}^{n_i} \left(y_{ki} - \overline{y}\right)^2. \tag{9}$$

The null hypothesis is tested by using the test statistic

$$F = \frac{N - I}{I - 1} \cdot \frac{SST - SSE}{SSE}. \tag{10}$$

If the null hypothesis holds, the test statistic follows F-distribution,

$$F \sim F(I - 1, N - I), \tag{11}$$

with $I - 1$ and $N - I$ degrees of freedom. The p-value of the test statistic can then be compared to critical limits of the F-distribution.

If the null hypothesis is rejected, it is evident that at least two of the population means differ from each other in a statistically significant way. The ANOVA itself does not tell which the differing populations are, but this can be found out using pair-wise t-tests.

3.3.2 Multi-Way ANOVA

The one-way ANOVA generalizes into a multi-way case, where the data is decomposed into populations with respect to more than one factor.

In a two-way case, there are two factors, A and B, and the statistical model can be parametrized as

$$y_{kij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{kij}, \tag{12}$$

where μ again is the grand mean, α_i is the population i-specific A-effect, β_j the population j-specific B-effect, $(\alpha\beta)_{ij}$ the interaction effect AB of the populations i and j, and ε_{kij} the residual term of observation k. Additionally, there is a constraint that all effects sum to zero over all populations:

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0. \tag{13}$$

The two-way ANOVA tests three things: whether there is an effect of treatment A, of treatment B, or of the interaction AB. The null hypotheses are formulated as:

$$\begin{aligned} H_{AB}: &\quad \text{No interaction effect} \\ H_A: &\quad \text{No effect of treatment A} \\ H_B: &\quad \text{No effect of treatment B.} \end{aligned}$$

For statistical testing, the following square sums need to be calculated: the total square sum

$$SST = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (y_{kij} - \overline{y}_{...})^2, \tag{14}$$

the square sum of treatment A

$$SSA = JK \sum_{i=1}^{I} (\overline{y}_{.i.} - \overline{y}_{...})^2, \tag{15}$$

the square sum of treatment B

$$SSB = IK \sum_{j=1}^{J} \left(\overline{y}_{..j} - \overline{y}_{...}\right)^2, \tag{16}$$

and the square sum of the interaction of treatments A and B

$$SSAB = K \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\overline{y}_{.ij} - \overline{y}_{.i.} - \overline{y}_{..j} + \overline{y}_{...}\right)^2. \tag{17}$$

The ANOVA decomposition then becomes

$$SST = SSA + SSB + SSAB + SSE. \tag{18}$$

The existence of the interaction effect can be tested by calculating the test statistic

$$F_{AB} = \frac{IJ(K-1)}{(I-1)(J-1)} \cdot \frac{SSAB}{SSE} \sim F\left((I-1)(J-1), IJ(K-1)\right), \tag{19}$$

which follows F-distribution with $(I-1)(J-1)$ and $IJ(K-1)$ degrees of freedom, in the case that the null hypothesis H_{AB} holds. Single-treatment effects are tested with test statistics

$$F_A = \frac{IJ(K-1)}{I-1} \cdot \frac{SSA}{SSE} \sim F\left((I-1), IJ(K-1)\right) \tag{20}$$

for treatment A and

$$F_B = \frac{IJ(K-1)}{J-1} \cdot \frac{SSB}{SSE} \sim F\left((J-1), IJ(K-1)\right) \tag{21}$$

for treatment B, and also they follow F-distribution if the respective null hypotheses H_A and H_B hold. These three statistics can be interpreted as comparisons of variances within and between populations.

The least squares estimate of the two-way ANOVA model is obtained by minimizing the square sum

$$SS = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{kij} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij}\right)^2, \tag{22}$$

which leads to estimates

$$\begin{aligned}
\hat{\mu} &= \overline{y}_{...} \\
\hat{\alpha}_i &= \overline{y}_{.i.} - \overline{y}, i = 1, \ldots I \\
\hat{\beta}_j &= \overline{y}_{..j} - \overline{y}, j = 1, \ldots J \\
\widehat{(\alpha\beta)}_{ij} &= \overline{y}_{.ij} - \overline{y}_{.i.} - \overline{y}_{..j} + \overline{y}, i = 1, \ldots, I, j = 1, \ldots, J.
\end{aligned} \tag{23}$$

However, these values are only point estimates. The strength and significance of an effect is evaluated from the p-value of the corresponding F-test result. With small sample-size, the estimates are not reliable.

3.4 Multiple testing problem

As ANOVA and the t-test are univariate methods, similar testing needs to be performed separately for each variable of a higher-dimensional data matrix. However, there is a complication in such multiple testing because ANOVA tests are independent of each other, leading to an increase in the false positive discovery rate. The more tests are performed in parallel, the more likely it is that at least one of the test results is incorrect.

There are several simple by-passes for attempting to overcome this multiple testing problem, of which two most common techniques are presented here.

3.4.1 Bonferroni correction

Bonferroni correction attempts to keep the significance level of the test setting constant regardless of the number of individual tests being performed. This is simply done by dividing the original critical limit p-value by the number of individual tests. For instance, original critical p-value of 0.05 with 10 parallel tests would lead to a corrected p-value of 0.005.

In practice, the Bonferroni correction has been proven to be too conservative, meaning that the new critical p-value is too low and, thus, a null hypothesis is rejected too rarely.

3.4.2 False discovery rate

A more sophisticated way of dealing with false positives is a method called false discovery rate (FDR) [14]. The first step of the procedure is to compute the ordinary p-values of the tests but not to reject the corresponding null hypotheses yet. The number of p-values smaller than the critical value p^* is m. Then, these m p-values are sorted in an ascending order

$$p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)} \tag{24}$$

and the last p-value $i = k$ that fulfils the inequality

$$p_{(i)} \leq \frac{i}{m} q^* \tag{25}$$

is sought, when q^* is the chosen false discovery rate. In the last step, all null hypotheses $H_{(i)}$ for which $i = 1, \ldots, k$ are rejected, leading to a multiple test result with false discovery rate of q^*.

False discovery rate controls the expected proportion of incorrectly rejected null hypotheses. It is less conservative than the Bonferroni correction. This means that the null hypothesis is decided to be valid less often in the case, where the null hypothesis should be discarded.

4 Classical statistical multivariate methods

There are numerous multivariate methods that take into account the multivariate property of the data. Many of the widely used methods are devised to deal with large data matrices. Such methods help the statistician in finding relevant information from the data.

In the field of metabolomics, as in bioinformatics in general, multivariate methods are required, as the number of features in the data is very high. The new method presented in this thesis has similarities to several methods presented in this section. Reduced Bayesian factor analysis forms the dimensionality reduction module of the method. The entire method is based on normality assumption of the data and conditional normal distributions are calculated during the iterative steps of the algorithm. The properties of conditional normal distribution are presented in this section.

In this section, relevant multivariate methods are presented and their connection to the new Bayesian method is explained. In sub-section 5, competing methods developed for small sample-size data are presented.

4.1 Clustering

Clustering means unsupervised grouping of patterns into clusters. The patterns to be grouped usually are either observations or features. When clustering observations, the goal is to find groups of similar observations. The main application area of clustering methods is to find these groups when no class information is available beforehand, whereas supervised methods are used when the class labels of the observations are known. Clustering is a considerably harder task than classification.

There are various ways of performing clustering. The two most commonly used methods are k-means clustering and hierarchical clustering. In k-means clustering, the data is divided into k clusters. The cluster centers and cluster assignments are updated via an alternating iterative process. Hierarchical clustering yields a dendrogram, that is a tree, representing the nested groupings of observations and similarity levels at which groupings change.

In addition to selecting the clustering method, the choice of metric plays a critical role. The result of the clustering is very strongly dependent on how the distances between data points are measured. The two most common choice for the measures of the distance between two d-dimensional data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are the Euclidean distance, or more generally, the Minkowski metric

$$\begin{aligned} d_p(\mathbf{x}_i, \mathbf{x}_j) &= \left(\sum_{k=1}^{d} |x_{i,k} - x_{j,k}|^p \right)^{1/p} \\ &= \|\mathbf{x}_i - \mathbf{x}_j\|_p, \end{aligned} \tag{26}$$

where p decides the norm to be used. When $p = 2$ the metric becomes Euclidean distance.

Another widely used distance metric is the squared Mahalanobis distance

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{\Sigma^{-1}} (\mathbf{x}_i - \mathbf{x}_j), \tag{27}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the observed vectors and $\mathbf{\Sigma^{-1}}$ is the sample covariance or the known covariance matrix of the data generating process [22]. The Mahalanobis distance is also present in the density function of the multivariate normal distribution.

Clustering methods can also be used for finding similarly behaving features. When clustering features, clustering can be used to assist dimensionality reduction: similar features are grouped. These groupings can then be used in, for instance, reducing the number of statistical tests to be performed.

4.2 Multivariate normality assumption

It is a common assumption in statistics that the data is normally distributed, and this holds also for data with more than one features. The parameters of the normal distribution are its mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. In statistical literature, the fact that a random variable $\mathbf{X}$ follows normal distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$ is expressed as

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}). \tag{28}$$

Here, as with many methods that are based on the normality assumption, $\mathbf{\Sigma}$ is assumed to be a positive definite symmetric matrix. To make the computation of the inverse matrix $\mathbf{\Sigma}^{-1}$ possible, the matrix $\mathbf{\Sigma}$ is required to be of full rank, meaning that all the eigenvalues of the matrix are greater than zero. When $\mathbf{\Sigma}$ is estimated from data, the requirement sets a lower limit on the number of observations: there has to be at least as many observations as there are features in the data ($n \geq p$).

The probability density function (PDF) of the normal distribution

$$p(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right) \tag{29}$$

defines the shape of the distribution. It integrates to one over all space and an integration over a given domain gives the probability of an observation falling into that domain.

When a multi-dimensional variable

$$\mathbf{X} = \left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right) \tag{30}$$

is partly observed such that

$$\mathbf{X}^{(1)} = (X_1, X_2, \ldots, X_q) \tag{31}$$

is a q-dimensional unobserved vector and

$$\mathbf{X}^{(2)} = (X_{q+1}, X_{q+2}, \ldots, X_p) \tag{32}$$

is an $p-q$-dimensional observed vector, and the mean vector $\boldsymbol{\mu}$ is split into two in a similar manner, the conditional distribution of the unobserved part becomes

$$\mathbf{X}^{(1)} | \left(\mathbf{X}^{(2)} = \mathbf{x}^{(2)}\right) \sim \mathcal{N}\left(\boldsymbol{\mu}^{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\right), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right), \tag{33}$$

where the dependency of the unobserved part on the observed part is defined by the covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \tag{34}$$

where due to the symmetry of the covariance matrix, equality

$$\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^T \tag{35}$$

holds.

The properties of normal distribution are utilized in many methods. Next, several multivariate methods that assume the normality of the data, are presented.

4.3 Principal component analysis

Principal component analysis (PCA) [16] is a well-established method that is used, for instance, in the fields of exploratory data analysis, dimensionality reduction and information visualization. PCA is usually defined as the orthogonal projection of the data onto a lower-dimensional linear space, known as the principal subspace, such that the variance of the projected data is maximized.

The optimal projection directions are the eigenvectors of the covariance matrix $\mathbf{S}$ or correlation matrix $\mathbf{R}$ of the data. The covariance matrix $\mathbf{S}$ is estimated from the data as

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^T, \tag{36}$$

where $\mathbf{x}_n$ is nth sample of the data set and $\overline{\mathbf{x}}$ is the vector of feature means. The correlation matrix $\mathbf{R}$ is estimated one element i, j at a time as

$$\rho_{ij} = \frac{1}{N}\sum_{n=1}^{N}\frac{(x_{ni} - \overline{x}_i)}{\sigma_i}\frac{(x_{nj} - \overline{x}_j)}{\sigma_j}, \tag{37}$$

where x_{ni} is the scalar value of the feature i in sample n, $\overline{x}_i$ is the scalar mean of the feature i over all samples, and σ_i is the standard deviation of the feature i. In both estimates, the summation goes through all N samples.

The ith eigenvector $\mathbf{u}_i$ and the corresponding eigenvalue λ_i fulfil the eigenvector equation

$$\mathbf{S}\mathbf{u}_i = \lambda_i\mathbf{u}_i \tag{38}$$

A $p \times p$-dimensional square matrix has p eigenvalues of which, in the case of covariance or correlation matrix, all are non-negative. The magnitude of an eigenvalue describes the amount

$$\sigma_i^2 = \lambda_i \tag{39}$$

of variance that is explained by the direction of the corresponding eigenvector, or component.

