Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Lene Tillerli Omdal

Statistical Analysis of the Association between MicroRNAs in Breast Milk, Perinatal Probiotic Supplement and the Development of Atopic Dermatitis

Master's thesis in Applied Physics and Mathematics
Supervisor: Turid Follestad
Co-supervisor: Melanie Rae Simpson and Mette Langaas

June 2022

NTNU
Norwegian University of
Science and Technology

Lene Tillerli Omdal

Statistical Analysis of the Association between MicroRNAs in Breast Milk, Perinatal Probiotic Supplement and the Development of Atopic Dermatitis

Master's thesis in Applied Physics and Mathematics
Supervisor: Turid Follestad
Co-supervisor: Melanie Rae Simpson and Mette Langaas
June 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

The Probiotics in the Prevention of Allergy among Children in Trondheim (ProPACT) study showed a 40% reduction in the risk of developing atopic dermatitis in children whose mothers received a probiotic supplement before and whilst breastfeeding, compared to children whose mothers received a placebo alternative. The biological explanation for this risk reduction has not yet been fully understood. In this thesis, we analyze if microRNAs in breast milk, measured 10 days postpartum, are possible contributors to the risk reduction. We perform two analyses; one in which we examine the effect of the probiotic supplement on microRNAs and one in which we examine whether any microRNAs are associated with development of atopic dermatitis by 2 years of age. In addition, we perform a clustering analysis to explore patterns and groupings in the data.

The analysis in this thesis is based on data from 60 mother-child pairs which were semi-randomly selected from the ProPACT study. We use differential expression analysis to investigate if the probiotic supplement has an effect on the expression values of individual microRNAs. To perform this analysis we employ the statistical method *voom*. To investigate whether any microRNAs are associated with the development of atopic dermatitis, we perform variable selection using an elastic net model. We investigate two different methods, nested cross-validation and repeated cross-validation, to find the model parameters of the elastic net model. We proceed with repeated cross-validation to estimate confidence intervals of the coefficients by employing bootstrapping and the accelerated bias-corrected method. In the exploratory clustering analysis, we use hierarchical clustering with Euclidean and correlation based dissimilarity measures.

The probiotic supplement is associated with differential expression of one microRNA, *miR*-577, when taking into account the multiplicity of tests by controlling the false discovery rate at 10% using the Benjamini and Hochberg method. In total, 47 microRNAs have a raw p-value below 0.05, but except for *miR*-577, none of them have an acceptable false discovery rate adjusted p-value. The five microRNAs *miR*-342-3p, *miR*-3605-3p, *miR*-500a/b-5p, *miR*-625-3p and *miR*-6515-5p are associated with development of atopic dermatitis, i.e. they have an estimated 95% confidence interval that does not include zero. The microRNA *miR*-3605-3p is also one of the microRNAs with a raw p-value below 0.05 in the analysis of the effect of probiotics. In the cluster analysis *miR*-3605-3p, *miR*-6515-5p and *miR*-577 are grouped together, indicating that they are correlated. In conclusion, one microRNA is found to be affected by the probiotic supplement, and five microRNAs are found to be associated with atopic dermatitis in breast milk at 10 days postpartum. However, we found no conclusive evidence that probiotics affect the same microRNAs that are associated with atopic dermatitis. Two of the miRNAs associated with atopic dermatitis are, however, clustered with the single microRNA found to be affected by the probiotics. Further studies may consider focusing on the microRNAs that were the most promising.

Sammendrag

Studien Probiotics in the Prevention of Allergy among Children in Trondheim (ProPACT) viste en reduksjon på 40% i risikoen for å utvikle atopisk dermatitt for barn der mødrene fikk et probiotisk tilskudd før og mens de ammet, sammenlignet med barn der mødrene fikk et placebo alternativ. De biologiske mekanismene bak denne forebyggende effekten er ennå ikke kjent. I denne masteroppgaven undersøker vi om mikroRNA i morsmelk, målt 10 dager etter fødsel, er mulige bidragsytere til denne risikoreduksjonen. Dette gjør vi ved å utføre to analyser; en der vi undersøker effekten av det probiotiske tilskuddet på mikroRNAene i morsmelken og en der vi undersøker om noen av mikroRNAene i morsmelken er assosiert med utvikling av atopisk dermatitt hos barnet innen fylte 2 år. I tillegg utfører vi en klyngeanalyse for å undersøke mønstre og grupperinger i dataene.

Analysen er basert på data fra 60 mor-barn par utvalgt fra ProPACT-studien, delvis tilfeldig utvalgt og delvis basert på noen seleksjonskriterier. For å undersøke effekten av det probiotiske tilskuddet undersøker vi om individuelle mikroRNA er ulikt uttrykt hos mødrene som fikk probiotika sammenlignet med mødre som fikk en placebo. Denne analysen er basert på den statistiske metoden *voom*. For å undersøke om det er en assosiasjon mellom mikroRNAene og utvikling av atopisk dermatitt, bruker vi en elastic net regresjonsmodell. Vi undersøker to ulike metoder, nestet kryssvalidering (nested cross-validation) og repetert kryssvalidering (repeated cross-validation), for å bestemme modellparametrene i elastic net modellen. Videre bruker vi den repeterte kryssvalideringen for å estimere konfidensintervall for de estimerte koeffisientene. Dette gjør vi ved å bruke bootstrapping og en metode kjent som accelerated bias-corrected method for å korrigere for bias. I den eksplorative analysen bruker vi hierarkisk klyngeanalyse med både Euklidisk og korrelasjonsbaserte distansemål.

Det probiotiske supplementet er signifikant assosiert med ulike utrykksverdier for ett mikroRNA, *miR*-577, når vi kontrollerer andelen forventede falske positive til å være 10%, ved å bruke Benjamini og Hochberg metoden. Totalt har 47 mikroRNA en (ikke justert) *p*-verdi under 0.05, men med unntak av *miR*-577 har ingen av dem en akseptabel *p*-verdi etter justering der forventet andel falske positive kontrolleres til å være 10%. Fem mikroRNA er assosiert med utvikling av atopisk dermatitt, dette er *miR*-342-3*p*, *miR*-3605-3*p*, *miR*-500*a*/*b*-5*p*, *miR*-625-3*p* og *miR*-6515-5*p*. Det vil si de har estimerte 95% konfidensintervall som ikke inkluderer null. MikroRNAet *miR*-3605-3*p* er også et av mikroRNAene med en (ikke justert) *p*-verdi under 0.05 i analysen av effekten til probiotika. Klyngeanalysen grupperer *miR*-3605-3*p*, *miR*-6515-5*p* og *miR*-577 sammen, noe som indikerer at de er korrelerte. Vi finner altså ett mikroRNA som blir påvirket av det probiotiske tilskuddet og fem mikroRNA som er assosiert med atopisk dermatitt. Vi fant imidlertid ingen bevis for at det probiotiske tilskuddet påvirker de samme mikroRNAene som er assosiert med atopisk dermatitt. To av mikroRNAene assosiert med atopisk dermatitt er imidlertid gruppert sammen med det ene mikroRNAet signifikant påvirket av det probiotiske tilskuddet. Ytterligere studier kan vurdere å fokusere på de mikroRNAene som ser mest lovende ut.

Preface

This thesis finishes my Master of Science (M.Sc.) in Applied Physics and Mathematics with specialization in Industrial Mathematics and Statistics at the Norwegian University of Science and Technology (NTNU) in Trondheim. This thesis concludes the course TMA4900; Industrial Mathematics, Master's Thesis. The dataset and medical hypothesis were provided by Melanie Rae Simpson at St. Olavs Hospital in Trondheim. The topic of this thesis was to analyze if miRNAs in breast milk are contributors to the risk reduction in children developing atopic dermatitis associated with a perinatal probiotic supplement. The work in this thesis was carried out during the spring semester of 2022. However, it is a continuation of the introductory work done in my project thesis during the fall of 2021.

First, I would like to thank my supervisor Turid Follestad for excellent guidance, input and feedback throughout this last year. I would also like to thank Melanie Rae Simpson for all her guidance, especially regarding the biological aspects of this thesis and for letting me aid in their ongoing research. Additionally, I would like to direct a huge thanks to Mette Langaas for providing both valuable feedback and assistance along the way. I greatly appreciate the help all of you have given me while writing this thesis. In truth, I would not have been able to come this far without the support I have received from my friends and family. I would like to express my sincere gratitude to Arne Rustad for all the valuable discussions and motivational words this year and for always supporting me, and lastly I would like to thank my parents for their constant support and help throughout my entire education.

