Multi-Task Covariate Selection for Genome-Wide Gene Expression Data

Master's Thesis

Johan J. Lugthart

Supervisors:
Dr. T.A.L. van Erven and Dr. M.J. Jonker

January 19, 2017

Mathematical Institute
Leiden University

Abstract

We study covariate selection for genome-wide gene expression data, by comparing three methods, one of them introduced in this thesis. All methods apply the multi-task principle to the regression of the different genes.
The first method detects which covariates are important for all the genes and uses only these covariates for the regression. The second uses Akaike Information Criterion for each gene independently. The third and new method introduces a general trend over all the genes. This trend describes the averaged influence of each covariates on the different genes. In its original formulation this method is computationally intractable, but we derive a reduction to a LASSO problem, which van be solved efficiently.
We report experiments on simulated and real data sets which demonstrate that the new method gives the best results if there is a general trend over all the genes, where one of the other methods is preferable if there is not a trend or if there are several trends. If there are several trends the new method orders the covariates according to the mean influence, hereby different trends can cancel each other out. The first two methods on the other hand are insensitive to different trends. We analyse gene expression data of human embryos and find that there are five groups of genes with their own general trend.

Contents

Chapter 1

Introduction

1.1 Biological Relevance

The field of molecular biology aims at understanding the functioning of cells and multicellular organisms at the molecular level [AJL+14]. An important area of molecular biology studies how genetic information, stored in Deoxyribonucleic acid (DNA), is transformed into working functional biological systems such as cells. A gene is a region of DNA and is the molecular unit of heredity. The genetic information stored in the DNA is vital for organisms, because virtually all biological molecular processes are driven by genes. Humans have approximately 20,000 genes, which are together called the genome [EJR+14]. The genetic information in the genome is converted into cellular processes through a process that is called "gene expression" [AJL+14]. Briefly, genes are transcribed into messenger RNA (Ribonucleic acid), and the messenger RNA is translated into proteins. This is not the only way genes can be expressed, but transcription and translation are important processes, as proteins are often responsible for the characteristics and functions of a cell by regulating communication, production and breaking-down processes. The DNA is the in principle the same in all the individual cells of an organism. The expression of the genes and correspondingly the presence of the proteins explains the enormous amount of several functions of cells. Hence, the expression of genes is frequently studied in biomedical research, to investigate which genes are involved in diseases, or which genes are potential drug targets. Understanding the expression of the genes, and understanding whether and how internal and environmental factors are related to this expression is therefore very important. Examples of these factors are the age, the sex, and the size of the organism and the temperature or oxygen concentration of the environment. Research on the relationship between a factor and the expression of a gene is complex, therefore it is important to decide which factors are worth the effort of researching.

The expression of genes is measured with a microarray analysis or RNA-sequencing [Bum13, WGS09]. These technologies quantify the amount of messenger RNAs in a cell, or in a pool of cells. This results in a dataset with two sets of information for each sample: the values of the covariates and the level of expression of every gene. This is shown systematically in figure 1.1 on the following page.

An analysis of the relation between one covariate and one specific gene cannot be done

	Organism 1	Organism 2	$\cdots$	Organism n
Covariate 1	X_1^1	X_2^1	$\cdots$	X_n^1
Covariate 2	X_1^2	X_2^2	$\cdots$	X_n^2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Covariate P	X_1^P	X_2^P	$\cdots$	X_n^P
Expression of gene 1	Y_1^1	Y_1^2	$\cdots$	Y_1^n
Expression of gene 2	Y_2^1	Y_2^2	$\cdots$	Y_2^n
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Expression of gene G	Y_G^1	Y_G^2	$\cdots$	Y_G^n

Figure 1.1: The structure of the data. For each organism, we know the value of the covariates and the expression of the genes. This information is stored in the matrices X and Y respectively.

reliably, because of the limited set of samples in biological experiments, since a typical dataset has around thirty covariates and around the same number of samples. Instead of analyzing the change in the expression of one gene at a time, we will use the data quantifying the expression of multiple genes. We analyse these data with a statistical approach known as multi-task learning, first introduced by James and Stein [JS61]. In this case the separate regression problems of the different genes are the tasks. In multi-task learning the analysis of one specific dataset is improved by 'learning' from the other datasets.

Understanding the expression of the genes, and understanding whether and how internal and environmental factors are related to this expression is therefore very important. Examples of these factors are the age, the sex, and the size of the organism and the temperature or oxygen concentration of the environment. Research on the relationship between a factor and the expression of a gene is complex, therefore it is important to decide which factors are worth the effort of researching.

1.2 Aim of the Research

Analysing data sets by multi-task learning to select the most important factors can be done with several methods. The various procedures differ mainly in the way they use the information gained by the analysis of the other tasks. In this research we searched for a good method to analyse gene expression data. We compared two existing methods and introduced a new one. In this thesis we will give a survey of the performance of these methods on simulated data and on real gene expression data.

The first method is introduced by Obozinski et al. in [OTJ06] and is called 'Multi-Task Feature Selection' (MTFS). This method is developed to find the covariates which are important for all the tasks, they analysed for example handwriting and the classification of cancer types. It learns which covariates are important and estimates the coefficients of

these important covariates. More information over this method and the other methods is given in chapter 2.
The second method, which we call 'Akaike Information Criterion per Gene' (AICpG), is developed by M.J. Jonker in [MJW+16]. This method applies Akaike Information Criterion on the different tasks independently and generates a priority list by counting for how many tasks the covariates are included. This method is already used, but has not yet been extensively studied.
The new method that dr. T.A.L. van Erven and I introduce in section 2.4 is named 'General Trend Method' (GTM). It differs from the former two methods by using a general trend in the size of the coefficients of the covariates over all the tasks. This trend creates the possibility to make a better estimation of these coefficients if most of the genes are influenced in a similar way.

1.3 Related Work

There are other methods that are related to the previously-mentioned methods, but do not solve this exact problem. We will discuss two of them.

1.3.1 James-Stein Estimation

One classic example of Multi-Task Learning is Stein's Estimation, [S+56], or the improved James-Stein estimation, [JS61]. The James-Stein estimation improves the estimation of several means (in particular the mean of a multi-variate normal) by applying a shrinkage toward the mean of the means. This increases the bias of the estimator, but reduces its variance. This shrinkage works best if the different dimensions of the multi-variate are somehow related. The James-Stein estimation uses information from all other dimensions to get a better estimation. In this thesis we will use the information of all the genes together to get an estimation of the coefficients of the regression problem with a smaller variance. The general trend method will do this by shrinking the coefficients toward some kind of average. This is done for all the covariates separately.

1.3.2 Multi-Task Ridge

Evgeniou, Pontil and Toubia introduce a method in [EPT07]. They want to estimate the coefficients of several linear regression problems. They introduce a general trend and apply a shrinkage toward this vector of coefficients. For each covariate, the amount of shrinkage of their method is related to the variance of the estimated coefficients of this covariate; it shrinks a coefficient more if it varies a relatively large amount with respect to the other coefficients corresponding to this covariate. In contrast to the methods in this thesis, it does not have a selecting feature to select which covariates are important.

1.4 Outline of the Thesis

To discover a good method to analyse gene expression data we compare three statistical analysis methods. These three methods are introduced and analysed theoretically in chapter 2. In particular, an reduction of GTM to a LASSO problem is derived. With this reduction GTM can be implemented efficiently. Also, the underlying assumptions of the methods are described. In chapter 3 simulated data are used to investigate the performance of the models on three different types of data. With a single trend has the General Trend Method the most consistent ordering of the covariates according to importance. If there is an opposite effect of a covariate on different genes GTM finds this covariate unimportant, in contrast to MTFS and AICpG, where different trends can not cancel each other out. Chapter 4 applies the three models to a biological dataset of the gene expression of 37 human embryos. We compare the output of the models, analyse the stability of the outcomes, and show that there are five groups of genes, which are influenced differently by the covariates. A conclusion and discussion are given in chapter 5.

Chapter 2

Model and Methods

In this chapter we will describe the three methods and compare them theoretically. Before that we will introduce some notation and formalise the regression problem.

2.1 Model

The data that are analysed in this thesis are the gene expressions of a lot of genes of several organisms. For each organism we have some biological and clinical data, for example their age and method of treatment, and the activity of (part of) their genes. The expressions of the genes are the dependent variables, depending on the biological and clinical covariates. We model the expression of gene g as a linear function of the covariates. We call the number of organisms n and the gene expressions of gene g for all the organisms together Y_g. The values of the covariates form the design matrix X. The regression coefficients are written in β_g. So together, for the model of the expression of gene g, we have:

$$Y_g = X\beta_g + \text{noise}. \tag{2.1}$$

With a total number of G genes, this results in G linear regression problems. The number of covariates is called P.

Solving these regression problems independently results in unreliable results, because the number of organisms is small, and similar to the number of covariates. However, the regression problems for all the different genes are very similar. They are all about gene expression, and all the design matrices are exactly the same, because these are different genes from the same organism. So it is likely that a multi-task approach where the different regression problems can learn from each other will give better results than if we do all the regression problems independently from each other. Two methods in this thesis use a multi-task approach for the estimation of the coefficients of β_p; one method only orders the covariates according to importance.

To shorten the writing, we introduce some extra notation. We merge the gene expression vectors Y_g to a $G \times n$ gene expression matrix Y, where every row corresponds to the expression levels of one gene for all the organisms. All the regression coefficient vectors

β_g are merged into a $G \times P$ coefficients matrix B, with rows which are the coefficients of one gene. The columns of this coefficients matrix we call β^p, they are the coefficients of one covariate for all the genes. The p^{th} coefficient of gene g is β_g^p.

All three methods use an objective function to decide which coefficients matrix is the best fit. This objective function contains two parts, the squared error between the gene expression and the predicted value and a penalty term that differs between the three methods:

$$\hat{B} = \arg\min_B \left\{ \sum_{g=1}^{G} \|Y_g - X\beta_g\|_2^2 + \text{Pen}(B) \right\}. \tag{2.2}$$

2.2 Existing Method: MTFS

2.2.1 Definition

Multi-task Feature Selection (MTFS) is introduced by Obozinski et al. in [OTJ06]. This method is a multi-task covariates selection method. The idea of the method is that there are multiple related regression problems with the assumption that the different regression problems share a subset of relevant covariates from a large common set of covariates.

Obozinski et al. introduce an extended ℓ_1-norm for all the regression problems together. It is a norm for the $G \times P$ dimensional coefficients matrix defined as

$$\|B\|_{1,2} := \sum_{p=1}^{P} \sqrt{\sum_{g=1}^{G} (\beta_g^p)^2}, \tag{2.3}$$

$$= \sum_{p=1}^{P} \|\beta^p\|_2, \tag{2.4}$$

$$= \left\| \left(\|\beta^1\|_2, \cdots, \|\beta^P\|_2 \right) \right\|_1. \tag{2.5}$$

This norm is used in the penalty term and multiplied with a positive penalty parameter λ. This gives the following minimisation problem:

$$\hat{B} = \arg\min_B \sum_{g=1}^{G} \|Y_g - X\beta_g\|_2^2 + \lambda \|B\|_{1,2}. \tag{2.6}$$

They also give an algorithm to solve minimisation problem (2.6) for a path of decreasing penalty parameter λ.

2.2.2 Properties

The usage of the ℓ_1-norm in the 'Least Absolute Shrinkage and Selection Operator' (LASSO, [Tib96]) leads to sparse solutions. In this more general setting, the ℓ_1-norm of the ℓ_2-norms leads to a solution where $\hat{B}$ has columns with only zeros. In other words, for several covariates the coefficient is zero for all the genes. For the other covariates the ℓ_2-norms lead to coefficients of $\hat{B}$ around zero, but almost never zero. These coefficients do not have to be close to each other.

When the penalty parameter λ decreases from infinity to zero, the columns of $\hat{B}$ become non-zero one after the other. This gives an ordering for the importance of the covariates.