The eigenvectors are commonly called as the PCA coefficients. The value u_{ij} in the eigenvector $\mathbf{u}_i$ is the weight of the original variable j in the ith projection direction. The sign of the weight determines the direction to which the PCA projection is influenced by the original variable. The absolute value of the weight determines the importance of the original variable in the projection variable.

The projections, or PCA scores, $\mathbf{y}_j^{(r)}$, of the data vectors $\mathbf{x}_j$ on the first r components are obtained through linear projection

$$\mathbf{y}_j^{(r)} = \mathbf{U}_r^T \mathbf{x}_j, \tag{40}$$

where $\mathbf{U}_r$ is a $p \times r$-dimensional matrix with the first r scaled eigenvectors as its columns. The width of the projection matrix depends on the number of eigenvectors to be used in the projection.

There is a PCA solution when there are less samples than variables ($n < p$) [16]. In such case, the remaining $p - n$ eigenvalues are zero because the covariance matrix is not of full rank. Even though PCA is not based on a statistical model, there exists a probabilistic interpretation of the model.

Probabilisic PCA [16] is a latent variable model. In a latent variable model, the observations are assumed to be generated by latent variables, which are unobserved. In probabilistic PCA, the conditional distribution of the observed variable $\mathbf{x}$ given the latent variable $\mathbf{z}$ is defined as

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\mathbf{x}|\mathbf{U}_r^T\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right). \tag{41}$$

The scalar σ^2 is governing the variance of the conditional distribution. The marginal distribution of observations is

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \mathbf{U}_r^T\mathbf{U}_r + \sigma^2\mathbf{I}\right). \tag{42}$$

PCA is one of the most widely-used multivariate methods. It has similarities to several other multivariate methods with a latent variable interpretation, such as factor analysis.

4.4 Factor analysis

Factor analysis is a dimensionality reduction method that has many similarities to PCA. It is used for finding a small number of unobserved components called factors that describe high-dimensional data well.

Factor analysis is a linear-normal latent variable model. It is closely related to the probabilistic version of PCA. In factor analysis, a latent variable $\mathbf{z}$ is normally distributed with zero mean and unit variance:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{43}$$

The distribution of the observed variable $\mathbf{x}$ is conditioned on the value of the latent variable $\mathbf{z}$ and the distribution is again normal,

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Psi}), \tag{44}$$

in which the mean of $\mathbf{x}$ is a general linear function of $\mathbf{z}$ governed by the $p \times m$ projection matrix $\mathbf{W}$ and the p-dimensional vector $\boldsymbol{\mu}$. The projection matrix contains the factor coefficients. Covariance matrix $\boldsymbol{\Psi}$ is a $m \times m$-dimensional positive-definite diagonal matrix, where m is the number of latent components in use.

From a generative point of view, the p-dimensional observed variable $\mathbf{x}$ is defined by a linear transformation of the m-dimensional latent variable $\mathbf{z}$ plus additive normally distributed noise:

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \tag{45}$$

where

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}). \tag{46}$$

When integrating the latent variable $\mathbf{z}$ out, the marginal distribution of observed variable $\mathbf{x}$ becomes

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}\right). \tag{47}$$

In factor analysis, there is a rotational ambiguity in the projections $\mathbf{W}$. Usually this problem is overcome by using the varimax rotation, which maximizes the sum of the variances of the factor-wise squared coefficients. It has a tendency of producing high coefficient values for few variables and low coefficient values for the rest of the variables, leading to enhanced interpretability.

The definition of factor analysis differs from that of probabilistic PCA only in that the conditional distribution of the observed variable $\mathbf{x}$ given the latent variable $\mathbf{z}$ is taken to have a diagonal rather than isotropic covariance. In probabilistic PCA, the covariance matrix is a scaled identity matrix $\sigma^2\mathbf{I}$, whereas in factor analysis it is a positive definite diagonal matrix $\boldsymbol{\Psi}$. Because of a covariance matrix with higher number of degrees of freedom, factor analysis is not applicable when there are less samples than variables ($n < p$), while PCA is still solvable in such case.

4.5 MANOVA

MANOVA, or multivariate analysis of variance, is a multivariate statistical test of population effects analogous to the univariate ANOVA [15]. As shown in Subsection 3.4, also ANOVA can be extended to the multivariate case through multiple testing but it is not reliable when the features, or dependent variables, are correlated.

The test decides whether the two population means in multi-dimensional feature space are significantly different.

In the case of two-population setting, the MANOVA reduces to Hotelling's T-square test, which is related to linear discriminant analysis (LDA). The Hotelling's T-square statistic follows Hotelling's T^2 distribution, which is a non-central F-distribution. The statistic is defined as

$$t^2 = \frac{n_x n_y}{n_x + n_y}(\overline{\mathbf{x}} - \overline{\mathbf{y}})^T \mathbf{S}^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}}) \sim T^2(p, n_x + n_y - 2), \tag{48}$$

where n_x and n_y are the numbers of samples in the two populations, $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are the respective population mean estimates, and $\mathbf{S}$ is the pooled covariance matrix estimate

$$\mathbf{S} = \frac{\sum_{i=1}^{n_x}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T + \sum_{j=1}^{n_y}(\mathbf{y}_j - \overline{\mathbf{y}})(\mathbf{y}_j - \overline{\mathbf{y}})^T}{n_x + n_y - 2}. \tag{49}$$

The one-way MANOVA is a generalization of Hotelling's T-square test to a case of more than two populations. To separate k populations from each other, $k-1$ linear discriminants are required. Test statistic Wilk's λ determines, whether the whole set of discriminants significantly separates out the populations.

MANOVA is not suitable for analysis of data with less samples than variables, as then the sample covariance matrix is not invertible. Even with more samples, the results may not be trustworthy, because low population-specific sample-sizes may still lead to incorrect results.

MANOVA reveals whether some of the population means are significantly different but it does not provide information on which populations are the differing ones. Also, the test does not reveal from which of the features in the data the differences arise. To find out the differentially behaving features, further ANOVA tests are required to be performed on individual features.

The previously discussed ANOVA and MANOVA methods work for balanced experimental settings. Often, the experiment is somehow unbalanced, and this may happen in three ways [23]. Firstly, either the numbers of observations for the different treatment combinations may be unequal. Secondly, some of the treatment combinations may be missing altogether. Thirdly, in multivariate data, some of the experimental units may have been measured for only a subset of the response variables.

Unbalanced ANOVA methods have been proposed for these three cases of unbalanced design. The results, though, are highly dependent on the procedure that has been used. For univariate case, Bartlett's ANCOVA method attempts to deal with missing values. Also EM algorithm-based iterative methods have been developed for imputing missing values. Analysis of a set of balanced data using standard methods leads to correct estimates of the parameters but biased tests of significance.

5 The small sample-size problem

The straightforward multivariate generalization of ANOVA, MANOVA, is unfortunately useless when $n < p$, since the sample covariance matrix becomes singular. A further technical complication is that (M)ANOVA does not directly reveal the location or direction of the effect (up or down), and these have to be deduced by other methods. There are three common ways for tackling the small sample size problem: dimensionality reduction, regularization of the covariance matrix, and clustering of similarly behaving variables.

5.1 Dimensionality reduction

Dimensionality reduction methods are generally divided into two major approaches: feature selection and feature extraction. In the feature selection approach a subset of the original features are chosen to represent the data set and further analysis is carried out using only the selected features. In the feature extraction approach, the original features are transformed into hybrid features that are a combination of the original features. Projections, such as PCA and factor analysis, are typical examples of the feature extraction approach. Also the methods presented in sub-section utilize feature extraction.

The earliest attempts at performing ANOVA-like analysis with dimensionality reduction on multivariate datasets employ PCA as a dimensionality reduction method and perform ANOVA on the projected data [24]. The choice of number of components to use in the projection has been left for the user to decide. Analyzing the variances of several component scores still retains the multiple testing problem, even though the reduction in the number of tests might be considerable. Another fact to be criticized is that PCA performed on the pooled data does not necessarily find the directions of most variability that are related to the treatments. Still, these studies have built a foundation on more recent methods, like ASCA that is presented later in the section.

Multiple testing problem can be avoided by using MANOVA instead of ANOVA. Still, with datasets with large number of features, dimensionality reduction is required. The recently developed 50-50 F-test is one of the methods combining MANOVA and PCA [17]. The central idea of the 50-50 F-test is that the p-value of a test does not change under a linear transformation. The method divides the projection directions into three groups: first, the ones that represent important relationships to the treatments, second, the ones that do not represent these relationships, and third, the ones that are buffer directions between the former two groups. The division is based on the variance of the directions: the first group contains the directions of most variation and the second group the least. The difference between the ANOVA square sums of the group of most variation and the group of least variation is compared with an F-test.

The 50-50 F-test has turned out to be more reliable than MANOVA when there is

high collinearity between groups of variables. In addition, the method also works in the $n < p$ case where MANOVA fails to give any result. Still, this method does not reveal which of the metabolites differ between populations, so additional analysis following the test is required. Another weakness in the method is that PCA does not necessarily find projections relevant to the experimental design and the biological question.

Partial least-squares (PLS) is a commonly used method for regression and classification and can deal with collinear $n < p$ datasets [25]. However, PLS can overfit badly, and component scores found in a supervised manner are not necessarily reliable for interpretation [19].

Overfitting of PLS can to some extent be avoided with proper cross-validation. The same applies for most other methods too. The problem in many PLS experiments has been the mixing up of training and test sets. This leads to over-optimistic estimates in performance of the method.

For studying multivariate $n < p$ metabolomic datasets with a two-way experimental setup, a method called ANOVA-simultaneous component analysis or ASCA [18] has been proposed. The method performs dimensionality reduction with simultaneous component analysis (SCA), which is a PCA-like method useful in analyzing the simultaneous underlying variation in several related datasets. The data is split into populations according to the treatment covariates, and each treatment adds a specific projection corresponding to the effect of the treatment.

When PCA is used as a dimensionality reduction method, the choice of number of the PCA components has to be made in a heuristic way. Again, not even the strongest components necessarily provide biologically relevant directions. This drawback applies also to ASCA.

When the experimental design is balanced, that is, the number of samples in each population is equal, SCA procedure reduces to ordinary PCA. In ASCA, the ANOVA effects are solved independently of each other. Each effect is estimated from a separate simultaneous component projection. While this is a working solution, it involves major simplifications.

The example analysis presented in [18] does not include all of the two-way ANOVA terms. There is no technical limitation of including all the terms into the model but it is uncertain whether weak effects are found then. If the number of samples is low compared to the number of features, this simplification is justifiable for the sake of identifiability [26].

Because it is based on PCA, ASCA does not reduce the dimensionality by hard clustering but by assigning varying loadings on variables. Even though in ideal case the loadings matrix might be fairly sparse, in general the interpretability might not be so good.

A sparse Bayesian ANOVA model has been proposed for the $n < p$ case [27]. A linear four-way ANOVA model was applied to each gene, using a shared point-mass mixture prior to allow only a small fraction of effects to be non-zero. The

sparsity helps in controlling against false discoveries in multiple testing, and also in interpreting the results.

5.2 Covariance regularization

Regularization of the covariance matrix is another way to deal with $n < p$. The covariance matrix has to be made non-singular to use traditional statistical multivariate methods, such as Factor Analysis, MANOVA, Linear Discriminant Analysis, or Canonical Correlation Analysis (CCA). The simplest approach is to use a diagonal correlation matrix, which can be interpreted as assuming the variables to be (conditionally) independent. Lots of less drastic regularization methods have been proposed for shrinking the singular sample covariance matrix towards a positive definite matrix, usually a diagonal matrix; for instance [28, 29]. A usual procedure for restricting the projection matrix for Bayesian PCA [30] and FA [31] is by using an Automatic Relevance Determination prior (ARD). Recently sparsity has been imposed in Bayesian PCA and CCA [32], resulting in additional advantages in interpretability.

Bayesian sparse factor regression models [33], developed for gene expression data, are suitable for $n \ll p$ regression tasks. Sparsity is enforced by a heavy point-mass mixture prior allowing only a small fraction of regression coefficients to be non-zero. The method is useful in finding only the variables (genes) most strongly related to the external covariate and to infer relationships between the variables via common latent factors. The sparsity also helps in interpreting the components. The model was used for binary regression, corresponding to a one-way experimental setup.

