Eberhard Karls Universität Tübingen
Fakultät für Informations- und Kognitionswissenschaften
Wilhelm-Schickard-Institut für Informatik

Master Thesis Bioinformatics

Probabilistic Modelling of Expression Variation in Modern eQTL Studies

Max Zwießele

October 13, 2012

Reviewer

Karsten Borgwardt
University of Tübingen

Detlef Weigel
Max Planck Institute Tuebingen

Supervisor

Oliver Stegle
Max Planck Institute Tuebingen

Neil Lawrence, Nicoló Fusi
Sheffield Institute for Translational Neuroscience

Name, first name: Zwießele, Max

Probabilistic Modelling of Expression Variation in Modern eQTL Studies

Master Thesis Bioinformatics

Eberhard Karls Universität Tübingen

Period: 15.04.2012-15.10.2012

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Masterarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Ort, Datum Unterschrift

Abstract

Phenotypes of species exhibit vast differences between their individuals, but is there a link between these phenotypic differences and their genotypes? Recent studies have shown that genotypes of organisms can highly vary between different individuals of species. Each individual is therefore given an own tool set of genes and control mechanisms to handle exposure to stress factors. This can enable the individual to handle a stress situation better than its neighbours. We want to understand the mechanisms involved, which are controlled by the differing genotypes of an organisms. This is where association studies were invented. Association studies link genomic loci (genotype variants) to the differing phenotypes of a species of interest. These differences in phenotype were ultimately mediated from the genotype through transcriptional state of the cells. Thus, one approach of linking genotype to phenotype is to directly associate the genotype and gene expression as a phenotype. Second generation sequencing technologies allow us to both genotype and expression profile individuals on a genome-wide scale.

Such genome-wide association studies (GWAS) are becoming more and more popular and reveal connections between genotype and gene expression, called expression quantitative trait loci (eQTL) studies. Associating the genotype to gene expression is a hard problem, because of the shear number of genomic loci which

have to be mapped to all gene expression values of all genes. Which is, instead of only a few observable global phenotypes (such as e.g. hair colour), we now have 10s of thousands of gene expression phenotypes. Additionally, confounding variation is introduced through environmental conditioning and combinatorial genetic effects, such as population structure or differing gender. Such confounding variation is strongly amplified by each phenotype observed and has to be accounted for. In this thesis we focus on modelling such confounding variation in eQTL studies in a Bayesian setting, called variational Bayesian Gaussian process latent variable model (variational Bayesian GPLVM). Modelling confounding variation in a Bayesian setting allows us not only to estimate confounding factors themselves, but also the dimensionality of confounders necessary to explain confounding variation in gene expression. This gives us the novel opportunity to account for uncertainty in confounders and to explicitly find the number of confounding factors, which contribute to the confounding variation in eQTL studies. Here, we can exploit the dimensionality of gene expression to find hidden confounders, which are not directly observable.
Combining several variational Bayesian GPLVMs to one model, we further expand the power of this Bayesian approach to handle multiple eQTL studies. This extension enables us to find which genetic variants are private or shared between experiments. E.g. we find condition specific associations between genotype and phenotype when observing same individuals in differing conditions.

With our approach, we established a new way of estimating confounding variation in GWAS and went even a step further, finding shared and private associations between several GWAS. This gives a good foundation to build on in future approaches of finding associations between genotype and phenotype on a genome-wide scale.

Acknowledgements

First of all, I thank Karsten Borgwardt, Neil Lawrence and Oliver Stegle for giving me the opportunity to work on this thesis and for their indispensable feedback. Furthermore, I thank Nicoló Fusi for his innovative ideas, excellent motivation and virtuous patience.

I also thank the members of my work group in Sheffield - Andreas Damianou, Alfredo Kalaitzis, Arif Rahman, Ciira Maina, James Hensman, Jens Nielson, Nicolas Durrande and Ricardo Andrade Pacheco - for useful, inspiring discussion and proper proofreading.

Finally, I thank my parents Sibylle and Frieder for their patience and assistance in helping me finishing this thesis.

Contents

List of Figures

List of Tables

Nomenclature

$N \in \mathbb{N}$	Natural number
$x \in \mathbb{R}$	Real number
$\mathbf{x} \in \mathbb{R}^N$	Vector $\mathbf{x}$ of size $N \times 1$
$\mathbf{X} = (\mathbf{x}_d)_{1 \leq d \leq D} \in \mathbb{R}^{N \times D}$	Matrix $\mathbf{X}$ of size $N \times D$
$\mathbf{I}_N \in \mathbb{R}^{N \times N}$	Unit matrix of size $N \times N$ - Matrix of ones on first diagonal
$\mathbf{I}_{N \times D} \in \mathbb{R}^{N \times D}$	Unit matrix of size $N \times D$
$\mathbf{X}^\top$	Transpose of $\mathbf{X}$, such that $\mathbf{X}_{ij} = \mathbf{X}^\top_{ji}$ and $\mathbf{X}^\top \in \mathbb{R}^{D \times N}$
$\lvert\mathbf{X}\rvert$	Determinant of $\mathbf{X}$
$\mathbb{E}_{p(\mathbf{X})}\big(f(\mathbf{X})\big) = \langle f(\mathbf{X}) \rangle_{p(\mathbf{X})} = \int f(\mathbf{X}) p(\mathbf{X})\, d\mathbf{X}$	Expected value of $f(\mathbf{X})$ under the distribution $p(\mathbf{X})$

Chapter 1

Gene Expression as Quantitative Trait

In the past decade new technologies for revealing (i.e. sequencing) deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) of organisms made it possible to compare genomic data on a genome-wide scale. These technologies (namely, next-generation sequencing technologies [29]) opened the access to many different so-called "omics" fields in bioinformatics. One example of comparing "omics" is to relate gene expression to the genotype variation. The declared goal is to reveal functional relationships between genotype and phenotype of organisms on a genome-wide scale. In developing a certain phenotype an organism undergoes a complex procedure of expressing the information written down on a molecular level in form of DNA sequence and structure. The information gets translated into proteins, which are the workers of the individual cell(s) of an organism. A cell comprises several hundred thousand proteins carrying out the orders written in the DNA. We are interested in the effects of the genotype on the process of transcription of a DNA sequence into messenger RNA (mRNA), which is defined as gene expression. Gene expression is affected by genotype variation and greatly influenced by confounding factors. Confounding factors can be environmental or combinatorial effects of the genotype itself. To be able to elucidate the effect of genotype variation on gene expression one needs to study associations between genotype and differential gene expression of several different individuals, comprising differing genotype variants. Global acting effects in all experiments of such association studies can be assigned to confounding circumstances, such as exposure to heat, or chemicals or more globally, the habitat the organism occurs in. In this chapter we will first give a short introduction of how DNA is transcribed into mRNA and translated

to proteins, second, introduce mechanisms regulating the process of transcription, third, describe why accounting and searching for confounding factors is a step towards better detection of genotype-phenotype associations, and last, describe the methodology of genome-wide association studies and the advantages such studies bring.

1.1 From DNA to Protein

Most heritable information in a cell of an organism is coded in its DNA. One unit of heritable information coding for one or more proteins is called a gene (first announced by Mendel [22]). The set of active genes determines the specific function of a cell. In order to understand an organism as a whole, we need to understand the role and function of DNA in every cell of an organism.

1.1.1 Discovery and Structure of DNA

Before we look at the functionality of DNA, we briefly describe the discovery of DNA[1]. The discovery of DNA begins with Gregor Mendel. Mendel discovered in 1865, through breeding experiments with peas, that the different phenotypes of the peas were inherited based on specific laws (Mendel [22]). These laws were later called "Mendelian laws". In 1869 Friedrich Miescher first isolated DNA, but at this time he failed to realise the importance of his discovery. In 1909 Wilhelm Johannsen for the first time used the word gene to describe a unit of heredity. In the years from 1949 until 1953 the structure of DNA was revealed by several studies. DNA consists of four bases, connected by a sugar-phosphate backbone. The atoms are connected from `5'C` atom to `3'C` of the base, which gives the DNA strand it's direction (from `5'` to `3'`) . These bases, called nucleotides, build up two complementary linearly composed strands, which form a double helix. Two nucleotides respectively form hydrogen bonds, which stabilise the double helix. The four nucleotides are Adenine, Thymine, Guanine and Cytosine. A and T can bind to each other by forming two hydrogen bonds. Therefore, A and T are said to be complementary. G and C are also complementary: they form three hydrogen bonds. This complementary strand is called complementary DNA (cDNA). When two fitting cDNA strands bind to each other they build up the double helix, mentioned above. The binding of two complementary cDNA strands is called hybridisation, where

[1]Text taken from my Bachelor Thesis `http://people.kyb.tuebingen.mpg.de/maxz/GPTimeShift.pdf`

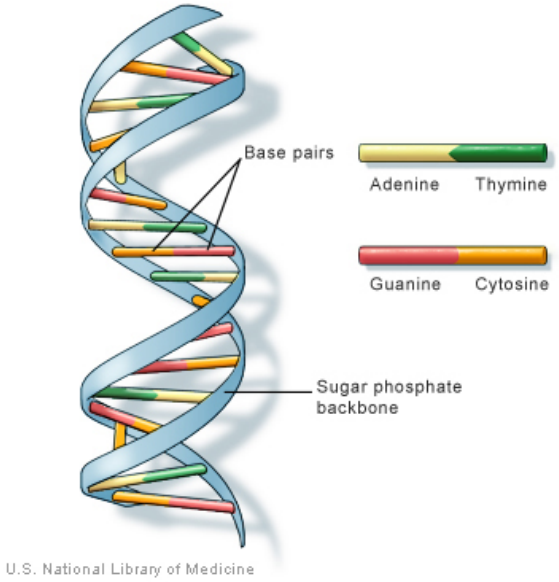


Figure 1.1: Double helix structure of the DNA. The double helix is built up by linearly linked nucleotides - connected by a sugar-phosphate backbone. Hydrogen-bonds form only between A and T and between G and C. U.S. National Library of Medicine [37]

the complementary strands are also directionally complementary. Inside the cell, DNA is always present in double helix form. Figure 1.1 depicts the double helix form with its backbone and nucleotides. In 1957, Francis Crick stated that the DNA is transcribed into RNA and then translated into a protein. In 1977, the first methods to sequence DNA were introduced by Frederick Sanger, Allan Maxam and Walter Gilbert. This was a big step towards understanding the genetics of an organism. The next years revealed many different properties of DNA. In the 1980's, the first microarray technologies were developed. By 2001, the human genome was sequenced as a whole and published in Nature by Lander et al. [17]. In the years since 1865 many more discoveries about DNA and its properties were made, but to describe all of them would go beyond the scope of this thesis.

1.1.2 Functional View on DNA – Genes, Expression and Proteins

DNA is divided into so-called coding regions and non-coding regions. Thus, DNA not only contains all genes (coding regions) of an organism, it also comprises a large percentage of regions which do not code for proteins (non-coding regions), which take effect in secondary processes (see more for example in Ahnert et al. [1]; Mercer et al. [23]). A coding region is a protein coding sequence (CDS for coding sequence)

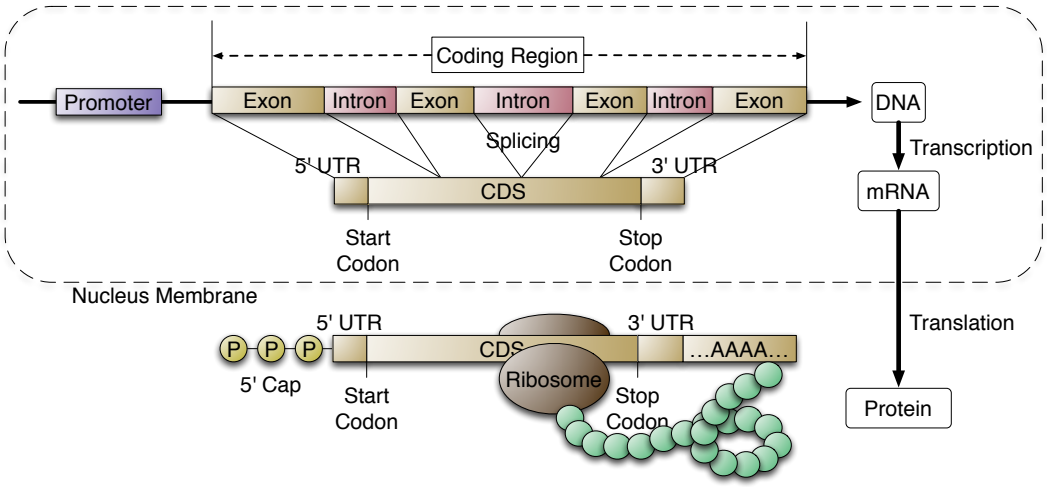


Figure 1.2: The first line shows the coding region of a gene and its promoter on the DNA. Introns of the coding region are transcribed into mRNA (second line), where introns are spliced away. The exons form the protein coding sequence (CDS), flanked by untranslated regions and the start and stop codon of the protein. The resulting mRNA gets chemically sealed by a phosphor cap and a poly-A end. Protected from reducing proteins it passes the nucleus membrane and gets translated into a protein in the cell lumen (depicted as green folding chain).

of DNA, flanked by specific start and stop codons (triplets of nucleotides), which in turn are flanked by untranslated regions (UTRs), one for each side. The coding region is interleaved by introns which are also not translated. The DNA is translated into pre-mRNA (pre-messenger ribonucleic acid). After splicing, mRNA only contains exons of the coding region, as depicted in Figure 1.2. To be save from reducing proteins mRNA gets chemically sealed by a poly-A end. It passes the nucleus membrane and gets translated into a protein by ribosomes.

The promoter of a gene precedes the coding region and controls the expression level of the controlled gene. It provides a binding site for transcription factors and RNA polymerase, which translates DNA to RNA.

A gene is called expressed if it is transcribed into mRNA. And the amount of mRNA determines the expression value of this particular gene at a given time point. Gene expression changes over time and is influenced by many factors. Some of these factors are nearby – other genes, proteins in the cell or hormones and metabolites – and others are from the outside – environmental influences, such as temperature, light or chemical exposure, drugs and gender. We will discuss such influences in section 1.3 in greater detail.

To understand the role of a certain proteins, we use expression levels of their corresponding genes. We assume the translated protein of a gene to be activated in the cell if the gene is expressed (transcribed into mRNA). Thus, if a gene is discovered

to be differentially expressed under certain (changes of) circumstances we can with high probability assign this gene (and its transcript) to respond and take part in the handling of this circumstance.
Non coding regions contain markers, which additionally influence the expression of genes. Even one single change in the genetic code - a single nucleotide polymorphism (SNP) - can give rise to a difference in gene expression of one or several genes. SNPs can occur in both coding and non-coding regions, but occur most frequently in non-coding regions. A SNP comes into play through either non-coding RNA (so called microRNA), which gets transcribed and can pass the nucleus membrane, or through transcription factors, which are able to control gene expression when binding to a SNP. Additionally, a SNP can influence a gene, which in turn codes for a transcription factor. This particular transcription factor then influences possibly a variety of genes. This leads to a high number of combinatorial effects of SNPs. Thus, finding such influences is a non trivial task. We need to account for combinatorial effects between SNPs, find transcription factor effects and explain environmental influences away.
We have given an overview of some aspects of genetics that are key to understand this thesis. A detailed view on genetics is beyond the scope of this thesis. We refer interested readers to the books of Hennig [14]; Seyffert and Balling [28]; Zien [40].