2.3 Applied Method: AICpG

2.3.1 Introduction and Definition

Akaike Information Criterion per Gene (AICpG) is the second method we will use for the comparison. The method is introduced by M.J. Jonker and used to analyse the gene expression data in [MJW$^+$16], but its properties have not yet been extensively studied. This method is based on the well-known Akaike Information Criterion (AIC) method introduced by H. Akaike in [Aka74]. AIC is a way to select between different models, by taking the goodness of fit and the complexity of the model into consideration. If the different models are linear regression models with all the different subsets of covariates, AIC chooses a model with only the important covariates. In this way AIC can select covariates.

In our case we can see the regression on the data as being different regression problems for all the genes. For all these G regression problems we have n samples and P covariates. AICpG is a method where AIC is applied to all these regression problems independently. For all the genes this leads to a linear fit where a subset of included covariates is chosen. From all these G different covariates selections, AICpG creates an ordering of the importance of covariates by counting for how many genes each covariate is selected. The number of times a covariate is selected, is a measure of the importance of a covariate. So, by analogy, ordering the covariates according to importance is like democratic voting, where each gene 'votes' on which covariates are important. Ordering the covariates by the number of votes gives the ordering of the covariates according to importance.

2.3.2 Properties

By approaching the regression problem as being G independent problems, AICpG does not learn from other genes for the estimation of the coefficients of B, so the estimation $\hat{B}$

is only as good as AIC can estimate all the β_g. In the setting of this research, AIC cannot do this estimation good, because the sample size is small in comparison with the number of covariates. A second issue is that AIC is based on an asymptotic approximation, and for a small sample size is this approximation not reliable. This leads to a large probability to overfit, [CH$^+$08] and it is likely that $\hat{B}$ has too many non-zero coefficients.

For the calculation of the ordering of the covariates according to importance, AICpG uses all the genes. This results in a more reliable ordering, in spite of the unreliable estimation of B.

2.4 New Method: GTM

2.4.1 Motivation and Definition

General Trend Method (GTM) is a new method, introduced in this thesis. It is created specially for data as described in section 2.1. This method is based on two main assumptions about the data: first, that the different genes have a similar dependence on the covariates, which is defined as a general trend μ, and second that only a few covariates are important for all the genes. GTM writes the regression coefficients of a specific gene as the sum of two parts to model the assumption of similarity between genes: a general trend that is equal for all the genes, and a fine-tuning part θ_g that is gene specific;

$$\beta_g = \mu + \theta_g \text{ for all } g \in \{1, \ldots, G\}. \tag{2.7}$$

All these G vectors θ_g combined gives the matrix $\Theta := (\theta_1, \ldots, \theta_G)^\intercal$. These θ_g are small because most of the influence is captured by the general trend. GTM uses the ℓ_2-norm to penalizes $\hat{\theta}_g$. This penalty is also used in the Ridge procedure and leads to small estimations of $\hat{\theta}_g$. In this thesis we restrict ourselves to the case without a penalty on the offset per gene. This is equivalent to a centralisation of the data for all the genes separately.

The second assumption is the sparsity of μ: most of the covariates will be unimportant for most of the genes. The ℓ_1-norm as used by the LASSO procedure is known to give a good selection of the covariates. GTM penalizes μ in the same way to get a sparse estimation of μ. The total regulation term GTM uses in the minimisation problem is the sum over these two different penalties:

$$P(B) = P(\mu, \Theta) = \sum_{g=1}^{G} \|\Gamma\theta_g\|_2^2 + \lambda_1 \|\mu\|_1 , \tag{2.8}$$

with $\Gamma = \sqrt{\lambda_2}(0, 1, \cdots, 1)^\intercal I_{P+1}$ a Tikhonov matrix that gives a Ridge penalty without a penalty of the offset. λ_1 and λ_2 are two positive penalty parameters. This combined

penalty term will give a sparse estimation for μ and small estimations of the θ_g. Recalling the general minimization problem from equation (2.2)

$$\hat{B} = \arg\min_{B} \sum_{g=1}^{G} \|Y_g - X\beta_g\|_2^2 + \text{Pen}(B), \tag{2.9}$$

and substituting $\beta_g = \mu + \theta_g$ and the penalty we get:

$$(\hat{\mu}, \hat{\Theta}) = \arg\min_{(\mu,\Theta)} \left\{ \sum_{g=1}^{G} \|Y_g - X(\mu + \theta_g)\|_2^2 + \sum_{g=1}^{G} \|\Gamma\theta_g\|_2^2 + \lambda_1 \|\mu\|_1 \right\}. \tag{2.10}$$

2.4.2 Reduction to LASSO problem

The general trend μ is a P dimensional coefficients vector, and the fine-tuning Θ is a $G \times P$ dimensional matrix of coefficients, so equation (2.10) is a $(G+1) \cdot P$ dimensional minimisation problem. GTM has to solve this problem. Given the fact that the number of genes, G, is around 30,000, equation (2.10) is hard to solve. In this section we will use some algebra and rewriting to reduce this $(G+1) \cdot P$ dimensional minimisation problem to a P dimensional LASSO regression problem.

All the coefficients are coupled because optimising some $\hat{\theta}_g^p$ will influence the optimal coefficients for the other covariates of this gene. The changes in these coefficients for this specific gene will interact with that optimal $\hat{\mu}_p$ for this coefficient. All these $\hat{\mu}_p$ interact with the coefficients of all the genes for these covariates. This is shown systemically in figure 2.1 where an arrow means the interaction/influence in the optimisation.

$$\hat{\theta}_g^p \rightarrow \hat{\theta}_g \rightarrow \hat{\mu} \rightarrow \hat{\Theta}$$

Figure 2.1: A scheme of the dependence between the coefficients of minimisation problem (2.10). The arrows show the dependences.

The dependence between $\hat{\mu}$ and the full $\hat{\Theta}$ matrix is the most challenging part, because the huge amount of genes that comes into play here. The optimisation of $\hat{\theta}_g^p$ only interact with the optimisation of the other coefficients of this gene for a fixed μ. So for a fixed μ, the minimisation problem reduces to G independent optimisation problems, each with

P coefficients. For a fixed μ we get:

$$(\hat{\Theta}|\mu) = \underset{(\Theta|\mu)}{\arg\min} \sum_{g=1}^{G} \|Y_g - X(\mu+\theta_g)\|_2^2 + \sum_{g=1}^{G} \|\Gamma\theta_g\|_2^2 + \lambda_1 \|\mu\|_1 , \tag{2.11}$$

$$= \underset{(\Theta|\mu)}{\arg\min} \sum_{g=1}^{G} \left(\|Y_g - X(\mu+\theta_g)\|_2^2 + \|\Gamma\theta_g\|_2^2 \right), \tag{2.12}$$

$$= \underset{(\Theta|\mu)}{\arg\min} \sum_{g=1}^{G} \left(\|(Y_g - X\mu) - X\theta_g\|_2^2 + \|\Gamma\theta_g\|_2^2 \right). \tag{2.13}$$

In this sum, each term can be minimized separately. Each term is a Ridge regression problems [HK70] on $Y_g - X\mu$, which is the error on the estimation of Y_g if we use only the general trend μ for the estimation. A Ridge regression problem can be solved analytically with solution

$$\hat{\theta}_g(\mu) = (X^\intercal X + \Gamma^\intercal\Gamma)^{-1} X^\intercal (Y_g - X\mu). \tag{2.14}$$

With this exact solution we know the optimal $\hat{\Theta}$ as function of μ.

With this partial result we can look at the mean minimisation problem of equation (2.10). We are interested in $\hat{\mu}$, because we know $\hat{\Theta}$ if we know $\hat{\mu}$. So for $\hat{\mu}$ we get:

$$\hat{\mu} = \underset{\mu}{\arg\min} \left\{ \min_{\Theta|\mu} \sum_{g=1}^{G} \left(\|(Y_g - X\mu) - X\theta_g\|_2^2 + \|\Gamma\theta_g\|_2^2 \right) + \lambda_1 \|\mu\|_1 \right\}, \tag{2.15}$$

$$= \underset{\mu}{\arg\min} \left\{ \sum_{g=1}^{G} \left[\|(Y_g - X\mu) - X\theta_g)\|_2^2 + \|\Gamma\theta_g\|_2^2 \right]_{\theta_g = \hat{\theta}_g(\mu)} + \lambda_1 \|\mu\|_1 \right\}. \tag{2.16}$$

Introducing $\mathscr{C} := (X^\intercal X + \Gamma^\intercal\Gamma)^{-1} X^\intercal$ and substituting in (2.14) gives:

$$\hat{\mu} = \underset{\mu}{\arg\min} \left\{ \sum_{g=1}^{G} \left(\|(Y_g - X\mu) - X\mathscr{C}(Y_g - X\mu)\|_2^2 + \|\Gamma\mathscr{C}(Y_g - X\mu)\|_2^2 \right) + \lambda_1 \|\mu\|_1 \right\}, \tag{2.17}$$

$$= \underset{\mu}{\arg\min} \left\{ \sum_{g=1}^{G} \left(\|(I_n - X\mathscr{C})Y_g - (I_n - X\mathscr{C})X\mu\|_2^2 + \|\Gamma\mathscr{C}(Y_g - X\mu)\|_2^2 \right) + \lambda_1 \|\mu\|_1 \right\}. \tag{2.18}$$

For every ℓ_p-norm we have that $\|a\|_p^p + \|b\|_p^p = \sum_{i=1}^{n} a_i^p + \sum_{i=1}^{m} b_i^p = \sum_{i=1}^{n+m} c_i^p = \|c\|_p^p$ where c is the merge vector $(a, b)^\intercal$. So we can write every term in the sum of equation (2.18) as one norm by stacking the two vectors:

$$\hat{\mu} = \underset{\mu}{\arg\min} \left\{ \sum_{g=1}^{G} \left\| \begin{pmatrix} I_n - X\mathscr{C} \\ \Gamma\mathscr{C} \end{pmatrix} Y_g - \begin{pmatrix} I_n - X\mathscr{C} \\ \Gamma\mathscr{C} \end{pmatrix} X\mu \right\|_2^2 + \lambda_1 \|\mu\|_1 \right\}. \tag{2.19}$$

We can do the same for the sum of G norms by making a merged vector of all these G vectors. To short the writing we define $\mathscr{D} := \begin{pmatrix} I_n - X\mathscr{C} \\ \Gamma\mathscr{C} \end{pmatrix}$.

$$= \arg\min_{\mu} \left\{ \left\| \begin{pmatrix} \mathscr{D}Y_1 \\ \vdots \\ \mathscr{D}Y_G \end{pmatrix} - \begin{pmatrix} \mathscr{D}X \\ \vdots \\ \mathscr{D}X \end{pmatrix} \mu \right\|_2^2 + \lambda_1 \|\mu\|_1 \right\}. \tag{2.20}$$

The minimisation problem in equation (2.20) is a LASSO regression problem [Tib96]. For the the estimation of Θ we only have to substitute this $\hat{\mu}$ in all the G expressions of $\hat{\theta}_g(\mu)$ given in (2.14).