5.3 Linear mixed models and clustering

It is common to assume that metabolites (as well as mRNAs) form strongly correlated groups, and then to study group-wise differential expression. Studying genes or metabolites one at a time results in a high risk of false positives when $n < p$, and the risk can be reduced by studying groups. This has been done on known groups of genes [34]; other usual approaches include clustering variables according to p-values or choosing only variables with a small enough p-value prior to doing multivariate analyses.

Several methods have been proposed for clustering gene-expression profiles with Linear Mixed Models, usually with a time-dependent experimental design [35, 36]. In a particularly interesting study [35], a model-based clustering algorithm was set up by assigning each cluster a subject and cluster-specific random effect common to genes in the cluster. The effect allows modelling correlations and clustering correlated genes, and the clustering solution was computed as the maximum likelihood estimate of the linear model additionally utilizing one-way covariate information as fixed effects. The primary interest of this method was clustering rather than the interpretation of the fixed effects, but it gives inspiration for us to progress to ana-

lyzing 2-way effects in a model regularized by assuming a cluster structure for the metabolites.

The current methods are not perfect for small sample-size data. Most methods do not estimate the uncertainty of the results, which is a major drawback in the data analysis of small sample-size data sets. Many dimensionality reduction methods fall short in the interpretability of the results as the hybrid features are a combination of several original features.

6 Bayesian approach

Methods that follow Bayesian data analysis paradigm have been developed for making inference from data using probabilistic models. The essential characteristic of Bayesian methods is their explicit use of probability for quantifying uncertainty in inferences.

The main benefits of Bayesian analysis are the abilities of estimating the uncertainty of the model, of explicitly incorporating prior knowledge into the model, and of being able to cope with datasets of small sample-size. Uncertainty of the model is presented explicitly through posterior distributions of individual variables and the joint posterior distribution of the model.

The Bayesian approach is suitable for analysis of small sample-size data because the approach provides a framework for estimating the uncertainty of the model. This is useful in controlling over-fitting of the model. Also the use of strong prior information is advantageous because the prior information can be used to limit the flexibility of the model.

In this section, the principles of Bayesian modelling are explained. Many Bayesian models utilize the normality assumption of the data. Bayesian treatment of the mean parameter of the normal distribution is demonstrated in Section 6.2. Also the concepts of hierarchical model, mixture model and sampling are explained in Section 6.

6.1 Bayesian modelling

Uncertainty is taken into account in Bayesian modelling such that, instead of a single model, the result of the modelling task is a distribution of possible models. Model family is a set of models that are of similar form. The members of a model family differ from each other by the values of the parameters of the model. Bayesian inference estimates a probability for each member of the model family given the observed data.

A Bayesian model consists of observed variables and unobserved parameters. For each of the model parameters, a prior probability distribution is assigned. The prior probability represents the expected distribution of the parametes before any observations of the random variables of the model have been made. Prior distributions are used by the modeller in order to incorporate prior knowledge, on the phenomenon that is being modelled, into the model. These priors affect the eventual result of the analysis and, thus, special attention will have to be paid on the choice of the priors.

As in general probabilistic modelling, one can calculate the probability, or likelihood, $p(x \mid \theta)$ of the observed data x given the values of the model parameters θ. Then, instead of finding the maximum likelihood estimate of the parameters, in Bayesian analysis one turns the likelihood function into a posterior probability function, which is a conditional distribution of parameters given the observed data. This is done by

using the Bayes' rule,

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}, \tag{50}$$

on the likelihood function $p(x \mid \theta)$ and the prior probability $p(\theta)$. The marginal likelihood of the data, $p(x)$, is usually hard to calculate but usually it is not needed because it remains constant over the analysis.

Using the Bayes' rule, the probability of observations given the model parameters can be expressed as a probability of model parameters given the observations. If the modeller is interested in obtaining a point estimate of the parameters, the so-called maximum *a posteriori* (MAP) estimate can be calculated. The MAP estimate is the mode of the probability function of parameters given the observations, $p(\theta \mid x)$. The MAP estimate gives the parameter values with the highest posterior probability for the model when certain observations have been made.

Conjugate prior is a type prior that is usually preferred due to the ease of computation of the posterior probability. In Bayesian analysis, the posterior distribution of a variable is calculated as a product of the likelihood and the prior divided by a constant normalization term. Prior $p(\theta)$ is conjugate to the likelihood function $p(x \mid \theta)$ if their product is a distribution of the same family as the likelihood function. Usually the evidence term can be left out because it is a normalization term that is not dependent on the parameter values.

6.2 Bayesian multivariate normal distribution

Numerous statistical methods make an assumption of normality of the model variables. In Bayesian analysis, the posterior distribution of the model parameters is calculated. For the multivariate normal distribution, equations of the posterior distribution of the mean parameter are presented in this sub-section.

In the case of multivariate normal distribution with known covariance matrix $\mathbf{\Sigma}$, the common conjugate prior for the mean parameter is normal distribution $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{\Lambda}_0)$. The posterior distribution of the mean parameter $\boldsymbol{\mu}$ is then

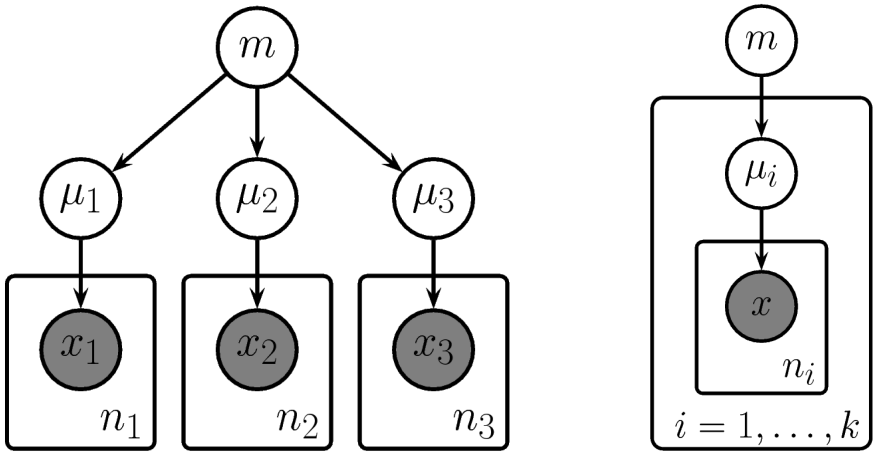
$$p(\boldsymbol{\mu}|\boldsymbol{x}, \mathbf{\Sigma}) \propto \exp\left(-\frac{1}{2}\left[(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\mathbf{\Lambda}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) + \sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right]\right) \tag{51}$$

which can be completed to a quadratic form,

$$\begin{aligned} p(\boldsymbol{\mu}|\boldsymbol{x}, \mathbf{\Sigma}) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_n)^T\mathbf{\Lambda}_n^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_n)\right) \\ &= \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_n, \mathbf{\Lambda}_n) \end{aligned} \tag{52}$$

where the mean and covariance parameters $\boldsymbol{\mu}_n$ and $\mathbf{\Lambda}_n$ are

$$\boldsymbol{\mu}_n = (\mathbf{\Lambda}_0^{-1} + n\mathbf{\Sigma}^{-1})^{-1}(\mathbf{\Lambda}_0^{-1}\boldsymbol{\mu}_0 + n\mathbf{\Sigma}^{-1}\overline{\boldsymbol{x}}) \tag{53}$$



(a) A model of three groups

(b) A model of k groups and n_i observations in group i

Figure 2: Plate diagrams of hierarchical models of groups of observations. In the models, each group i has a specific parameter μ_i. The group-specific parameters have a shared hyperparameter m. Multiple instances of a variable are encoded by a plate, where the number of instances is shown at the down-right corner of the plate.

and

$$\mathbf{\Lambda}_n^{-1} = \mathbf{\Lambda}_0^{-1} + n\mathbf{\Sigma}^{-1}. \tag{54}$$

The inverse of the covariance matrix is called the precision. This distribution is a weighted average of the data and the prior mean, with weights given by the data and prior precision matrices, $n\mathbf{\Sigma}^{-1}$ and $\mathbf{\Lambda}_0^{-1}$, respectively. The posterior precision is the sum of the prior and data precisions. [37]

6.3 Hierarchical models

Many statistical applications involve multiple parameters that can be regarded as related or connected. This implies that the joint probability model for these parameters should reflect the dependence among them.

It is natural to model many problems hierarchically, with observable outcomes modelled conditionally on certain parameters, which themselves are given a probabilistic specification in terms of further parameters, known as hyperparameters. Nonhierarchical models are usually inappropriate for hierarchical data: with few parameters, they generally cannot fit large datasets accurately, whereas with too many parameters, they tend to overfit.

In a typical hierarchical model, multiple observations arise indirectly from a shared latent factor. For instance, when measuring the same feature from several individuals, each individual may have a personal expected value for the observation. These

individual-specific expected values then may come from one distribution, the prior distribution of the expected values. The expected value of the individual-specific expected values is then called a hyperparameter.

A typical example of a hierarchical model is a set of grouped observations, with a group-specific mean parameter. Each observation x_{ij} is part of one group i and, thus, follows the group-specific mean μ_i. The group-specific means then arise from a shared hyperparameter m.

In Figure 2(a), a graphical model of the simple model of three observation groups is shown. Group-specific means μ_1, μ_2 and μ_3 are generated by the hyperparameter m. The plates indicate that there are several, in fact n_1, n_2 and n_3, observations x in groups 1, 2 and 3, respectively. Figure 2(b) represents a generalization of the three-group model into a k-group model

If there is no other information than the observations, no ordering or grouping of the individual-specific parameters can be made. Then, these parameters are assumed to be symmetric in their prior distribution, that is, they are independent and identically distributed and share a common hyperparameter. The symmetry is represented probabilistically by exchangeability of the parameters in their joint distribution.

The symmetry assumption can, though, be broken if additional information on the phenomenon is available. For instance, if it is known that some of the individuals have been exposed to a treatment, the exchangeability between the treated and untreated individuals is lost, unless this information is somehow included into the model.

The unobserved hyperparameter ϕ affects observations x only through the unobserved parameters θ, which is why the joint posterior distribution of this hierarchical model becomes

$$\begin{aligned} p(\phi, \theta|x) &\propto p(\phi, \theta)p(x|\phi, \theta) \\ &= p(\phi, \theta)p(x|\theta) \\ &= p(\phi)p(\theta|\phi)p(x|\theta), \end{aligned} \tag{55}$$

where $p(\phi)$ is the marginal probability of the hyperparameter, $p(\theta|\phi)$ is the conditional probability of the parameter θ given the hyperparameter ϕ, and $p(x|\theta)$ is the likelihood of the observations x.

6.4 Clustering with Gaussian mixture model

Mixture distributions arise in practical problems when the measurements of a random variable are made in several unknown conditions. Often, it is a better idea to model the data with a mixture of multiple simple unimodal distributions than with a single multimodal distribution. Mixture modelling is a good approach to a problem where it is reasonable to expect that the populations of observations or

sampled parameters consist of several sub-populations, each of which is relatively simple to model separately.

As well as k-means clustering, the mixture model has a pre-selected number of centroids, on each of which a statistical distribution is then centered. Normal distribution is a common choice as the unimodal distribution. The marginal distribution of an observation $\mathbf{x}$ is a sum of these k distributions, and instead of a hard clustering of k-means clustering, each point in the data space has a continuous probability of belonging to any of the k clusters, depending on the width and distance of the distributions.

When setting up and computing a mixture model, an unobserved binary indicator variable $\mathbf{z}$ is introduced. This indicator variable specifies the mixture component from which the corresponding observation is drawn. Each mixture component k is assigned the prior probability

$$p(z_k = 1) = \pi_k \tag{56}$$

that sums to one over all components.

When the mixture component that an observation arises from is known, the indicator variable $\mathbf{z}$ becomes fixed, so that only the kth value corresponding to the known component k is one in the vector, while other values of the vector are zero. The conditional probability of the observation given the known mixture component k then simply becomes

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{57}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and covariance of the mixture component k, respectively. In practice, the closer an observation is to the given component centroid, the higher conditional probability the observation gains.