Contents

Chapter 1

Introduction

Atopic dermatitis (AD) is the most common type of eczema and occurs in approximately 20% of children and approximately 10% of 18 year old’s (Lyngra 2015). AD is a chronic, inflammatory skin condition where the skin appears red, itchy and dry. At times the skin may develop rashes with swelling and exuding blisters. About 50% of those with AD will also develop other allergic diseases, such as asthma and allergic rhinoconjunctivitis (hayfever).

The study done by Asher et al. (2006) is one of many studies showing that the prevalence of allergic diseases, including asthma and AD, is increasing. The hygiene hypothesis has been proposed as an explanation for this increase. The hygiene hypothesis initially proposed that early exposure to infections may strengthen the immune system. More recently, a greater emphasis has been placed on a reduced diversity of microbial exposures in early life as a driving factor behind the increased rates of allergic disease, rather than lower rates of infectious disease per se. Breastfeeding or nursing is known to protect against infection and inflammation, and assists with the development of the infants’s immune system. On the contrary, alternatives to breast milk, such as infant formula, or short breastfeeding duration have been associated with an increased risk of developing type 1 diabetes, celiac disease, some childhood cancers, inflammatory bowel disease and allergic diseases (Nuzzi et al. 2021). This shows that breast milk not only has short term effects, but may also have long term consequences by playing a role in the development of the immune system.

In the study by Dotterud et al. (2010), a Trondheim based research team investigated whether giving a probiotic supplement to the mother during the later stages of pregnancy and whilst breastfeeding reduces the risk of AD in the child in a randomized trial, called the Probiotics in the Prevention of Allergy among Children in Trondheim (ProPACT) study. Probiotics are living microorganisms, or bacteria, that in adequate amounts, according to the World Health Organization, can be beneficial to the health of the consumer (Sirevåg 2021). The ProPACT study showed a 40% risk reduction for development of AD in the child after 2 years when the mother were given a probiotic supplement compared to those provided with a placebo alternative. Although the positive effect of a perinatal probiotic supplement is established, the explanation behind it is still unknown.

Many components in breast milk have been suggested as contributors to its long term effect, including growth factors, immunoglobulins, chemokines, cytokines and recently microRNAs. Breast milk contains large amounts of total RNA and a large proportion of microRNA (miRNA) in breast milk are considered "immune related" (Simpson et al. 2015). In a previous study from the same Trondheim based research group, the role of breast milk miRNAs as a mediating factor in the preventative effect of probiotics, was investigated using a subset of breast milk samples collected 3 months after birth from women participating in the ProPACT study (Simpson et al. 2015). Of the 415 women who participated in the ProPACT study, some were lost to follow-

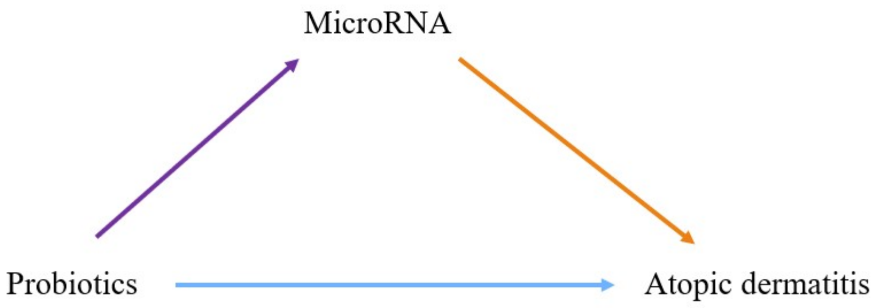


Figure 1.1: Figure describing the two analyses in this thesis. The arrow from probiotics to atopic dermatitis represents the established effect of probiotics; that probiotics reduce the risk of the child developing AD (Dotterud et al. 2010). The arrows from probiotics to miRNA, and from miRNA to atopic dermatitis represents the two different analyses performed in this thesis.

up and not all 3 month samples were available. From the eligible mother-child pairs 54 were semi-randomly selected, based on the selection criteria described in Simpson et al. (2015). For each of the 54 mother-child pairs the breast milk samples were analyzed to quantify their miRNA profile. After filtering out lowly expressed miRNAs they were left with 125 miRNAs. For each of the 125 miRNAs they fit a linear model using both the *limma* (Smyth 2004) and *voom* (Law et al. 2014) methods implemented in the R (R Core Team 2021) package *limma*. The purpose was to analyze differential expression between mothers who received a probiotic supplement and a placebo, but also between those with a child who developed AD and those whose child did not. In addition they examined a model adjusting for maternal atopy and the presence of older siblings. It was the largest sequencing analysis of human breast milk small RNA to that date and they were able to characterize the general profile of breast milk miRNA at 3 months postpartum. They did not observe conclusive evidence of differential expression between the probiotic and placebo groups, nor between the AD groups. They concluded that individual breast milk miRNAs at 3 months postpartum are unlikely to play a major role in the prevention of AD.

In this thesis, the aim is to further investigate miRNAs in breast milk 10 days postpartum as possible contributors to reducing the risk of developing AD. We choose to separate the research question into two distinct analyses; one where we analyze the effect of a perinatal probiotic supplement on miRNAs in breast milk and one where we analyze if any miRNAs in breast milk are associated with the child developing AD. The link between these two analyses and the previously established effect of the perinatal probitic supplement is shown in Figure 1.1. In addition to these two analyses, we perform an exploratory cluster analysis to study patterns in the data and provide an overview of the similarities between measured miRNA counts and between samples.

Some previous work was conducted on the effect of perinatal probiotic ingestion on the expression of breast milk miRNAs 10 days postpartum in a project thesis during the fall semester of 2021. This analysis is improved and expanded upon in this thesis. The part of the analysis examining the effect of the probiotic supplement on miRNAs considers only the effect on individual miRNAs. As in the study by Simpson et al. 2015, the statistical methods and R packages *limma* and *voom* are used to carry out the analysis.

limma is a widely used method and package for analyzing differential expression of RNA-sequences between conditions. However, it only provides a way to test for differential expression of individual miRNAs and does not take the correlation and cooperation between them into account. In the second part of the analysis, when examining whether any miRNAs are associated with the development of AD in the child, we perform a multivariable analysis instead. By using a multivariable analysis we are able to perform joint identification of miRNAs. As in many medical analyses, we are dealing with a high dimensional dataset and we want a way to select important miRNAs in breast milk with respect to development of AD. In this analysis we use elastic net (Zou and Hastie 2005) for this purpose as it has proven to be an effective way to perform variable selection of correlated variables and is applicable for high dimensional problems (Cho et al. 2010; Giglio and Brown 2018; Pavlou et al. 2016). Elastic net was introduced by Zou and Hastie (2005) and they were the first to demonstrate its ability to identify all variables associated with the response, even if they are correlated. As miRNAs usually are highly correlated and we wish to perform joint detection, we aim to perform variable selection by fitting an elastic net model to the data. The elastic net model was only applied for its automatic variable selection and not to create a predictive model for atopic dermatitis.

The structure of this thesis is as follows: In Chapter 2, we present relevant biological theory, in particular, we explain what miRNAs are and the methodology for quantifying their expression values. In Chapter 3, we describe the theory underlying the statistical methods and concepts used in the analyses. This includes hierarchical clustering, generalized linear models, multiple hypotheses testing, *limma*, elastic net regularization and resampling methods. Then, in Chapter 4, we present the dataset of the breast milk samples collected 10 months postpartum, explain the filtering procedure and present the two normalization methods used to perform pre-processing. In Chapter 5, we present the methods of the two main analyses and the exploratory cluster analysis. This includes an improved analysis of the probiotics effect on miRNAs using *limma* and the elastic net model used for selecting miRNAs associated with AD. We propose a procedure using cross validation to determine the model parameters of an elastic net model, and compare this to a more standard method. Additionally, we present a modification of the computation of the bias correction factor in the bias-corrected accelerated method for computing confidence intervals that makes it more applicable for the case of coefficient estimates of an elastic net model. Lastly, in Chapter 7, we present a discussion of the results, the methods used, ideas for further work and a conclusion.