Theorem 1. *The $P \cdot (G+1)$ dimensional minimisation problem*

$$(\hat{\mu}, \hat{\Theta}) = \arg\min_{(\mu,\Theta)} \left\{ \sum_{g=1}^{G} \|Y_g - X(\mu + \theta_g)\|_2^2 + \sum_{g=1}^{G} \|\Gamma\theta_g\|_2^2 + \lambda_1 \|\mu\|_1 \right\}$$

is equivalent to the P dimensional LASSO regression problem

$$\hat{\mu} = \arg\min_{\mu} \left\{ \left\| \begin{pmatrix} \mathscr{D}Y_1 \\ \vdots \\ \mathscr{D}Y_G \end{pmatrix} - \begin{pmatrix} \mathscr{D}X \\ \vdots \\ \mathscr{D}X \end{pmatrix} \mu \right\|_2^2 + \lambda_1 \|\mu\|_1 \right\},$$

$$\text{with } \mathscr{D} := \begin{pmatrix} I_n - X\mathscr{C} \\ \Gamma\mathscr{C} \end{pmatrix} \text{ and } \mathscr{C} := (X^\intercal X + \Gamma^\intercal\Gamma)^{-1} X^\intercal.$$

The estimation $\hat{\Theta}$ is given by

$$\hat{\theta}_g(\hat{\mu}) = \mathscr{C}(Y_g - X\hat{\mu}).$$

We started with a $P \cdot (G+1)$ dimensional regression problem written in a non-standard format. We solved $G \cdot P$ dimensions exactly as a function of the last P using the known solution of a Ridge regression problem. We rewrote it to a LASSO regression problem. The LASSO format at the end has a data vector of length $G \cdot (n+P)$ and a $G \cdot (n+P) \times P$ dimensional design matrix.

2.4.3 Properties

GTM gives one ordering of the covariates for every parameter λ_2. By decreasing parameter λ_1 from infinity to zero, the coefficients of $\hat{\mu}$ will become non-zero consecutively, which implies an ordering of the covariates from most important to least important. This can be done by using the LARS algorithm [EHJ+04] which solves the LASSO problem directly for all λ_1.

Due to the ℓ_1-norm in the penalty of μ we get a sparse estimation vector $\hat{\mu}$. With this $\hat{\mu}$ are the $\hat{\theta}_g$ solutions of Ridge regression, which results in a $\hat{\Theta}$ matrix that is not sparse, so also the $\hat{B}$ matrix will have all non-zero coefficients.

2.5 Similarities and Differences

There are three main differences between the methods. For the selection of covariates there is a difference between GTM and MTFS on one hand and AICpG on the other hand. The selection by GTM and MTFS works similarly to the selection by the LASSO procedure. The ℓ_1-norm in GTM and MTFS prefers a sparse vector. So, for a fixed value penalty parameter λ_1 for GTM or λ for MTFS the result is a vector with some coefficients zero and some non-zero. The order in which the coefficients become non-zero if the penalty parameter λ_1 or λ decreases from infinity, where all coefficients are zero, to zero, where all coefficients are non-zero, gives a ordering of the importance of the covariates. AICpG has a different approach, it does not have a changeable penalty parameter, but works with a voting system. The regression problems for the genes vote independently on with coefficients they want to include in the model, where the regression problems can vote on more than one covariate. Counting the votes gives an ordering on the importance of the covariates.

The second difference, also between AICpG and the other methods, is how the regression problems of the genes are related. AICpG estimates the β_g independently of each other and only combines the different genes for the ordering of the covariates. GTM and MTFS forces the $\hat{\beta}_g$ to be similar in some way. MTFS decides, using all the genes, which covariates are important and therefore have non-zero coefficients. GTM does not only select the covariates, using all the genes, but also estimates the value of the coefficients roughly for all the genes by using a general trend μ that is equal for all the genes. So for these two methods the genes can learn from each other by the estimation of B.

A third difference is between GTM and the other methods. GTM fits a general trend for all the genes with the shrinkage of $\hat{B}$ toward $\hat{\mu}$, and also uses this general trend for the ordering of the covariates. MTFS and AICpG do not have a shrinkage toward something else then zero.

The structure of the estimated $\hat{B}$ differs between the methods. GTM gives a $\hat{B}$ where all coefficients are non-zero, but with similar looking rows $\hat{\beta}_g$, so for a column $\hat{\beta}^p$, all the elements are around the same number $\hat{\mu}_p$. MTFS gives a sparse $\hat{B}$, where part of the columns $\hat{\beta}^p$ are totally zero, and the other columns have only non-zero elements. The elements of such a non-zero column are not related to each other. AICpG also gives a sparse $\hat{B}$ matrix, but without a structure.

Chapter 3

Experiments with Simulated Data

In this chapter we will perform experiments with the three methods introduced in chapter 2. The goal of the experiments is to get a better understanding what kind of answers the methods give for different kinds of data. By using simulated data, we know the real structure of the data, so we can look at the difference between the real structure and the structure estimated by the models. We can investigate the influence of noise and the stability of answers by running several iterations.

In total we will do three experiments. In experiment 1 the data structure contains a general trend. In the second experiment the genes are split in two groups, each group with its own general trend. The last experiment uses data without a general trend.

We will compare the methods mainly on two points, the ordering of the covariates and the difference between the estimated coefficients and the real coefficients. Each method orders the covariates by importance. During the introduction of the methods we explained how the ordering of the covariates is determined from the solution of the minimisation problem of the method. Important questions are 'What makes covariates important for the different models?' and 'How much does the order differ for different iterations?'

For the error of the estimated $\hat{B}$ on the real B matrix as well as for the error of the estimated $\hat{\mu}$ on μ we use the mean squared error:

$$B\text{-estimation error} := \frac{1}{GP} \sum_{g=1}^{G} \left\| \hat{\beta}_g - \beta_g \right\|_2^2, \tag{3.1}$$

$$\mu\text{-estimation error} := \frac{1}{P} \left\| \hat{\mu} - \mu \right\|_2^2. \tag{3.2}$$

To create data, we first create the coefficients matrix B and the design matrix X. The design matrix is $n \times P$ dimensional, which we fill with independent draws from a standard normal distribution. Second, we create the dependent variables, which are the gene expression levels, matrix Y, by multiplying B and X and adding independent, normal distributed noise with mean 0 and variance σ_Y^2 to all the elements of the matrix.

3.1 Experiment 1: One General Trend

In this first experiment we use data with a structure as described in section 2.1; we have a general trend and for each gene there is a correction. In particular, we will look how the sizes of the general trend and the gene-wise corrections influence the ordering and the estimation.

3.1.1 Setup

We make data with 30 covariates, divided in five blocks of six. For each block we take a constant size of the general trend and different sizes for the gene-wise corrections. All the corrections θ_g^p are independent normally distributed with mean zero and a standard deviation that differs between the covariates. The general trend μ and the standard deviation of the θ_g^p are shown in figure 3.1. We will generate 100 datasets with the same

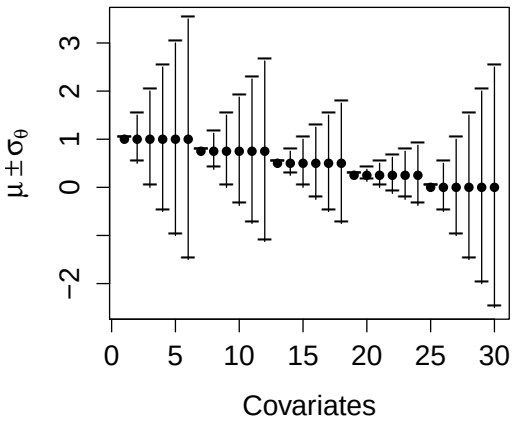


Figure 3.1: The size of the coefficients of μ and the standard deviation of the gene-wise corrections.

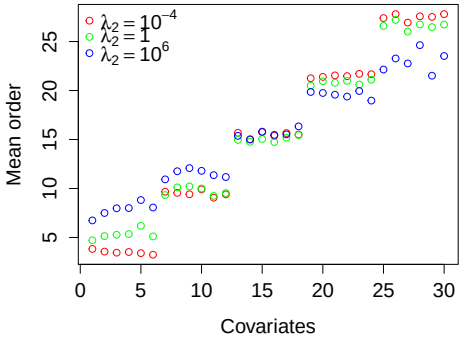
μ and σ_θ.

3.1.2 Results

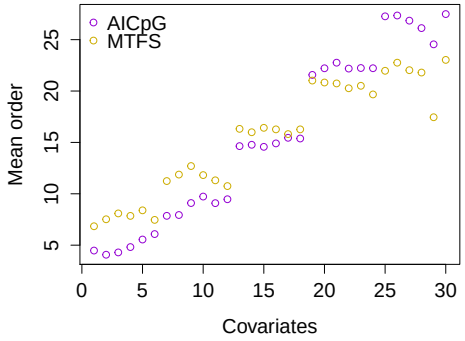
Order of the covariates

We calculate the order for 100 iterations and calculate the mean order over these 100 iterations for all methods. GTM gives for each value of λ_2 its own ordering. We will take

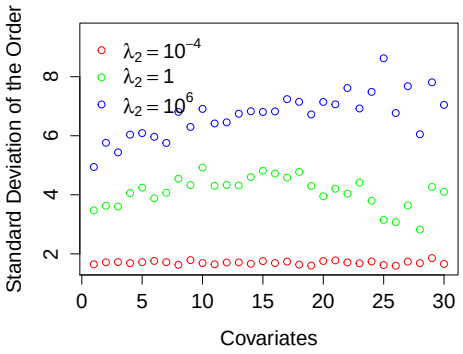
a small, medium and large value for λ_2. These values correspond to a consistent ordering over the iterations. a relative small estimation error on B and a relative estimation error of μ. The mean order and the standard deviation of these orderings are shown in figure 3.2. We see that for all the methods the mean order of the covariates is constant



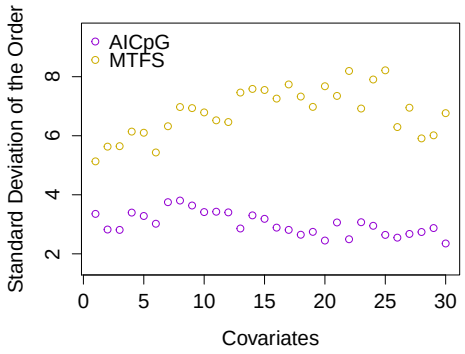
(a) The mean order of the covariates using GTM.



(b) The mean order for AICpG and MTFS.



(c) The standard deviation of the order using GTM.

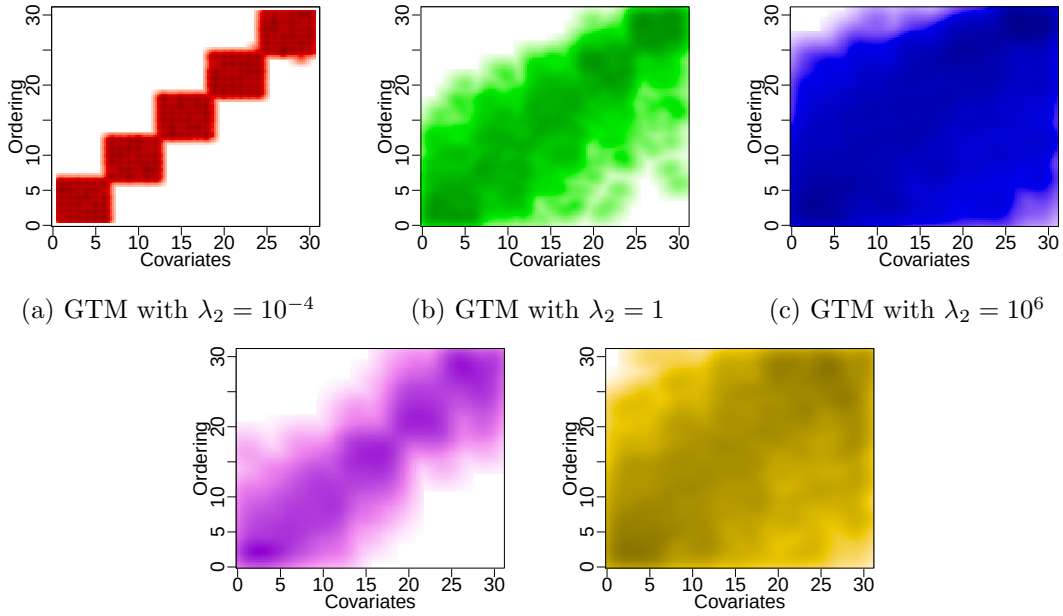


(d) The standard deviation for AICpG and MTFS.