As each observation $\mathbf{x}$ is assumed to be drawn from one of the components, the marginal distribution of the observation is calculated from the joint distribution of the observation and the indicator variable $\mathbf{z}$ by summing over all possible states of the indicator variable:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \tag{58}$$

Using the Bayes' rule the probability of a mixture component given the data can be calculated as

$$\begin{aligned} p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \end{aligned} \tag{59}$$

In clustering, each observation can then be assigned to the component j that has the highest probability given the observation, $p(z_j = 1|\mathbf{x})$. The cluster assignments can

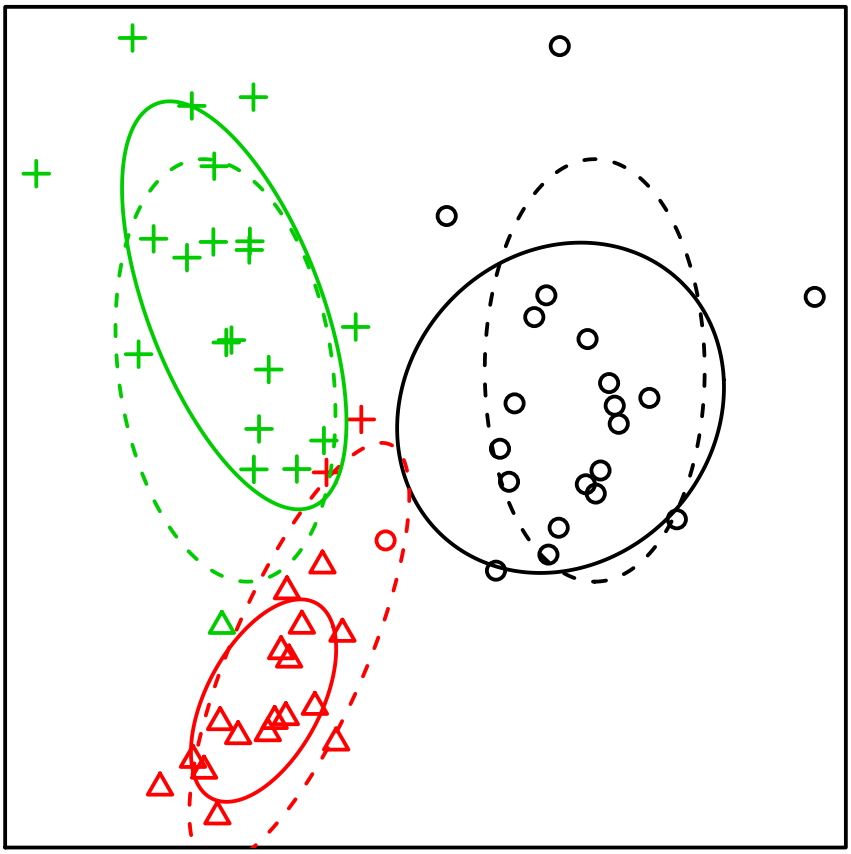


Figure 3: An example clustering with mixtures of Gaussians. The original three classes are encoded into colors of the data points, whereas the found classes are encoded into shapes of the points. The original 67 % contours of the distributions are shown with dashed lines and the found contours with solid lines.

be updated iteratively in turn with updating the centroid positions. This method is a version of the expectation-maximization (EM) algorithm [38], which finds a local optimum for the mixture component positions.

An example is presented in Figure 3. The example consists of data points on a two-dimensional plane. Each point has been drawn from one of three bivariate normal distributions. When a mixture of Gaussians algorithm with three centroids is applied to the problem, most of the observations are classified correctly. The shapes of the distributions resemble the original distributions but are not equal to them, as the sample size is finite.

Gaussian mixture model is usually used to cluster samples. In the new model presented in this thesis, it is used as a part of a hierarchical model to cluster features.

6.5 Inference

For most probabilistic models of practical interest, exact inference is intractable, meaning that the joint distribution of the model is not computable in a closed form. This leads to a need of approximation methods. There are two widely-used ways for approximating the joint distribution. The first approach is based on deterministic inference algorithm-based techniques, and the second approach utilizes numerical sampling-based stochastic techniques.

The inference algorithms make use of simplified deterministic approximations of the underlying real distribution. The result is quick to compute once good approximating distributions have been found. Methods such as variational Bayes and expectation propagation fall into this category.

The other category includes methods that approximate the joint distribution $p(\mathbf{z})$ by taking samples from simpler conditional distributions. The sampling methods and especially Gibbs sampling are presented in the next sub-sections.

6.5.1 Sampling methods

The general idea behind sampling methods is to obtain a set of parameters $\mathbf{z}^{(l)}$ drawn independently from the distribution $p(\mathbf{z})$. Sampling then allows the expectation

$$E[f] = \int f(\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} \tag{60}$$

to be approximated by a finite sum

$$\hat{f} = \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{z}^{(l)}). \tag{61}$$

For graphical models with directed graph structure, the joint distribution is straightforward to approximate with an ancestral sampling approach, where the conditional distribution of an unobserved variable is dependent on the variables within its Markov blanket. The blanket includes the parents and the children of the variable, and the co-parents of the children. The joint distribution is then a product of all the conditional distributions in the model.

Markov chain Monte Carlo (MCMC) is the most common family of methods used in sampling. Markov chain is a series of random variables, where the value of a new member of the chain only depends on the current value, and not on the former values. This conditional independence is called the Markov property.

The Metropolis algorithm is a method for jumping from a set of variable values to another. The set of values at one instance is called the state of the algorithm. The acceptance of the new state depends on the probability of the old and new state. A jump to more probable state is accepted with probability one, whereas the

probability of jumping to a less probable state is a quotient of the probabilities of the new and old state. Generally, the jump to a new state is accepted with probability

$$A(\mathbf{z}^*, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\widetilde{p}\,(\mathbf{z}^*)}{\widetilde{p}\,(\mathbf{z}^{(\tau)})}\right), \tag{62}$$

where $\mathbf{z}^*$ is the new candidate state and $\mathbf{z}^{(\tau)}$ is the current state.

The Metropolis-Hastings algorithm is a generalization of the Metropolis algorithm to the case where the proposal distribution is no longer a symmetric function of its arguments. Then, the probability of jumping from a state to another is different from the probability of performing the jump into the opposite direction between the same states.

Gibbs sampling is a special case of the Metropolis-Hastings algorithm. It is used to obtain approximate inference in the new method presented in this thesis. In the next sub-section, Gibbs sampling is explained in more detail.

6.5.2 Gibbs sampling

Gibbs sampling is a simple MCMC algorithm. It uses the fact that in many cases it is straightforward to approximate the joint distribution step-by-step through conditional distributions, even though the joint distribution itself is not directly computable in closed form.

In Gibbs sampling, a new value is drawn for one variable at a time from a conditional distribution, where all the other variables of the model are given the current value of the state. In practice, many of the given variables drop out of the conditional distribution, as they are outside the Markov blanket of the conditioned variable. Gibbs sampling is especially useful when the marginal distributions of neighboring variables are conjugate to each other, leading to conditional distributions of similar form as the marginal distributions.

The Gibbs sampling procedure for M variables is:

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$
   - Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$
   - $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$
   - $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$ (63)

In the algorithm, $z^{(\tau)}$ represents the previous state of the Markov chain and $z^{(\tau+1)}$ is the new state that is sampled on iteration round τ. The Gibbs sampling is known to

converge to the joint distribution of the model. Because only one variable is updated at a time, the convergence is slow, however.

When one iteration round is complete, one new sample with all updated variables has been obtained. Gibbs sampling differs from the standard Metropolis-Hastings algorithm in the sense that one variable is updated at a time, and that the update is accepted in any case. Because of this property, there is strong correlation between successive samples, which is called auto-correlation of the chain.

The issue of auto-correlation can be overcome by thinning the chain. Then, only every nth sample is included into the final distribution estimate. The integer n is selected so that there are enough samples between the two samples that are to be collected to fade out the auto-correlation.

Also because of auto-correlation, the beginning part of the chain is highly dependent on the initial values of the sampled variables. Burn-in is a procedure, where the initial part of the chain is discarded because they do not represent the underlying joint distribution as the chain has not converged yet. The burn-in phase is not shown in the algorithm above.

The convergence of a Markov chain to the underlying joint distribution can be guarded with several techniques. One way is to run several parallel Markov chains and compare the within and between variances of them, as is done when computing the potential scale reduction factor (PSRF) [39]. It is widely used but also has its drawbacks: running parallel chains is computationally costly and inefficient [40].

When the estimated model includes several components or clusters, label switching problems arise. In label switching, the components mutually switch places. The phenomenon is normal because a good sampler algorithm traverses through all possible states. Label switching causes problems in inference because the posterior distributions of the components become mixed and then do not correspond to a single component, as is assumed. This is a major problem of sampling methods and only partly heuristic solutions are available to overcome it.

The label switching problem arises also when analyzing parallel Markov chains. However, aligning components of parallel Markov chains is not as difficult task as that of detecting label switching within one Markov chain.

The most important downside of Gibbs sampling is the amount of computation required for obtaining a reliable estimate of the joint distribution. In case of a complex hierarchical model, the estimation of the joint distribution may require even tens of thousands of Gibbs samples. The longer the sampling is continued and the more parallel chains are used, the better the estimate asymptotically is.

In the new method presented in this thesis, Gibbs sampling is being used. As it is a complex hierarchical model, the joint distribution is not computable in closed form. Variational approximation would lead to faster computation but its implementation is more complex than that of the Gibbs sampler algorithm.

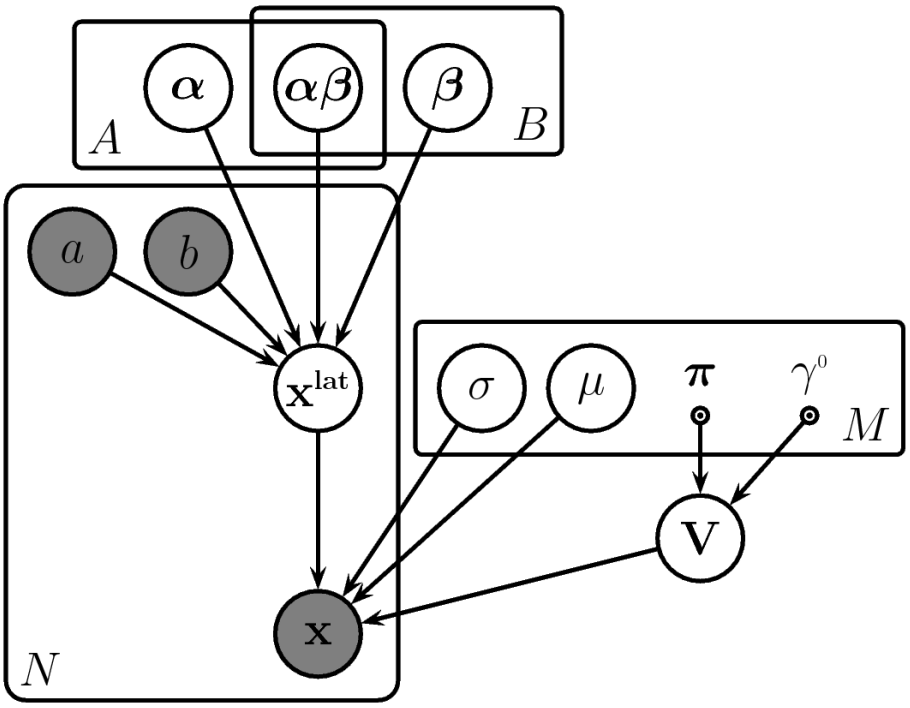


Figure 4: Plate diagram of the two-way clustering factor analysis model.

7 The new method

In this section, a new model is formulated for two-way analysis. It is a combination of two-way analysis and factor analysis models, where the ANOVA-type two-way effect terms are assigned as hyperparameters of the latent factors. To deal with the small sample size, the projection matrix is formed as a sparse clustering matrix containing only one non-zero element for each feature. This is particularly reasonable under the assumption that metabolomics data contains strongly correlated groups of variables. The projection matrix is now non-singular even in the $n \ll p$ cases. The joint posterior distribution is evaluated with Gibbs sampling procedure. The new method was published in the journal Data Mining and Knowledge Discovery [41].

In effect the model shown in Figure 4 consists of a factor analyzer and an ANOVA-type two-way structure. In the factor analyzer, the loadings assume cluster memberships multiplied with scales. The population-specific priors assume a two-way structure.

7.1 Modelling metabolomic datasets

Metabolomic data has certain properties that are important for the model to take into account. Even after this customization, the resulting model still is a reasonably general multi-way factor analysis model.

Due to the existing biochemical pathways where metabolites are converted to one another by chemical reactions, metabolomics data contains correlations caused by tiny fluctuations in metabolic concentrations being transmitted through the pathway. Groups of metabolites are strongly correlated even over biological replicates having the same experimental treatment, a feature not apparent in for instance gene expression data where the correlations mainly result from responses of the genes to the external perturbations [6]. Another peculiar feature of metabolomics data is that mean concentrations and scales of different metabolites vary by orders of magnitude; they can be modelled by metabolite-specific means and scales of, say, a healthy control group.