Some parts of this thesis are inspired by work from the project thesis. These are Chapter 2, Section 3.3, 3.4, 4.2.1 and some of the analysis in Section 5.2.

Chapter 2

Biological Background

2.1 MicroRNAs and Gene Expression

MiRNAs are small molecules that aid in the process of regulating gene expression. Genes are segments of deoxyribonucleic acid (DNA), known as the hereditary material in almost all organisms. This is where all the information needed to maintain and build organisms is stored.

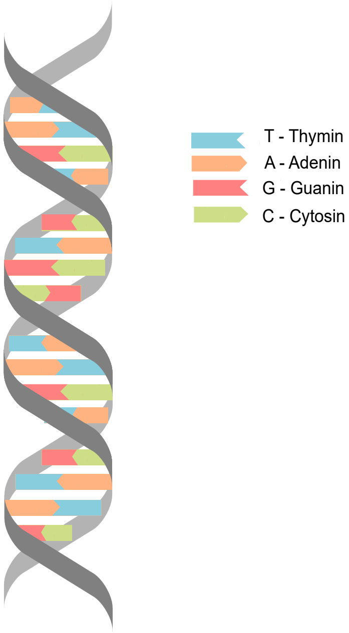


Figure 2.1: Figure showing the basic helix structure of the DNA including the four bases; thymin, adenin, guanin and cytosin.

The DNA is a nucleic acid, which is a class of molecules found in cells and viruses, that store and control the expression of genetic information. DNA has a helix structure and an illustration of this structure is shown in Figure 2.1. The information contained inside the DNA is stored using only four chemical bases; adenine (A), guanine (G), cytosine (C), and thymine (T), and a sugar phosphate backbone. As shown in the figure, these bases pair up with each other and form two long strands of base pairs that coil around each other in a double helix structure. Some sequences of base pairs in the DNA code for proteins, these sequences are what we call genes. The process of making proteins from DNA is known as protein synthesis and consists of two steps: the first is transcription and the second is translation. During transcription, sequences of the DNA are "read", and the information is copied to messenger RNA (mRNA). The mRNA works, as the name suggests, as a messenger and carries the information out of the cell nucleus to the ribosome where the translation step occurs. During translation, the order of the base pairs in the mRNA is used to guide the formation of proteins. However, only 1% of the DNA sequences in a human cell code for proteins, leaving about 99% as non-coding.

All cells contain the same DNA and thus the same genes. However, cells are able to have different structures and functions because different cells transcribe different genes and thus will make different proteins. Once a gene is transcribed, we say that the gene is expressed. MiRNA is a type of non-coding RNA that regulates gene expression post transcriptionally. They inactivate mRNAs, which are necessary for translating a gene into a protein. The miRNA binds to the mRNA and inhibits translation either by staying bounded or cutting the molecule, ultimately leading to the cell destroying it. In both cases, the gene cannot be translated. The miRNAs can decrease gene expression to different degrees, from slightly downregulated to completely turned off. Additionally, miRNAs are small molecules and one type of miRNA are able to bind multiple mRNAs, thus one miRNA can affect multiple genes.

2.2 Quantifying Expression Values

To quantify the expression level of RNA, miRNA or any other nucleic acid we use expression analysis. We use the words expression value or expression level to describe how much miRNA is present in a sample. Performing microarray experiments is one way of quantifying these expression levels, and has driven many statistical advances in high-dimensional data analysis. However, in recent years, RNA sequencing or RNA-seq has emerged as a new way of measuring the expression levels (Datta and Nettleton 2014). RNA-seq is the method used to obtain the data used in this analysis. However, since *limma*, which is described in Section 3.4, was developed for data resulting from microarray experiments, we also give a brief description of this experiment here.

2.2.1 Microarrays

Microarrays are laboratory tools used for measuring expression levels in thousands of DNA sequences at the same time. These DNA sequences can be any kind of genomic feature such as genes or miRNAs. Microarrays are usually made of glass and on the scale of a microscopic slide. The glass is covered with thousands of small spots, called probes, each containing a small amount of a specific and unique DNA sequence. There are two different types of microarray experiments, one-color or two-color. In one-color microarrays, sometimes called single-channel microarrays, each sample interacts separately with a microarray. In two-color microarrays, two samples interact with the same microarray at the same time. Two-color microarrays are rarely used anymore, so we consider only the single-color microarray experiment here.

The mRNA strands are first converted to complementary DNA (cDNA) and marked with a fluorescent dye (Bumgarner 2013). Then we let the cDNA strands in the sample freely interact with the probes. The strands are then able to bind to their corresponding DNA-sequence on these probes. When they do bind, the fluorescent dye is released. Therefore, measuring the intensity of the color at each probe is a way to measure the expression value of each DNA-sequence. Since the resulting measurements are of intensities, the measured expression values are necessarily continuous. We can replicate this experiment using different samples to obtain a matrix that contains the expression value of each RNA in each sample. These data are then often used to test for differential expression between a predefined grouping of the samples.

2.2.2 RNA-sequencing

RNA-seq is a recently developed technique and has become the preferred approach for quantifying the amount of RNA present in a sample. The explanation of RNA-seq in this section is mainly based on the theory in Chapter 2 in the book by Datta and Nettleton (2014). One of the most profound differences between RNA-seq and microarray experiments is the type of data they produce. As previously stated, expression levels are quantified on a continuous scale in microarray experiments, while in RNA-seq the expression levels are quantified as discrete counts. A disadvantage of microarrays is that we need to have knowledge of the target sequences in advance to construct the probes. Since this is not needed for RNA-seq, it is more suitable for the discovery of new transcripts. Overall, RNA-seq offers an accurate and cost effective way of studying expression value of all RNA molecules. To perform RNA-seq, there are different types of platforms that can be used. Regardless of which platform is used, most RNA-seq experiments consist of the same main steps; preparing a sequencing library, sequencing and data analysis.

Firstly, the sequencing library is prepared. A sequencing library is a collection of millions of DNA fragments to be used in the sequencing step. When preparing a sequencing library the RNA is isolated, and then fragmented into smaller molecules. This is due to the machine (used to perform the sequencing) only being able to sequence small fragments and the RNAs can be thousands of bases long. In small RNA sequencing however, like the sequencing of miRNAs, the molecules are already so small that the fragmenting step is not needed. The commonly used platform Illumina only sequences DNA, so the single stranded RNAs are reverse transcribed to cDNAs, which are then complemented to double stranded DNAs. Double stranded DNA is more stable and is easier to amplify and modify. Sequencing adaptors are added to the ends of the fragments. This is done mainly for the machine to recognize the fragments but also to make it possible to sequence multiple samples at the same time.

Following the amplification is the sequencing step, where base pairs from ends of the DNA fragments are read. The DNA library is amplified using the polymerase chain reaction (PCR). The polymerase chain reaction is a method for rapidly making thousands to millions of copies of a DNA sequence. This enables small samples of DNA to be studied in detail (Datta and Nettleton 2014, Chapter 2). Still considering the Illumina platform, the library is loaded onto what is called a flow cell where the DNA fragments are laid out vertically in a grid. The machine has fluorescent probes which are color coded according to base type. These probes attach to the first base of each DNA sequence on the flow cell. The fluorescent light is then imaged, and the probes move to the second base. This process continues until the machine has fully determined each sequence. As a result, we are able to count the instances of each sequence.

The resulting data are usually pre-processed before statistical analysis, typically by normalization and filtering. This is to accommodate for both technical and biological variation. Exactly how pre-processing is performed in this data analysis is explained in Chapter 4.2 , but we discuss some of the biological and technical reasoning behind it here. When analyzing RNA-seq data, it is essential to consider the sequencing depth. Jiang et al. (2019) defines sequencing depth as the ratio of the total number of bases obtained by sequencing to the size of the genome or the average number of times each base is measured in the genome. In other words, sequencing depth can be seen as a measure of how long we keep counting the number of bases. If we have a large sequencing depth we are able to count more of the bases, and thus more of the miRNAs, and we will have a higher total count of miRNAs in that sample compared to a sample with a smaller sequencing depth. The sequencing depth thus affects sequencing cost, genome or sample coverage, expression levels and more. Since the sequencing depth can vary between samples it is suggested to scale the counts to library size as a form of normalization. The library size is the total number of reads in one sample and is dependent on the sequencing depth. It makes intuitive sense to normalize according to library size, as it is expected that sequencing a sample to half the depth will give, on average, half the number of reads (Robinson and Oshlack 2010).