Figure 3.2: The mean order and the standard deviation of the covariates, taken over 100 iterations.

for groups of covariates with the same size of the coefficient μ_p, and the covariates enter the model from a large μ_p to a small. There is no clear relation between the fluctuations around this mean and the size of μ_p or σ_θ^2. So for the order of the covariates the size of the general trend is important, but the size of the gene specific addition is unimportant.

The densities of the orderings over the 100 iterations are shown in figure 3.3 on the following page, a darker color corresponds to a higher density.



(a) GTM with $\lambda_2 = 10^{-4}$ (b) GTM with $\lambda_2 = 1$ (c) GTM with $\lambda_2 = 10^6$

(d) AICpG (e) MTFS

Figure 3.3: The densities of the orderings of the 100 iterations.

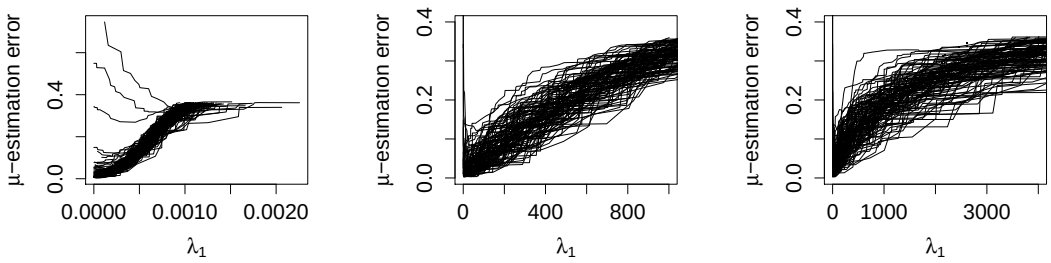
The separation between the groups is best for GTM with a small λ_2, in only one of the 100 iterations two covariates from two successive groups are interchanged. When λ_2 increases the separation becomes worse. AICpG orders as good as GTM with $\lambda_2 = 1$ and MTFS is comparable with GTM with the largest value of λ_2. For GTM with the small and middle λ_2 and for AICpG we can see the block structure in the density, for the large λ_2 and MTFS there is still a higher density around the diagonal, but the block structure is not visual.

Estimation of μ and B

Another criterion is the error on the estimation of B and for GTM on μ (The other methods do not give an estimation of the general trend). For the measure of this error we use the mean squared error. In each iteration, GTM gives for the three fixed values of λ_2 the estimation errors on μ and B as functions of λ_1. The three values of λ_2 are chosen such that the first one gives a good estimation of the ordering, the second the best estimation for B and the last the best estimation of μ. AICpG gives one single number for the B-estimation error and MTFS gives the estimation error of B as function of λ.

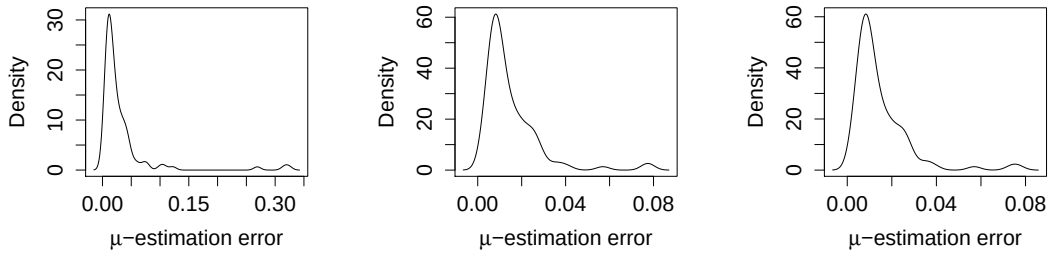
Figure 3.4 on the next page shows the μ-estimation error as function of λ_1 for three values of λ_2 for all the 100 iterations. Besides this full error path as function of λ_2 we also look at the minimal estimation error. For all the iterations we take the minimal

estimation error, and made a density plot of these 100 minimal errors in figure 3.5.



(a) $\lambda_2 = 10^{-5}$ (b) $\lambda_2 = 20$ (c) $\lambda_2 = \infty$

Figure 3.4: The error on the estimation of μ by GTM as function of λ_1, for three values of λ_2.



(a) $\lambda_2 = 10^{-5}$ (b) $\lambda_2 = 20$ (c) $\lambda_2 = \infty$

Figure 3.5: A density plot of the minimal μ-estimation error for three values of λ_2.

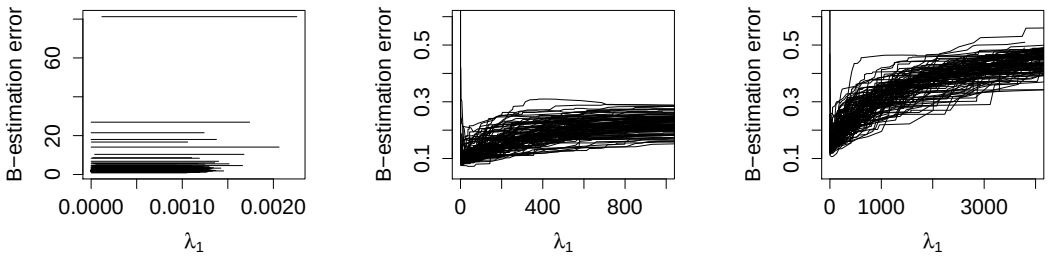
This minimal estimation error is somewhat to optimistic, because normally we do not know at which λ_1 value this minimum is achieved. It would be better to take one specific value of λ_1 for all the iterations, but I do not have the data for this.

The relation between the μ-estimation error and λ_1 has the standard structure from overfitting to underfitting with in between the smallest estimation error. For the smallest value of λ_2 is the estimation of the general trend the worst, especially for some iterations. The range of useful λ_1 values strongly depends on the size of λ_2; if λ_2 is small and λ_1 not, the total penalty is smaller if the full structure is captured by $\hat{\Theta}$ with $\hat{\mu} = 0$ because $\hat{\Theta}$ is hardly penalized. If λ_2 is large; λ_1 can be large also, because the large penalty on $\hat{\Theta}$ forces the general structure to $\hat{\mu}$ for the minimal total penalty.

All methods give an estimation of B, so we can compare all three methods. For none of the methods we expect a good estimation of B; there are simply too many coefficients, roughly the same number as there are data points. But we still can learn some things

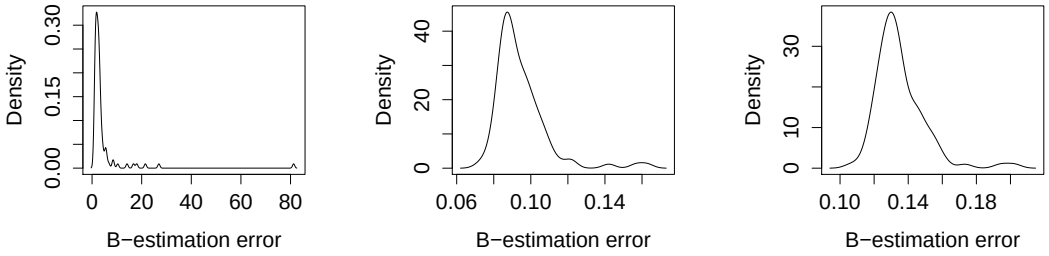
about the methods by looking at this B-estimation error.

The estimation error of the total B matrix for GTM and the density of the minimal B-estimation errors are shown in figures 3.6 and 3.7. Just as for the μ-estimation error, this is done for all the iterations and with the same values of λ_2. We take a different vertical axis for figure 3.6a because the estimation error is much larger for this small value of λ_2.



(a) $\lambda_2 = 10^{-5}$ (b) $\lambda_2 = 20$ (c) $\lambda_2 = \infty$

Figure 3.6: The estimation error on B by GTM as function of λ_1, for three values of λ_2. The first graph has a different vertical axis scale.

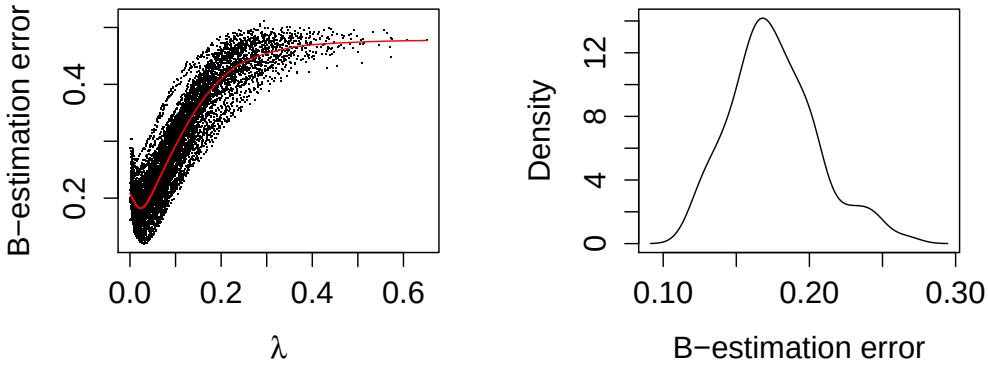


(a) $\lambda_2 = 10^{-5}$ (b) $\lambda_2 = 20$ (c) $\lambda_2 = \infty$

Figure 3.7: A density plot of the minimal B-estimation error for three values of λ_2.

The estimation error on B is large if λ_2 is small. As already mentioned, the estimation of B is unreliable. With a larger λ_2 we add some shrinkage of $\hat{\Theta}$ to $\hat{\mu}$. This gives some bias, but, similar to the use of Ridge regression in a standard regression problem, the variance reduction is larger. This results in a smaller estimation error and a more consistent error over the iterations.

MTFS gives a B-estimation as function of λ. The error as function of λ and the density of the minimal error for all the iterations are plotted in figure 3.8 on the next page.



(a) The B-estimation error for MTFS as function of λ for all the 100 iterations. In red a kernel estimation.

(b) The density of the minimal B-estimation error over the 100 iterations.

Figure 3.8: The B-estimation error for MTFS.

AICpG gives one estimation of B and the mean estimation error over the 100 iterations is given in table 3.1, together with the means of the minimal estimation error for GTM and MTFS. The B-estimation error of AICpG is similar to the error of GTM with the

Table 3.1: The minimal μ and B-estimation errors averaged over the 100 iterations.

	GTM			AICpG	MTFS
	$\lambda_2 = 10^{-5}$	$\lambda_2 = 20$	$\lambda_2 = \infty$		
Mean minimal μ-estimation error	0.032	0.015	0.015	·	·
Mean minimal B-estimation error	4.3	0.09	0.14	4.1	0.18

small value of λ_2. As already mentioned in section 2.3 this is caused by overfitting.

Discussion

The minimal estimation errors of GTM and MTFS are overly optimistic, because if we have real data, we cannot determine what are the best choices for the penalty parameters. With for example cross-validation we can estimate the best amount of penalty.

One striking finding is the difference in effect between the ordering and the estimation of B and μ for GTM if we increase the penalty on $\hat{\Theta}$. The ordering becomes worse (Figure 3.3 on page 18) where the estimation errors on μ and B become better (Figures

3.4 and 3.6 on pages 19 and 20). This second effect is easy to explain because a larger penalty reduces the variance of the estimation, and this variance reduction is larger than that increase of the bias. The deterioration of the ordering if λ_2 increases, is harder to explain. The ordering is based on the estimation of μ, and the estimation error of μ decreases. Probably the introduction of bias has a more negative effect on the ordering than on the estimation of μ and B. Another option is that for the selection of covariates using GTM, the overfitting of Θ does not matter. A small λ_2 gives more overfitting, but a coefficient of $\hat{\mu}$ becomes only non-zero if it is really needed.