Factor analysis models where latent factor(s) fluctuating around zero are assumed to generate correlated fluctuations around the variable-specific means, fit well the above assumptions. To solve the $n < p$ limitation of factor analysis and to simplify the interpretation of the results, each variable is assumed to be generated by exactly one factor. The factor analysis task can now be interpreted to include model-based clustering of features as a subtask. Biologically the task is related to finding subparts of linear pathways, which is a current research trend in bioinformatics [42].

The effects of covariates, such as disease, are assumed to be visible in the same factors that describe the activity of parts of the biochemical network. These effects appear as up- or down-regulations of the factors. The healthy control biological replicates are assumed to fix the "coordinate basis" of the problem, from which the up- and down-regulations deviate the means of factor values.

Multi-way modelling in high-dimensional metabolomics data, with grouping assumptions made to regularize the problem, is a new approach for generative modelling of measurement data. Sparse latent factor models [33], being regression-type approaches, can only be used to discover variation of the data that is explained by external covariates. This is reasonable for gene expression data also considering that it has been claimed [6] that for gene expression data, correlations between variables arise mainly due to responses to external variation. However, in metabolomics, fluctuations due to biochemical pathways themselves are another important source of variation that can be useful for instance in finding biological pathways not responding to external covariates. This motivates to construct a model for the whole dataset with hierarchical generative modelling.

The clustering methods based on linear mixed models [35, 36] have so far not been used for multi-way experimental setups, and because their main goal is clustering they have not considered estimation of statistical significance of the effects.

As PCA in general, ASCA, the only currently existing method addressing the multi-way generative $n < p$ metabolomics data, can only be considered as an exploratory visualization of PCA scores of one effect at a time. It does not estimate the statistical significance of the effects, although an approach based on permutation tests was later proposed [43].

In summary, a method combining central aspects needed to model metabolomic datasets in a single, hierarchical generative model, is introduced. The two-way

experimental setup of the research problem is included as population-specific priors on the latent variables. As a projection matrix, a scaled clustering matrix enabling $n < p$ cases is used. It allows an easy interpretation of the clusters related to the different latent factors. Inference on the statistical significance of the effects of external covariates is done by studying the confidence intervals of the posterior distribution. The method is additionally capable of finding clusters of correlated metabolites that are not related to external covariates, but can be interpreted as sub-parts of biochemical pathways. The method generalizes directly to a general multi-way analysis, but for simplicity of presentation the model is introduced in the two-way case.

7.2 Dimensionality reduction

Factor analysis (FA) model [44] for n exchangeable replicates of the control group is

$$p(\mathbf{x}_j|\mathbf{V}, \mathbf{x}_j^{lat}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \mathcal{N}(\mathbf{x}_j|\boldsymbol{\mu} + \mathbf{V}\mathbf{x}_j^{lat}, \boldsymbol{\Psi}), \tag{64}$$

where the latent variables $\mathbf{x}_j^{lat}$ are normally distributed:

$$\mathbf{x}_j^{lat} \sim \mathcal{N}(0, \mathbf{I}). \tag{65}$$

Here $\mathbf{x}_j$ is a p-dimensional observation vector, $j = 1, \ldots, n$, and $\mathbf{V}$ is the projection matrix that is assumed to generate the data vector $\mathbf{x}_j$ from the latent variable $\mathbf{x}_j^{lat}$. The $\mathbf{V}\mathbf{x}_j^{lat}$ models such common variance of the data around the feature means $\boldsymbol{\mu}$ that can be explained by factors common to all or many features, effectively estimated from the sample covariance matrix of the dataset. The sample covariance becomes decomposed into $\hat{\boldsymbol{\Sigma}} = \mathbf{V}\mathbf{V}^T + \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is a diagonal residual variance matrix with diagonal elements σ_i^2 modelling the feature-specific noise not explained by the latent factors. The elements of the latent variable vector $\mathbf{x}_j^{lat}$ are known as factor scores. Following the discussion on unidentifiability problems in [44], the covariance matrix of $\mathbf{x}^{lat}$ is set to be the identity matrix.

At this point, the covariates are not yet assumed to induce any special effects, and when $n < p$, $\mathbf{V}$ cannot be estimated due to the singularity of the sample covariance matrix.

Now, the model is extended in two complementary directions. First, $\mathbf{V}$ is restricted to be a non-singular sparse clustering matrix, suitable for data containing highly correlated groups of variables. Secondly, the model is extended to include a two-way experimental setup in the latent variable space.

The data are assumed to contain strongly correlated groups of metabolites, each of which is governed by one latent variable. The projection matrix $\mathbf{V}$ is a positive-valued clustering matrix where each row has one non-zero element corresponding to

the cluster assignment of the variable:

$$\mathbf{V} = \begin{bmatrix} \gamma_1 & 0 & 0 \\ 0 & 0 & \gamma_2 \\ \vdots & \vdots & \vdots \\ 0 & \gamma_j & 0 \\ 0 & \gamma_{j+1} & 0 \\ \vdots & \vdots & \vdots \end{bmatrix}. \tag{66}$$

The positive-valued clustering matrix leads to clusters with positively correlated features. Including negative correlations by not limiting the matrix would be possible but is not reasonable considering the metabolic pathway structure.

The location of the non-zero value on row i is encoded by indicator v_i. It follows a categorical distribution, that is, multinomial distribution with a single observation, with an uninformative prior distribution $\boldsymbol{\pi}_i$ that does not depend on the size of the cluster. However, $\boldsymbol{\pi}_i$ could be used to encode prior information on the known grouping of variables.

The variation of each feature within a cluster is assumed to be modeled by the same latent variable, but the scales may differ. The scales γ_i are assigned heavy empirical priors γ_i^0 that keep them close to the values of the control group, to make the γ_i and the population prior effects identifiable. The distribution of γ_i is parametrized as a scaled inv-χ^2 distribution with a degrees-of-freedom weighted sum of empirical prior and data scale, as done in Gelman [37].

The feature-specific residual variances σ_i^2, that are the diagonal elements of $\boldsymbol{\Psi}$, follow a scaled inv-χ^2 with an uninformative prior.

In summary, the covariance matrix is regularized by assuming that the main correlations are positive correlations between features belonging to the same cluster. This correlation is mediated through a common latent variable, which is a reasonable assumption for metabolomics data.

7.3 Two-way analysis

For two-way analysis it is assumed that the samples have been classified into two sets of classes, $a = 0, \ldots, A$ and $b = 0, \ldots, B$. A traditional two-way (M)ANOVA model would be

$$\mathbf{x}_j|_{\text{class}(j)=(a,b)} = \boldsymbol{\mu} + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab} + \boldsymbol{\epsilon}_j, \tag{67}$$

where "class(j)" denotes the class labels of sample j, $\boldsymbol{\mu}$ is the grand mean over all samples, $\boldsymbol{\alpha}_a$ and $\boldsymbol{\beta}_b$ are the main effects of the two directions and $(\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}$ are the interaction effects for $a = 0, \ldots A$ and $b = 0, \ldots B$.

It is assumed that the ANOVA-type effects act on the latent variable space, which makes sense both in terms of the interpretation of the latent variables as activities of

metabolic pathway parts, and in making it possible to estimate the model for small sample sizes. In the K-dimensional latent variable space the (M)ANOVA model becomes

$$\mathbf{x}_j^{lat}|_{\text{class}(j)=(a,b)} = \boldsymbol{\mu}^K + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab} + \boldsymbol{\epsilon}_j^K, \tag{68}$$

where "class(j)" denotes the class labels of sample j, and K denotes lower dimensionality.

The ANOVA effects are set as population priors to the latent variables, which in turn are given Gaussian priors

$$\boldsymbol{\alpha}_a, \boldsymbol{\beta}_b, (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{69}$$

To simplify the interpretation of the effects, the model now deviates from the standard ANOVA convention. A similar choice has been done successfully in other ANOVA studies [27], and it does not significantly sacrifice generality. The parameter vector $\boldsymbol{\mu}$ describing feature-specific means is set to the mean of one class, the control group, instead of the grand mean. One group now becomes the baseline to which other classes are compared by adding main and interaction effects. The terms $\boldsymbol{\alpha}_0$, $\boldsymbol{\beta}_0$, $(\boldsymbol{\alpha}\boldsymbol{\beta})_{00}$, $(\boldsymbol{\alpha}\boldsymbol{\beta})_{a0}$ and $(\boldsymbol{\alpha}\boldsymbol{\beta})_{0b}$ therefore become zero. The difference between the classes is now modelled directly with $\mathbf{x}^{lat}$ and hierarchically by the main effects $\boldsymbol{\alpha}_a$, $\boldsymbol{\beta}_b$ and $(\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}$.

As a simple example, consider 2×2 ANOVA analysis. The classes are now $(a, b) = (0, 0), (1, 0), (0, 1), (1, 1)$. The ANOVA terms for samples belonging to different classes are

$$\begin{aligned}
&\mathbf{x}_j^{lat}|_{(a,b)=(0,0)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\
&\mathbf{x}_j^{lat}|_{(a,b)=(1,0)} \sim \mathcal{N}(\boldsymbol{\alpha}_1, \mathbf{I}), \\
&\mathbf{x}_j^{lat}|_{(a,b)=(0,1)} \sim \mathcal{N}(\boldsymbol{\beta}_1, \mathbf{I}), \text{ and} \\
&\mathbf{x}_j^{lat}|_{(a,b)=(1,1)} \sim \mathcal{N}(\boldsymbol{\alpha}_1 + \boldsymbol{\beta}_1 + (\boldsymbol{\alpha}\boldsymbol{\beta})_{11}, \mathbf{I}).
\end{aligned} \tag{70}$$

There is no effect estimated for the control class $(a, b) = (0, 0)$. The terms $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ now directly model the difference of the samples in the two directions compared to the control group, and the interaction term $(\boldsymbol{\alpha}\boldsymbol{\beta})_{11}$ models the interactions of the two directions. In standard ANOVA, four main effects and four interaction effects would have to be estimated and compared. The inference on the statistical significance of the ANOVA effects now reduces to inferring whether the posterior distribution of these effects is above (up-regulation) or below (down-regulation) zero with, say 95 % probability. Each component of the terms, representing different clusters, is estimated individually. It is worth noting that having only one class would reduce the problem to factor analysis.

The hierarchical model is summarized as

$$\begin{aligned}
\boldsymbol{\alpha}_0 = 0, \boldsymbol{\beta}_0 = 0, (\boldsymbol{\alpha}\boldsymbol{\beta})_{a0} = 0, (\boldsymbol{\alpha}\boldsymbol{\beta})_{0b} = 0, (\boldsymbol{\alpha}\boldsymbol{\beta})_{00} = 0 \\
\boldsymbol{\alpha}_a, \boldsymbol{\beta}_b, (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab} \sim \mathcal{N}(0, \mathbf{I}) \\
\mathbf{x}_j^{lat}|_{\text{class}(j)=(a,b)} \sim N(\boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}, \mathbf{I}) \\
\mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu} + \mathbf{V}\mathbf{x}_j^{lat}, \boldsymbol{\Psi}).
\end{aligned} \tag{71}$$

7.4 Empirical prior

To fix the means of the control group as the baseline of the ANOVA-type analysis, strong empirical priors are used for the grand mean $\boldsymbol{\mu}$ and the scale parameter γ_i. The γ_i^0 is the standard deviation of the control group, and n_0 controls the strength of the prior. The total number of samples in the data set, $n_0 = n$, is used as the prior strength. The $\boldsymbol{\mu}$ is the mean vector calculated over the control group. For simplicity and following the results of [45], $\boldsymbol{\mu}$ is subtracted from the whole data and is not sampled, corresponding to the centering discussed in section 7.3.

7.5 Gibbs sampling

In this sub-section, the Gibbs sampling equations for the new model are presented. Thus, even though it is not stated explicitly, the following equations are conditional distributions of the model variables. In the equations, samples are indexed by $j = 1, \ldots, n$, features by $i = 1, \ldots p$, and clusters by $k = 1, \ldots, K$.