Chapter 3

Statistical Methods

In this chapter we present the statistical theory underlying the methods used to analyze the data.

3.1 Hierarchical Clustering

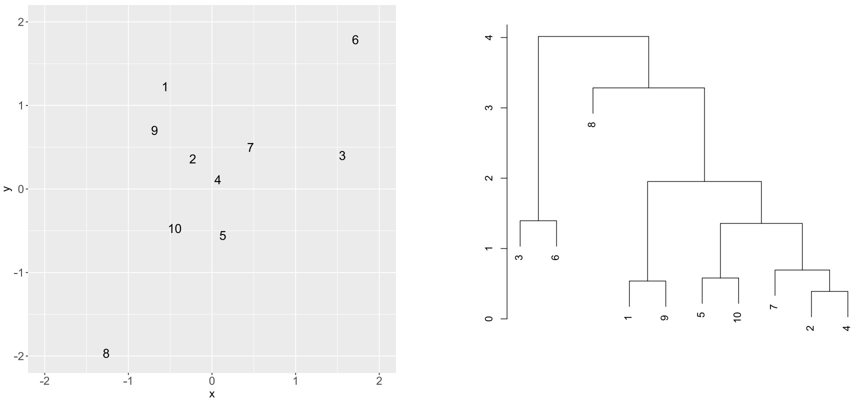
Cluster analysis, or simply clustering, refers to a set of methods for grouping similar objects. Clustering is an unsupervised learning method, and is often used as part of an exploratory analysis as it allows for visualization of the data and can reduce computational complexity in further analysis. Clustering is often used in medical applications to look for subgroups within the patients or within variables in order to better understand a disease.

Hierarchical clustering is a type of cluster analysis and is commonly used to find patterns in expression data of genes or other RNA sequences. It is important to note that clustering methods aim to group the data and will do so even if there is no real grouping present. We first describe the hierarchical clustering algorithm in Section 3.1.1. This section is based on Chapter 12 in the book by James et al. (2013). Secondly, in Section 3.1.2 we define three dissimilarity measures that can be used for hierarchical clustering. Lastly, we define the Ward linkage and explain the functionality of a linkage function in Section 3.1.3.

3.1.1 The Hierarchical Clustering Algorithm

We consider the most common type of hierarchical clustering, called agglomerative or bottom-up clustering. This type of clustering can be represented by an upside-down tree, called a dendrogram. A dendrogram provides an interpretable illustration of the recursive merging done by the algorithm and can be viewed as a graphical summary of the data (Hastie, R. Tibshirani and Friedman 2009). An example of a dendrogram obtained from simulated data is shown in Figure 3.1. Also shown is a plot of the simulated data points. The dendrogram consists of what we call branches and leaves. The leaves are objects we want to cluster, in this case the simulated data points, and the branches represent a merging of the observations into a cluster, or a merging of clusters.

In the example above there are two variables and 10 observations. In other cases there might be many variables as well. Depending on what we are interested in, we can perform clustering on the observations or the variables. For the continuing explanation, we consider clustering the variables. The hierarchical clustering algorithm starts off by considering each variable as its own separate cluster. With N variables, we start with N clusters. Then the two most similar clusters are merged into one new cluster. We now have $N-1$ clusters. Next, the two most similar clusters in this new set are merged. This process continues until there is



(a) Plot of 10 simulated data points from a bivariate standard normal distribution.

(b) Hierarchical clustering denrogram of the siulated data using the Euclidean distance and the Ward linkage method.

Figure 3.1

only one big cluster left, containing all the variables. The hierarchical clustering algorithm is described in Algorithm 1. As seen in the dendrogram in Figure 3.1 (b), the two points 2 and 4 are merged first. These are also the two closest points in Figure 3.1 (a). However, dendrograms can be misleading. As seen in Figure 3.1 (a), there does not seem to be any clear grouping of the data points, which makes sense as all the points are simulated from the same bivariate standard normal distribution. However, in the dendrogram in Figure 3.1 (b) it might look like the two observations 3 and 6 are different from the rest, and that there are two or three natural clusters in the data.

Algorithm 1: Hierarchical clustering algorithm

```
Input  : Data set of $N$ observations, dissimilarity measure and linkage function
Let each variable be treated as a singleton cluster    /* Start off with $N$ clusters */
while There are more than one cluster left do
    Calculate the distance matrix between clusters
    Merge the two most similar clusters
end
Output: Dendrogram
```

Hierarchical clustering differs from other clustering methods in that we do not need to specify the number of clusters beforehand. To identify clusters we "cut" the tree horizontally, after the clustering is performed. For example, if we cut over three branches we get three clusters. The height of a branch indicates how different the two clusters merged by this branch are. Thus, a higher branch indicates a greater difference between the two. The visual representation of the dendrogram can be used to determine the best place to cut the tree. There exist many other methods and statistics for determining the number of clusters such as the Gap Statistic, Silhouette Coefficient or the elbow method (R. Tibshirani et al. 2001; Rousseeuw 1987; Thorndike 1953). However, we will not consider those methods in this thesis.

One downside to hierarchical clustering is that there is an underlying assumption of a hierarchical structure of the clusters. This indicates that cutting a tree at a greater height will result in merging of the clusters that would be obtained by cutting at a lower point. In some settings this assumption may be unrealistic. Consider a dataset that contains information about a population of people where the best splitting of people into two groups is by gender, but into three groups is by working status (in school, working or retired). Then since these groups are not nested they will not be well represented by a hierarchical structure. In other words, the algorithm behaves in a way such that if two observations have been clustered at a lower level they will never be separated, even though this might be optimal at a higher level.

3.1.2 Dissimilarity Measures

To determine which variables are similar we define a dissimilarity measure. The Euclidean distance, between two variables, $\mathbf{x}_1$ and $\mathbf{x}_2$, measured over n observations is defined as

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{n} (x_{j,1} - x_{j,2})^2},$$

where $x_{j,1}$ and $x_{j,2}$ are the values of $\mathbf{x}_1$ and $\mathbf{x}_2$ for observation j.

It is also common to define a dissimilarity measure based on correlation. This allows for variables with similar shaped variable profiles across different samples to be considered similar, rather than only comparing magnitudes. Since we want highly correlated variables to be considered similar, let the correlation based distance be defined as

$$d(\mathbf{x}_1, \mathbf{x}_2) = 1 - \mathrm{Corr}(\mathbf{x}_1, \mathbf{x}_2). \tag{3.1}$$

If we want to consider negatively correlated variables to be as similar as positively correlated variables, we can use the absolute value of the correlation. Then the absolute correlation based distance can be defined as

$$d(\mathbf{x}_1, \mathbf{x}_2) = 1 - |\mathrm{Corr}(\mathbf{x}_1, \mathbf{x}_2)|. \tag{3.2}$$

3.1.3 The Ward Method

A linkage function determines how we measure the dissimilarity between clusters containing multiple variables. Common types of linkages include complete, average, single, centroid and Ward.

The Ward method was introduced by Ward (1963), and differs from the other linkage functions in that it does not use the distance between clusters to group them. Instead, it merges the two clusters that result in the minimum increase in total within cluster variation. The within cluster variation can be measured using the error sum of squares. The error sum of squares of a cluster with p variables, $\mathbf{x}_1, \ldots, \mathbf{x}_p$, with n observations is defined as

$$\mathrm{ESS} = \sum_{k=1}^{p} ||\mathbf{x}_k - \mathbf{m}||^2, \tag{3.3}$$

where $\mathbf{m}$ is a vector of the mean for each observation over all variables, or the center of the cluster and $\mathbf{x}_k$ is a vector of observations of variable k. If we consider two clusters, A and B, then the change in error sum of squares is defined by Ward (1963) as

$$\Delta \text{ESS}_{A\cup B} = \sum_{i\in A\cup B} ||\mathbf{x}_i - \mathbf{m}_{A\cup B}||^2 - \sum_{i\in A} ||\mathbf{x}_i - \mathbf{m}_A||^2 - \sum_{i\in B} ||\mathbf{x}_i - \mathbf{m}_B||^2.$$

This is also known as the Ward distance. If we have a total of n clusters, choosing which clusters to merge in the next step requires an evaluation and comparison of the Ward distance of each of the $n(n-1)/2$ possible unions.