3.2 Experiment 2: Two Different Trends

In this second experiment we will look at data where there are two groups of genes, both groups have a general trend, but these trends are different. If, for example, some genes will become more active by aging and other genes will become less active, we will get this structure. This data structure is not assumed by GTM, so we expect that GTM will give different answers.

3.2.1 Setup

To create the data, we will create two datasets with half of the genes and with the same design matrix, but one group with general trend μ^1 and the other group with general trend μ^2. For the creation of the Θ matrices we draw from a normal distribution with $\sigma_\theta^2 = 0.3$ for all the genes and covariates. At the end we merge the two datasets to one normal sized data set of G genes. The two μ vectors are shown in figure 3.9. The coefficient-wise mean of the two μ is added in black. For both μ the magnitude

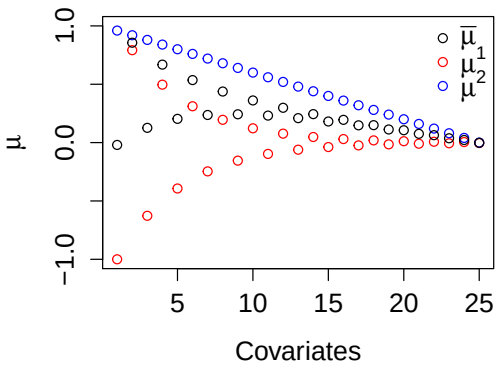


Figure 3.9: The two different μ vectors in red and blue for the covariates. The coefficient-wise mean of the two μ's is shown in black.

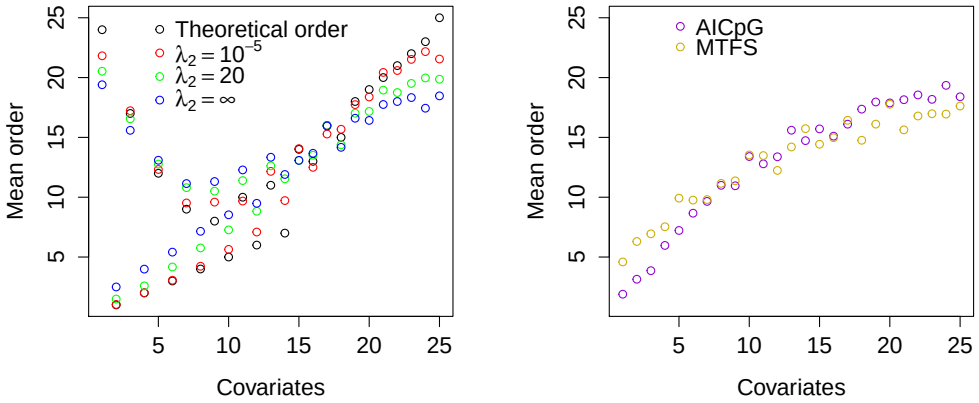
is decreasing in the covariates, so the first covariates are more important than the last covariates. But for the odd covariates the signs of the two μ are opposite to each other. This gives a $\bar{\mu}$ that is around zero for the first odd covariates, because the two μ cancel each other out.

In this test we will look at the ordering of the covariates given by the methods and an option to detect if the data has a structure with a general trend or if there are different trends for different subsets of genes. To do this, we simulate data 100 times.

3.2.2 Results

Ordering of the Covariates

We simulate the data 100 times and for each iteration we calculated the ordering. The mean orders over the 100 iterations are shown in figure 3.10. The order of GTM for three values of λ_2 are shown in figure 3.10a and the order of AICpG and MTFS are shown in figure 3.10b.



(a) GTM and in black the order of $\bar{\mu}$ ordered according to the size.

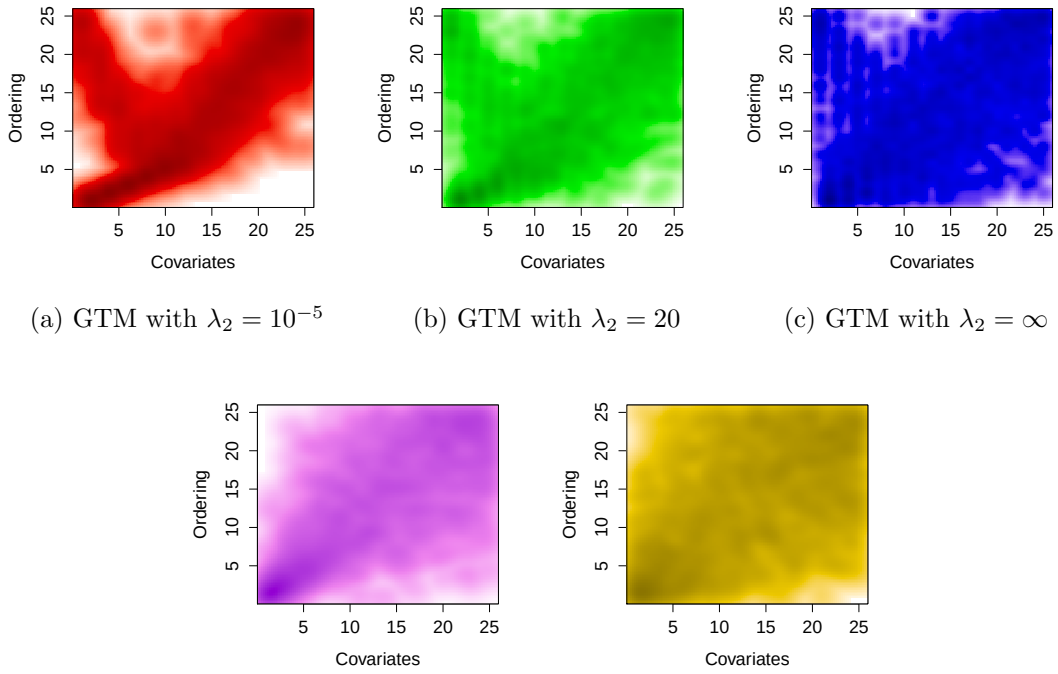
(b) AICpG and MTFS

Figure 3.10: Mean order of the covariates over 100 iterations.

GTM fits one $\hat{\mu}$ for all the genes, in this case there are two μ, and the calculated ordering follows the order of the mean $\bar{\mu}$ ordered from large to small in absolute value. This ordering is the black ordering in figure 3.10a. If the different μ cancel out, for example covariates 1,3 and 5, or if both μ are small, covariates 20 to 25, GTM assigns a low importance to the covariate and if the two μ have the same sign the covariates are ordered first, for example covariates 2, 4 and 6.

AICpG and MTFS are not based on a general effect over all the genes and the different μ do not cancel out by the calculation of the order, as shown in figure 3.10b on the previous page. They order on the magnitude of $\left|\mu^1\right| + \left|\mu^2\right|$ instead of $\left|\mu^1 + \mu^2\right|$ Both methods give an ordering where the first covariates are the most important. Similarly to experiment 1, AICpG gives a better ordering than MTFS.

Similarly to the first experiment we calculate the densities of the 100 orderings, these are shown in figure 3.11. We see that for GTM the consistence of the orderings decreases



(a) GTM with $\lambda_2 = 10^{-5}$ (b) GTM with $\lambda_2 = 20$ (c) GTM with $\lambda_2 = \infty$

(d) AICpG (e) MTFS

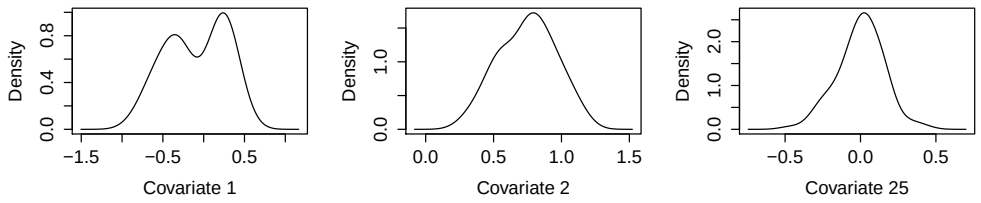
Figure 3.11: The densities of the orderings of 100 iterations.

if λ_2 increases, as well as in the first test. The orderings are less consistent for AICpG and MTFS.

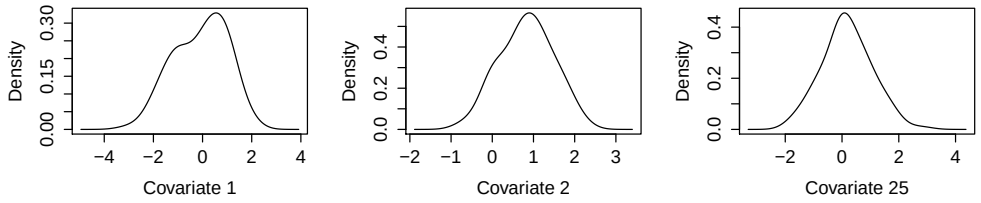
Detection of Different Trends

For the interpretation of the ordering given by the methods, it is important to know if the underlying structure of the data has one μ or that there are groups of genes that interact in different ways. The $\hat{\mu}$ estimation of GTM gives only the mean so $\hat{\mu}$ does not tell if there are different trends. For this we need to look at all the different genes. For some specific covariate p we have the estimation $\hat{\beta}_g^p$ for all the G genes. If there is one

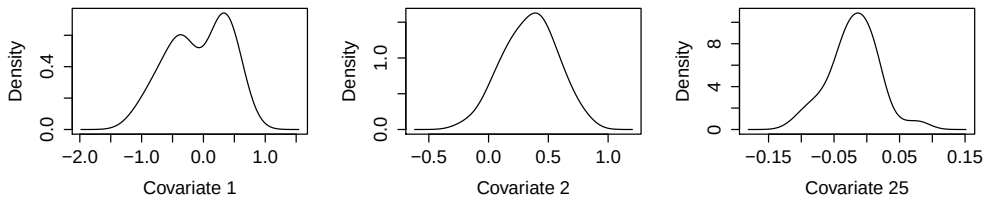
general trend we would assume that for all these genes, these $\hat{\beta}_g^p$ are around μ_p. If there are two trends, one part of the $\hat{\beta}_g^p$ is around μ_p^1 and another part around μ_p^2. The density plots of these G coefficients in $\hat{\beta}^p$ are shown in figure 3.12 for three covariates. Above are the estimations from GTM, in the middle from MTFS and below from AICpG. The



(a) GTM with $\lambda_1 = 32$ and $\lambda_2 = 20$



(b) AICpG



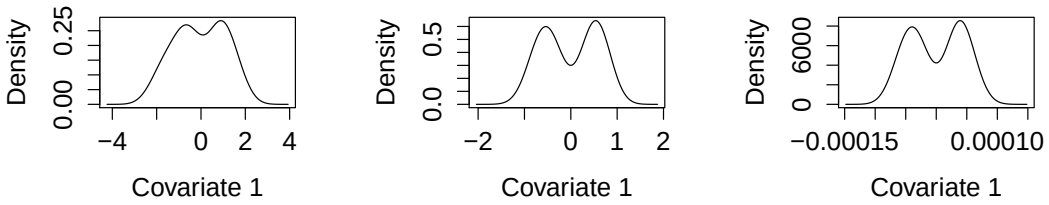
(c) MTFS with $\lambda = 0.041$

Figure 3.12: The densities of the estimated coefficients over all the 100 genes for three different covariates and one specific iteration. The first covariate shows two peaks, the second one peak at a positive number and the last covariate shows one peak at zero.

two distinct peaks for covariate 1 show that there are two groups of genes which are influenced differently by this covariate. For covariate 2 there is one peak around 0.5, so all the genes are influenced similar for covariate 2 on a positive way. Covariate 25 is not important because its peak is around zero. GTM gives the clearest separation between the two groups of genes, and AICpG has only a vague separation between the groups of genes.