1.2 Genome-wide Association Studies

Translation and transcription is a complex interaction of DNA, mRNA, ribosomes and many other proteins and molecules. To elucidate the underlying processes of such complex interactions we apply methods which are able to capture the "whole picture". Based on microarray and RNA-seq technologies [6; 29; 38] we are able to analyse such interactions on a genome-wide scale, see e.g. Hirschhorn and Daly [15]. This is where genome-wide association studies have been developed. In genome-wide association studies (GWAS) we try to associate genotype and phenotype on a genome-wide scale. For example in expression quantitative trait loci (eQTL) studies (See e.g. Gilad et al. [13]), we measure association between gene expression and SNPs, which are one form of quantitative trait loci. Such that we can assign the regulatory effects of a quantitative trait locus on the expression of one or more genes. It is a hard problem to find such associations on a genome-wide scale, first, because there are ten thousands of genes and SNPs in one genome, and second, because there are many (yet) unknown factors involved (see section 1.3). In this section we introduce the most common genomic variants and SNPs in greater

detail.

1.2.1 Genetics of Gene Expression

In genome-wide association studies we search for associations between genotype and phenotype. That means that, when there is a correlation between a genotype (variant) and a phenotype (gene expression pattern, disease etc.) we assume that this genotype causes the phenotype to change. The expression is, thus, seen as a quantitative trait [15], varying due to genomic variants.
Genomic variants are essential for certain phenotypes to arise. By figuring which genomic variants cause a certain phenotype we can assess risks for diseases or other changes in phenotype. The simplest type of variant results from a single base mutation within the genomic sequence. This mutation gives rise to the most common form of variation, single nucleotide polymorphisms (SNPs). Insertions or deletions (INDELs) are another form of variation, where a single nucleotide is inserted or deleted during replication of the DNA. If a sequence of more then one nucleotide is deleted or inserted we speak of variable number tandem repeat polymorphisms (VNTRs). Such insertions or deletions are caused by repetitive sequences, which expand or contract as a result of insertion or deletion events [3].
A set of common genomic variants are often referred to as a genomic allele, changing the phenotype due to their combinatorial effects. In an organism comprising two sets of chromosomes (as humans are) there are two alleles – one per chromosome (See [14; 28] for details). In this thesis we are mostly interested in SNPs and their alleles. A SNP is called common if the less common allele (of the SNP), the minor allele occurs more often then $20\,\%$. Such alleles comprise a large percentage of the genome in humans and combine to the huge quantity of 5 to 10 million SNPs across the genome [2]. Some of such alleles are pathogenic (cause disease, when present or absent) and are essential for establishing genetic tests in clinical use. We search for connections between SNPs and the phenotype we are interested in (here gene expression). A SNP can be nearby the gene (often chosen to be between $\pm 500\,\mathrm{bp}$ and $\pm 15{,}000\,\mathrm{bp}$ away from the gene) of which it changes the expression level. Such an association is called a *cis* association. Every association in which the SNP is farther away from the gene it is associated with is called a *trans* association (Figure 1.3).
The shear number of SNPs contained in a genome makes it a hard problem to study associations between them and the phenotype. Fortunately several SNPs are highly locally correlated to the SNPs in their vicinity (they are in linkage disequilibrium) allowing for testing only on some proxy agents of such groups of SNPs. Hirschhorn

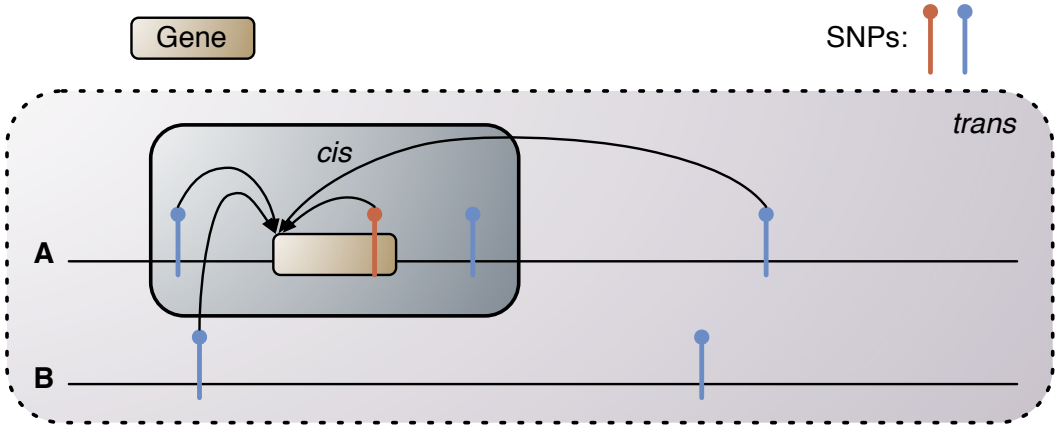


Figure 1.3: Association variants of SNPs associated with a gene. The red SNP indicates a SNP within a gene, and blue SNPs indicate SNPs in non-coding areas of the genome. All associations which are within *cis* range (on the same chromosome – here **A**) are called *cis* associations. Every association exceeding the window of *cis* range or is located on another chromosome (e.g. **B**) is called a *trans* association.

and Daly [15] even state that a few hundred well chosen SNPs should be enough to provide information about the most common genetic variants. As the costs for sequencing ([10; 29]) are steadily decreasing we might be able to afford frequent genome-wide association studies in the human genome. The HapMap[2] project tries to map the whole genome for genes affecting health, disease, and respond to drugs and environmental factors. They use GWAS for creating this map. Another project, the dbSNP[3] database collects SNPs of common organisms.

1.3 Confounders in Genome-wide Association Studies

Gene expression is a complex process in that it is controlled and influenced by many internal and external factors. The genetic variants around an expressed gene (which can be far away within the nucleus) influence the expression level either directly or through a signalling pathway within the cell and are named internal factors. Additionally, the outside world influences the organism's gene expression, as well. Such factors trigger long cascades of protein-protein interactions, eventually reaching through the cell membrane and nucleus to influence the DNA directly. These factors are called external factors. As the influences of such factors are not driven by the genotype of an organism we want to account for the confounding effects they might introduce in a genome-wide association study (more on GWAS

[2] `http://hapmap.ncbi.nlm.nih.gov`
[3] `http://www.ncbi.nlm.nih.gov/projects/SNP/`

in Section 1.2). We call these confounding effects confounding factors, confounding variation or simply confounders.
In this section we will introduce the non-genetic effects in gene expression. First, we will introduce the internal, and second, the external factors effecting gene expression of an organism.

1.3.1 Genetically driven confounding Variation

As seen in Section 1.1.2, transcription is a carefully controlled process. It is controlled by transcription factors, which in turn are controlled by internal molecules of an organism. Metabolites and hormones influences gene expression through cascades of protein-protein interactions, reaching through the nucleus membrane to the chromosomes and DNA within. Genes can code for transcription factors which influence gene expression of other genes and possibly themselves. More generally all kinds of combinatorial effects of gene-gene, gene-protein and protein-protein interactions can change the expression of many genes.
Population stratification [15] can also influence gene expression between sub groups of a population. Some sub groups can share similar gene expression patterns, but between those groups the expression patterns deviate. This behaviour is also discovered when looking at races of species, populations cut off from one another or "cliques" within a population. If for example all individuals in a subgroup share one genetic variant this genetic variant will be significantly correlated with the gene expression and will create a bias towards being correlated. The genetic variant is over-represented in the genotype and will therefore influence global confounding variation detection significantly.
As last confounder we need to mention gender effects on gene expression. Clearly different genders have different phenotypes, but share them within their gender. Thus, gender effects can be seen as population stratification effects creating a bias if an unequal distribution of genders are chosen to be analysed.
For more details on confounders in GWAS see e.g. Hirschhorn and Daly [15].

1.3.2 Environmentally driven confounding Variation

Gene expression is not only influenced by the genotype, but by a variety of additional effects. Most of which are environmental - caused by the surrounding of the organism. Exposure to different temperature conditions, different lighting conditions or chemicals are examples. However, there are many more external environmental effects one can think of, some of which are physical: altitude, pressure,

humidity etc; human made: smoking, air conditioning, nutrients in cities, change in sea water quality etc; or other organisms in the environment: viral and bacterial infection, nutrient starving due to over-representation etc. All these influences create a mixture of changing effects on the phenotype of an organism. And, therefore, create confounding variation in the phenotype (gene expression) which might not be known, nor their combinatorial effects are predictable in advance.

Confounding Variation and Probabilistic Modelling

In the two previous sections we saw the huge variety of effects controlling and influencing gene expression of an organism. We are only interested in associations between genotype (SNPs) and phenotype (gene expression) of the organism, and not in confounding variation. Unfortunately, the combination of all internal and external hidden factors (Section 1.3) creates confounding variation of unknown dimension and abundance. Thus, we need not only a method of estimating confounding variation itself, but also of estimating the dimension of such variation. It is uncertain which of the confounding factors have a real effect on gene expression and how much their effect influences gene expression. To take account for this uncertainty in estimating confounding variation, we build a fully Bayesian model to predict the most probable confounding factors (abundance and dimension) under the uncertainty of combinatorial effects. A Bayesian probabilistic approach would consist in putting a prior on parameter variables and integrate them out. With that, all possible values of these parameters are taken into account (represented by uncertainty) and the most probable one is chosen (which is the mean/expected value in most cases). In the following chapter we will introduce a variational Bayesian approach on estimating lower dimensional latent variables, creating the data observed and will use this model to account for confounding variation. The model learns latent variables as shared underlying input, creating the data observed. Thus, if there is shared variation between the gene expression of all samples collected, this will be taken as a latent variable. And this is exactly what a confounding factor is. In a genome-wide association study we compare different genomic variants of the same organism, thus effects of genomic variants should not be shared across samples.

Chapter 2

Probabilistic Modelling of Variation in Gene Expression

In association studies our aim is to associate the genotype with the phenotype. Thus, we want to find a function f^{asso} mapping the genotype $\mathbf{S}$ to the phenotype $\mathbf{Y}$;

$$\mathbf{Y} = f^{\text{asso}}(\mathbf{S}) \ , \tag{2.1}$$

where the genotype $\mathbf{S}$ are K genomic variants over N samples of an organism; $\mathbf{S} \in \mathbb{R}^{N \times K}$, and the phenotype are D gene expression values over the same N samples; $\mathbf{Y} \in \mathbb{R}^{N \times D}$. As soon as we have found a function f^{asso} as described above, we can find how and how much the genotype $\mathbf{S}$ effects the phenotype $\mathbf{Y}$. As an example suppose f^{asso} a linear mapping

$$\mathbf{Y} = f^{\text{asso}}(\mathbf{S}) = \mathbf{SW} \tag{2.2}$$

between phenotype and genotype with weights $\mathbf{W} \in \mathbb{R}^{K \times D}$. This weight matrix will be very sparse, because typically only a few SNPs of the genome will have an effect on the phenotype and the entry w_{kd} encodes for the effect of the kth genomic variant on the dth gene. This, however, is not enough in genome wide association studies, because in such studies there are confounding variation contributing to gene expression variation as discussed in Section 1.3. Thus, the function f^{asso} should also be able to find and account for confounding factors $\mathbf{X}$. We denote these confounding factors as $\mathbf{X} \in \mathbb{R}^{N \times Q}$, where there are Q confounding factors over all N samples collected. Thus, the function f^{asso} gets the hidden factors as another

parameter and the full function we are searching for takes the form

$$\mathbf{Y} = f^{\text{asso}}(\mathbf{S}, \mathbf{X}) \ . \tag{2.3}$$

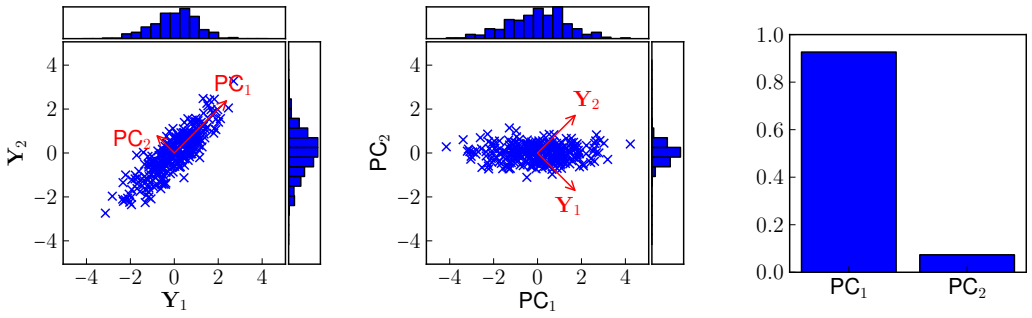
Finding confounding factors $\mathbf{X}$ is to estimate the effect of such factors (the matrix $\mathbf{X}$ itself) and the dimension Q of the these factors. In other words we want to estimate how and how many confounding effects are shared in creating the phenotype observed. In the following, we will introduce how to estimate confounding factors and their dimensionality. First we, will introduce the well known principal component analysis and second a probabilistic model of principal component analysis in a Gaussian process setting, called Gaussian process latent variable model. We will then extend the Gaussian process latent variable model to a fully Bayesian variational approximation to be able to estimate the dimensionality Q of confounding factors $\mathbf{X}$.

2.1 Principal Component Analysis

One well known approach of estimating confounding factors $\mathbf{X}$ of lower dimensionality is the so called principal component analysis (PCA). PCA transforms the phenotype expression matrix $\mathbf{Y}$ onto a basis, such that the directions of the axes of the new basis correspond to the directions of the highest variance of the phenotype. In order to find lower dimensional confounding factors we then cut off low variance components. Suppose you have a phenotype $\mathbf{Y} \in \mathbb{R}^{N \times D}$ of $N = 300$ samples over $D = 2$ genes correlated as depicted in Figure 2.1(a). PCA chooses a new basis to maximise the variance explained. We will possibly find the basis shown in red in the figure, annotated as $\text{PC}_{\{1,2\}}$. This basis corresponds to the direction of the eigenvectors $\mathbf{V}^{-1} \in \mathbb{R}^{D \times D} = (\mathbf{v}_d)_{1 \le d \le D}$ of the covariance $\mathbf{Y}^\top \mathbf{Y}$ of the phenotype. And the fraction of variance explained is the fraction of eigenvalues of this covariance. We keep both matrices sorted descending according to the eigenvalue fractions. As we can clearly see, the second principal component does not explain much variance and can be cut off as noise.

We can use PCA to find Q dimensional latent factors in the phenotype. As already mentioned the direction of highest variance of the phenotype correspond to the direction of the eigenvector corresponding to the highest eigenvalue. Thus, to transform the phenotype $\mathbf{Y}$ into the full principal component space $\mathbf{X}^{\text{PCA}} = (\text{PC}_d)_{1 \le d \le D}$, we have to multiply it by the eigenvectors $\mathbf{V} \in \mathbb{R}^{D \times D}$ of the covariance matrix $\mathbf{Y}^\top \mathbf{Y}$

$$\mathbf{X}^{\text{PCA}} = \mathbf{Y}\mathbf{V} \ . \tag{2.4}$$



(a) Scatter plot for both genes. Directions of principal components are shown as red arrows. Histograms show distribution of gene expression values.

(b) Transformed $\mathbf{Y}$ plotted in full principal component space. PC_1 histogram comprises much more variance, then the one of PC_2.

(c) Fractions of principal component variance. The second principal component PC_2 can be neglected compared to the first component PC_1.

Figure 2.1: Principal component analysis on a small example gene expression matrix $\mathbf{Y}$, such that $\mathbf{Y}_1$ corresponds to all samples of the first gene and $\mathbf{Y}_2$ corresponds to the samples of the second gene. (a) shows a scatter plot of the gene expressions. Directions of principal components are shown in red. The histograms show the variance of genes. In PCA space (b) the variance is differently distributed. As shown in (c) the first principal component explains most of the variance, whereas the second principal component explains almost none. Thus, in order to describe the data approximately, we can neglect the second principal component and only use the first as an approximation of the data.