3.2 Linear and Generalized Linear Models

Regression models aim to model the effect of a set of explanatory variables, also called covariates, $\{x_1, x_2, \dots, x_p\}$ on a response variable y. The purpose can be either prediction of the response variable or to gain insight into how the explanatory variables influence the response. Different models are needed for different types of response and explanatory variables as they both can be either continuous, binary, categorical or discrete. We first explain regression in its simplest form, which is linear regression. Then, the generalized linear model framework is presented and lastly we describe logistic regression, which is a generalized linear regression model for binary response variables.

3.2.1 Linear Regression

The theory presented in this section is based on Chapter 2 in the the book *Regression Models, Methods and Applications* by Fahrmeir et al. (2007) unless otherwise stated. Using linear regression models we assume that the response is a linear function of the explanatory variables. Consider n observation pairs $(\mathbf{x}_j, y_j)$, for $j = 1, 2, \dots, n$, where $\mathbf{x}_j = [x_{j,1}, \dots, x_{j,p}]^T$ is a vector of p explanatory variables. An intercept is included in the model by letting one of $x_{j,k}$ be equal to 1 for all $j = 1, 2, \dots, n$. We assume that the response variable of observation j can be written on the form

$$\begin{aligned} y_j &= \text{E}(y_j | x_{j,1}, \dots, x_{j,p}) + \varepsilon_j \\ &= \beta_1 x_{j,1} + \dots + \beta_p x_{j,p} + \varepsilon_j \\ &= \boldsymbol{\beta}^T \mathbf{x}_j + \varepsilon_j, \end{aligned}$$

where $\boldsymbol{\beta} = [\beta_1, \dots, \beta_p]^T$ is a vector of unknown coefficients. The random component of the response variable, ε_j, is a stochastic variable with expectation $\text{E}[\varepsilon_j] = 0$ and variance $\text{Var}[\varepsilon_j] = \sigma^2$, which we assume to be independent and identically distributed. Thus, the relationship between the response and the covariates is not deterministic and implies that y_j also is a stochastic variable whose distribution depends on the covariates. For one of the most common forms of multiple regression, normal multiple regression, we additionally assume that the error terms follow a normal distribution. This is the regression model we will use in the *limma* method which is presented in Section 3.4. As the response y_j given $\mathbf{x}_j$ is a linear combination of a constant, $\boldsymbol{\beta}^T \mathbf{x}_j$, and a random variable, ε_j, the response is normally distributed with mean $\boldsymbol{\beta}^T \mathbf{x}_j$ and variance σ^2. Although we consider the expected value of the response given the covariates, we often omit this in the notation and just write $\text{E}[y_j]$.

The model including all observations can be written on matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = [y_1, \ldots, y_n]^T$ is the $n \times 1$ vector of responses, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ is the $n \times p$ design matrix, $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients and $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]^T$ is the $n \times 1$ error term vector. The assumptions for the error vector can be written as $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. This results in $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{Var}[\mathbf{y}] = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

3.2.2 The Generalized Linear Model

A generalized linear model (GLM) is, as the name suggests, a generalization or expansion of a linear model. The linear regression model, described in the previous section, and regression models for non-normal response variables can be described using a unified framework. This is what we call the GLM-framework. The GLM framework consists of three main elements:

1. a random component equal to the response variable y with a distribution belonging to the exponential family,
2. a systematic component equal to the linear predictor η and
3. a link function g, which connects the mean of the response, $\mu = \mathrm{E}(y)$, to the linear predictor η.

We will now further elaborate on these three elements. The exponential family is a family of distributions including the Gaussian, Poisson, binomial, multinomial, gamma, beta and many other distributions. The density of a univariate exponential family for the response variable y is defined by Fahrmeir et al. (2007) as

$$f(y|\theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} w + c(y, \phi, w)\right).$$

The parameter θ is called the canonical parameter and the second parameter, ϕ, is a dispersion parameter. The value w is usually known. The functions $b(\theta)$ and $c(y, \phi, w)$ are also known, and it is required that the first and second derivatives, $b'(\theta)$ and $b''(\theta)$, exist. The parameter θ is the parameter of main interest and is connected to the linear predictor, while ϕ is often of secondary interest. It can be shown that θ is related to the mean and the variance of the distribution by

$$\mathrm{E}(y) = \mu = b'(\theta), \qquad \mathrm{Var}(y) = \frac{\phi b''(\theta)}{w}. \tag{3.4}$$

The linear predictor is defined as

$$\eta = \mathbf{x}^T\boldsymbol{\beta},$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_p]^T$ is the vector of p explanatory variables or $p - 1$ variables plus intercept, and $\boldsymbol{\beta}$ is the $p \times 1$ vector of the corresponding unknown coefficients. If we again consider the n observation pairs, $(\mathbf{x}_j, y_j)$, for $j = 1, 2, \ldots, n$, the linear predictor including all observations can be written as

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

where $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_n]^T$ and $\mathbf{X}$ is a $n \times p$ design matrix with $\mathbf{x}_j^T$ as the rows.

The link function connects the linear predictor to the mean of the distribution by

$$\eta = g(\mu).$$

The canonical link function is a function that connects the mean, μ, to the canonical parameter of the exponential family, that is $\theta = g(\mu)$. We further elaborate on the logit link function, the canonical link function for logistic regression in Section 3.2.3. Other link functions than the canonical can also be used, however, the domain of the link function should match the domain of the mean, μ. The linear regression model can be formulated as a GLM by using the identity link function, linking the linear predictor directly to the mean, and using the normal distribution as the probability distribution.

Estimating Regression Coefficients

The coefficients, $\boldsymbol{\beta}$, are unknown and need to be estimated. The standard way to do this is by maximizing the likelihood function, that is, maximizing the likelihood of observing the data we have, given that our assumptions of a GLM are satisfied.

Consider n independent observations $\mathbf{y} = [y_1, \ldots, y_n]^T$ and corresponding covariate matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$. Let y_j be a realization of the stochastic variable Y_j with a distribution belonging to the exponential family with canonical parameter θ_j and dispersion parameter ϕ and let the linear predictor $\eta_j = \mathbf{x}_j^T\boldsymbol{\beta}$ be linked to the mean, μ_j, of the distribution of Y_j by the link function g. Then, the likelihood function is defined as

$$\begin{aligned} L(\boldsymbol{\theta}, \phi|\mathbf{y}, \mathbf{X}) &= \prod_{j=1}^{n} f_{Y_j}(y_j|\theta_j, \phi) \\ &= \prod_{j=1}^{n} \exp\left(\frac{y_j\theta_j - b(\theta_j)}{\phi}w + c(y_j, \phi, w)\right). \end{aligned} \tag{3.5}$$

Since log is a monotonically increasing function, then maximizing the logarithm of the likelihood and maximizing the likelihood will result in the same estimates. Maximizing the log likelihood is usually easier than maximizing the likelihood directly. Taking the log of Expression (3.5) results in

$$\begin{aligned} l(\boldsymbol{\theta}, \phi|\mathbf{y}, \mathbf{X}) &= \ln L(\boldsymbol{\theta}, \phi|\mathbf{y}, \mathbf{X}) \\ &= \sum_{j}^{n} \frac{y_j\theta_j - b(\theta_j)}{\phi}w + c(y_j, \phi, w). \end{aligned}$$

The maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$, often denoted $\hat{\boldsymbol{\beta}}$, is the value of $\boldsymbol{\beta}$ that maximizes the likelihood function L or the log likelihood function l. However, the way the likelihood is defined above, $\boldsymbol{\beta}$ is not included in the expression. If the canonical link function is used, then $\boldsymbol{\beta}$ is easily entered into the equation by inserting $g(\mathbf{x}_j^T\boldsymbol{\beta})$ for θ_j. However, for other link functions $\boldsymbol{\beta}$ can be inserted into the equation by utilizing the relation between θ_j and μ_j as described in Expression (3.4).