Conditional distribution of the latent variable $\mathbf{x}^{lat}$ is normal,

$$\mathbf{x}_j^{lat} \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_j^{lat}, \hat{\boldsymbol{\Sigma}}^{lat}\right), \tag{72}$$

with mean

$$\hat{\boldsymbol{\mu}}_j^{lat} = \hat{\boldsymbol{\Sigma}}^{lat}\left(\mathbf{V}^T\boldsymbol{\Psi}^{-1}\mathbf{x}_j + \boldsymbol{\alpha}_a + \boldsymbol{\beta}_b + (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab}\right) \tag{73}$$

and covariance

$$\hat{\boldsymbol{\Sigma}}^{lat} = \left(\mathbf{V}^T\boldsymbol{\Psi}^{-1}\mathbf{V} + \mathbf{I}\right)^{-1}. \tag{74}$$

The ANOVA effects are sampled as

$$\boldsymbol{\alpha}_a \sim \mathcal{N}\left(\frac{1}{n_a + 1}\sum_{j \in a}\left(\mathbf{x}_j^{lat} - \boldsymbol{\beta}_{b_j} - (\boldsymbol{\alpha}\boldsymbol{\beta})_{ab_j}\right), \frac{1}{n_a + 1}\mathbf{I}\right), \tag{75}$$

$$\boldsymbol{\beta}_b \sim \mathcal{N}\left(\frac{1}{n_b + 1}\sum_{j \in b}\left(\mathbf{x}_j^{lat} - \boldsymbol{\alpha}_{a_j} - (\boldsymbol{\alpha}\boldsymbol{\beta})_{a_j b}\right), \frac{1}{n_b + 1}\mathbf{I}\right), \tag{76}$$

$$(\boldsymbol{\alpha}\boldsymbol{\beta})_{ab} \sim \mathcal{N}\left(\frac{1}{n_{ab} + 1}\sum_{j \in ab}\left(\mathbf{x}_j^{lat} - \boldsymbol{\alpha}_{a_j} - \boldsymbol{\beta}_{b_j}\right), \frac{1}{n_{ab} + 1}\mathbf{I}\right), \tag{77}$$

where n_a, n_b and n_{ab} denote the number of samples belonging to group a, b, and both a and b, respectively. In practice, the influence of all other effects is subtracted from the latent variable when the effect in question is being sampled.

Finally, each feature i is assigned to one cluster using multinomial distribution with the probability of cluster k being

$$p\left(v_i = k\right) = \frac{\pi_k \Pi_j p\left(x_{ji} | \mu_i + \gamma_i x_{jk}^{lat}, \sigma_i\right)}{\Sigma_k \pi_k \Pi_j p\left(x_{ji} | \mu_i + \gamma_i x_{jk}^{lat}, \sigma_i\right)}. \tag{78}$$

The residual variance and the scale parameter are feature-specific variables. They are sampled each element separately from inv-χ^2 distributions

$$\sigma_i^2 \sim \text{Inv-}\chi^2\left(n, \Sigma_j \left(x_{ij} - \mu_i - \gamma_i z_{jk}\right)^2\right), \tag{79}$$

and

$$\gamma_i^2 \sim \text{Inv-}\chi^2\left(n + n_0, \frac{n\hat{\gamma}_i^2 + n_0 {\gamma_i^0}^2}{n + n_0}\right), \tag{80}$$

where

$$\hat{\gamma_i} = \frac{\Sigma_j \left(x_{ji} x_{jk}^{lat}\right)}{\Sigma_j \left(x_{jk}^{lat}\right)^2}. \tag{81}$$

7.6 Model selection using predictive likelihood

The model complexity depends on the number of clusters in the dimensionality reduction part. The optimal number of clusters can be found with a cross-validation scheme by comparing the predictive likelihoods of different models on test data.

In cross-validation scheme, the complete data is split into training and validation parts. The training part of the data is used for learning the model, which in this case means performing the sampling of unobserved variables in a Bayesian way. The learned model is then used for computing the likelihood of the validation data. The estimate for the new data likelihood is obtained by calculating the likelihood using the hidden variable values of one posterior sample at a time.

As sampling gives an approximation on the distribution of unobserved variables of the model given the observations, computing predictive likelihood gives an approximation on the likelihood of a new observation $\tilde{x}$ given the previous observations x:

$$\begin{aligned} p(\tilde{x}|x) &= \int p(\tilde{x}, \theta|x) \mathrm{d}\theta \\ &= \int p(\tilde{x}|\theta, x) p(\theta|x) \mathrm{d}\theta \\ &= \int p(\tilde{x}|\theta) p(\theta|x) \mathrm{d}\theta. \end{aligned} \tag{82}$$

This integral of predictive likelihood is approximated by computing the product of the posterior and predictive probabilities for every set of unobserved variable values θ. The estimate of the expectation value of the predictive likelihood is then simply the mean of the new data likelihood given one posterior sample:

$$p(\tilde{x}|x) \approx \frac{1}{L} \sum_{l=1}^{L} p(\tilde{x}|\theta^{(l)}). \tag{83}$$

In addition to the estimate of the expectation value, the Bayesian approach also enables the estimation of the distribution of the predictive likelihood value. This means that the approach gives an estimate on any quantile of the distribution, leading to a possibility of showing interpretable and credible confidence intervals.

A 5-fold cross-validation scheme was used, meaning that the cross-validation was repeated five times with different divisions of data into training and validation sets. The distribution of the predictive likelihood value can be estimated over all these repeats. Then, a model complexity that has highest predictive likelihood can then be selected for use in further analysis of new, independent, data.

7.7 Inference from posterior distributions

The posterior distribution of each model parameter is approximated by the set of posterior samples obtained via the sampling scheme. Inference bases on these posterior distributions. The interesting parameters are the clustering matrix $\mathbf{V}$ and the ANOVA effects, which in the case of a two-way analysis are, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $(\boldsymbol{\alpha\beta})$. Additionally, the metabolite-specific parameters scale γ, mean μ and residual σ may provide interesting information for the modeller.

Together, the clustering matrix and the ANOVA effects show which metabolites act in a similar way over the treatments. The set of posterior samples of the clustering matrix $\mathbf{V}$ contains the information on which metabolites are clustered together at each sampling step. The clustering information is more interpretable, when the cluster to which the metabolite most often belongs is calculated. This is called the mode of the clustering. Usually the mode is rather stable in the sense that in most of the posterior samples, the metabolite is assigned to its mode cluster. Thus, the mode clustering is a fairly good binary representation of whether two metabolites are behaving in a similar fashion or not.

Each ANOVA effect is an estimate on the difference between the corresponding treatment and its control group with all other effects averaged out. When the mode of the clustering is stable, the connection between the metabolites of a cluster and the corresponding effects is strong. Then, it is evident that especially the metabolites of the cluster mode are responsible for producing the ANOVA effects of that cluster.

It is obvious that many, or even most, of the measured metabolites are not affected by the treatments of the experiment. Thus, it is reasonable that clusters containing such metabolites have treatment effects around zero. Zero, which is the baseline,

is usually covered by the posterior distribution of an effect. For clusters containing metabolites affected by the treatment, the posterior distribution of the corresponding effect is outside zero.

The inspection of a posterior distribution is analogous to hypothesis testing in classical statistics, where the p-value of a test statistic is computed. Now, in the Bayesian analysis, one finds out, whether the baseline value zero is part of the inspected distribution. Also now a p-value can be computed for the hypothesis that zero is within the distribution and this p-value can be compared to a critical limit, e.g. 0.05, to decide whether the effect is significant.

The metabolite-specific scale parameter γ_i is a univariate variable. While the clustering matrix V decides which metabolites act in a similar fashion and the treatment effects model common up- and down-regulations, γ_i decides the magnitude of the variation of an individual metabolite within the cluster in the direction of the estimated effect. Thus, the metabolites within one cluster are expected to have a similar pattern over the treatments but the extent, or amplitude, of the pattern is accounted for by the scale parameter.

The metabolite-specific mean parameter $\boldsymbol{\mu}$ is used to explain the baseline concentrations of different metabolites. When this is modelled by a particular parameter, it does not bias other parameters, such as the ANOVA effects. The baseline concentrations as such are not of the primary interest in two-way analysis. Another way of dealing with the baseline problem is to normalize the data by removing the feature-specific mean before the actual analysis. This way, though, is more prone to error when dealing with small sample size data sets.

Like the scale and mean parameters, the residual parameter σ is a vector with a scalar value corresponding to each metabolite. The posterior distribution of the residual parameter is critical because it is a measure of how well the data fits into the model. If the data suits the model well, the value of the residual is around one.

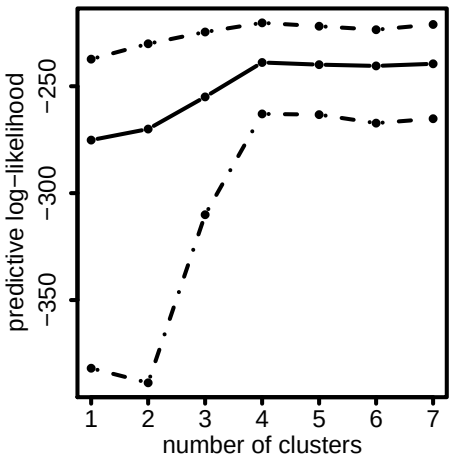


Figure 5: The correct number of clusters $K = 4$ is found for generated data in model complexity selection. Average predictive likelihood of left-out data is shown as a function of number of clusters. Increasing the number of clusters after $K = 4$ does not increase the likelihood.

8 Experiments

In this section, the new method is applied to data sets. First, performance of the method is demonstrated with simulated data both as a function of sample size and as a function of noise level. Then, the method is used to analyze metabolomic data from a diabetes study both in a two-level two-way setting and in a multi-level two-way time-series setting.

8.1 Performance as function of sample size

In this experiment with simulated data, it is demonstrated how well the method finds up-regulation effects as a function of the number of samples. The data is generated with the following parameters: There are four classes within a 2-way experimental setup as in Eqn (70). There are $K = 4$ clusters in which the following effects are generated: $\boldsymbol{\alpha}_1 = (+2, 0, 0, 0)$, $\boldsymbol{\beta}_1 = (0, +2, 0, 0)$ $(\boldsymbol{\alpha\beta})_{11} = (0, 0, +2, 0)$. Dimensionality of the dataset is $p = 200$. The optimal number of clusters is chosen by predictive likelihood, recovering the correct number of clusters $K = 4$ (Fig. 5).

The sample size now varies from $n = 20$ to $n = 1000$, such that the four classes have an equal number of samples (e.g., $n = 20$ means 5 samples in each class). The noise parameters are set to $\sigma_i = 1$, scale parameters to $\gamma_i = 1$, and mean parameters to $\mu_i = 0$ for $i = 1, \ldots, p$. The prior n_0 is fixed to $n_0 = 20$. In each run, 1000 Gibbs samples are collected after 1000 burn-in iterations. For each sample size, 10 independent datasets with the same parameters are generated and Gibbs sampling repeated for each.

The posterior intervals and means of the pooled posterior distributions of the effects

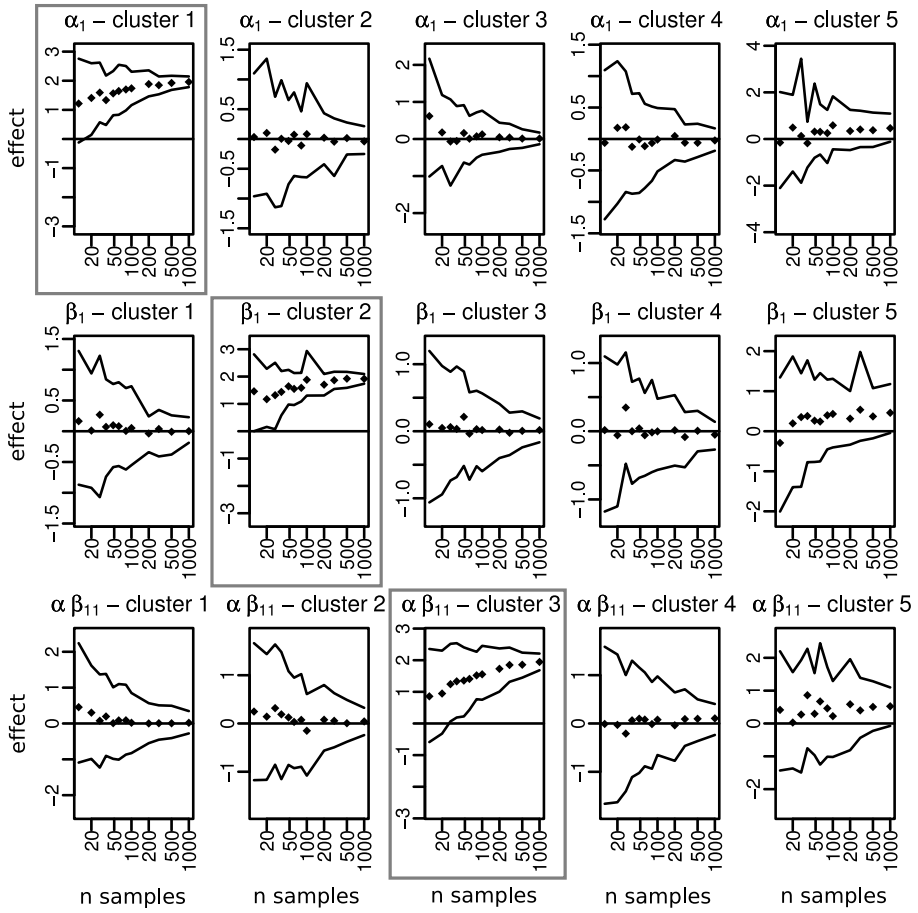


Figure 6: The method finds the generated effects $\alpha_{cluster1} = +2$, $\beta_{cluster2} = +2$, $(\alpha\beta)_{cluster3}$=+2. In the other clusters, no effects are found. The 95% posterior intervals of the main and interaction effects are plotted for each cluster.

are plotted for each found cluster in Figure 6. In the experiment, the model is intentionally computed with a slightly misplaced number of clusters to demonstrate effects of minor misspecification, having $K = 5$ clusters instead of the optimal $K = 4$.