Here we can see the connection to the linear generative model (2.2), which can be recovered by right-multiplying the inverse of the eigenvector matrix $\mathbf{V}$ to the PCA solution

$$\mathbf{Y} = \mathbf{X}\mathbf{V}^{-\mathbf{1}} \quad . \tag{2.5}$$

Figure 2.1(b) shows PC_1 against PC_2, which corresponds to the columns of $\mathbf{X}^{\text{PCA}}$. To reduce the dimensionality to a value $Q < D$, we take only the first Q eigenvectors $\mathbf{V}_Q$, instead of the full eigenvector matrix. This corresponds to the Q highest eigenvalues of the covariance and we get a lower dimensional representation

$$\mathbf{X} = \mathbf{Y}\mathbf{V}_Q \quad . \tag{2.6}$$

In our toy example it is easy to say that we do not need the second principal component to explain data. But when we look at a real experiment comprising $2{,}000$ and more genes, it is not as easy to determine the number Q of dimensions to keep. There are some heuristics to find the number of dimensions. One of which finds the kinks in plotting the fractions of variance against the principal components and cuts off at the first kink.

We can get over the problem of finding the right dimension of latent factors by building a Bayesian probabilistic model on the problem. The Bayesian probabilistic view on functions is to choose the most probable setting out of all possible ones. Thus, by modelling the thoughts of PCA in a Bayesian probabilistic setting

$$p(\mathbf{Y}) = \int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})\, d\mathbf{X} \ , \tag{2.7}$$

we can determine the number of dimensions Q of latent factors by the most probable solution to Q. There have been many approaches on probabilistic modelling of gene expression data accounting for latent factors [7; 12; 15; 31; 38]. We will focus on an approach of Lawrence [18]. Lawrence [18] calls the model Gaussian process latent variable model (GPLVM). This model is, in the linear case, equivalent to PCA and can be extended to non-linear mappings $f(\mathbf{X}) = \mathbf{Y}$. In the following chapter we will have a closer look to GPLVM.

2.2 Gaussian Process Latent Variable Model (GPLVM)

In this section we will look at the core method of this thesis; the Gaussian process latent variable model (GPLVM) by Lawrence [18, 19]. Considering multidimensional phenotypes (e.g. gene expression series) $\mathbf{Y} = (\mathbf{y}_d)_{1\leq d\leq D} \in \mathbb{R}^{N\times D}$ of N samples, each of which has dimensionality D, we search for a low dimensional input space $\mathbf{X} \in \mathbb{R}^{N\times Q}$ and function $f(\mathbf{X})$ generating higher dimensional outputs $f(\mathbf{X}) = \mathbf{Y}$. Not to be mistaken by the association function f^{asso} (2.1). For simplicity we will consider the linear mapping $\mathbf{Y} = \mathbf{XV}$ with weights $\mathbf{V} \in \mathbb{R}^{Q\times D}$, weighting the individual influences of latent features and release that assumption later. In a usual latent variable model setting we would put a prior on the latent inputs $\mathbf{X}$ and integrate them out. Then we would learn the weights for a given input $\mathbf{X}$ to create the data observed, which is exactly the approach of probabilistic principal component analysis proposed by Tipping and Bishop [33]. In GPLVM we, instead, put a prior on the mapping from $\mathbf{X}$ to $\mathbf{Y}$ (which in the linear case corresponds to the weights $\mathbf{V}$), and integrate them out. Thus, we learn the most probable input $\mathbf{X}$ for a given output $\mathbf{Y}$ created by the generative model $\mathbf{Y} = \mathbf{XV}$. (This turns out to be equivalent to principal component analysis with linear mappings [4; 34]). The prior for the weights takes the form

$$p(\mathbf{V}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{v}_d|\mathbf{0}, \alpha^{-1}\mathbf{I}_Q) \ , \tag{2.8}$$

where $\mathbf{v}_d$ is the dth column (i.e. weighting dimension d) of $\mathbf{V}$. By integrating over $\mathbf{V}$ the likelihood for the phenotype $\mathbf{Y}$, given an input $\mathbf{X}$ can be written as

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{d=1}^{D} p(\mathbf{y}_d|\mathbf{X}) \ , \tag{2.9}$$

where

$$p(\mathbf{y}_d|\mathbf{X}) = \mathcal{N}(\mathbf{y}_d|\mathbf{0}, \mathbf{K}_{NN} + \beta^{-1}\mathbf{I}_N) \ . \tag{2.10}$$

Here $\mathbf{K}_{NN}$ is the covariance matrix created by the covariance function $k(\mathbf{x}, \mathbf{x}')$, evaluated at every sample (i.e. row) $\mathbf{x}$ and $\mathbf{x}'$ of the inputs $\mathbf{X}$. In the linear case this corresponds to

$$\mathbf{K}_{NN} = \mathbf{X}\mathbf{A}\mathbf{X}^\top \ , \tag{2.11}$$

where $\mathbf{A} \in \mathbb{R}^{Q\times Q}$ is a diagonal matrix with parameters $\boldsymbol{\theta} = \mathrm{diag}(\mathbf{A})$. We can now estimate the distribution of the data (2.7) by learning both the parameters of the model and the latent inputs jointly by finding the maximum a posteriori (MAP) estimate of $\mathbf{X}$. This estimate can be found by maximising the log-likelihood of the model given $\mathbf{X}$

$$\ln p(\mathbf{Y}|\mathbf{X}) = -\tfrac{DN}{2}\ln(2\pi) - \tfrac{D}{2}\ln|\mathbf{K}| - \tfrac{1}{2}\mathrm{tr}(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^\top) \ , \tag{2.12}$$

where $\mathbf{K} = \mathbf{K}_{NN} + \beta^{-1}\mathbf{I}_N$. Thus, the set of parameters

$$\{\hat{\mathbf{X}}, \hat{\boldsymbol{\theta}}, \hat{\beta}\}_{\mathrm{MAP}} = \underset{\{\mathbf{X},\boldsymbol{\theta},\beta\}}{\arg\max} \ \ln p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}, \beta) \ , \tag{2.13}$$

which maximises the log-likelihood (2.12) contains the (a posteriori) most probable inputs for the given phenotype $\mathbf{Y}$. Here we intentionally added the parameters $\boldsymbol{\theta}$ of the covariance function and the precision parameter β of each Gaussian process (2.10). This set also contains the most probable set of inputs $\mathbf{X}$, generating the phenotype under the generative model $\mathbf{Y} = \mathbf{X}\mathbf{V}$. This model is generated by the covariance function one chooses to use (which we assumed to be linear).
To find the most probable set of parameters (2.13) we apply gradient based optimisation, because in most cases there is no fixed point solution to maximise (2.12). However in the linear case the fixed point solution can be written as the well known PCA (Section 2.1). The Gradient of (2.12) can be written as

$$\frac{\partial \ln p(\mathbf{Y}|\mathbf{X})}{\partial \mathbf{X}} = \alpha\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^\top\mathbf{K}^{-1}\mathbf{X} - \alpha D\mathbf{K}^{-1}\mathbf{X} \ , \tag{2.14}$$

where α is the noise parameter for the prior over $\mathbf{V}$ (2.8). Solving this gradient for $\mathbf{X}$ leads to

$$\frac{1}{D}\mathbf{Y}\mathbf{Y}^\top\mathbf{K}^{-1}\mathbf{X} = \mathbf{X} \ . \tag{2.15}$$

With some algebraic manipulation this solution can further be simplified [19] to

$$\mathbf{X} = \mathbf{U}_Q\mathbf{L}\mathbf{\Sigma}^\top \ , \tag{2.16}$$

where $\mathbf{U}_Q$ is a $N \times Q$ matrix, whose columns are eigenvectors of $\mathbf{Y}\mathbf{Y}^\top$, $\mathbf{L}$ is a diagonal $Q \times Q$ matrix, whose diagonal entries $\boldsymbol{\ell} = (l_q)_{1\leq q\leq Q}$ are $l_q = \left(\frac{\lambda_q}{\alpha D} - \frac{1}{\beta\alpha}\right)^{-\frac{1}{2}}$, while λ_q is the qth eigenvalue of $\mathbf{Y}\mathbf{Y}^\top$, and $\mathbf{\Sigma}$ is an arbitrary $Q\times Q$ orthogonal matrix. This is the PCA (Section 2.1) solution to the problem (Lawrence [19]). In Section 2.1 we used the eigenvectors $\mathbf{V}_Q$ of the covariance matrix $\mathbf{Y}^\top\mathbf{Y}$. These can be transformed to the eigenvectors $\mathbf{U}$ of the matrix $\mathbf{Y}\mathbf{Y}^\top$. (see e.g. Bishop [5]; Tipping and Nh [34]). This transformation and a corresponding normalisation leads to the solution above.

Thus, PCA can be seen as a special case of the GPLVM model, assuming a linear mapping f between phenotype and genotype. Now let us relax this assumption. We can see that (2.9) is a product of D independent Gaussian processes with linear covariance function $k(\mathbf{x}, \mathbf{x}')$. Thus, it is natural to extend this model to non-linear mappings between hidden factors $\mathbf{X}$ and phenotype $\mathbf{Y}$ by introducing any non-linear covariance function k, for example the ARD squared exponential (introduced in Section 2.5). With that, we are able to find non-linear input spaces, generating observed data. Still, we have a problem with applying this model to high dimensional data, because we have to compute the inverse $\mathbf{K}^{-1}$ of the $N \times N$ covariance matrix $\mathbf{K}$. In the next chapter we will see how to introduce $M \ll N$ inducing inputs for $\mathbf{K}$, such that $\bar{\mathbf{K}}$ is a $M \times M$ matrix, corresponding to the rank M form of $\mathbf{K}$. As $\bar{\mathbf{K}}$ is a rank M form of $\mathbf{K}$, the underlying Gaussian process prior still is able to approximate the true posterior of the likelihood.

2.3 Variational Bayesian GPLVM

In GPLVM we have successfully integrated out the mapping variables $\mathbf{V}$ from hidden factors $\mathbf{X}$ to the phenotype $\mathbf{Y}$. By putting a prior on $\mathbf{V}$ and integrating them out, we learn the most probable hidden factors $\mathbf{X}$, creating observed data $\mathbf{Y}$. To estimate the right dimensionality of $\mathbf{X}$, though, we need to treat the hidden factors

as random variables, as well. That means, we also put a prior on $\mathbf{X}$

$$p(\mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\mathbf{0}, \mathbf{I}_{Q\times Q}) \tag{2.17}$$

and integrate them out. Such that, we can compute the marginal likelihood of the model in a fully Bayesian way (Equation 2.7)

$$p(\mathbf{Y}) = \int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})\, d\mathbf{X} \ .$$

This integral, however, is intractable as the hidden variables appear non-linearly in the covariance matrix of the likelihood $p(\mathbf{Y}|\mathbf{X})$ (2.9). We, therefore, use a variational approximation to the model and estimate the true posterior of the hidden variables $p(\mathbf{X}|\mathbf{Y})$ by a variational distribution $q(\mathbf{X})$ which factorises across the hidden factor dimensions

$$q(\mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_n, \mathbf{S}_n) \ , \tag{2.18}$$

where $\{\boldsymbol{\mu}_n, \mathbf{S}_n\}_{1\leq n\leq N}$ are the variational parameters and $\mathbf{S}_n$ is a diagonal (variance) matrix. Now we have an approximated distribution for the hidden factors which can be used to make the corresponding integral (2.7) analytically tractable. We introduce the variational distribution in the model as follows (compare (2.7)):

$$\log p(\mathbf{Y}) = \log \int q(\mathbf{X}) \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{q(\mathbf{X})} d\mathbf{X} \ . \tag{2.19}$$

By applying Jensen's inequality

$$f(\mathbb{E}(\mathbf{X})) \geq \mathbb{E}(f(\mathbf{X})) \qquad \text{(if } f \text{ convex)} \tag{2.20}$$

we get Jensen's lower bound on the log marginal likelihood of the model

$$\begin{aligned}
\log p(\mathbf{Y}) \geq \mathcal{F}(q) &= \int q(\mathbf{X}) \log \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{q(\mathbf{X})} d\mathbf{X} && (2.21)\\
&= \int q(\mathbf{X}) \log p(\mathbf{Y}|\mathbf{X}) d\mathbf{X} - \int q(\mathbf{X}) \log \frac{p(\mathbf{X})}{q(\mathbf{X})} d\mathbf{X} && (2.22)\\
&= \int q(\mathbf{X}) \log p(\mathbf{Y}|\mathbf{X}) d\mathbf{X} - \mathrm{KL}(q||p) && (2.23)\\
&= \tilde{\mathcal{F}}(q) - \mathrm{KL}(q||p) \ , && (2.24)
\end{aligned}$$

where $\mathrm{KL}(q||p)$ is the Kullback–Leibler (KL) divergence between q and p. Thus, we minimise the KL divergence between the variational distribution $q(\mathbf{X})$ and the true conditional distribution $p(\mathbf{X}|\mathbf{Y})$ by maximising this lower bound. This, in turn,

maximises the probability for the model to describe the observed gene expression values. We can compute the KL divergence analytically (q and p are both Gaussian), but the left side $\tilde{\mathcal{F}}(q)$ is intractable, because the hidden factors $\mathbf{X}$ are highly non linear in the likelihood $p(\mathbf{Y}|\mathbf{X})$. Thus, we need a further approximation for the likelihood. Titsias and Lawrence [36] use a variational sparse GP approximation proposed in Titsias [35] and integrate it into the model. The concept of variational GP is to introduce M additional (auxiliary) gene expression values $\mathbf{Y}^M \in \mathbb{R}^{M\times D}$ which, if sufficiently many, still approximate the true posterior of the observed gene expression values $\mathbf{Y}$. Due to the Gaussian process assumption, these (auxiliary) gene expression values are only samples from the GP prior at corresponding inputs $\mathbf{X}^M \in \mathbb{R}^{M\times Q}$. The number M of inducing inputs (additional samples i.e. rows) $\mathbf{X}^M$, thereby, determines the accuracy of the approximation, as depicted in figure 2.2. These so called inducing inputs are not additional parameters in the model, but are variational parameters which still are determined from data. The sparse implementation of the method allows Titsias and Lawrence [36] to calculate another approximation to the Jensen's lower bound $\tilde{\mathcal{F}}(q)$ and with that calculate the full approximation analytically

$$\begin{aligned}\tilde{\mathcal{F}}(q) \geq \sum_{d=1}^{D} \log\left(\left\langle \exp\{\langle \log \mathcal{N}(\mathbf{y}_d|\boldsymbol{\alpha}_d, \beta^{-1}\mathbf{I}_N)\rangle_{q(\mathbf{X})}\}\right\rangle_{p(\mathbf{Y}_d^M)}\right) \\ - \tfrac{\beta}{2}\mathrm{Tr}\left(\langle \mathbf{K}_{NN}\rangle_{q(\mathbf{X})}\right) \\ + \tfrac{\beta}{2}\mathrm{Tr}\left(\mathbf{K}_{MM}^{-1}\langle \mathbf{K}_{MN}\mathbf{K}_{NM}\rangle_{q(\mathbf{X})}\right) \quad , \end{aligned} \tag{2.25}$$

where $\mathbf{K}_{MM}$ is the covariance matrix evaluated at the inducing inputs $\mathbf{X}^M$ and $\mathbf{K}_{NM} = \mathbf{K}_{MN}^\top$ is the covariance matrix evaluated between $\mathbf{X}$ and $\mathbf{X}^M$, β is the precision for the model and $\boldsymbol{\alpha}_d = \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{y}_d^M$ is the mean of the GP prior under the assumption of only having observed the inducing inputs $\mathbf{X}^M$. The last two terms of this equation can be seen as correcting terms of the sparse approximation of the GP posterior and the expression inside the logarithm is the expected value of the GP posterior, having observed the outputs at the optimal inputs $\mathbf{X}$, under the assumption of having observed the inducing inputs $\mathbf{X}^M$. Please see the details of this approximation in the papers of Damianou et al. [8]; Titsias and Lawrence [36].
As the hidden factors $\mathbf{X}$ are variational parameters the analytical solution to the log marginal likelihood of the model contains the expected values of the covariance matrices under the approximate variational distribution q. These expected values $\psi_0 = \mathrm{Tr}\left(\langle \mathbf{K}_{NN}\rangle_{q(\mathbf{X})}\right) \in \mathbb{R}, \boldsymbol{\Psi}_1 = \langle \mathbf{K}_{NM}\rangle_{q(\mathbf{X})} \in \mathbb{R}^{N\times M}$ and $\boldsymbol{\Psi}_2 = \langle \mathbf{K}_{MN}\mathbf{K}_{NM}\rangle_{q(\mathbf{X})} \in \mathbb{R}^{M\times M}$ are embedded within the covariance functions and can be computed for a number of standard covariance functions, such as the linear and squared expo-