3.2.3 Logistic Regression

Logistic regression is a special case of generalized linear models used for for binary response variables. For the n observation pairs $(\mathbf{x}_j, y_j)$, assume the binary response y_j is coded by 0 and 1, such that $y_j \in \{0, 1\}$ are independent and follow the Bernoulli distribution with probability π_j. The Bernoulli probability mass function is defined as

$$f(y|\pi) = P(Y = y) = \pi^y(1-\pi)^{1-y} \qquad \text{for } y \in \{0,1\},$$

where $P(Y = 1) = \pi = \mathrm{E}(y) = \mu$ and $\mathrm{Var}(Y) = \pi(1-\pi)$. This can be rewritten on exponential form as

$$f(y|\pi) = \exp\left(y\ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)\right).$$

If we define $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$ as the canonical parameter we can obtain the density of the Bernoulli distribution in the form of an exponential family

$$f(y|\theta) = \exp(y\theta + \ln(1+\exp(\theta)),$$

where $b(\theta) = \ln(1-\pi) = -\ln(1+\exp(\theta))$, $\phi = 1$, $w = 1$ and $c(y,\pi,w) = 0$. Further, it can be shown that

$$\mathrm{E}(y) = b'(\theta) = \pi \quad \text{and} \quad \mathrm{Var}(y) = b''(\theta) = \pi(1-\pi)$$

holds.

We want to estimate the probability for $y_j = 1$ conditional on the covariates. The linear predictor in this case is defined as

$$\eta_j = \beta_0 + \beta_1 x_{j,1} + \cdots + \beta_p x_{j,p}.$$

Since we are estimating the probability for $y_j = 1$, the predictions need to lie between 0 and 1 for all $\mathbf{x}_j$. This is achieved by using an appropriate link function $g(\pi_j) = \eta_j$, and the logit, probit and complimentary log-log (cloglog) are commonly used. We will consider only the logit link function here. The logit link function is defined as

$$g(\pi_j) = \log\left(\frac{\pi_j}{1-\pi_j}\right) = \eta_j = \beta_0 + \beta_1 x_{j,1} + \cdots + \beta_p x_{j,p}, \tag{3.6}$$

and the corresponding response function is defined as

$$\mathrm{E}(y_j) = P(y_j = 1) = \pi_j = h(\eta_j) = \frac{\exp(\eta_j)}{1+\exp(\eta_j)}.$$

For binary response variables the logit link function is the canonical link function, as the link function is defined as $g(\pi_j) = \theta$. Further transforming Expression (3.6) of the log odds gives an easy way to interpret the model coefficients. We explain how using the example below.

Example log odds

Assume that we are interested in understanding why some people develop a disease. Further, assume that the development is only dependent on two factors, age and weight. Age is a continuous variable and weight is a categorical variable with 4 categories: underweight, normal weight, overweight, obese. Since the weight variable is categorical we use dummy variable encoding and let normal weight be the reference category. Using a dataset containing this information about a population, we can fit a logistic regression model. Since only the log odds, and not the expected value of the response, is linear in the covariates, the interpretation of these results is not straightforward. However, as seen in Expression (3.6), we are able to obtain a linear model for the logarithmic odds. If we rewrite this expression of the log-odds we get

$$\frac{P(Y_j = 1)}{P(Y_j = 0)} = \frac{\pi_j}{1 - \pi_j} = \exp(\eta_j) = \exp(\mathbf{x}_j \boldsymbol{\beta}),$$

which gives a convenient expression for interpreting the effect of the covariates in a logistic regression model. The quantity $\frac{P(Y_j=1)}{P(Y_j=0)}$ is the odds, and it is the probability of $y_j = 1$ relative to the probability of $y_j = 0$. In other words, this is the odds of developing the disease. A unit increase in one of the covariates, while keeping all other covariates the same, leads to a multiplicative change in the odds by that covariate's coefficient. If we, for example, consider a change in age from $x_{j,age} = 60$ to $x_{j,age} = 61$, while all other covariates are kept constant, then

$$\begin{aligned}\frac{P(Y_j = 1|x_{j,age} = 60 + 1)}{P(Y_j = 0|x_{j,age} = 60 + 1)} &= \exp(\beta_0) \cdot \exp(\beta_{age} x_{j,age}) \cdots \exp(\beta_{obese} x_{j,obese}) \\ &= \exp(\beta_0) \cdot \exp(\beta_{age}(60 + 1)) \cdots \exp(\beta_{obese} x_{j,obese}) \\ &= \frac{P(Y_j = 1|x_{j,age} = 60)}{P(Y_j = 0|x_{j,age} = 60)} \exp(\beta_{age}).\end{aligned}$$

Consequently, if $\beta_{age} > 0$ the odds increases with increasing age, if $\beta_{age} < 0$ the odds decreases with increasing age, and if $\beta_{age} = 0$ age has no effect. In a similar way, we can examine the change in odds with changing weight. The odds of developing the disease is $\exp(\beta_{overweight})$ times greater for people in the overweight category compared to people in the normal weight category.

Estimating Regression Coefficients of a Logistic Model

As we have assumed a Bernoulli distribution for the response $\mathbf{y} = [y_1, \ldots, y_n]^T$, the corresponding likelihood function is

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{j=1}^{n} \exp(y_j \theta_j + \ln(1 + \exp(\theta_j)).$$

Using the relationship defined above, $\theta_j = \ln\left(\frac{\pi_j}{1-\pi_j}\right)$ where π_j is linked to $\boldsymbol{\beta}$ as given in Expression (3.6), we get

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{j=1}^{n} \exp\left(y_j \ln\left(\frac{\pi_j}{1 - \pi_j}\right) + \ln\left(1 - \pi_j\right)\right).$$

The log likelihood is then defined as

$$\begin{aligned}l(\boldsymbol{\beta}|\mathbf{y}) &= \sum_{j=1}^{n} y_j \ln\left(\pi_j\right) + (1 - y_j) \ln\left(1 - \pi_j\right) \\ &= \sum_{j=1}^{n} y_j(\mathbf{x}_j^T \boldsymbol{\beta}) - \ln(1 + \exp(\mathbf{x}_j^T \boldsymbol{\beta})).\end{aligned}$$

From this expression the gradient and Hessian matrix with respect to the vector $\boldsymbol{\beta}$ can easily be computed by simple matrix differentiation rules in combination with the chain rule. The gradient and Hessian are then entered into the iteratively reweighted least squares algorithm. See Chapter 5 in the book by Fahrmeir et al. (2007) for more details.

The Deviance of a Logistic Model

The deviance is a godness-of-fit statistic, and often used for model selection. The deviance is defined as two times the difference between the log likelihood of the saturated model and the log likelihood of the model (Fahrmeir et al. 2007, Chapter 5). The saturated model is a model that has as many covariates as observations and therefore fits the data perfectly. For logistic regression the log likelihood of the saturated model is zero. Thus the binomial deviance for n observations is defined as

$$\mathrm{Dev} = \sum_{j=1}^{n} -2\left(y_j \ln \hat{p}_j + (1-y_j)\ln(1-\hat{p}_j)\right),$$

where y_j is the observed response for observation j and $\hat{p}_j$ is the prediction of observation j. A smaller deviance indicates a better fit.

3.3 Multiple Hypotheses Testing

In this section we discuss a challenge arising when testing multiple hypotheses at the same time and a method for dealing with this challenge. The theory in this chapter is mainly based on Halle et al. (2017). Multiple testing refers to testing multiple hypotheses at the same time, maybe even using the same data.

3.3.1 Probability of a Type I error

When performing a hypothesis test we first define the null and the alternative hypothesis. There are both one-sided and two-sided hypotheses tests, but we will only consider the two sided hypothesis test here. Consider a hypothesis test defined as

$$H_0 : \beta_k = 0 \qquad\qquad H_1 : \beta_k \neq 0,$$

where β_k is the k'th coefficient in a regression model. There are two types of errors we can make, called type I and type II errors. A type I error is to reject the null hypothesis when the null hypothesis is true. A type II error is to fail to reject the null hypothesis when the null hypothesis is false. Testing many hypotheses simultaneously creates some unwanted problems; the chance of rejecting a true null hypothesis grows undesirably large.