The results show that the model finds the generated effect in each cluster and does not find false-positive effects in clusters where none were generated (although there is a fair measure of uncertainty in the estimates for small sample sizes). Uncertainty of the effects, that is, the width of the posterior interval, diminishes as the number of samples grows, as expected. Correct clustering is found from the posterior of $\mathbf{V}$ each time.

In metabolomics experiments, usually 20 to 60 samples are available. These sample-sizes are on the borderline. At larger sample-sizes, the posterior distribution of the generated effects are completely above the baseline of zero. At smaller sample-sizes, all generated effects are estimated to have overlap with the baseline and, thus, are not considered as significant effects.

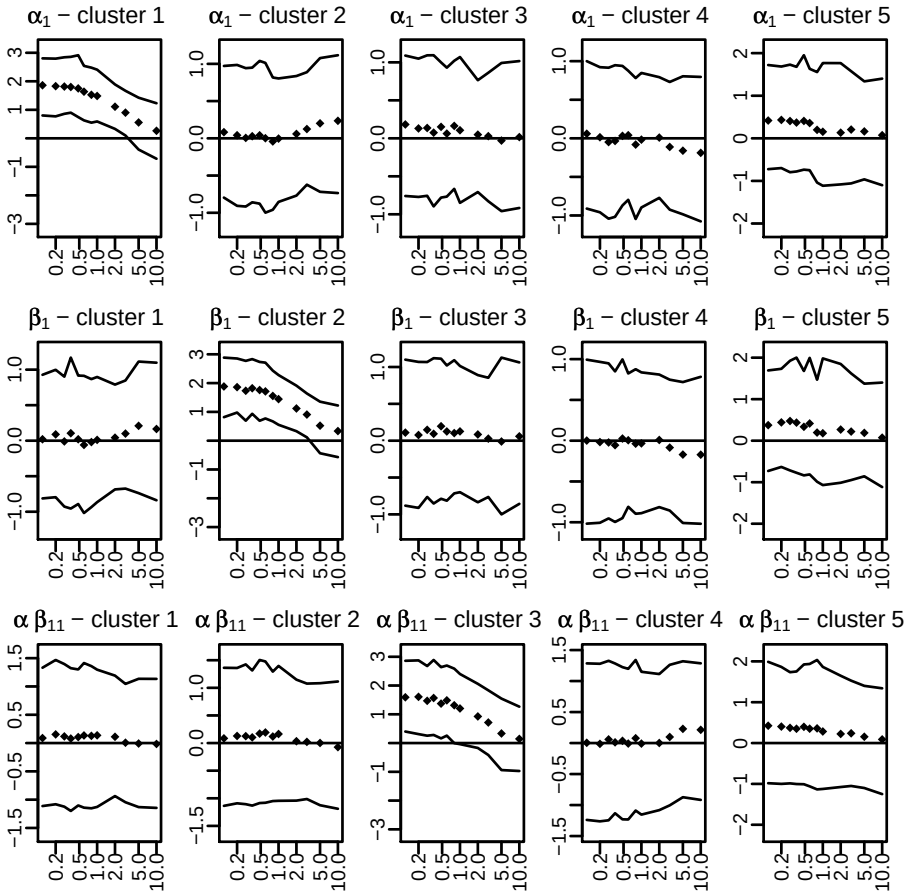


Figure 7: The method finds the generated effects $\alpha_{cluster1} = +2$, $\beta_{cluster2} = +2$, $(\alpha\beta)_{cluster3}$=+2 at reasonable but not at extremely high noise levels. In the other clusters, no effects are found. The 95% posterior intervals of the main and interaction effects are plotted for each cluster.

8.2 Performance as function of noise level

The study in Section 8.1 evaluated the performance of the method as a function of the sample size. In this section, the model performance is estimated as a function of the data noise level. Again, $K = 4$ clusters were generated with one effect in each of the first three clusters: $\boldsymbol{\alpha}_1 = (+2, 0, 0, 0)$, $\boldsymbol{\beta}_1 = (0, +2, 0, 0)$ $(\boldsymbol{\alpha\beta})_{11} = (0, 0, +2, 0)$. Dimensionality of the data set was $p = 150$ and sample size $n = 40$.

For each learning, a new data set was generated. During the experiment, the noise level of the data was gradually increased from $\sigma_i = 0.12$ to $\sigma_i = 10$ for all dimensions $i = 1, \ldots, p$ simultaneously. Again, scale parameters were set to unity $\gamma_i = 1$ and mean parameters to zero $\mu_i = 0$. The posterior intervals and means of the pooled posterior distributions of the effects are plotted for each cluster in Figure 7.

The results show that the generated effects are found very clearly when the noise level is low. It is particularly attractive that no false positive effects are found.

When the noise level increases, the found effect simply fades away. The increase in the noise level does not affect the estimated uncertainty of the effect estimate, which can be seen in the unchanged width of the confidence intervals. This behaviour is somewhat unintuitive in the sense that the increased uncertainty in the data should also come across to the uncertainty in the estimates of the unobserved variables. However, the upper end of the noise scale is extremely high and it is unlikely that any other method could find the generated effects of this size from a noisy data set with as few samples. The problem that is encountered at the high noise level is likely due to the strong priors on the effects. The hidden ANOVA variables $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{(\alpha\beta)}$ are given a $\mathcal{N}(\mathbf{0}, \mathbf{I})$ normal prior.

8.3 Lipidomic diabetes data set

Data from the DIPP diabetes follow-up study was analyzed using the new model. The origin of the data is described in more detail in Section 2.5.

In the first analysis below, one time point from the data set is selected for a two-way analysis to find the illness- and gender-related effects in the data. The second analysis contains more data, as six time points are included. Also this analysis is a two-way setting but now the time direction has six levels. The other direction, again, is the health status.

8.3.1 Cross-sectional study

The two-way experimental setup of a single time point (avg. time 750) in subjects who later progress to type 1 diabetes [1] was studied. The classes are healthy female (18 samples, subjects who have not progressed to diabetes, chosen as the control group), healthy male (17), diabetic female (11 who have later progressed to diabetes), diabetic male (8). The data contains 53 lipids as features. Following the notation of the example of equation (70), the disease effect is estimated with the parameter $\boldsymbol{\alpha}_1$ for each cluster, $\boldsymbol{\beta}_1$ models the gender-effect and $(\boldsymbol{\alpha\beta})_{11}$ models the interaction of these two effects.

The optimal number of clusters was found to be $K = 6$. According to the results shown in Figure 8, there is a positive, statistically significant gender effect found for clusters 1 and 4, signifying that males have a higher concentration for 18 and 4 lipids, respectively. A negative disease effect is found for cluster 3, signifying that diabetic patients have a lower concentration for 5 lipids. It is worth mentioning that even though the other effects are not statistically significant, clusters of strongly correlated lipids are still found.

8.3.2 Time-development study

Finally, the performance of the model for a simple time-series analysis of the human diabetes is demonstrated. The time indices are treated as independent values of the

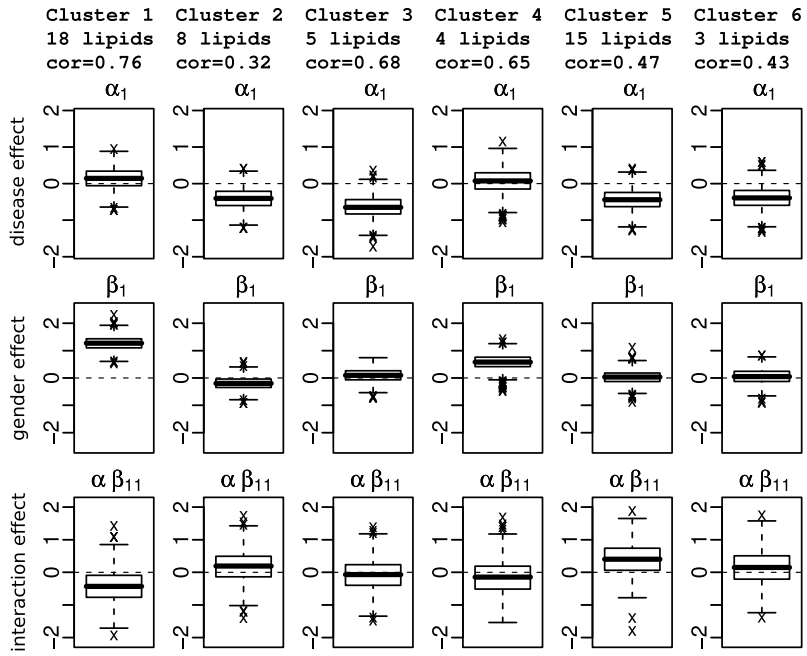


Figure 8: The method finds statistically significant effects for the human diabetes cross-sectional healthy-diseased, male-female comparison. Effects are found for $\alpha_{cluster3}$, $\beta_{cluster1}$ and $\beta_{cluster4}$. The figure shows posterior intervals of the main and interaction effects for each cluster. In addition, average correlation coefficients between lipids within each cluster are given.

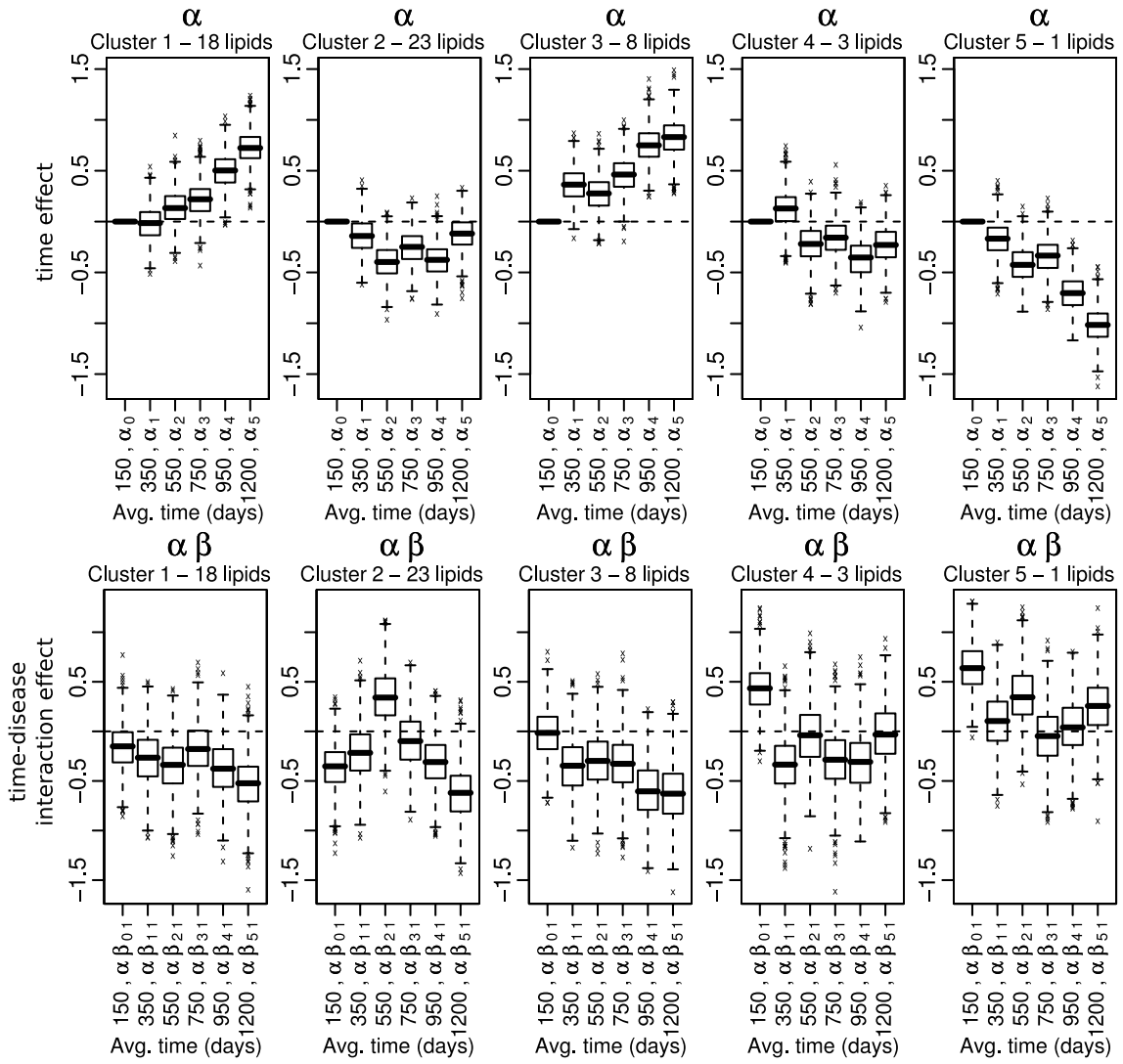


Figure 9: Statistically significant time-varying behavior is found for each cluster in the human diabetes data (above). Time-disease interaction effects are found as well for clusters 3, 4 and 5 (below). Posterior intervals of the main effect (time) and interaction effects (time, disease state) are plotted.

covariate; later the model will be extended by taking the time order into account, for instance by assuming a hidden Markov Model structure [21, 46].