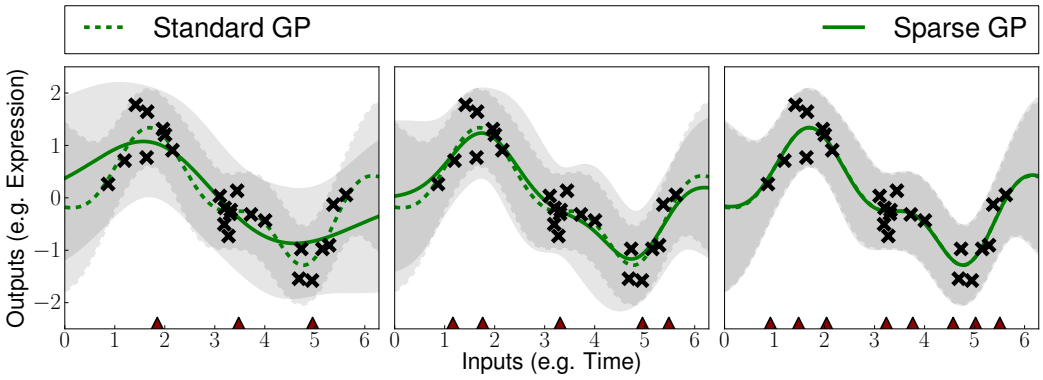


Figure 2.2: Sparse GP approximation to a Standard GP with differing numbers of inducing inputs $\mathbf{X}^M$. From left to right: $3, 5$ and 8 inducing inputs. The learnt positions of inducing inputs are indicated as red triangles at the bottom of each figure. Note how the inducing inputs always being learnt at points of high interest, because of presence of data. You can see that the ability of the sparse GP to approximate the true distribution of the observed gene expression values increases with the number of inducing inputs. The sparse approximation with 8 inducing inputs is similar to the Standard GP, and thus, it is enough to have 8 inducing inputs approximating the full covariance matrix of all $N = 20$ samples. The number of inducing inputs still has to be chosen sufficiently large to be able to approximate the true covariance structure in a more realistic setting.

nential covariance function, analytically. With these so called ψ-statistics the lower bound takes the form

$$\tilde{\mathcal{F}}(q) \geq \log\left\{\frac{\beta^{\frac{N}{2}}|\mathbf{K}_{MM}|^{\frac{1}{2}}}{(2\pi)^{\frac{N}{2}}|\beta\boldsymbol{\Psi}_2+\mathbf{K}_{MM}|^{\frac{1}{2}}}\exp\left\{-\tfrac{1}{2}\mathbf{y}_d^\top\mathbf{W}\mathbf{y}_d\right\}\right\} - \tfrac{\beta}{2}\psi_0 + \tfrac{\beta}{2}\mathrm{Tr}\left(\mathbf{K}_{MM}^{-1}\boldsymbol{\Psi}_2\right) \quad , \tag{2.26}$$

where $\mathbf{W} = \beta\mathbf{I}_N - \beta^2\boldsymbol{\Psi}_1\left(\beta\boldsymbol{\Psi}_2+\mathbf{K}_{MM}\right)^{-1}\boldsymbol{\Psi}_1^\top$ is the approximation of the full covariance, given the inducing inputs $\mathbf{X}^M$. We are now able to maximise the lower bound on the log marginal likelihood of the model following Equation (2.13) with the model parameters $\{\beta, \boldsymbol{\theta}\}$ and the variational parameters $\{\{\boldsymbol{\mu}_n, \mathbf{S}_n\}_{1\leq n\leq N}, \mathbf{X}^M\}$. This is similar to the MAP approach of the GPLVM (Section 2.2), with the important addition of the variational parameters of the hidden factors. This addition allows us to variationally determine the number of dimensions Q necessary to describe the hidden factors of the gene expression automatic relevance determination (ARD). ARD determines the necessary number of dimensions by "switching off" unnecessary dimensions of the hidden variables as will be described in more detail in Section 2.5.

2.3.1 Computation of the ψ-Statistics

The ψ-statistics are the expected values of the covariance matrix evaluated at the hidden factors and inducing inputs, respectively, under the variational distribution $q(\mathbf{X})$. ψ_0 can be written as [36]

$$\psi_0 = \sum_{n=1}^{N} \int k(\mathbf{x}_n, \mathbf{x}_n) \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_n, \mathbf{S}_n) \, d\mathbf{x}_n \ , \tag{2.27}$$

where $\mathbf{x}_n$ is the nth sample (i.e. row) of the hidden factors $\mathbf{X}$.
$\boldsymbol{\Psi}_1$ is an $N \times M$ matrix with [36]

$$(\boldsymbol{\Psi})_{nm} = \int k(\mathbf{x}_n, \mathbf{x}_m^M) \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_n, \mathbf{S}_n) \, d\mathbf{x}_n \ , \tag{2.28}$$

where $\mathbf{x}_n$ as above and $\mathbf{x}_m^M$ is the mth sample (i.e. row) of the inducing inputs $\mathbf{X}^M$.
And $\boldsymbol{\Psi}_2$ is an $M \times M$ matrix $\boldsymbol{\Psi}_2 = \sum_{n=1}^{N} \boldsymbol{\Psi}_2^n$ with [36]

$$(\boldsymbol{\Psi}_2^n)_{mm'} = \int k(\mathbf{x}_n, \mathbf{x}_m^M) k(\mathbf{x}_{m'}^M, \mathbf{x}_n) \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_n, \mathbf{S}_n) \, d\mathbf{x}_n \ . \tag{2.29}$$

Squared Exponential ψ-statistics

Titsias and Lawrence [36] computed the ψ-statistics for the squared exponential covariance function

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left\{ \tfrac{1}{2} \boldsymbol{\alpha}^\top ||\mathbf{x} - \mathbf{x}'||^2 \right\} \tag{2.30}$$

as follows: ψ_0 is

$$\psi_0 = N\sigma^2 \ .$$

$\boldsymbol{\Psi}_1$ is written as

$$(\boldsymbol{\Psi}_1)_{nm} = \sigma^2 \prod_{q=1}^{Q} \frac{\exp\left\{ -\frac{1}{2} \frac{\alpha_q (\mu_{nq} - x_{mq}^M)^2}{\alpha_q S_{nq} + 1} \right\}}{(\alpha_q S_{nq} + 1)^{\frac{1}{2}}} \ . \tag{2.31}$$

And $\boldsymbol{\Psi}_2 = \sum_{n=1}^{N} (\boldsymbol{\Psi}_2^n)$ is the sum over

$$(\boldsymbol{\Psi}_2^n)_{mm'} = \sigma^4 \prod_{q=1}^{Q} \frac{\exp\left\{ -\frac{\alpha_q (x_{mq}^M - x_{m'q}^M)^2}{4} - \frac{\alpha_q \left(\mu_{nq} - \bar{x}_q^M\right)^2}{2\alpha_q S_{nq} + 1} \right\}}{(2\alpha_q S_{nq} + 1)^{\frac{1}{2}}} \ , \tag{2.32}$$

where $\bar{x}_q^M = \frac{x_{mq}^M + x_{m'q}^M}{2}$ is the mean of the qth hidden factor dimension between the mth and m'th sample of the inducing inputs.

Linear covariance ψ-statistics

The ψ-statistics for the linear covariance function

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{\Lambda} \mathbf{x}' \quad , \tag{2.33}$$

with the ARD parameters $\boldsymbol{\alpha}$ in the diagonal matrix $\mathbf{\Lambda}$ are as follows:

$$\psi_0 = \sum_{n=1}^{N} \mathrm{Tr}\{\mathbf{\Lambda}(\boldsymbol{\mu}_n \boldsymbol{\mu}_n^\top + \mathbf{S}_n)\} \quad , \tag{2.34}$$

$$(\mathbf{\Psi}_1)_{nm} = \boldsymbol{\mu}_n^\top \mathbf{\Lambda} \mathbf{x}_m^M \quad , \tag{2.35}$$

$$(\mathbf{\Psi}_2)_{mm'} = \sum_{n=1}^{N} \mathbf{x}_m^M \mathbf{\Lambda}(\boldsymbol{\mu}_n \boldsymbol{\mu}_n^\top + \mathbf{S}_n) \mathbf{\Lambda} \mathbf{x}_{m'}^M \quad . \tag{2.36}$$

2.3.2 Model Complexity

The model complexity of the variational Bayesian GPLVM scales as $\mathcal{O}(NM^2)$, because all operations on the variational parameters can be done in sequence and only the sparse approximation $\mathbf{K}_{MM}$ has to be inverted. The inversion of the covariance matrix is based on the sparse approximation which is the most costly part of the algorithm. Still optimisation of parameters may proceed slowly, because there are $(2N + M) \cdot Q$ variational parameters and the (hyper-) parameters $\boldsymbol{\theta}$ for the model itself. Also the choice of the optimiser can influence the quality and speed of convergence (based on personal experience, we encourage the usage of the BFGS-algorithm, see e.g. Fletcher [11]). And last the initialisation of the model plays a huge role. So if the model is not learning as efficiently as you would like or learnt optima are far from expected regions you should try to make initialisation more sensible. E.g. setting the inducing variables $\mathbf{X}^M$ to a multivariate random initialisation, instead of a grid etc.

2.4 Manifold Relevance Determination [9]

As seen we now have a tool to estimate global confounding variation in gene expression. Gene expression is a highly non absolute process. Therefore, we need to be able to compare several gene expression experiments do each other. This approach is commonly done in differential gene expression determination, where the gene expression of one organism is compared between exposure to stress and a normal state. Thus, most gene expression experiments have several exposures to stress

sources with the same samples of population or species. We want to be able to capture confounding variation in such settings, as well. Damianou et al. [9] developed a method which enables variational Bayesian GPLVM to learn confounders shared between experiments, as well as determine confounders, which are private for each experiment. Damianou et al. [9] call this method manifold relevance determination. In manifold relevance determination we have one confounder space $\mathbf{X}$ shared between all experiments $\mathbf{Y}^i$. For each experiment we learn a set of ARD parameters $\boldsymbol{\alpha}^i$, which determine the relevance of each of the learnt confounders for the particular gene expression experiment $\mathbf{Y}^i$ (see 2.5 for ARD). Thus, in the end we have shared confounding factors between gene expression experiments and confounding factors, which are private for each experiment (or other combinations of shared and private if more then 2 experiments are analysed). If, for example, $\boldsymbol{\alpha}_k^1 = 0$ and $\boldsymbol{\alpha}_k^2 \neq 0$ we can deduce that confounder k is private in gene expression experiment $\mathbf{Y}^2$ (See results Section 3.2 for figures). In (2.19) we defined the evidence for one gene expression experiment $\mathbf{Y}$, but we now introduce I experiments $\mathbf{Y}^{1\leq i\leq I}$. Thus, we want to compute the evidence on the gene expression

$$p(\mathbf{Y}^{1\leq i\leq I}) = \int p(\mathbf{Y}^{1\leq i\leq I})p(\mathbf{X})\, d\mathbf{X} \ , \tag{2.37}$$

where $\mathbf{X}$ is the confounder space. This is intractable in $\mathbf{X}$, because $\mathbf{X}$ is nonlinear in the covariance of the founding GP priors. Thus, we will apply the variational trick as seen in (2.19) again. By introducing the variational approximation we obtain a variational bound on the evidence of the model:

$$\mathcal{F}(q) = \int q(\Theta)q(\mathbf{X}) \log \left(\frac{\prod_{i=1}^I p(\mathbf{Y}^i|\mathbf{X})}{q(\Theta)} \frac{p(\mathbf{X})}{q(\mathbf{X})} \right) d\mathbf{X} \tag{2.38}$$

$$= \sum_{i=1}^I \tilde{\mathcal{F}}^i(q) \quad - \mathrm{KL}(q||p) \ , \tag{2.39}$$

where $q(\Theta)$ is a variational distribution which factorises over the experiments $q(\Theta) = \prod_{i=1}^I q(\Theta^i)$. Here Θ^i are variational parameters for experiment i and

$$\tilde{\mathcal{F}}^i(q) = q(\Theta^i)q(\mathbf{X}) \log \frac{p(\mathbf{Y}|\mathbf{X})}{q(\Theta^i)}\, d\mathbf{X} \tag{2.40}$$

is the variational bound on the intractable part of $\mathcal{F}(q)$. By introducing a sparse approximation on the intractable variational bound $\tilde{\mathcal{F}}^i(q)$ Damianou et al. [9] overcome the intractability of (2.38) in $\mathbf{X}$ and by intelligently choosing $q(\Theta)$, it cancels out in the bound $\mathcal{F}(q)$ (see details in Damianou et al. [9]). The full variational bound ind MRD becomes a sum over variational Bayesian GPLVM bounds (2.25)

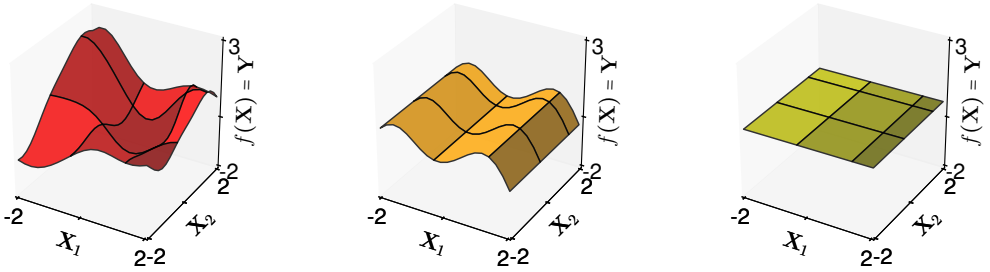
over the different gene expression experiments $\mathbf{Y}^{1\leq i\leq I}$, and the KL-divergence has to be accounted for only once, since $\mathbf{X}$ is shared.