		Results		
		Not reject H_0	Reject H_0	Total
Truth	H_0 true	U	V	m_0
	H_0 false	T	S	$m-m_0$
	Total	$m-R$	R	m

Table 3.1: Table summarizing results of m hypotheses tests. The number m_0 is the number of true null hypotheses, V is the number of type I errors, T is the number of type II errors, U is the number of true null hypotheses not rejected and S is the number of correctly rejected null hypotheses.

The outcome of m tests can be summarized as shown in Table 3.1. Denote the number of true null hypotheses by m_0, the number of type I errors by V, and the number of type II errors by T. Further, the number of true null hypotheses not rejected is denoted by U and the number of correctly rejected null hypotheses by S. While performing the hypotheses tests, only the total number of tests, m, and number of rejected hypotheses, R, are known.

As mentioned, the chance of a type I error grows undesirably large when preforming many tests simultaneously. Consider a multiple testing problem where we want to test m hypotheses and let the local significance level be α_{loc}. This means that the probability of falsely rejecting each individual null hypothesis is controlled at α_{loc}, if we consider a method producing exact p-values. A p-value p is called an exact p-value if $P(p \leq \alpha) = \alpha$ for all $\alpha \in [0,1]$. When considering all m tests and assuming all null hypotheses are true, the expected number of type I errors becomes $m\alpha_{loc}$. We are interested in the probability of committing at least one type I error among the m tests. If we continue to assume all null hypotheses to be true and also assume independence between p-values, then the probability of at least one false positive result among the m tests is

$$\begin{aligned} P(\text{At least one type } I \text{ error}) &= 1 - P(\text{No type } I \text{ errors}) \\ &= 1 - P(V = 0) \\ &= 1 - P(p_1 > \alpha_{loc} \cap p_2 > \alpha_{loc} \cap \cdots \cap p_m > \alpha_{loc}) \\ &= 1 - (1 - \alpha_{loc})^m, \end{aligned}$$

where p_i is the p-value of test i. This shows that the probability of getting at least one type I error grows rapidly with increasing number of tests.

In differential expression analysis we often analyze the effect of some variables on hundreds to thousands of genes or miRNAs, leading to hundreds to thousands of hypotheses tests. As we have seen, this might lead to the probability of at least one type I error to be higher than we are willing to accept. However, there are ways to address this problem by taking the multiplicity of tests into account. This is done by controlling an appropriate error rate. If we want to focus on preventing type I errors rather than type II errors, controlling the family-wise error rate (probability of at least one type I error) using the Bonferroni method may be useful (Holm 1979; Hochberg 1988). However, in exploratory analysis, and not confirmatory analysis, we often consider controlling the false discovery rate (FDR) and not the FWER as it is not as strict a criterion and offers higher power.

3.3.2 False Discovery Rate

The FDR is the expected proportion of falsely rejected null hypotheses. When considering if a coefficient is significant, we use α as a threshold for the p-value to control the probability of a type I error. In the same way, we can control the false discovery rate. The FDR is defined by Halle et al. (2017) as

$$\text{FDR} = \begin{cases} \mathbf{E}\Big(\frac{V}{V+R}\Big) & \text{for } R > 0 \\ 0 & \text{else.} \end{cases}$$

The FDR can be controlled using the Benjamini and Hochberg procedure (Benjamini and Hochberg 1995). To explain how this is done, consider testing m hypotheses based on the corresponding m p-values. Let $p_{(1)} \leq p_{(2)} \leq p_{(3)} \leq \cdots \leq p_{(m)}$ be the ordered p-values, and denote $H_{(i)}$ as the null hypothesis corresponding to the p-value $p_{(i)}$. Then define k such that

$$k = \arg\max_i \left\{ p_{(i)} \leq \frac{i}{m}\alpha^* \right\}.$$

where α^* is the rate at which we want to control the FDR. The Benjamini and Hochberg procedure works by rejecting all null hypotheses $H_{(i)}$ for $i = 1,2,3,\ldots,k$. This is also true for any configuration of false null hypotheses (Benjamini and Hochberg 1995). J. J. Goeman and Solari (2014) states that this procedure controls the FDR at level α^* under some dependence structure of the p-values known as the stronger positive dependence through stochastic ordering. This is further explained by J. J. Goeman and Solari (2014) on page $1953-1954$ and $1961-1963$.

The Benjamini and Hochberg procedure can be used to create adjusted p-values using the following formula

$$\tilde{p}_i = \min\left\{\alpha | \mathrm{H}_{(i)} \text{ is rejected at FDR level } \alpha\right\}.$$

3.4 limma and voom

limma is a R software package using linear models and empirical Bayes for differential expression analysis (Smyth 2004). It was developed while microarrays were the most widely used experiment to quantify RNA sequences, meaning it was specifically developed to analyze continuous expression data. What has made *limma* a popular method is that it deals with one of the main challenges of doing inference on expression data; the small number of observations. Gene expression data, or expression data in general, are often high dimensional with $n \ll p$. This imposes serious challenges when performing inference about each RNA sequence in the dataset. *limma* exploit the similarities between RNA sequences by borrowing information from all the other sequences in the dataset when doing inference about one of them. This way of borrowing information increases power to detect differentially expressed miRNAs, and allows for miRNAs to have individual variances (Law et al. 2014).

The *voom* method, introduced by Law et al. (2014), makes the borrowing information concept of *limma* available for discrete RNA-seq data. Typically, the variance of discrete expression counts increase with increasing expression value. Before any analysis is performed, the data are pre-processed to account for both biological and experimental variations. This usually transforms the data to a continuous scale. Thus one could argue that the data could be used directly in the *limma* pipeline. However, this ignores the fact that the variability depends on the expression value. By estimating a mean variance trend, *voom* incorporates this trait of discrete distributions into the *limma* pipeline.

Although there exist many different methods for differential expression analysis of RNA-seq data, using different distributions and approaches, we only consider *limma* and *voom* in this thesis. We first briefly explain the methods used in the original *limma* package, then we proceed to explain how *voom* make these methods available for discrete RNA-seq data.

3.4.1 limma

limma fits a separate linear model to each miRNA, and uses an empirical Bayes method to borrow information between models. The methods and theory in this section are based on the paper that originally introduced *limma* (Smyth 2004). The response of the linear models in *limma* are the $\log_2$ expression values of the miRNAs. These expression values are usually normalized to counts per million (cpm) before any analysis. This normalization procedure is explained in Section 4.2.

Linear model

Consider n observation pairs $(\mathbf{x}_j, y_{i,j})$, for $j = 1, 2, \ldots, n$ of a RNA sequence i, where $\mathbf{x}_j = [x_{j,1}, \ldots, x_{j,p}]^T$ is a vector of p explanatory variables. A linear model is fitted to each miRNA

i as

$$\mathrm{E}\big[\mathbf{y}_i\big] = \mathrm{E}\begin{bmatrix} y_{i,1} \\ y_{i,2} \\ \vdots \\ y_{i,n} \end{bmatrix} = \mathbf{X}\boldsymbol{\beta}_i$$

where $\mathbf{y}_i = [y_{i,1}, \ldots, y_{i,n}]^T$ are the $\log_2$ transformed cpm values, $\boldsymbol{\beta}_i = [\beta_{i,1}, \ldots, \beta_{i,p}]$ is the coefficient vector and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ is the design matrix with the covariate vector $\mathbf{x}_j^T = [x_{j,1} \ldots, x_{j,p}]$ as row j. Further assume that the variance of the response can be written as

$$\mathrm{Var}(\mathbf{Y}_i) = \mathbf{W}_i \sigma_i^2,$$

where $\mathbf{W}_i$ is a known diagonal weight matrix with elements $w_{i,1}, w_{i,2}, \ldots, w_{i,n}$ and σ_i^2 is the unknown variance of miRNA i.

Assume that we are interested in testing for differential expression between two groups. We then include that group variable as a binary categorical covariate, $x_{i,k}$, in the linear model and test by performing a hypothesis test defined as

$$H_0 : \beta_{i,k} = 0, \qquad H_1 : \beta_{i,k} \neq 0.$$

In cases with multiple experimental conditions, where we want to test for differential expression between multiple groups, we need to define a contrast matrix. This can be defined as $\boldsymbol{\beta}_i^* = \mathbf{C}^T \boldsymbol{\beta}_i$, where $\mathbf{C}^T$ is the contrast matrix and $\boldsymbol{\beta}_i^*$ is the difference between the experimental conditions of interest, called the contrasts. There might be more or fewer contrasts than coefficients.