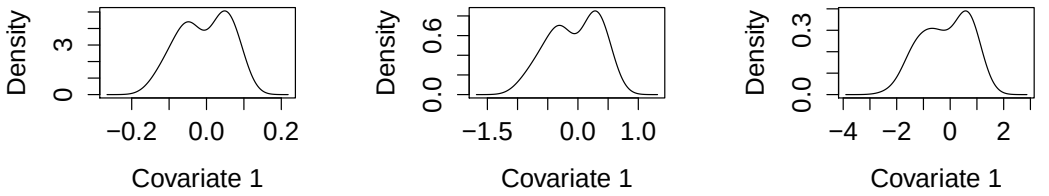
The specific settings for the penalty parameters are not important for the shape of the

density plots. A larger value of λ_2 gives smaller coefficients of $\hat{\Theta}$ by GTM. The result is that the peaks are closer together, but that the width of the peaks are smaller also. Very small values of λ_2 give a less clear separation of the peaks, but beside that, there is a wide range of λ_2 that results in the same shapes. λ_1 influences primary $\hat{\mu}$ and the different peaks are caused by $\hat{\Theta}$ so this parameter is even less influential. With a linear scaling of the horizontal and vertical axis we get the same shape, as shown in figure 3.13. For MTFS this is also true. λ needs to be small enough to have non-zero coefficients for this specific covariate, but if this is the case, a smaller value of λ does not matter, as shown in figure 3.14.



(a) $\lambda_2 = 10^{-5}$ and $\lambda_1 = 10^{-5}$ (b) $\lambda_2 = 20$ and $\lambda_1 = 22$ (c) $\lambda_2 = 10^6$ and $\lambda_1 = 44$

Figure 3.13: The density plot of covariate 1 using GTM for different values of λ_2. The shapes are similar for all the values.



(a) $\lambda = 0.13$ (b) $\lambda = 0.058$ (c) $\lambda = 0.0016$

Figure 3.14: The density plot of covariate 1 using MTFS for different values of λ. The shapes are similar for all the values.

3.3 Experiment 3: No General Trend

The following experiment uses data with some structure in the covariates, but without a general trend. For this some covariates are used for more genes than other covariates.

3.3.1 Setup

In this experiment each gene is influenced by a different subset of covariates. To create differences in importance of the covariates some covariates are more frequently of influence on the activity of a gene. We include covariate p for gene g with probability $\Pr(p)$. The important covariates are the covariates corresponding to a high probability $\Pr(p)$ and the less important covariates have a small probability. We take for $\Pr(p)$ an exponential decaying function in p. If the covariate is included, we draw the coefficient β_g^p from the normal distribution with mean zero and variance 25.

$$\beta_g^p = \begin{cases} a_{g,p} & \text{with probability } \Pr(p) \\ 0 & \text{with probability } 1 - \Pr(p) \end{cases}, \tag{3.3}$$

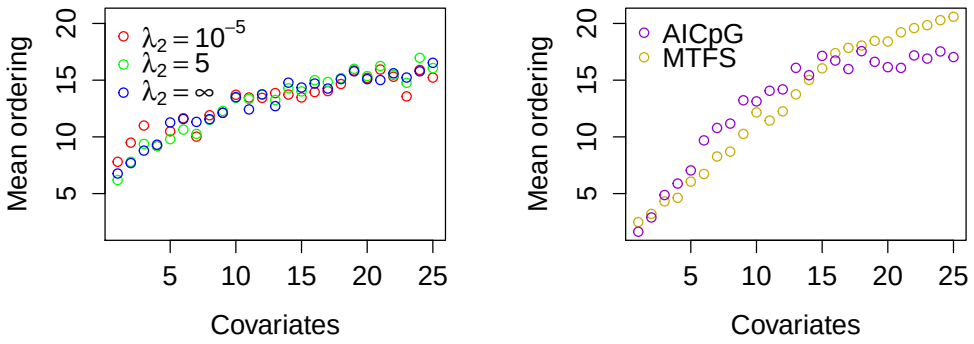
$$\text{where } a_{g,p} \sim N\left(0, \sigma^2 = 25\right) \text{ and } \Pr(p) \propto e^{-\tau p}. \tag{3.4}$$

Note that $\mathbb{E}\left[\beta_g^p\right] = 0$ for all the genes and covariates, and that the expected number of non-zero elements in β^p decreases with p.

3.3.2 Results

Order of covariates

We calculate the order for 100 iterations and calculate the mean order over these iterations for all methods. We take λ_2 equal to 10^{-5}, 5 and infinity for GTM. The mean orders are shown in figure 3.15. In figure 3.15a the orders for GTM and in figure 3.15b the orders of AICpG and MTFS. We see that for all the methods the important covari-

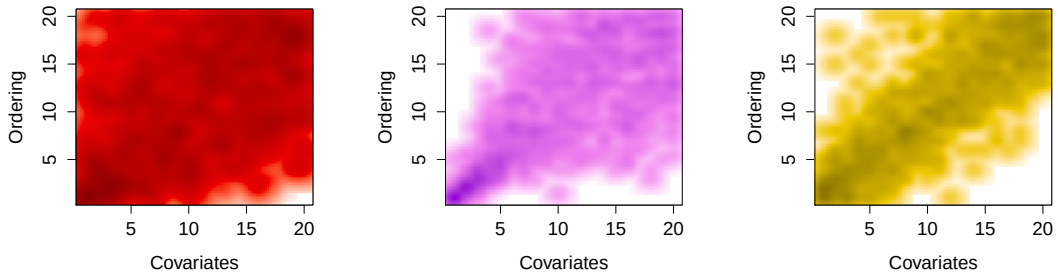


(a) GTM for different values of λ_2 (b) AICpG and MTFS

Figure 3.15: The mean ordering of the covariates over 100 iterations.

ates are the first covariates. AICpG and MTFS gives a better order estimation than GTM. GTM do not give a difference in the accuracy of the ordering for the different values of λ_2.

The densities of the orderings over the 100 iterations are shown in figure 3.16, a darker colour means a higher density. For GTM we only included one value of λ_2 because the other values gave similar plots. The orderings of the different iterations are totally



(a) GTM with $\lambda_2 = 10^{-5}$ (b) AICpG (c) MTFS

Figure 3.16: The densities of the orderings of 100 iterations.

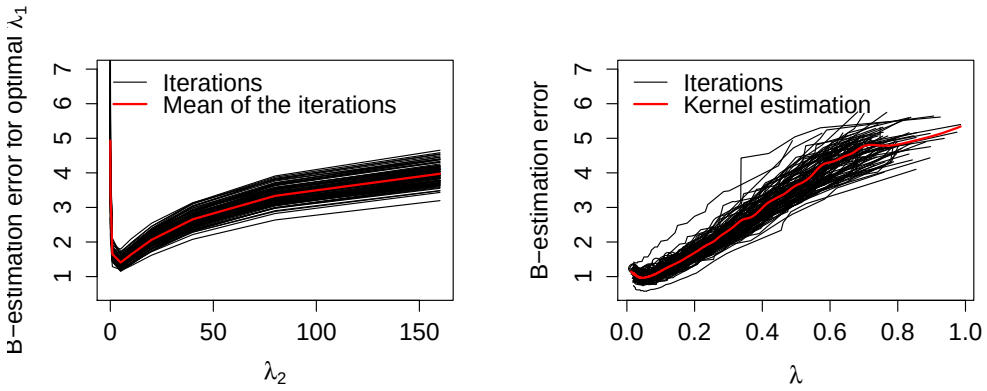
spread out for GTM. AICpG has ordered the first few covariates very well but for the other covariates it is spread out. MTFS has a clear dark diagonal over all the covariates, but is not that good as AICpG for the first few covariates.

B-Estimation Error

That there is no general trend does not only influence the quality of the ordering negatively for GTM but has also an influence on the estimation of B. GTM uses the λ_1 penalty parameter for the penalty of the estimation of the general trend. $\hat{\mu}$ will be small for all the values of λ_1 and the estimation of the gene-wise correction will be more important. In the experiment we saw that for a given λ_2 value, the influence of λ_1 on the estimation error is minimal; the differences between iterations is much larger.

Figure 3.17 on the facing page shown the estimation error for GTM and MTFS. For the estimation error of GTM we used the optimal values of λ_1. A less optimal λ_1 changes the estimation error of GTM by averaged less than 0.05, so this would hardly change the graph.

The mean estimation errors are shown in table 3.2 on the next page. For GTM and MTFS is the optimal amount of penalty used. The fluctuation between iterations is added in the second row.



(a) GTM; The B-estimation error as function of λ_2 for the optimal values of λ_1.

(b) MTFS; The B-Estimation error as function of λ.

Figure 3.17: The B-Estimation errors for all the iterations for GTM and MTFS.

Table 3.2: The minimal B-estimation errors averaged over the 100 iterations.

	GTM with $\lambda_2 = 5$	AICpG	MTFS
Mean of minimal B-estimation errors	1.4	5.2	1.0
Standard deviation around this mean	0.14	5.2	0.13

3.3.3 Discussion

Despite the fact that we did not create B with a general trend over the genes, GTM still gives a positive correlation between the covariates and the order, which is clear in figure 3.15a on page 27. In experiment 1 we saw that the methods order on the magnitude of μ, but actually μ is impossible to detect because Θ can ruin this structure, such that B does not have this structure. So the methods can only detect B and not the underlying μ, and to explain the ordering we need to take all the different genes into account. In test 2 we saw that GTM orders the covariates on the sum of the two trends and AICpG and MTFS on the sum of the absolute value of the trends. Similar to experiment 1, this is a simplified situation. It is better to say that GTM orders on $\sum_g \beta_g^p$, where AICpG and MTFS orders on $\sum_g |\beta_g^p|$. The partial sums create some random walk with steps β_g^p, and the general trend μ can be seen as an drift. If there is a general trend, this drift dominates the random walk if we have enough genes and the size of Θ does not matter. In this third experiment there is no drift for GTM because $\mathbb{E}[\beta_g^p] = 0$. But the first covariates have more non-zero β_g^p, so the random walk takes more actual steps. Taking more steps, means a higher variance, so it is more likely that the first covariates has a larger value for $\sum_g \beta_g^p$. For AICpG and MTFS there is still a

drift because $\mathbb{E}\,|\beta_g^p| > 0$, so for these methods it is more easy to detect the important covariates.

That AICpG orders the first few covariates much better than the other covariates is striking. An explanation can be that $\Pr(p)$ decreases exponentially, so the absolute difference between the successive values of $\Pr(p)$ is larger for the first covariates, but we do not see this effect for the other methods.

We saw that the B-estimation error of GTM is hardly influenced by the size of λ_1. λ_1 regulates the estimation of the general trend, but because there is not such trend, this penalty parameter is unimportant; $\hat{\mu}$ will be around zero for all λ_1.

3.4 Summary of Findings

In this chapter we have seen how the three methods work by applying them to generated data. The analysis of data with different trends leads to the conclusion that the ordering of the covariates is done in two different ways. AICpG and MTFS order the covariates by the sum of the absolute values of the coefficients: $\sum_{g=1}^{G} |\beta_g^p|$. GTM orders the covariates by using the sum of the coefficients itself: $\sum_{g=1}^{G} \beta_g^p$. Data with a general trend give a similar ordering for all methods, figure 3.2 on page 17, because this trend will be dominant and $\sum_{g=1}^{G} |\beta_g^p| \approx |\sum_{g=1}^{G} \beta_g^p| \approx G \cdot |\mu_p|$. In cases where there are several trends or no trends at all, the coefficients can cancel each other out for GTM, but not for AICpG or MTFS, figures 3.10 on page 23 and 3.15 on page 27.

The reliability of the ordering differs between the methods and for GTM also between different values of λ_2. GTM needs a net trend to give a consistent ordering. If this is the case, GTM with a small λ_2 penalty parameter is most consistent over the iterations, followed by AICpG, MTFS and GTM with larger λ_2, see figures 3.3 on page 18 and 3.11 on page 24. AICpG and MTFS are more reliable if there is no net trend, figure 3.16 on page 28. All methods can detect if the data contain multiple trends. GTM gives the best detection followed by MTFS and AICpG detects the different trends least clear, figure 3.12 on page 25.