2.5 Automatic Determination of Active Dimensions

In Gaussian process based models we assume a multivariate Gaussian distribution over observed data $\mathbf{Y}$ with covariance matrix $\mathbf{K}$. This covariance matrix $\mathbf{K}$ is generated using a covariance function $k(\mathbf{x}, \mathbf{x}')$ and evaluating it for each pair of samples (i.e. rows) of inputs $(\mathbf{x}, \mathbf{x}') \in \mathbf{X} \times \mathbf{X}$, where $\mathbf{X} \times \mathbf{X}$ corresponds to all possible pairs of rows of $\mathbf{X}$. The choice of covariance function determines the underlying (generative) function of the Gaussian process [26]. In this thesis we are interested in "switching off" dimensions of latent variables. To achieve such behaviour, we apply the so called automatic relevance determination (ARD) method (see e.g. Rasmussen [26, Chapter 5]). Looking at the ARD squared exponential covariance function (Equation (2.30))

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left\{\tfrac{1}{2}\boldsymbol{\alpha}^\top \|\mathbf{x} - \mathbf{x}'\|^2\right\} \ ,$$

we can see that the parameters $\boldsymbol{\alpha} \in \mathbb{R}^D$ create a relevance score for each dimension. If α_d is large, dimension d of $\mathbf{x}$ is relevant to the covariance. If α_d is small, dimension d gets irrelevant for the covariance and the corresponding dimension gets flattened out in the generative function space. To get an intuition on how these parameters work we plot three examples of an Gaussian process with two dimensional input (which for example corresponds to time) $\mathbf{X}$ and one dimensional output (for example expression value) $\mathbf{Y}$ with an ARD squared exponential covariance function in Figure 2.3. In the left plot you see both dimensions being relevant for the function ($\alpha_1 = \alpha_2 = 1$). In the middle plot we "switched off" dimension 2 by setting $\alpha_2 = 0.000{,}1$. The right most plot shows both dimensions being "switched off", where $\alpha_1 = \alpha_2 = 0.000{,}1$. Thus, the model is able to decide whether a dimension d is relevant or not by setting the value α_d of the corresponding dimension to a value of choice. The model has to be driven to choose sparse parameters $\boldsymbol{\alpha}$ to achieve switching off dimensions now. This is usually done by introducing penalisation terms over the size of the ARD parameters, but in a Bayesian GP setting we integrate out parameters, and thus, penalisation cannot be done explicitly in the model. We, therefore, introduce penalisation implicitly through giving the model the opportunity to choose sparsity of parameters and the penalisation term falls out naturally of the integrals of the model (compare $\mathrm{Tr}(\mathbf{K}_{MM}^{-1}\boldsymbol{\Psi}_2)$ in Equation (2.25)). This



(a) Both dimensions of the input $\mathbf{X}$ are relevant.

(b) Only dimension 1 of the input $\mathbf{X}$ is relevant.

(c) Both dimensions of the input $\mathbf{X}$ are "switched off".

Figure 2.3: Three examples of a Gaussian process with ARD squared exponential covariance function. From left to right you can see that (a) both input dimensions are relevant, (b) only $\mathbf{X}_1$ is relevant and (c) both input dimensions are irrelevant. Here $\mathbf{X}_i$ denotes the ith dimension (i.e. column) of the input $\mathbf{X}$.

is called Bayesian model selection [26]. Integrating over model parameters (in a GP setting) always introduces a penalisation term, which contains the trace over the covariance matrix (or its inverse) and for that reason a model always searches for a balance between complexity (many $\alpha_d \neq 0$) and generalisation (almost all $\alpha_d = 0$). That means the model chooses reasonable complexity for a good generalisation of stochastic non-significant variation.
The linear covariance function (2.33)

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{\Lambda} \mathbf{x}' \quad ,$$

with ARD parameters $\boldsymbol{\alpha}$ in the diagonal matrix $\mathbf{\Lambda}$ has also the property of ARD, similar to the above. In this thesis, we mostly work with the linear ARD covariance function, because in the linear case we have closed forms for the posterior of the mapping variables $\mathbf{V}$ and know the underlying basis function f is linear.

Chapter 3

Applying Variational Bayesian GPLVM

In this chapter we will apply variational Bayesian Gaussian process latent variable models to biological data to first show its capability of detecting confounding factors in causal relationships between genotype and phenotype of organisms. And second, find shared/private information between *trans* associations of datasets. As already discussed in previous chapters, we focus on detecting connections between genetic variation and differential gene expression.

3.1 Modelling Confounding Factors in eQTL Studies

Expression quantitative trait loci (eQTL) studies aim at linking genetic variation in specific trait loci (mostly SNPs) to gene expression variation (Section 1.2). Though the success on eQTL studies to detect genetic reasoning on gene expression variation, it is well known that such interactions are prone to (unknown) environmental and combinatorial confounding factors (confounders) as discussed in Section 1.3. Several methods to account for confounding factors in eQTL studies have been developed on a variety of algorithmically underlying bases [24; 27; 31]. In this context studies have shown, that genetic variation is prone to confounding factors, which have to be accounted for. One method which accounts for confounding variation in eQTL studies is the PANAMA algorithm (Fusi et al. [12]), which is a GPLVM (Section 2.2) approach for modelling confounding factors and genotypes jointly.
PANAMA assumes the genotype $\mathbf{S} = (\mathbf{s}_k)_{1\leq k\leq K} \in \mathbb{R}^{N\times K}$ of K SNPs and confounding factors $\mathbf{X} = (\mathbf{c}_q)_{1\leq q\leq Q} \in \mathbb{R}^{N\times Q}$ of Q latent variables to additively contribute to

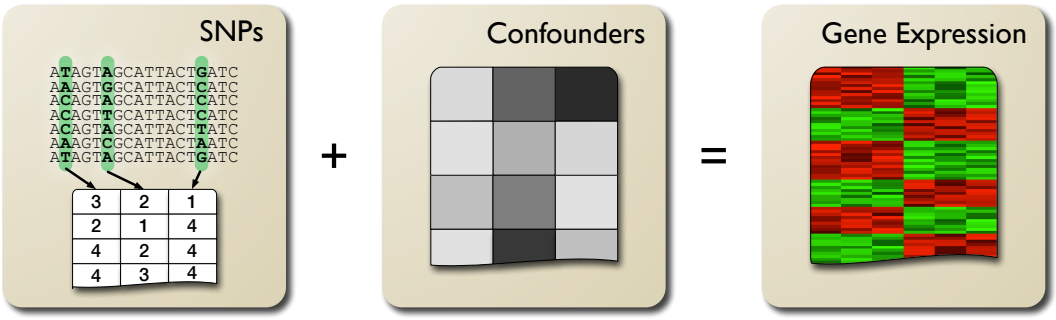


Figure 3.1: PANAMA model: SNPs (genotype) and hidden factors (confounders) contribute additive to the phenotype (gene expression). In this figure we omitted the weight matrices $\mathbf{V}, \mathbf{W}$ for clarity.

the phenotype $\mathbf{Y} \in \mathbb{R}^{N \times D}$ of N samples of D gene expression levels.

$$\mathbf{Y} = \boldsymbol{\mu}\mathbf{1}^\top + \mathbf{S}\mathbf{V} + \mathbf{X}\mathbf{W} + \boldsymbol{\varepsilon} \ , \tag{3.1}$$

where $(\mu_d)_{1 \le d \le D}$ is a mean expression level per gene d, ε_d is a gene specific noise and $\mathbf{V} = (\mathbf{v}_d)_{1 \le d \le D}$ and $\mathbf{W} = (\mathbf{w}_d) 1 \le d \le D$ are $K \times D$ and $Q \times D$ dimensional weight matrices, weighting the specific contribution of the corresponding genotype and confounding latent variables, respectively. See Figure 3.1 for an illustration of this model. Fusi et al. [12] developed an iterative approach, adding genotypes (SNPs) highly correlated with confounding factors to estimation, during fitting of the model. With that the model is able to decide which contribution is more descriptive, the genotype or the confounder. Additionally, this procedure allows for automatic determination of the number Q of confounders needed to explain confounding variation not explained by the genotype.

In variational Bayesian GPLVM we are able to detect the confounding factors and their dimensionality through the Bayesian approximation of the marginal distribution of latent variables (See section 2.3) in combination with ARD parameters (Section 2.5). This procedure finds consistent variation over the gene expression matrix and treats this as confounding variation, not explainable by the genotype, which is non-consistent over the samples collected. To be able to detect the dimensionality we only have to provide a sufficient number of dimensions to fit and variational Bayesian GPLVM will "switch off" all dimensions not needed, setting the ARD parameters of the respective dimensions near to zero. In the following, we will first introduce the experimentour applications are applied to [30]. Second, we will introduce the variational Bayesian GPLVM variant of the PANAMA model. Third, we will apply this model to a simulated experiment as a proof of concept: that we, indeed, are able to find the right dimensionality of the confounding factors. Last,

we will apply the model to a realistically simulated experiment, created and used by Fusi et al. [12] in order to have a ground truth. This simulated experiment is created using the glucose part of the experiment by Smith and Kruglyak [30].

3.1.1 Genome-Wide eQTL Dataset [30]

We used a genome-wide association study (GEO accession number GSE9376), consisting of 109 samples of crosses of two parental strains of *Saccharomyces Cerevisiae* (yeast) grown in two conditions as carbon source, namely glucose and ethanol. As we know the genotype of the parental strains (BY and RM, the crosses are genotyped as well) this data set allows for association studies on environmental condition. The gene expression pattern in yeast changes drastically between glucose and ethanol as carbon source. When the cells run low on glucose they switch to a primarily respiratory state in which they metabolise ethanol (Smith and Kruglyak [30]). The experiment consists of $N = 109$ samples of $D = 5493$ genes profiled with a genotype of $K = 2956$ SNPs. The genotype was provided by Fusi et al. [12]. It is the same for both experiments, such that only the environment changes between the experiment. In this setting only the genomic variant can influence the behaviour of the gene expression and, thus, we find gene-environment interactions, as the environment changes.

3.1.2 Bayesian Estimation of Confounding Variation

As already mentioned in the previous section, we want to apply variational Bayesian GPLVM to find confounding variation in genome wide association studies (GWAS). Variational Bayesian GPLVM is able to detect the dimensionality of confounding factors through the ARD property of the covariance function, which we apply. Thus, we do not use the iterative approach from Fusi et al. [12] of adding genotype to the model, while fitting. We learn confounders (confounding factors) $\mathbf{X}$ by variational Bayesian GPLVM on gene expression levels $\mathbf{Y}$. Variational Bayesian GPLVM, therefore, does not need to see the genotype while fitting the model.
We assume linear confounding variation and choose the linear ARD covariance function for dimensionality deduction as described in Section 2.5. With these learnt confounders we then can learn the weights for the linear mixed model (3.1) by taking learnt confounders into account. Thus, we first learn confounding factors $\mathbf{X}$ by applying variational Bayesian GPLVM on the gene expression matrix $\mathbf{Y}$. Afterwords, we use the learnt hidden factors and the true genotype $\mathbf{S}$ to apply the same linear mixed model as in PANAMA (3.1). The output of the linear mixed model is

an association matrix $\mathbf{A}^p \in \mathbb{R}^{D\times K}$, where a_{dk} indicates whether there is an association between the dth gene and the kth SNP in form of a p-value. p-values denotes the statistical significance of the test deviating from the null-hypothesis of there being no association. The association matrix $\mathbf{A}^p$ is then adjusted for multiple testing and the p-values contained are converted into q-values resulting in the q-value association matrix $\mathbf{A}^q$ (see more on statistical analysis of GWAS in e.g. Storey and Tibshirani [32]). Thus, if an adjusted q-value a_{dk} is smaller than a false discovery threshold of e.g. 0.05 the association between gene d and SNP k is called significant - the null hypothesis of there being no association gets rejected.

3.1.3 Finding the Right Dimensionality

To be able to assess the quality of variational Bayesian GPLVM to find the right dimensionality of a latent input space, we simulated a experiment, consisting of $K = 800$ SNPs and $D = 1000$ gene expression values for $N = 100$ samples. We simulated $Q = 10$ confounding factors, contributing linearly to gene expression (Figure 3.1). Thus, we generated samples from the linear mixed model (3.1), which underlies our model. The number Q of confounding factors should be learnt properly on a simple simulation, sampled from the underlying model. We initialised the model with 23 dimensions. As you can see in Figure 3.2 the variational Bayesian GPLVM model "switched off" 13 of the 23 ARD parameters $\boldsymbol{\alpha}$, resulting in $Q = 10$ learnt dimensions. Thus, the model is not only able to extract confounding variation in a gene expression matrix, but also depicts the right dimensionality of the confounding factors. As a comparison we additionally plotted the learnt ARD parameters of a standard GPLVM model. You can see that the standard GPLVM learns more than 10 confounding factor dimensions, because it is not able to deduce the right dimensionality in a Bayesian way. This simple simulation shows the ability of the

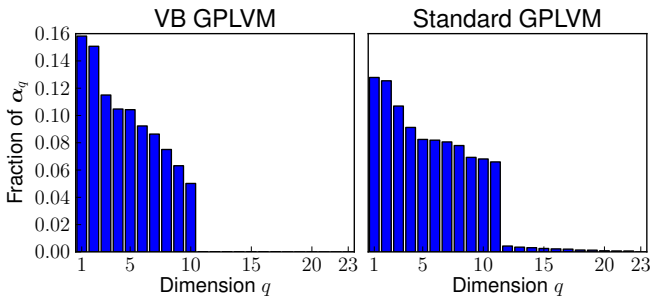


Figure 3.2: Learnt ARD parameters $\boldsymbol{\alpha}$ of variational Bayesian (VB) GPLVM and GPLVM, respectively. Both methods were applied to a simulated experiment containing 10 confounding factors. For comparability, we plotted fractions of the ARD parameters and sorted them descending. The left side shows the $Q = 10$ learnt dimensions by variational Bayesian GPLVM, whereas the standard GPLVM model (right) learns around 20 hidden factor dimensions.

variational Bayesian GPLVM to "switch off" dimensions properly and treat additional noise as such and does not give them any weight (dimensions $Q > 10$ in Figure 3.2). In the next section, we will apply this same methodology to a more realistic simulation of confounding variation in gene expression data, created by Fusi et al. [12].

3.1.4 Realistic Simulation

The next simulation we apply our method to is a more realistic simulation of confounders, based on the glucose part of the real differential gene expression experiment by Smith and Kruglyak [30], introduced in Section 3.1.1. The simulation takes the (known) genotype of the yeast population used in the experiment and creates a gene expression matrix through mixing confounding variation into the gene expression created by the genotype. In this procedure Fusi et al. [12] carefully adjust the simulation to be as similar to the gene expression discovered as possible (see the paper of Fusi et al. [12] for in-depth details on this simulation). The simulation does not change the number of samples $N = 109$ of $D = 5493$ genes profiled nor the number of $K = 2956$ SNPs (genotype), but it provides a ground truth for the association matrix $\mathbf{A}$ learnt by the linear mixed model (see 3.1.2) and, thus, allows for model comparison. Following models are applied to the simulation:

Linear In the linear model we apply the linear mixed model to the gene expression, without any correction for confounding variation. We expect this method to associate too much genetic variation to genes. Therefore, this model deals as a upper limit of finding associations between genotype and phenotype (gene expression). It is the least conservative model of the compared models.

Surrogate Variable Analysis (SVA) This model was created by Leek [20]; Leek and Storey [21]. It models confounders as so-called surrogate variables, which are assumed to be linear combinations of the true confounding variation and can be estimated using assumptions on the noise variance of the data. It is a statistical test on confounding variation. The learnt surrogate variables behave in a similar way as confounders do and are used to correct for consistent confounding variation in observed data. See Leek [20]; Leek and Storey [21] for a more detailed description of the method.

GPLVM For comparison we apply GPLVM to find confounders. This method does find too many confounders and is therefor prone to explain genetic variation by confounding factors. It will explain away genetic signal and is therefore the most conservative model.