In the diabetes data, lipidomic profiles of healthy human patients and patients developing into type 1 diabetes had been measured at variable intervals. The measurements were aligned to six time-points. The two-way setup now contains time effects and a healthy-diseased categorization. Now it is assumed that there is no static disease effect, but instead disease effects change in time. Therefore, only time effects $\boldsymbol{\alpha}_{1,\ldots,5}$ and time-disease interaction effects $(\boldsymbol{\alpha\beta})_{(0,\ldots,5)1}$ are estimated. The latter now indicate, for each time point, the deviation caused by the disease from the normal time-development.

The optimal number of clusters was found to be $K = 5$. The results shown in Figure 9 reveal clear time-dependent behavior, estimated by the $\boldsymbol{\alpha}$, that is distinct for all clusters. Statistically significant interactions of time and disease ($\boldsymbol{\alpha\beta}$) are found at timepoint 0 for clusters 4 and 5 (disease up-regulation), at timepoint 4 for cluster 3 (disease down-regulation) and at timepoint 5 for clusters 2 and 3 (disease down-regulation).

Further analysis of the results focusing on the metabolites grouped into affected clusters could provide interesting biological knowledge. In the experiment, metabolites from same functional group were clustered together. Many of the clusters show strong temporal shift in the concentrations. Further biological analysis of the interesting findings, however, is outside of the limited scope of this thesis.

9 Discussion

In this thesis, a new Bayesian method for analyzing data with a two-way experimental design was presented. The method has been designed to work with high-dimensional and small sample-size data. It is especially suited for metabolomic data, where metabolite concentrations are highly correlated.

The method is used for finding differences between populations of data samples. The samples are divided into populations according to covariates, which in bioinformatics describe, for instance, whether the sample is healthy or diseased.

The method performs dimensionality reduction by clustering features and finds ANOVA-type effects for these clusters. The method provides a complete Bayesian treatment for the data and gives posterior distributions of all model parameters as output and, thus, enables the assessment of uncertainty of the model.

The method performs well on normally distributed simulated data with a moderate noise level. Population effects are found already when the sample-size is small and no false positive effects are found. At extremely small sample-sizes, the level of uncertainty in the estimates is high but the estimates are correct. Identifiability of an effect is mostly dependent on the number of samples in the corresponding population.

The method presented in this thesis is a two-way model but it can without complications be extended to the multi-way case, where the number of covariates is greater than two. The estimation of a multi-way model requires more samples as the data set is divided into more populations than in the two-way case. Thus, the effects of multiple covariates and multiple covariate levels are not found as easily as the two-way effects.

Dimensionality reduction in the model does perform extremely well. For simulated data, the features are clustered exactly in correct clusters. For actual bioinformatics data, validation is not as straightforward. However, in many cases metabolites from previously known functional groups are grouped into the same cluster and no questionable results are found. The results both from simulated and biological experiments are promising and further development work of the model is planned.

The current version of the method does not take temporal correlations into account. From the time series experiment presented in Section 8.3.2, it is evident that the metabolite concentrations change during development of the individual. As an extension to the model, the linear time-development could be taken into account, leading to an improvement in the performance of the method.

In addition to small sample-size, a common problem in bioinformatics is the level of noise in the data. For instance, the difference between healthy and diseased samples is often smaller than the difference between repeats of measurements from one individual at several time instances. This sets limits on the identifiability of the covariate effects.

The presented model is a linear model. There are also non-linear approaches to

various problems in bioinformatics, which is reasonable because many biological processes are not linear by their nature.

Inference in the new Bayesian model is achieved by Gibbs sampling, which provides a reliable estimate on the joint distribution of the model. A common problem of sampling methods is that they are slow to run and there is a possibility of the model ending up in a local optimum if the sampler is not given time to converge. On long runs of the sampler algorithm, label switching becomes a major problem. With increasing data sets, computational cost of the sampling algorithm grows significantly. Even though the problems of local optimum, label switching and computational cost are important, they were not studied in detail in the diploma work and, thus, were not discussed in detail in the thesis.

The presented Bayesian model is a successful novel multi-way modelling approach to overcome the problems of small sample-size and high noise level. The model was presented in European Conference on Machine Learning in 2009 and was also published in a special issue of the journal Data Mining and Knowledge Discovery [41].

References

[1] Matej Orešič, Satu Simell, Marko Sysi-Aho, Kirsti Nanto-Salonen, Tuulikki Seppänen-Laakso, Vilhelmiina Parikka, Mikko Katajamaa, Anne Hekkala, Ismo Mattila, Päivi Keskinen, Laxman Yetukuri, Arja Reinikainen, Jyrki Lähde, Tapani Suortti, Jari Hakalax, Tuula Simell, Heikki Hyöty, Riitta Veijola, Jorma Ilonen, Riitta Lahesmaa, Mikael Knip, and Olli 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.

[2] Martin Kussmann, Frédéric Raymond, and Michael Affolter. Omics-driven biomarker discovery in nutrition and health. *Journal of Biotechnology*, 124(4):758 – 787, 2006. Highlights from ECB12 - Bringing genomes to life, 12th European Congress on Biotechnology (ECB12).

[3] Eoin Fahy, Shankar Subramaniam, H. Alex Brown, Christopher K. Glass, Jr. Merrill, Alfred H., Robert C. Murphy, Christian R. H. Raetz, David W. Russell, Yousuke Seyama, Walter Shaw, Takao Shimizu, Friedrich Spener, Gerrit van Meer, Michael S. VanNieuwenhze, Stephen H. White, Joseph L. Witztum, and Edward A. Dennis. A comprehensive classification system for lipids. *Journal of Lipid Research*, 46(5):839–862, 2005.

[4] Andrew D. Watson. Thematic review series: Systems Biology Approaches to Metabolic and Cardiovascular Disorders. Lipidomics: a global approach to lipid analysis in biological systems. *Journal of Lipid Research*, 47(10):2101–2111, 2006.

[5] K. Morgenthal, W. Weckwerth, and R. Steuer. Metabolomic networks in plants: Transitions from pattern recognition to biological interpretation. *Biosystems*, 83(2-3):108 – 117, 2006. 5th International Conference on Systems Biology - ICSB 2004.

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

[7] Xianlin Han and Richard W. Gross. Shotgun lipidomics: Electrospray ionization mass spectrometric analysis and quantitation of cellular lipidomes directly from crude extracts of biological samples. *Mass Spectrometry Reviews*, 2005.

[8] Tuulikki Seppänen-Laakso and Matej Orešič. How to study lipidomes. *Journal of Molecular Endocrinology*, 42(3):185–190, 2009.

[9] E. M. Lenz and I. D. Wilson. Analytical strategies in metabonomics. *Journal of Proteome Research*, 6(2):443–458, 2007.

[10] Julian L Griffin. Metabonomics: NMR spectroscopy and pattern recognition analysis of body fluids and tissues for characterisation of xenobiotic toxicity

and disease diagnosis. *Current Opinion in Chemical Biology*, 7(5):648 – 654, 2003.

[11] Mikko Katajamaa, Jarkko Miettinen, and Matej Orešič. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics*, 22(5):634–636, 2006.

[12] Silvia Wagner, Karoline Scholz, Maximilian Sieber, Marco Kellert, and Wolfgang Voelkel. Tools in metabonomics: An integrated validation approach for lc-ms metabolic profiling of mercapturic acids in human urine. *Analytical Chemistry*, 79(7):2918–2926, 2007.

[13] J. Susan Milton and Jesse C. Arnold. *Introduction to probability and statistics*. McGraw-Hill, 4th edition, 2003.

[14] Yoav Benjamini and Yosef 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.

[15] *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press, 1st edition, 1979.

[16] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

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

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

[19] Johan Westerhuis, Huub Hoefsloot, Suzanne Smit, Daniel Vis, Age Smilde, Ewoud van Velzen, John van Duijnhoven, and Ferdi van Dorsten. Assessment of plsda cross validation. *Metabolomics*, 4(1):81–89, 2008.

[20] Mark A. Atkinson and George S. Eisenbarth. Type 1 diabetes: new perspectives on disease pathogenesis and treatment. *The Lancet*, 358(9277):221 – 229, 2001.

[21] Janne Nikkilä, Marko Sysi-Aho, Andrey Ermolov, Tuulikki Seppänen-Laakso, Olli Simell, Samuel Kaski, and Matej Orešič. Gender-dependent progression of systemic metabolic states in early childhood. *Molecular Systems Biology*, 4(197), 2008.

[22] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[23] Ruth G. Shaw and Thomas Mitchell-Olds. ANOVA for unbalanced data: An overview. *Ecology*, 74(6):1638–1645, 1993.

[24] J. N. R. Jeffers. Principal component analysis of designed experiment. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):230–242, 1962.

[25] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection Techniques*, Lecture Notes in Computer Science, pages 34–51. Springer, 2006.

[26] Jeroen Jasper Jansen. *ASCA*. PhD thesis, Universiteit van Amsterdam, 2005.

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

[28] Guangzhi Cao and Charles A. Bouman. Covariance estimation for high dimensional data vectors using the sparse matrix transform. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 225–232. MIT Press, 2009.

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

[30] Christopher M. Bishop. Bayesian PCA. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, pages 382–388, Cambridge, MA, USA, 1999. MIT Press.

[31] Zoubin Ghahramani and Matthew J. Beal. Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems 12*, pages 449–455. MIT Press, 2000.

[32] Cédric Archambeau and Francis Bach. Sparse probabilistic projections. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 73–80. MIT Press, 2009.

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

[34] Lily Wang, Bing Zhang, Russell D. Wolfinger, and Xi Chen. An integrated approach for the analysis of biological pathways using mixed models. *PLoS Genetics*, 4(7):e1000115, 2008.

[35] 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.

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

[37] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, 2003.

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[39] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[40] Mary Kathryn Cowles and Bradley P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91:883–904, 1996.

[41] 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, 2009.

[42] Guido Sanguinetti, Josselin Noirel, and Phillip C. Wright. MMG: a probabilistic tool to identify submodules of metabolic pathways. *Bioinformatics*, 24(8):1078–1084, 2008.

[43] Daniel Vis, Johan Westerhuis, Age Smilde, and Jan van der Greef. Statistical validation of megavariate effects in ASCA. *BMC Bioinformatics*, 8(1):322, 2007.

[44] Sam Roweis and Zoubin Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345, 1999.

[45] Daniel B. Rowe. On estimating the mean in Bayesian factor analysis. In *Social Science Working Paper 1096, Division of Humanities and Social Sciences, Caltech, Pasadena, CA 91125*, 2000.

[46] Matthew Beal and Praveen Krishnamurthy. Gene expression time course clustering with countably infinite hidden markov models. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Arlington, Virginia, 2006. AUAI Press.