The estimation of B cannot be done without a substantial error. A shrinkage on the coefficients reduces the variance of the estimation, resulting in a smaller estimation error. GTM and MTFS have more shrinkage than AICpG resulting in a better estimation, tables 3.1 on page 21 and 3.2 on the previous page.

Chapter 4

Gene Expression of Human Embryos

In the previous chapter we have seen how the three methods work and perform by analysing simulated data. In this chapter we will analyse a data set with the gene expression of human embryos. First we will give some biological information about gene expression in general and about this data set in particular. The analysis consists of three parts; the ordering of the covariates, the detection of different trends and the reliability of the ordering.

4.1 Biological Explanation of the Data

The data we will analyse is the activity of genes from preimplantation human embryos which were donated to research. This was done by couples who had had an in vitro fertilisation treatment and fulfilled their child wish. The data is used for researching which covariates affect the gene expression levels.

The gene expression levels are measured by the microarray procedure. Gene expression is the process where the genetic information of the gene is used to produce a product, most of the time a protein. The production of this protein consists of various steps. The two most important steps are the transcription and the translation [AJL+14]. An mRNA copy of the part of the DNA strand corresponding to the protein is made, the transcription. This mRNA molecule is modified and exported out the nucleus of a cell. The mRNA is now used to produce the protein; the translation. So the amount of expression of a gene is related to the amount of mRNA molecules corresponding to this gene in the cell. The microarray procedure is based on this principle and is explained in [Dră03]. I will give a brief explanation. The microarray is an array with for all genes an entry where the genetic information of this gene binds. First, all the different mRNA molecules corresponding to all the different genes are isolated from the cell. Then the mRNA is copied to fluorescent complementary DNA and multiplied. These complementary DNA molecules are put on the array and binds to the entries of the corresponding gene. The entries corresponding to the most expressed genes are the brightest because there is more fluorescent complementary DNA that light up. The

multiplication of the complementary DNA is needed to increase the brightness. The multiplication is equal for the different complementary DNA molecules in the sample, so the proportion is constant. However, the multiplication factor is not constant for different samples, so we cannot compare the absolute activity of a gene for different embryos. Instead of the absolute activity, the relative activity is quantified; becomes the gene more or less active as function of the conditions.

To analyse the data with linear regression, we need to transform the data. We normalise the microarray data such that all embryos have the same total activity, because of the independent multiplication of the cDNA between embryos. We apply a log transformation to get the data more linear dependent on the covariates.

The specific dataset that we will analyse, comes from Mantikou et al. in [MJW+16]. The experimental setup is described in this paper and here I will only describe the data shortly. The gene expression levels of 35,927 genes are measured for 37 embryos. These embryos were six days old and grown in vitro. The independent variables or covariates are three biological factors: the developmental state of the embryo, the maternal age and the kind of treatment the parents got. The two environmental factors are the medium in which the embryo grows and the oxygen concentration of the air around the embryo. There are three developmental states: morula, 'early blastocyst' and blastocyst. The maternal age is split into three ranges: '≤ 35', '$36 - 38$' and '≥ 39' years. The two infertility treatments that are used are the conventional in vitro fertilisation (IVF) and intracytoplasmic sperm injection (ICSI). The two tested media in the test tubes where HTF and G5. There were two oxygen concentrations used in the experiment: high, 20%, and low, 5%.

All these covariates are factors with different levels, and do not have the rich structure of the real numbers. To analyse these covariates, we need to create some dummy variables. Each level of a covariate needs to have their own dummy variable, indicating if the embryo had this level of the factor. With this dummy variables we have twelve covariates. For all of the factors we can remove one dummy variable because we can add the effect of this level to the offset. This reduces the number of covariates, but can influence the results, because it is not symmetric and there is a penalty term in the model that does not penalise the offset. For the main analysis we will use all the covariates, but for the analysis about the stability of the order, where we have to calculate several iterations, we will remove the level with the lowest number of embryos. This reduces the number of covariates from twelve to seven, which results in saving computer time by the analysis.

4.2 Results

To analyse the data, we will look at three things. First we will look at the ordering of the covariates for all genes together and all the methods, second we will see that the genes can be classified in different groups of genes, each group with their own general

trend, and thirdly we will investigate the stability of the ordering.

4.2.1 Ordering of the Covariates

We calculate the ordering of the covariates using all the twelve dummy covariates. To create an ordering of only the five real covariates we order these five covariates by the first time that a dummy covariate of this covariate is used. The results for all the three methods are shown in table 4.1. We see that the ordering differs between the

Table 4.1: The order of importance of the covariates. The first three columns are orderings given by GTM for three values of λ_2 and the last two columns are the orderings according to AICpG and MTFS. GTM gives treatment as the most important covariate and the developmental stage as the least important. AICpG and MTFS put the developmental stage on top.

	GTM			AICpG	MTFS
	$\lambda_2 = 10^{-6}$	$\lambda_2 = 20$	$\lambda_2 = \infty$		
1^{st}	Treatment	Treatment	Treatment	Dev. stage	Dev. stage
2^{nd}	Maternal age	Medium	Maternal age	Maternal age	Treatment
3^{rd}	Medium	Maternal age	Medium	Treatment	Medium
4^{th}	Dev. stage	Oxygen con.	Oxygen con.	Medium	Oxygen con.
5^{th}	Oxygen con.	Dev. stage	Dev. stage	Oxygen con.	Maternal age

methods. According to GTM for all values of λ_2 the treatment is the most important covariate and the developmental stage is least or second least important. For AICpG and MTFS, however, the developmental stage is the most important covariate. The largest difference between AICpG and MTFS is the importance of the maternal age. The differences between the orderings of GTM with different λ_2 are small.

4.2.2 Different Trends

In chapter 3 we have seen that GTM orders in a different way than AICpG and MTFS, and that this results in different orderings if there is not one general trend, but several trends, or no trend at all, because GTM is not built for data without the structure of a general trend. For the analysis we use some kind of random walk. The time in the random walk are the different genes from gene/time 1 to gene/time G. For each covariate we define the random walk rw_p by:

$$rw_p : \{0, 1, \ldots, G\} \to \mathbb{R}, \tag{4.1}$$

$$g \mapsto \sum_{k=1}^{g} \hat{\beta}_g^p. \tag{4.2}$$

So we have

$$rw_p(0) = 0 \text{ and steps } rw_p(g) - rw_p(g-1) = \hat{\beta}_g^p. \tag{4.3}$$

This random walk rw_p is an approximation of the random walk of the real B that we used in the discussion of experiment 3 on page 29.

For this random walk we expect several things. If there is no trend, we will see a normal random walk without a drift. If there is a general trend μ we have that $\mathbb{E}[\beta_g^p] = \mu_p$. $\hat{B}$ is an estimation of B so $\hat{\beta}_g^p \approx \beta_g^p$, and so there will be a drift in the direction of the sign of μ_p with a speed a little smaller than the magnitude of μ_p. The speed is not equal to the magnitude of μ_p because of the shrinkage of $\hat{B}$. If there is no general trend, we do not have a drift and $\mathbb{E}[rw_p(g)] = 0$ for all genes. If there are two groups of genes, each with their own trend, our expectation depends on whether the genes are arranged per group. If the first genes are from one group, we expect that the random walk drifts away from zero according to the first trend. If the genes of the first group are passed, we get the genes of the second group. But this group has a different trend, so the drift will be in another direction. If the genes are not ordered, we expect to see a random walk with a drift equal to the mean of the two trends and with a little bit more noise.

Figure 4.1 shows the random walks for all the twelve dummy covariates and the offset. For this plot $\hat{B}$ is used from GTM with $\lambda_2 = 20$ and $\lambda_1 = 1.4 \cdot 10^{-9}$. The other methods

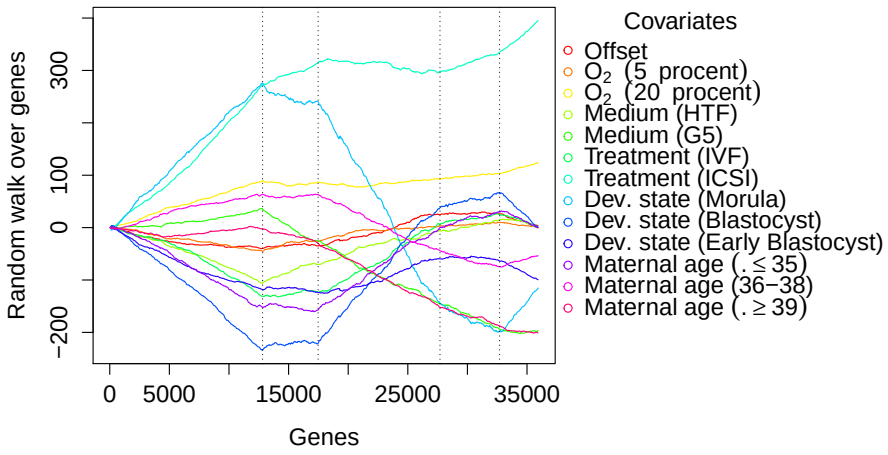


Figure 4.1: Random walks of the twelve dummy covariates. The dashed lines are the borders between different groups of genes. The $\hat{B}$ of GTM with $\lambda_2 = 20$ and $\lambda_1 = 1.4 \cdot 10^{-9}$ is used.

give comparable paths, but the size of the steps depends on the amount of shrinkage. From the graph it is clear that there is some drift for all the covariates, and that the speed of the drift differs between the covariate. The direction and speed of the drift change four times, and these changes are simultaneously for all the covariates around the genes 13,000, 17,500, 28,000 and 33,000. So we can conclude that there are five groups of genes and that the genes are ordered by these groups already.

With these different trends we can understand the differences between the order of GTM and the orders of AICpG and MTFS. GTM sets treatment on top, and the corresponding dummy covariate 'Treatment (IVF)' has the largest net drift over all the genes. According to AICpG and MTFS the developmental state is the most important. The dummy covariate 'Developmental stage (Morula)' has a large drift for most of the groups of genes, but the different trends compensate each other and there is not a large net trend over all the genes.

If we now analyse the groups separately from each other for GTM with $\lambda_2 = 20$ we get the orderings given in table 4.2. We see that with this separation in groups GTM

Table 4.2: The orderings for the five groups of genes using GTM with $\lambda_2 = 20$. The developmental stage is most important for three of the five groups and the treatment is less important.

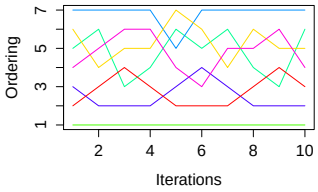
	The different groups of genes				
	1-12,822	12,823-17,477	17,478-27,700	27,701-33,000	33,001-35,927
1st	Dev. stage	Medium	Dev. stage	Maternal age	Dev. stage
2nd	Treatment	Dev. stage	Maternal age	Dev. stage	Treatment
3rd	Maternal age	Oxygen con.	Medium	Medium	Oxygen con.
4th	Medium	Maternal age	Treatment	Treatment	Maternal age
5th	Oxygen con.	Treatment	Oxygen con.	Oxygen con.	Medium

puts the developmental stage in three of the five groups on top and the treatment is less important in most groups than it was for all the genes combined. These separated orderings are more in line with the orderings of AICpG and MTFS.