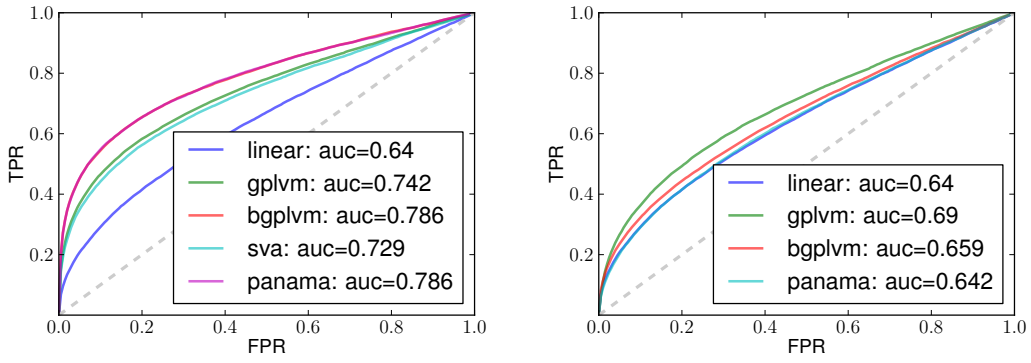
Variational Bayesian GPLVM We apply our method with sufficient ARD parameters ($Q = 50$) to fit confounding variation and "switch off" any surplus latent variables. Variational Bayesian GPLVM indicates that the number of confounders in this simulation is $Q \sim 43$ (data not shown). That is, $Q = 43$ ARD parameters are larger than a threshold of 5×10^{-4}.

PANAMA Lastly, we apply Panama as a comparison. Note here, that PANAMA uses an iterative approach, including genetic variation into estimation of confounding variation in each step. This means, that the genotype is needed to account for non-genetic signals. In our Bayesian way of estimating confounding variation we do not need the genotype for estimation.

Table 3.1: The four statements about classification, contrasting the biological truth (Truth) versus the classification outcome (Prediction). Where P=Positive, N=Negative, T=True and F=False.

		Prediction	
		P	**N**
Truth	**P**	*TP*	*FN*
	N	*FP*	*TN*

To assess the quality of variational Bayesian GPLVM in modelling confounding variation, we compare all methods through their ability to find the true associations. There are four classes of hits in a classification study. True positives, which are actual positives classified positive; false positives, actual negatives, which are classified positive; true negatives, rightly negative classified negatives; and false negatives, which are actual positives, but classified as negative. Table 3.1 shows these classes in an overview. We show the ability of the different algorithms to assign the right classes in a receiver operating characteristic (ROC) curve, which plots the false positive rate (FPR; number of false positives out of all positives) against the true positive rate (TPR; number of true positives out of all positives). A perfect classification would have $100\,\%$ in both FPR and TPR. Thus, the closer the area under the ROC curve (AUC) to 1, the better the quality of the algorithm in finding true associations. A completely random classification would have consistently a FDR and TPR of $50\,\%$, which is a straight line trough the origin and $(x = 1, y = 1)$.



(a) (b)

Figure 3.3: Receiver operating characteristic curves for all five methods applied to the real world simulation described in 3.1.4. **(a)**: Applied to the full glucose experiment. Variational Bayesian GPLVM (bgplvm) does as well as PANAMA does, although there is no genetic information used to estimate confounding variation in gene expression. All methods do classify better than the linear method, which does not include any estimation of confounding variation. Thus, we see, that accounting for confounding variation helps to find genotype to phenotype association in such studies. **(b)**: Applied to a subset of fewer genes ($D = 20$). GPLVM does best, although explaining away genetic signal. Variational Bayesian GPLVM still seems to find confounding variation, as it does a bit better than linear. SVA is not shown, because it needs more genes than samples to estimate confounders.

Thus, everything below this line, would indicate at a inverse classification and one would have to turn the prediction result around (which in most of the times means multiplying it by -1).

We apply all five methods proposed above and plot their respective assignment quality in form of ROC curves in Figure 3.3. You can see that the variational Bayesian GPLVM does as good as PANAMA does, although we do not include genetic information to learn confounding factors. To further investigate this result we provide p-value histograms over the p-values found for associations between genotype and phenotype for all five methods. A p-value histogram shows the count of p-values for a given number of bins spanning from zero to one. As we want the method to find clear cut associations between genotype and phenotype, we want the distribution of p-values to be dense towards zero and have a uniform shape afterwards (see [25] for details). Variational Bayesian GPLVM seems to give the most clear cut decisions for calling associations in all five compared methods (Figure 3.4(a)), while being the most conservative method overall. All other methods have a reasonably inflated p-value distribution, while being more generous in calling associations.

Next we investigate *cis* and *trans* associations of all methods. If a SNP is in the

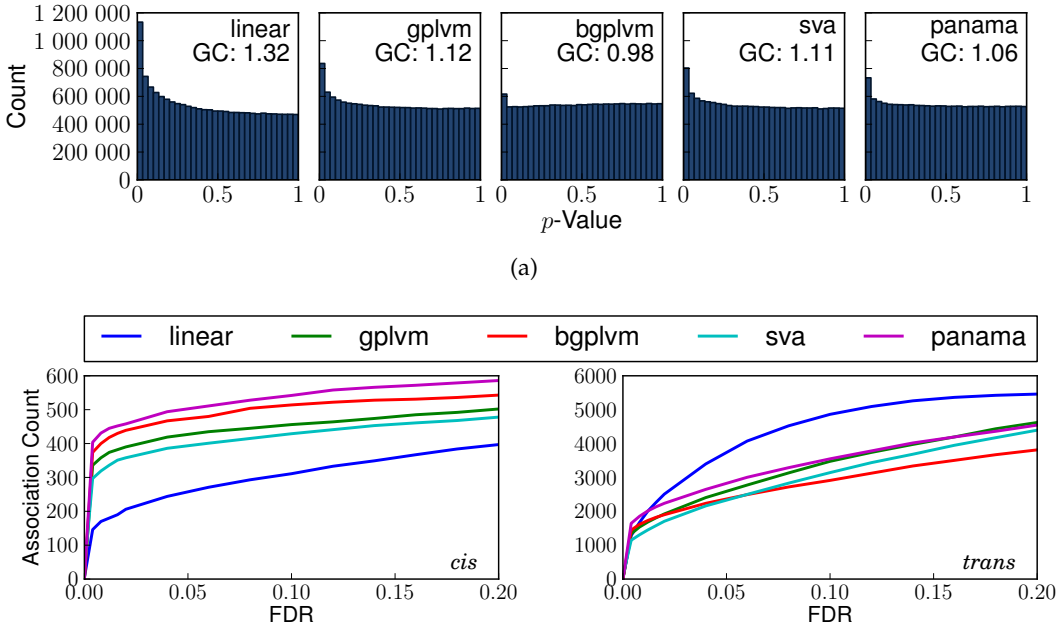


Figure 3.4: (a): p-value histograms for all five methods. Variational Bayesian GPLVM (bgplvm) seems to find the most clear cut signal in associations, while maintaining a conservative call rate of associations. Linear, GPLVM, SVA and PANAMA all are more generous in calling associations and are, therefore, not as clear cut as variational Bayesian GPLVM. The genomic control (GC) measures departure of the distribution from a uniform distribution. GC > 1 indicates inflation, where GC < 1 indicates deflation. **(b)**: *cis* and *trans* association counts as a function of the false discovery rate for each method. Variational Bayesian GPLVM prevents *trans* associations to be called, while maintaining a high *cis* calling rate.

proximity of it's gene by $\pm 500\,\mathrm{bp}$, we call the association a *cis* association. If it is farther away, the association is called *trans*. This follows the approach of Smith and Kruglyak [30]). Figure 3.4 shows the *cis* and *trans* associations as a function of the false discovery rate. Variational Bayesian GPLVM calls less associations than PANAMA, but the non-called associations appear to be false positives considering the performance shown in the ROC plot in Figure 3.3(a). Variational Bayesian GPLVM is conservative in calling *trans* associations, depleting associations caused by global confounding variation in gene expression. Taking confounders into account leads to overall decreased *trans* and increased *cis* association counts (compare confounder-less association "Linear"). This behaviour is exactly what is expected, as the confounding variation is assumed to be a global effect, creating false *trans* associations. As a overall theme, variational Bayesian GPLVM seems to cause less false positives overall.

This is not the only strength of a Bayesian Approach. Another strength lies within the ability to cope with uncertainty in confounding variation. Introducing uncer-

tainty into data should give an insight into how well it is able to handle uncertainty.

3.1.5 Realistic Simulation on Fewer Genes

We saw that variational Bayesian GPLVM is able to estimate the confounders themselves, as well as the number of confounders needed to explain confounding variation in gene expression. Another strength is that it is able to incorporate uncertainty over the confounders exact behaviour. Thus, this method should still be able to estimate confounding variation, when much uncertainty is added to gene expression. We can introduce uncertainty into gene expression by reducing the amount of genes the method has to estimate confounding variation. Thus, we sub sample the realistic simulation gene expression matrix $\mathbf{Y}$ and extract $D = 20$ genes per run creating ≈ 275 gene expression matrices $\mathbf{Y}_{1 \leq i \leq 275}$; predict confounding variation on these sub samples and then validate on the back substituted and corrected for multiple testing association matrix $\mathbf{A}^q$. Models which do not incorporate uncertainty into confounder prediction are expected to have difficulties in finding globally introduced variation in such a setting. And indeed variational Bayesian GPLVM is still able to find confounding variation, while not explaining all variation in the phenotype (Figure 3.3(b)). The ROC curve shows, that GPLVM does best in this setting, but the simulation seems to not be designed to work on a low amount of genes, because GPLVM learns most of genetic signal as confounding variation and does not call any *cis* or *trans* with low FDR (Figure 3.5). We can see that variational Bayesian GPLVM is even in such a low dimensional setting able to detect *trans* associations in gene expression. It is very conservative in calling *trans* associations (Figure 3.5). As we can see in the ROC curve of variational Bayesian GPLVM (Figure 3.3(b)) it finds some confounding variation and performs better than without correcting for confounders (compare "Linear"). Panama is not able to find confounding variation in gene expression in such a low dimensional setting, but it does not worse than linear, which means it does not over fit like GPLVM.

3.1.6 Glucose Smith Yeast

Now let us see how variational Bayesian GPLVM does in a real world experiment. Variational Bayesian GPLVM finds $Q \sim 38$ confounders in gene expression (data not shown), which is almost the same as the number of confounders found in the realistic simulation (where it was $Q \sim 43$). Unfortunately, we do not have a ground truth for associations between genes and SNPs in yeast. For that reason, we supply p-value histogram and *cis*/*trans* calling of the methods to compare their overall

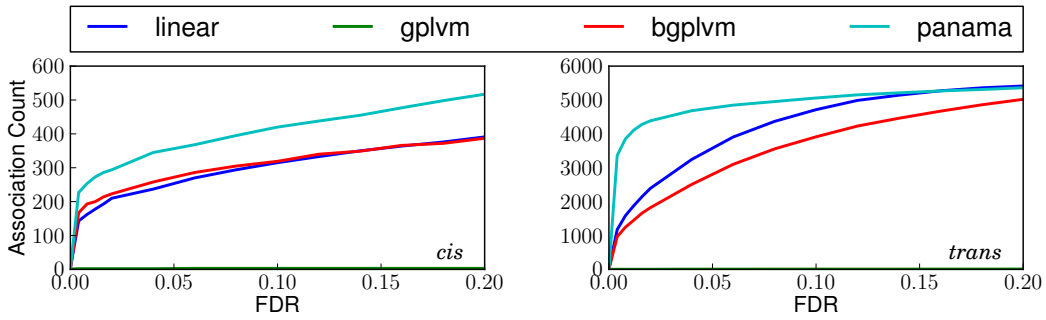


Figure 3.5: *cis* and *trans* association counts for fewer genes experiment on Smith GWAS experiment as a function of the false discovery rate for each method. The counts of GPLVM are zero for the shown FDRs, and therefore little reliable, because we know there are associations in the realistic simulation. Variational Bayesian GPLVM is most conservative in calling associations, but has less false positives overall (compare Figure 3.3(b)).

ability to extinguish *trans* associations, while maintaining *cis* associations. This is the preferred behaviour of a method which accounts for confounding variation, because the confounders are assumed to influence gene expression on a global scale. In Figure 3.6 we summarise the p-value histograms and association counts for *cis* and *trans*, respectively. Variational Bayesian GPLVM still gives the most clear signal in finding associations. GPLVM alone seems not to suit for detecting confounding variation in gene expression experiments. For this reason Fusi et al. [12] introduced genomic signal in an iterative approach to make up for explaining genetic signal in confounding variation.

The *cis* calling rate of variational Bayesian GPLVM is almost as high as PANAMA, but it deletes *trans* associations. SVA calls less *cis* associations and more *trans* associations. This indicates a high false positive rate in *trans* associations. GPLVM seems to account for *trans* associations, but as the p-value histogram indicates, it does not only fit confounding *trans* associations, but also explains genetic signal. In summary, variational Bayesian GPLVM has a high detection rate of confounding *trans* associations, without the need for introducing genetic signal in confounder prediction. SVA and PANAMA have a reasonable genomic control and reasonably high *cis* calling rates. Their *trans* calling rate is higher than the one of variational Bayesian GPLVM, which indicates that variational Bayesian GPLVM is more consistent in extracting true confounding variation, while maintaining genetic signal in gene expression.

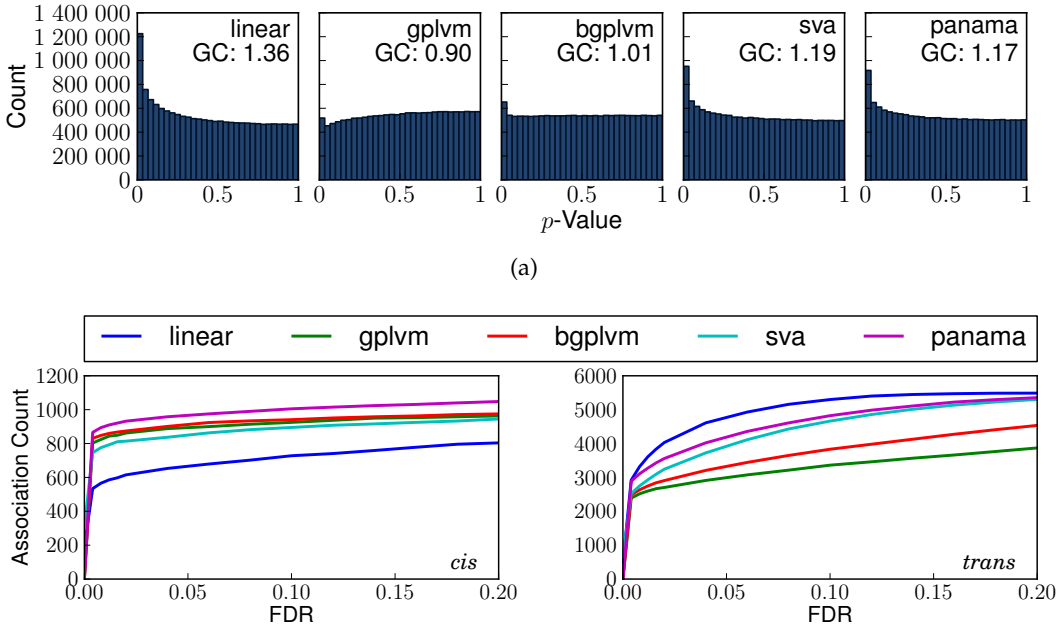


Figure 3.6: (a): p-value histograms for all five methods applied to the glucose part of the yeast experiment. Variational Bayesian GPLVM (bgplvm) still finds the most clear cut signal in associations, while maintaining a conservative call rate of associations. GPLVM does not seem to suit for confounder estimation, as the p-value distribution has a bump in middle. Linear, SVA and PANAMA all are still more generous in calling associations and are, therefore, not as clear cut as variational Bayesian GPLVM. **(b)**: *cis* and *trans* association counts as a function of the false discovery rate for each method. Variational Bayesian GPLVM prevents *trans* associations to be called, while maintaining a high *cis* calling rate.