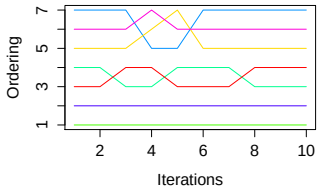
4.2.3 Stability of the Ordering

To analyse the reliability of the ordering in table 4.1 on page 33, we can look at the stability of the ordering between random subsets of genes. We take ten different subsets of 30,000 genes and for each subset we calculate the ordering according to all the methods. To reduce the computing time, we remove for each real covariate the dummy covariate corresponding to the level with the lowest number of embryos. So instead of twelve covariates we have seven covariates. This can be done because there is an offset that can include the influence of these levels. These ten orderings for all methods are plotted in figure 4.2 on the following page. Each line corresponds to a covariate, and y-value 1 means that this covariate is most important. If a line is constant it means that the corresponding covariate is of the same importance for all the ten iterations.

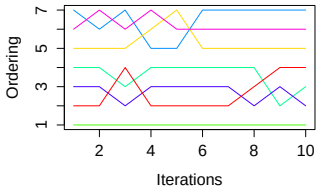
By taking subsets of genes we see that for $\lambda_2 = \infty$ in figure 4.2c on the next page that there are three groups of covariates with comparable importance. In these groups the order fluctuates, but between the groups, the order is stable. The most important groups of covariates contains only the covariate Treatment, and the second group contains the covariates Medium, 'Oxygen concentration' and 'Maternal age (≤ 35)'. The least important group contains 'Developmental stage (Blastocyst)', 'Developmental stage (Morula)'



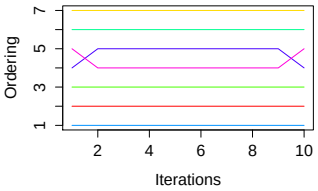
(a) Ordering according to GTM with $\lambda_2 = 0.001$.



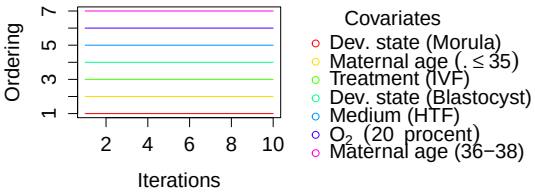
(b) Ordering according to GTM with $\lambda_2 = 20$.



(c) Ordering according to GTM with $\lambda_2 = \infty$.



(d) Ordering according to AICpG



(e) Ordering according to MTFS.

Figure 4.2: The order of the covariates for ten subsets of genes. Each line corresponds to a specific covariate, a constant line means that this covariate has the same importance over the ten iterations. Each subset contains 30,000 of the 35,927 genes.

and 'Maternal age $(36 - 38)$'. The ordering in table 4.1 on page 33 is in line with this ordering of groups of covariates.

The ordering given by AICpG is more stable as shown in figure 4.2d. Only two covariates switch position in the ten iterations. The ordering of the covariates given all the genes as shown in table 4.1 on page 33 for the six most important covariates is in line with the ordering from iteration two to nine. Only the two least important covariates have switched positions. The ordering given by MTFS is the same for all ten iterations and is equal to the ordering that we get if we use only these seven covariates and all the genes.

The instability of GTM can be understood by looking at the net trends. The net trend of several covariates is exact zero, so GTM cannot distinguish the importance of these covariates. So for different subsets of genes the order can be different. That the ordering is least stable for small λ_2 is because there is least shrinkage, so the $\hat{\beta}_g^p$ are farthest from zero, so there is more variance and different subsets are more likely to get different orderings.

4.3 Conclusions

The ordering of the covariates is different for all the methods, table 4.1 on page 33; according to GTM the treatment is the most important and for AICpG and MTFS this is the developmental stage of the embryo. For GTM however, the developmental stage is unimportant. These differences between AICpG and MTFS on one hand and GTM on the other hand indicate that the data contain different trends, because GTM orders differently if there is more than one trend. Further analysis using random walks, showed that there are five groups of genes, each of which have their own trend, figure 4.1 on page 34. A separated analysis of these different groups of genes with GTM showed that the developmental stage is important in these groups for GTM, table 4.2 on page 35.

So for further research of the gene expression of human embryos it is important to take the developmental state of the embryo and the treatment into account. Experimental covariates are less important, especially the oxygen concentration.

4.4 Discussion

By the analysis of the microarray data we found that there are five groups of genes, and even more strikingly, that the genes are grouped. We do not have a biological explanation, because the microarray does not have a special order for the genes and the processing of the data is nothing more than a normalisation. An experimental reason as a changing offset by the measurements is also not likely, because the random walks use the estimation of B and not the gene expressions. Because all methods show this behaviour it is not the failure of a method neither.

Probably related to this not understood behaviour, is that for GTM the net trend of several covariates is exact zero, and that the sums of the random walks of the dummy variables of a covariate is a straight line without noise and this line does not change between the groups of genes. But the other methods do not show these results.

All of this leads to the necessity of further research to reach a better understanding of the causes of these behaviours.

Chapter 5

Conclusion and Discussion

5.1 Summary and Conclusion

We will conclude this thesis by recalling the differences between the methods and their strengths and weaknesses. With at the end a recommendation of a method depending on the structure of the data.

MTFS The ordering of the covariates is by the average size of the effect of the covariates by the individual genes, so covariates with a large but opposite effect on groups of genes are important according to MTFS, section 3.2. The reliability of the ordering is ambiguously. If there is one trend or two trends the other methods give a more reliable ordering, figures 3.3 on page 18 and 3.11 on page 24. However, in the setting of experiment 3, where there is no trend, MTFS gives the most reliable ordering, figure 3.16 on page 28. With the gene expression data in chapter 4 is the ordering the most stable, figure 4.2 on page 36.
The estimation of B has a shrinkage toward zero, which gives a bias and reduces the variance. This reduction of the variance leads to a relative good, but still not good because of the lack of data points, estimation of B in comparison with AICpG and GTM, regardless of the presence or absence of a general trend, tables 3.1 on page 21 and 3.2 on page 29. The coefficients of $\hat{B}$ are only non-zero for the important covariates, section 2.2.

AICpG AICpG is like MTFS insusceptible for different trends and orders the covariates in a similar way as MTFS, but more reliable if there is at least one general trend, figures 3.3 on page 18, 3.11 on page 24 and 3.16 on page 28.
The estimation of the coefficients matrix B is worst of the three methods, due to overfitting, tables 3.1 on page 21 and 3.2 on page 29.

GTM The covariates are ordered by the mean effect on all the genes, so a covariate is not important if groups of genes are effected opposite, figure 3.10 on page 23. The ordering is the most reliable of the methods if there is at least one trend and if the λ_2 penalty parameter is small. The reliability is smaller for larger λ_2 and if there is no net trend, figures 3.3 on page 18, 3.11 on page 24 and 3.16 on page 28. The ordering of the real data is less reliable, because several covariates do not have a net trend, figure 4.2 on page 36.

Similar to MTFS, GTM has a shrinkage on the estimation of B. The estimation of B is the best of the three methods if there is one single trend and λ_2 is not small, table 3.1 on page 21, but MTFS is better if there is no trend, table 3.2 on page 29. GTM also estimates the general trend, section 2.4, and gives the clearest separation between groups of genes if there are two trends, figure 3.12 on page 25.

If there is a strong presumption that the data has a structure with one general trend, the most suitable method for the analysis is GTM. In the case that there is only an interest in the ordering, it is best to use a small value of the λ_2 parameter. If the estimation of B is also important, a larger λ_2 is more appropriate.

If it is not likely that there is one trend for all the genes, GTM is not suitable, unless the analysis is used to fit one model for all the genes. If the estimation of B is unimportant, the best method is AICpG. The use of MTFS is preferable if the estimated $\hat{B}$ will be used.

5.2 Future Research

The covariates are factors with different levels instead of some property with values in $\mathbb{R}$, dummy variables are needed as we have seen in chapter 4. In the analysis we treated these dummy variables as unrelated to each other. It would be better to treat them as a group of covariates that are all important or all unimportant. For MTFS this can be obtained by changing the algorithm slightly. For AICpG we need a group wise AIC method. A search in the literature did not give any results. If brute force is used to analyse all the possible subsets of covariates, a simple exclusion of subsets, that do not match with the groups, is enough. If an algorithm is used to reduce the number of subsets that needs to be analysed, it is probably harder to include the group structure. For GTM this means the usage of a groups wise LASSO as described by Yuan and Lin in [YL06], with no other changes.

The coefficients of $\hat{\beta}^p$ of an unimportant covariate p are not zero for GTM. So even if the covariate is unimportant it is still included in the model at the end. The method can be changed such that the gene-wise correction $\hat{\theta}^p$ can only be non-zero if the general trend of this covariate $\hat{\mu}_p$ is non-zero. This change will result in a minimisation problem that cannot be reduced to a LASSO problem, because $\hat{\theta}^g$ is not a linear function of $\hat{\mu}_p$ anymore.

An improvement on AICpG can be made with the use of the finite sample size correction of AIC. This AICc penalize complex models more to prevent that almost all the covariates are included if the sample size is small.

The analysis of the gene expression data in chapter 4 shows different groups of genes, with the genes already ordered in these groups. Further research is needed to understand if this is caused by the way we analysed the data or if it is caused by biological or

experimental factors.

All the methods can be extended to be suitable for more general multi-task data. Microarray data gives one single design matrix for all the regression problems of the different genes because for each gene we have the same n organisms where each gene corresponds to a task. This can be generalised to regression problems with multiple tasks where each task still has the same covariates but with their own design matrix. MTFS is introduced in this generality and Obozinski et al. used it to recognise handwritten characters. Instead of different genes they have people, and the people did not write every character just as often, so the sample size differs between the tasks. AICpG solves the regression problems of the different genes/tasks independent and the only thing we have to do is to change the weight of the votes of the tasks such that tasks with a larger sample size have more influence. For GTM only a minor change is needed in the reduction to a LASSO problem because for a given $\hat{\mu}$ are the estimations of $\hat{\theta}_g$ independent of each other.

Acknowledgements

I am grateful to my research supervisors Dr. T.A.L. van Erven and Dr. M.J. Jonker, who gave advice and guidance, and shared their knowledge and experience with me. It was my great pleasure to work with them and after all meetings I was full of ideas how to continue.

My grateful thanks are also extended to my sister Elza. She motivated me and helped me focus. In this way, she was my great help with the final writing of the thesis.
I wish to acknowledge the study support group for their assistance in scheduling and other ways of working.
Finally, I would like to thank all my family and friends, in particular my parents and my aunt Corry, who helped me with their ideas, sociability, and support.

Bibliography

[AJL+14] Brunce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science: New York, 6th edition, 2014.

[Aka74] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[Bum13] Roger Bumgarner. Overview of dna microarrays: types, applications, and their future. *Current protocols in molecular biology*, pages 22–1, 2013.

[CH+08] Gerda Claeskens, Nils Lid Hjort, et al. *Model selection and model averaging*, chapter 8.3, pages 232–235. Cambridge University Press Cambridge, 2008.

[Dră03] Sorin Drăghici. *Data analysis tools for DNA microarrays*, chapter 2, pages 15–26. CRC Press, 2003.

[EHJ+04] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[EJR+14] Iakes Ezkurdia, David Juan, Jose Manuel Rodriguez, Adam Frankish, Mark Diekhans, Jennifer Harrow, Jesus Vazquez, Alfonso Valencia, and Michael L Tress. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human molecular genetics*, 23(22):5866–5878, 2014.

[EPT07] Theodoros Evgeniou, Massimiliano Pontil, and Olivier Toubia. A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. *Marketing Science*, 26(6):805–818, 2007.

[HK70] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[JS61] William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

[MJW+16] Eleni Mantikou, Martijs J Jonker, Kai Mee Wong, Aafke PA van Montfoort, M de Jong, TM Breit, S Repping, and S Mastenbroek. Factors affecting the gene expression of in vitro cultured human preimplantation embryos. *Human Reproduction*, 31(2):298–311, 2016.

[OTJ06] Guillaume Obozinski, Ben Taskar, and Michael Jordan. Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, 2006.

[S+56] Charles Stein et al. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206, 1956.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[WGS09] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[YL06] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.