3.2 Finding Shared versus Private Information between Experiments

In this section we apply manifold relevance determination to two genome wide association study experiments. Here, the latent space is seen as genotype, such that we learn the genetic variants them selves. We will show correlation between the genotype and latent variables corresponds to genotype control for each environment we test in. We will apply the method to two gene expression experiments. One of which is the yeast data set from Smith and Kruglyak [30] described in Section 3.1.6. The other GWAS experiment was done in another lab by [6] with the same yeast population and thus the same genotype as of Smith and Kruglyak [30]. The experiment of Brem et al. [6] consists of $N = 112$ samples over $D = 7{,}084$ genes with the same $2{,}956$ genotyped SNPs grown in a glucose environment. We first apply manifold relevance determination between the glucose part of Smith and

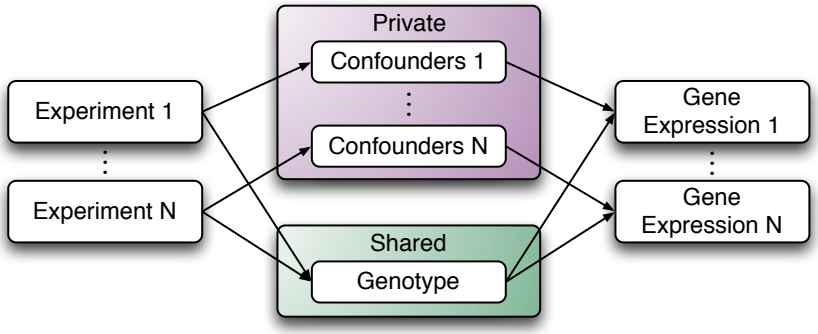


Figure 3.7: Idealised outcome of manifold relevance determination on eQTL studies. The genotype is known to be shared, and thus the shared factors found correspond to the genotype. Private factors, however, are introduced due to differing labs, time, experimenter etc and are therefore confounds. Note that this is idealised, and combinatorial effects can still happen in either space and introduce or extinguish (artificial) confounding variation.

Kruglyak [30] and the Brem et al. [6] GWAS experiment. We then determine the shared and private latent variation within the smith GWAS experiment (explained in Section 3.1.1) to determine the differences of latent variables when the environment changes.

3.2.1 Genomic Variation between Similar Gene Expression Experiments

Applying manifold relevance determination to similar experiments should give a insight in which confounding variation the different handling (different lab) and experimenters introduce to data. In a perfect world the outcome of such experiments should be the same, using the same samples. There is much confounding variation included, through, due to differing labs and experimenters. An ideal outcome of such an experiment would show that the shared space captures the biological driven input (genotype). And the private spaces should capture the specific confounders introduced due to the differing handling (Figure 3.7). The shared space should also be high dimensional, because the samples are the same and expression should be similar in the same environment, and the private spaces should show low dimensionality, because confounding variation should be little. We, however, found the shared space to be low dimensional, but the ARD weights for the shared space were high (Figure 3.8). This indicates a very high confounding variation due to the lab change they made [30]. The number of q-value corrected significant associations in the shared space (575) is lower than the number of significant hits in the private spaces (Brem= 601, Smith= 974). The fact that more significant associations are found in the Smith experiment indicates a cleaner experiment, because we can

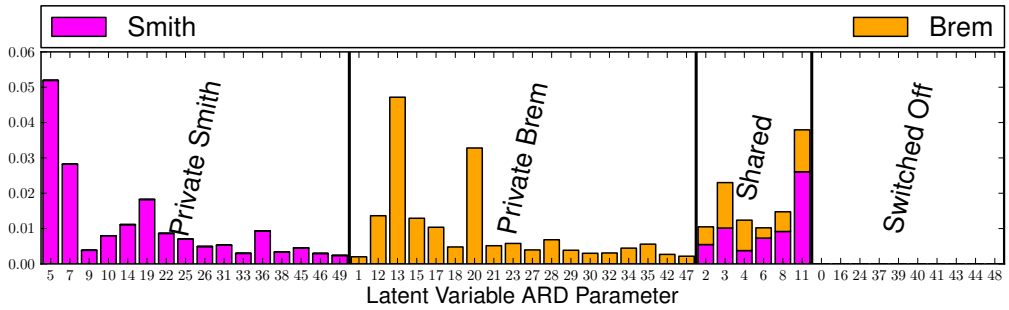


Figure 3.8: ARD weights for applying manifold relevance determination between Brem and Smith yeast grown on glucose as carbon source. The latent variables are sorted, such that they group up in their respective spaces (private or shared).

find more genotype phenotype associations here. The more convoluted Brem experiment, made 4 years earlier, leads to less hits and thus, the effects of SNPs onto gene expression are cancelled out by the confounder of changing the lab.

3.2.2 Private Spaces capture Condition Specific Information

We now want to capture information specific to the condition the organism is exposed to. We apply manifold relevance determination on both parts of the eQTL experiment of Smith and Kruglyak [30]. This gives an insight into differences of variation in latent dimensions of yeast between different environments, and shows which parts of the genotype corresponds to either processing of glucose or ethanol or both. Most shared latent variation between the experiments is expected to be biologically driven (As shown idealised in Figure 3.7). This means that the shared latent dimensions can be significantly correlated to the genotype (Figure 3.9). Private variation, however, is expected to be external confounding variation.
Thus, the shared latent variation can likely be assigned to biological reasoning in *trans* activity of the genotype. E.g. Cell cycle processes are known to be handled globally and act in a comprising way. All life cycle genes have a cyclic behaviour, which is likely captured by latent variable models. In fact, latent variable models like PCA have been applied to detect cell cycle genes [39]. In PCA, though, we would have to find heuristically how many dimensions are important and we would not know how sure we are of the latent spaces captured (Section 2.1).
We find correlations between latent variables of the shared space and SNPs, which are known to be associated to cell cycle processes (Table 3.2). Furthermore, growing yeast in ethanol does not hinder the cells from growing, and causes the cell cycle to shift, as a study by Kubota et al. [16] suggests. We can confirm finding correlation between the private space of ethanol and SNPs which regulate cell cycle and are in-

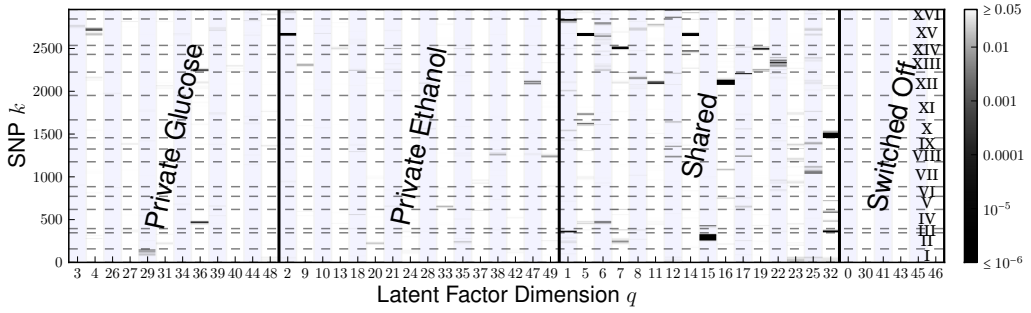


Figure 3.9: Correlations between SNPs and latent variables found by manifold relevance determination applied on Smith experiment introduced in section 3.1.1). The correlations are given in form of p-values. Note on the right the scale of p-values is below $p \leq 0.05$. Thus everything darker then plain white shown in the plot is a significant correlation and everything else has no significant correlation. You can see that the correlation between the genotype (SNPs) and the shared space is much higher than between genotype and private spaces.

volved in cell growth function (Table 3.2). The correlations between genotype and private latent variables of glucose as a carbon source are found to be not high and the correlated genes are involved in macromolecular and RNA processing pathways, which indicate a normal activity of the cell (Table 3.2).

Smith and Kruglyak [30] found a large hot spot of distant condition specific correlation (*trans* association) at chromosome XI (160 – $220\,\mathrm{cM}$). We find this locus correlated with the latent variable $q = 2$ of the private space in ethanol and with the latent variable $q = 4$ of the private space in glucose. The number of correlations are much higher in ethanol private space. Thus, we can confirm this locus being a condition specific distant association in ethanol, as Smith and Kruglyak [30] stated before. Additionally, we can assign different types of loci, e.g. chromosome XII shows a strong shared correlation within 600 – $800\,\mathrm{cM}$ (latent factors $q \in \{11, 16\}$), whereas that association is switched off completely in glucose (no latent factor) and heavily decreased in ethanol (latent factor $q = 47$). This means that most of this genomic variation is shared, but in ethanol there have to be some minor adjustments done.

Setting the threshold for a correlation to be significant to 0.05, we find that out of $14{,}729$ significant correlations $3{,}695/14{,}729$ ($25\,\%$) show an effect only in the ethanol condition, $3{,}201/14{,}729$ ($22\,\%$) show an effect only in the glucose condition and $7{,}833/14{,}729$ ($53\,\%$) show an effect in both conditions. These numbers are very similar to the results Smith and Kruglyak [30] found in their analysis of distant associations.

Table 3.2: Gene ontology terms found for significant hits (p-value $\leq 0.000{,}1$) in manifold relevance determination between yeast grown on glucose and ethanol as carbon source (Section 3.1.1). You can see that the shared space captures cell cycle and general regulatory processes; glucose private space captures metabolic and catabolic processes; and ethanol private space captures cell cycle regulatory processes, catalytic activity and hyphic growth.

Private Glucose	#Hits	GO
cellular macromolecule catabolic process	7	0044265
macromolecule catabolic process	7	0009057
ncRNA processing	6	0034470
ncRNA metabolic process	6	0034660
tRNA metabolic process	4	0006399
Private Ethanol	**#Hits**	**GO**
response to abiotic stimulus	6	0009628
regulation of cell cycle	5	0051726
pseudohyphal growth	4	0007124
cell growth	4	0016049
growth of unicellular organism as a thread of attached cells	4	0070783
regulation of cell cycle process	4	0010564
Shared	**#Hits**	**GO**
cell cycle	42	0007049
cell cycle process	34	0022402
tRNA metabolic process	14	0006399
cytokinesis	13	0000910
DNA replication	13	0006260
coenzyme metabolic process	13	0006732

Chapter 4

Discussion and Future Work

The variational Bayesian GPLVM has demonstrated its potential for accounting for confounding variation in gene expression experiments, without including the genotype into prediction. It finds the dimensionality, as well as the confounding variation (3.1.3) itself without having to introducing the genotype into prediction (3.1.2).
Variational Bayesian GPLVM finds associations in a simulated genotype-phenotype association experiment as well as PANAMA does (Section 3.1.4), while calling less *trans* associations. This is a clear indication on a lower false positive rate for this model compared to PANAMA. Additionally, the elimination of *trans* associations is more closely related to biological reasoning, as *trans* associations are likely to be explained by combinatorial effects, instead of real correlation.
Compared to SVA, variational Bayesian GPLVM seems to strictly be more stable in finding confounding variation. It calls more *cis* associations, while eliminating *trans* associations, which are consistent throughout the gene expression matrix. Thus, the assumption of confounding variation being global and effecting the whole gene expression matrix is highly probable. One should continue with the approach of taking uncertainty over the latent variables of interest into account, because with that the method even copes with little input dimensions, where other methods fail to compute (Section 3.1.5).

Introduce Genotype into Confounder Learning In order to further improve variational Bayesian GPLVM, we think of introducing the genotype into prediction. This would mean to introduce the genotype $\mathbf{S}$ into the covariance structure of the

model, such that

$$p(\mathbf{Y}|\mathbf{X}) = \prod \mathcal{N}(\mathbf{Y}_d|\mathbf{0}, \mathbf{K}_{NN} + \mathbf{K}_{\mathbf{S}} + \beta^{-1}\mathbf{I}) \ , \tag{4.1}$$

where $\mathbf{K}_{NN}$ is the covariance matrix evaluated at the confounding factors $\mathbf{X}$ and $\mathbf{K}_{\mathbf{S}}$ is the covariance matrix evaluated at the genotype $\mathbf{S}$. Note that the likelihood of the model after applying sparse GP includes the sparse approximation $\mathbf{K}_{MM}$, such that the likelihood of the model becomes

$$p(\mathbf{Y}|\mathbf{X}) = \prod \mathcal{N}(\mathbf{Y}_d|\mathbf{0}, \mathbf{K}_{NM}\mathbf{K}_{MM}^{-1}\mathbf{K}_{MN} + \mathbf{K}_{\mathbf{S}} + \beta^{-1}\mathbf{I}) \ . \tag{4.2}$$

This likelihood of the model would capture both effects, the effect of the confounders and the effect of the genotype onto gene expression. Thus, there would be no need for a linear mixed model to be applied to find associations and associations would have a probability assigned, as well as a lower number of input dimensions should not effect prediction, because the model could include genetic information to find confounding variation.
Having fit the confounders with this model, one could even turn around the dependency and predict the genotype changes in different confounding conditions.
By introducing a ARD covariance matrix $\mathbf{K}_S = \mathbf{SAS}^\top$, with ARD parameters $\alpha^{\mathbf{S}}$ in the diagonal matrix $\mathbf{A}$, we are able to learn the association strengths of individual SNPs onto respective gene expressions. This gives the opportunity to assign strong correlated SNPs to a subset of test SNPs, which can act as representers of all SNPs to reduce computational costs of subsequent tests and analysis.

Manifold Relevance Determination in GWAS Applying manifold relevance determination on genome wide association study experiments gives deep insights into which processes are needed for the one or the other (or even more, when applied to a higher amount of gene expression experiments) circumstance. We were able to determine which processes are exclusively involved in either of growing yeast on ethanol or glucose (Section 3.2.2) and even infer from which genes are involved in the one environment to be not necessary in the other environment. The next step is to apply this methodology to more then two experiments, to be able to infer not only private and shared (of all) latent processes, but also infer which processes are shared between two (or more) of the processes and are private for others. This approach gives rise to a whole new way of looking at genome wide association studies, to not only infer associations but also figure where associations belong. Additionally, this gives the opportunity to include different kinds of data

into latent variable learning, such as adding other phenotypes and search for commonalities between gene expression, SNPs and "visible" genotypes, such as disease and others.
The next step of applying manifold relevance determination is to include the genotype into learning of confounders and then use the learnt private latent spaces to account for confounding variation in association studies, as the private spaces encode for the specific environmental conditions and will explain away less genetic signal overall. Or include the genotype into determining shared and private spaces, to see what non-genetic signals are in the data and filter environmental latent variables to predict behaviour of such spaces for studies with different organisms, which are not as well studied as yeast is.

Appendix A

Mathematical Appendix

This appendix shows more detailed views on the most significant mathematical steps I got through during the work on the thesis.

A.1 Derivatives and technical details of Bayesian GPLVM

To be able to compute log marginal likelihood and derivatives of it we need to compute ψ statistics, because latent variables are given as distributions.

A.1.1 Squared exponential Ψ statistics

Computation of kernel:

$$k(x_q, x'_q) = \sigma_f^2 \cdot \exp\left\{-\frac{\alpha_q}{2}\left\|x_q - x'_q\right\|^2\right\} \tag{A.1}$$

Computation of Ψ statistics:

$$
\begin{aligned}
\psi_0 &= N\sigma_f^2 \\
(\Psi_1)_{nm} &= \sigma_f^2 \prod_{q=1}^{Q} \frac{e^{-\frac{1}{2}\frac{\alpha_q(\mu_{nq}-z_{mq})^2}{\alpha_q S_{nq}+1}}}{(\alpha_q S_{nq}+1)^{\frac{1}{2}}} \\
&= \sigma^2 \exp\left\{\sum_{q=1}^{Q} -\frac{1}{2}\left(\frac{\alpha_q(\mu_{nq}-z_{mq})^2}{\alpha_q S_{nq}+1} + \log(\alpha_q S_{nq}+1)\right)\right\} \\
(\Psi_2^n)_{mm'} &= \sigma_f^4 \prod_{q=1}^{Q} \frac{e^{-\frac{\alpha_q(z_{mq}-z_{m'q})^2}{4} - \frac{\alpha_q(\mu_{nq}-\bar{z}_q)^2}{2\alpha_q S_{nq}+1}}}{(2\alpha_q S_{nq}+1)^{\frac{1}{2}}} \\
&= \sigma_f^4 \exp\left\{\sum_{q=1}^{Q} -\frac{1}{2}\left(\alpha_q\left(\frac{(z_{mq}-z_{m'q})^2}{2} + \frac{(\mu_{nq}-\bar{z}_q)^2}{\alpha_q S_{nq}+\frac{1}{2}}\right) + \log(2\alpha_q S_{nq}+1)\right)\right\}
\end{aligned}
$$

For clarity the partial derivative is zero when not stated differently.
All other derivatives w.r.t hyperparameters σ_f^2, α_q and variational parameters $\mathbf{z}, S, \boldsymbol{\mu}$:

Amplitude parameter σ_f^2

$$
\begin{aligned}
\frac{\partial \psi_0}{\partial \sigma_f^2} &= N \\
\frac{\partial (\Psi_1)_{nm}}{\partial \sigma_f^2} &= \exp\left\{\sum_{q=1}^{Q} -\frac{1}{2}\left(\frac{\alpha_q(\mu_{nq}-z_{mq})^2}{\alpha_q S_{nq}+1} + \log(\alpha_q S_{nq}+1)\right)\right\} \\
\frac{\partial (\Psi_2^n)_{mm'}}{\partial \sigma_f^2} &= 2\sigma_f^2 \exp\left\{\sum_{q=1}^{Q} -\frac{1}{2}\left(\alpha_q\left(\frac{(z_{mq}-z_{m'q})^2}{2} + \frac{(\mu_{nq}-\bar{z}_q)^2}{\alpha_q S_{nq}+\frac{1}{2}}\right) + \log(2\alpha_q S_{nq}+1)\right)\right\}
\end{aligned}
$$

ARD parameters α_q

$$
\begin{aligned}
\frac{\partial (\Psi_1)_{nm}}{\partial \alpha_q} &= (\Psi_1)_{nm} \cdot -\frac{1}{2}\left(\frac{(\mu_{nq}-z_{mq})^2}{(\alpha_q S_{nq}+1)^2} + \frac{S_{nq}}{\alpha_q S_{nq}+1}\right) \\
\frac{\partial (\Psi_2^n)_{mm'}}{\partial \alpha_q} &= (\Psi_2^n)_{mm'} \cdot -\frac{1}{2}\left(\frac{1}{2}\frac{(\mu_{nq}-\bar{z}_q)^2}{(\alpha_q S_{nq}+\frac{1}{2})^2} + \frac{(z_m - z_{m'})^2}{2} + \frac{2S_{nq}}{2\alpha_q S_{nq}+1}\right)
\end{aligned}
$$

Latent Variables z_{mq}

$$\frac{\partial(\Psi_1)_{nm}}{\partial z_{mq}} = (\Psi_1)_{nm} \cdot -\frac{1}{2}\frac{\alpha_q(\mu_{nq} - z_{mq})}{\alpha_q S_{nq} + 1}$$
$$\frac{\partial(\Psi_2^n)_{mm'}}{\partial z_{mq}} = (\Psi_2^n)_{mm'} \cdot -\frac{\alpha_q}{2}\left((z_m - z_{m'}) + \frac{(\mu_{nq} - \bar{z}_q)}{\alpha_q S_{nq} + \frac{1}{2}}\right)$$

Variational Means $\boldsymbol{\mu}_{nq}$

$$\frac{\partial(\Psi_1)_{nm}}{\partial \mu_{nq}} = (\Psi_1)_{nm} \cdot -\frac{\alpha_q(\mu_{nq} - z_{mq})}{\alpha_q S_{nq} + 1}$$
$$\frac{\partial(\Psi_2^n)_{mm'}}{\partial \mu_{nq}} = (\Psi_2^n)_{mm'} \cdot -\frac{\alpha_q(\mu_{nq} - \bar{z}_q)}{\alpha_q S_{nq} + \frac{1}{2}}$$

Variational Variances S_{nq}

$$\frac{\partial(\Psi_1)_{nm}}{\partial S_{nq}} = (\Psi_1)_{nm} \cdot \frac{1}{2}\left(\frac{\alpha_q^2(\mu_{nq} - z_{mq})^2}{(\alpha_q S_{nq} + 1)^2} - \frac{\alpha_q}{\alpha_q S_{nq} + 1}\right)$$
$$\frac{\partial(\Psi_2^n)_{mm'}}{\partial S_{nq}} = (\Psi_2^n)_{mm'} \cdot \frac{1}{2}\left(\frac{\alpha_q^2(\mu_{nq} - z_{mq})^2}{(\alpha_q S_{nq} + \frac{1}{2})^2} - \frac{\alpha_q}{2\alpha_q S_{nq} + 1}\right)$$

A.1.2 Linear Ψ statistics

Kernel computation:

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T A \mathbf{x}' \tag{A.2}$$

Ψ statistic computation:

$$\psi_0 = \mathrm{tr}[A(\boldsymbol{\mu}_n \boldsymbol{\mu}_n^T + S_n)] \tag{A.3}$$

$$(\Psi_1)_{nm} = \boldsymbol{\mu}_n^T A \mathbf{z}_m \tag{A.4}$$

$$\sum_n (\Psi_2^n)_{mm'} = \sum_n \mathbf{z}_m^T A(\boldsymbol{\mu}_n \boldsymbol{\mu}_n^T + S_n) A \mathbf{z}_{m'} \tag{A.5}$$
$$= ZA(U^T U + (S^T \mathbf{1})\mathbf{I})AZ^T \tag{A.6}$$

Where inducing inputs, means and variances held as matrices $Z = \left(\mathbf{z}_m^T\right)_{m=1}^M \in \mathbb{R}^{M \times Q}, U = \left(\boldsymbol{\mu}_n^T\right)_{n=1}^N \in \mathbb{R}^{N \times Q}, S = \left(\mathrm{diag}(S_n)^T\right)_{n=1}^N \in \mathbb{R}^{N \times Q}$. And $S^T \mathbf{1} = \sum_{n=1}^N S_{n\cdot}$.

Derivatives w.r.t parameters:

$$\frac{\partial \Psi_1}{\partial A_q} = UZ^T \tag{A.7}$$

$$\begin{aligned}\frac{\partial \Psi_2}{\partial A_q} &= Z(U^TU + (S^T\mathbf{1})\mathbf{I})A_qZ^T + ZA_q(U^TU + (S^T\mathbf{1})\mathbf{I})Z^T \\ &= 2ZA_q(U^TU + (S^T\mathbf{1})\mathbf{I})Z^T \quad \text{, bc. A is diagonal.}\end{aligned} \tag{A.8}$$

$$\frac{\partial \Psi_1}{\partial Z_{mq}} = UA\,\mathbf{J}^{qm} \tag{A.9}$$

$$\frac{\partial \Psi_2}{\partial Z_{mq}} = \mathbf{J}^{mq}A(U^TU + (S^T\mathbf{1})\mathbf{I})AZ^T \tag{A.10}$$

Appendix B

Software Appendix

B.1 Python

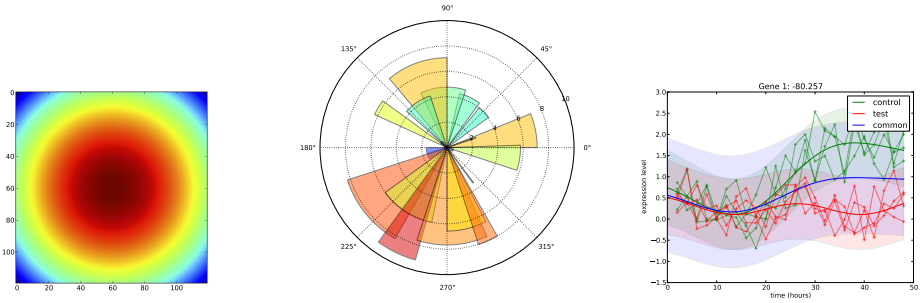
This thesis was implemented using the python script language. For employing the methods one has to make sure the *python2.7* package is installed on the corresponding machine. A quick how-to can be found on the python homepage `http://python.org/`, as well as the documentation, some libraries and more.
To use python appropriate to mathematical and plotting issues, we want to recommend the following libraries for the python language, as they are essential for using the methods introduced in this thesis.

B.1.1 SciPy

"SciPy (pronounced "Sigh Pie") is open-source software for mathematics, science, and engineering. It is also the name of a very popular conference on scientific programming with Python. The SciPy library depends on NumPy, which provides convenient and fast N-dimensional array manipulation. The SciPy library is built to work with NumPy arrays, and provides many user-friendly and efficient numerical routines such as routines for numerical integration and optimization. Together, they run on all popular operating systems, are quick to install, and are free of charge. NumPy and SciPy are easy to use, but powerful enough to be depended upon by some of the world's leading scientists and engineers. If you need to manipulate numbers on a computer and display or publish the results, give SciPy a try!"
For more information see their homepage `http://www.scipy.org/`.

B.1.2 Matplotlib

"matplotlib is a python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. matplotlib can be used in python scripts, the python and ipython shell (ala MATLAB® or Mathematica®†), web application servers, and six graphical user interface toolkits.*
matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc, with just a few lines of code. For a sampling, see the screenshots, thumbnail gallery, and examples directory"
Some example visualizations of the plotting library matplotlib are shown in figure B.1. See the whole example library, documentation, downloads and more on their



(a) Colormap, showing a gaussian distribution

(b) Polar map example

(c) Example of drawing a time series

Figure B.1: Some examples drawn from matplotlib library

homepage `http://matplotlib.sourceforge.net/`.

B.2 Implementation

You can download the scripts I wrote during this thesis from my website `http://people.kyb.tuebingen.mpg.de/maxz/`. The code has strong dependencies on the following packages:

- The pygp packages, which can be found at `http://pypi.python.org/pypi/pygp`.
- The PANAMA package, ready for download at `http://ml.sheffield.ac.uk/qtl/panama/`

References

[1] S. Ahnert, T. Fink, and A. Zinovyev. How much non-coding DNA do eukaryotes require? *Journal of Theoretical Biology*, 252(4):587–592, 2008.

[2] D. Altshuler, J. Hirschhorn, M. Klannemark, C. Lindgren, M. Vohl, J. Nemesh, C. Lane, S. Schaffner, S. Bolk, C. Brewer, et al. The common pparγ pro12ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature genetics*, 26(1):76–80, 2000.

[3] M. R. Barnes and I. C. Gray. *Bioinformatics for Geneticists*. John Wiley & Sons, Ltd., 2003.

[4] C. Bishop. Bayesian PCA. *Advances in Neural Information Processing Systems*, pages 382–388, 1999.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.

[6] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–5, Apr 2002. doi: 10.1126/science.1069516.

[7] A. Corvin, N. Craddock, and P. F. Sullivan. Genome-wide association studies: a primer. *Psychological Medicine*, 40(7):1063–77, Jul 2010. doi: 10.1017/S0033291709991723.

[8] A. Damianou, M. Titsias, and N. Lawrence. Variational Gaussian Process Dynamical Systems. *Neural Information Processing System (NIPS)*, 2011.

[9] A. C. Damianou, C. H. Ek, M. K. Titsias, and N. D. Lawrence. Manifold relevance determination. *ICML*, 2012.

[10] K. Davies. The future of next-gen sequencing (NGS). *Bio*IT World*, 2011.

[11] R. Fletcher. *Practical methods of optimization, Volume 1*. Wiley, 1987.

[12] N. Fusi, O. Stegle, and N. D. Lawrence. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput Biol*, 8(1):e1002330, Jan 2012. doi: 10.1371/journal.pcbi.1002330.

[13] Y. Gilad, S. A. Rifkin, and J. K. Pritchard. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet*, 24(8):408–15, Aug 2008. doi: 10.1016/j.tig.2008.06.001.

[14] W. Hennig. *Genetik*. Springer Berlin, 1998.

[15] J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6(2):95–108, Feb 2005. doi: 10.1038/nrg1521.

[16] S. Kubota, I. Takeo, K. Kume, M. Kanai, A. Shitamukai, M. Mizunuma, T. Miyakawa, H. Shimoi, H. Iefuji, and D. Hirata. Effect of ethanol on cell growth of budding yeast: genes that are important for cell growth in the presence of ethanol. *Bioscience, biotechnology, and biochemistry*, 68(4):968–972, 2004.

[17] E. Lander, L. Linton, B. Birren, C. Nusbaum, M. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[18] N. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16:329–336, 2004.

[19] N. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, November 2005.

[20] J. Leek. *Surrogate variable analysis*. PhD thesis, University of Washington, 2008.

[21] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9), September 2007.

[22] G. Mendel. *Versuche über Pflanzen-Hybriden (1865)*. Arkana-Verl., 1865.

[23] T. Mercer, M. Dinger, and J. Mattick. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*, 10(3):155–159, 2009.

[24] J. J. Michaelson, S. Loguercio, and A. Beyer. Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*, 48(3):265–76, Jul 2009. doi: 10.1016/j.ymeth.2009.03.004.

[25] S. Pounds. Estimation and control of multiple testing error rates for microarray studies. *Briefings in bioinformatics*, 7(1):25–36, 2006.

[26] C. E. Rasmussen. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[27] E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs, and A. J. Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*, 37(7):710–7, Jul 2005. doi: 10.1038/ng1589.

[28] W. Seyffert and R. Balling. *Lehrbuch der Genetik*. Spektrum Akademischer Verlag, 2003.

[29] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nat Biotechnol*, 26 (10):1135–45, Oct 2008. doi: 10.1038/nbt1486.

[30] E. N. Smith and L. Kruglyak. Gene-environment interaction in yeast gene expression. *PLoS Biol*, 6(4):e83, Apr 2008. doi: 10.1371/journal.pbio.0060083.

[31] O. Stegle, L. Parts, R. Durbin, and J. Winn. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol*, 6(5):e1000770, May 2010. doi: 10.1371/journal.pcbi.1000770.

[32] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *PNAS*, 100(16):9440–9445, August 2003.

[33] M. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[34] M. E. Tipping and C. C. Nh. Sparse Kernel Principal Component Analysis, 2001.

[35] M. Titsias. Variational learning of inducing variables in sparse gaussian processes. *Artificial Intelligence and Statistics*, 5:567–574, April 2009.

[36] M. K. Titsias and N. D. Lawrence. Bayesian Gaussian process latent variable model. *Artificial Intelligence and Statistics*, 2010.

[37] U.S. National Library of Medicine. What is DNA? Internet, September 2012. URL `http://ghr.nlm.nih.gov/handbook/basics/dna`.

[38] L. Wodicka, H. Dong, M. Mittmann, M.-H. Ho, and D. J. Lockhart. Genome-wide expression monitoring in saccharomyces cerevisiea. *Nature*, 1997.

[39] K. Yeung and W. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[40] A. Zien. *A Primer on Molecular Biology*, chapter 1, pages 3–34. The MIT Press, 2004.