Empirical Bayes method

A moderated t-statistic is derived using a hybrid of a classical and Bayesian approach. Bayesian statistics is a theory fundamentally different from the classical frequentist statistics (Casella and Berger 2002). Consider a random sample $X_1, X_2, \ldots, X_n$ drawn from a population with probability distribution $f(\mathbf{x}|\theta)$, where θ is a parameter of the distribution. In the classical approach θ is considered unknown but fixed. However, the Bayesian approach is based on a hierarchical model, where θ is considered a random variable whose variation can be described by a probability distribution. This distribution, called the prior distribution, is based on the experimenters initial beliefs and is decided prior to seeing the data. The prior distribution is denoted by $\pi(\theta)$. Given θ the observations have a probability distribution which is often referred to as the likelihood and is denoted by $f(\mathbf{x}|\theta)$. After collecting the observations, the prior is updated with the new information obtained by the observations using Bayes theorem. The updated version of the prior is what we refer to as the posterior distribution. The posterior distribution is defined as

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{f(\mathbf{x})},$$

where the denominator, $f(\mathbf{x})$, is the marginal distribution of $\mathbf{x}$ given as $f(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta$. The variable θ is not necessarily just one variable, but may be a vector of variables that influence the distribution. In a two-stage hierarchical Bayes model, the variables $\boldsymbol{\theta}$ can also be considered to be samples from a distribution characterized by hyperparameters, $\boldsymbol{\eta}$, given as

$f(\boldsymbol{\theta}|\boldsymbol{\eta})$, where the hyperparameters, $\boldsymbol{\eta}$, are believed to be draw from a distribution $\pi(\boldsymbol{\eta})$. As explained by Casella and Berger (2002), the empirical Bayes method differs from the standard Bayes method by not specifying the values of the hyperparameters $\boldsymbol{\eta}$ a priori. Instead these parameters are estimated from the data and the prior distribution is no longer formulated before seeing the data.

Estimating the Posterior Distribution

In the *limma* method an empirical Bayes approach is used to estimate a posterior variance for each miRNA and this is then inserted into the t-statistic in the place of the usual sample variance. This results in the moderated t-statistic.

The coefficients in each of the linear models are estimated using least squares. Denote the least squares coefficient estimator by $\hat{\boldsymbol{\beta}}_i$ and the residual sample variance by s_i^2. Then, $\mathrm{Var}(\hat{\boldsymbol{\beta}}_i) = \mathbf{V}_i s_i^2$ where $V_i = (\mathbf{X}^T \mathbf{W}_i^{-1} \mathbf{X})^{-1}$.

We assign a scaled chi-square distribution with d_0 degrees of freedom and scaling parameter s_0^2 as a prior for σ_i^{-2} such that

$$\frac{1}{\sigma_i^2} \sim \frac{1}{d_0 s_0^2} \chi^2_{d_0}. \tag{3.7}$$

Furthermore, we assume that the likelihood, $s_i^2|\sigma_i^2$, is assumed to approximately follow a scaled chi-squared distribution with d_i degrees of freedom defined as

$$s_i^2|\sigma_i^2 \sim \frac{\sigma_i^2}{d_i} \chi^2_{d_i}.$$

Plugging this and the prior distribution for $1/\sigma_i^2$ into Bayes theorem we find the posterior distribution of $1/\sigma_i^2$. We have that

$$\begin{aligned} f\left(\frac{1}{\sigma_i^2}\middle| s_i^2\right) &= cf\left(s_i^2|\sigma_i^2\right)\pi\left(\frac{1}{\sigma_i^2}\right) \\ &= c\frac{1}{2^{d_i/2}\Gamma(d_i/2)}\left(\frac{s_i^2 d_i}{\sigma_i^2}\right)^{d_i/2-1} \exp\left\{\frac{-s_i^2 d_i}{2\sigma_i^2}\right\}\left|\frac{d_i}{\sigma_i^2}\right| \\ &\times \frac{1}{2^{d_0/2}\Gamma(d_0/2)}\left(\frac{s_0^2 d_0}{\sigma_i^2}\right)^{d_0/2-1} \exp\left\{\frac{-s_0^2 d_0}{2\sigma_i^2}\right\}|d_0 s_0| \\ &= c'\left(\frac{1}{\sigma_i^2}\right)^{\frac{d_i+d_0}{2}-1} \exp\left\{-1/\sigma_i^2\left(\frac{s_i^2 d_i + s_0^2 d_0}{2}\right)\right\} \\ &\sim \mathrm{Gamma}\left(\frac{d_i+d_0}{2}, \frac{2}{s_i^2 d_i + s_0^2 d_0}\right), \end{aligned} \tag{3.8}$$

where c and c' are normalizing constants. For readability we do not write them out as they are not necessary for recognizing the distribution to be $\mathrm{Gamma}\left(\frac{d_i+d_0}{2}, \frac{2}{s_i^2 d_i + s_0^2 d_0}\right)$ distributed.

Estimation of Hyperparameters

The hyperparameters s_0^2 and d_0, in Expression (3.7), are estimated from the data. The hyperparameter s_0^2 can be viewed as a pooled variance of all RNA sequences in the dataset. This will become clearer once the posterior mean, $\tilde{s}_i^2$, is derived. We consider $\log(s_i^2)$ instead of s_i^2 while deriving the estimates because this follows a distribution closer to the normal distribution.

Smyth (2004) shows that the marginal distribution of s_i^2 is a scaled F- distribution with parameters (d_i, d_0), where d_i is the degrees of freedom of the chi-square distribution assumed on $s_i^2|\sigma_i^2$ and d_0 is the degrees of freedom of the chi-square prior assigned to σ_i^{-2}. Let z_i be defined as

$$z_i = \log(s_i^2).$$

Then z_i follows a constant plus Fisher's Z distribution (Smyth 2004).

The *limma* method estimates the two hyperparameters, s_0^2 and d_0, by matching up the theoretical mean and variance of z_i with the empirically estimated mean and variance of z_i. The theoretical mean and variance of z_i can be defined as

$$\begin{aligned} \mathrm{E}[z_i] &= \log(s_0^2) + \psi(d_i/2) - \psi(d_0/2) + \log(d_0/d_i) \\ \mathrm{Var}[z_i] &= \psi'(d_i/2) + \psi'(d_0/2), \end{aligned}$$

where ψ and ψ' are the diagamma and trigamma functions. By letting e_i be defined as

$$e_i = z_i - \psi(d_i/2) + \log(d_i/2),$$

we can obtain an unbiased estimator of $\mathrm{Var}(z_i)$: $\frac{1}{n-1}\sum_i (e_i - \bar{e})^2$ where $\bar{e} = \frac{1}{n}\sum_i e_i$. Matching this with the theoretical variance gives

$$\mathrm{E}\left[\frac{1}{n-1}\sum_i (e_i - \bar{e})^2\right] = \psi'(d_0/2) + \psi'(d_i/2).$$

We can then estimate d_0 by solving the following expression with respect to d_0

$$\psi'(d_0/2) = \frac{1}{n-1}\sum_i (e_i - \bar{e})^2 - \psi'(d_i/2). \tag{3.9}$$

Given the estimate of d_0, we use this to estimate s_0^2. By using the definition of e_i and the theoretical mean of z_i we have that

$$\begin{aligned} \mathrm{E}[e_i] &= \mathrm{E}[z_i] - \psi(d_i/2) + \log(d_i/2) \\ &= \log(s_0^2) - \psi(d_0/2) + \log(d_0) - \log(d_i) + \log(d_i/2) \\ &= \log(s_0^2) - \psi(d_0/2) + \log(d_0/2). \end{aligned}$$

To obtain an estimate of s_0^2, we solve the following expression

$$s_0^2 = \exp\left\{\bar{e} + \psi(d_0/2) - \log(d_0/2)\right\}.$$

In the case where the right hand side of expression (3.9) is less than 0, the equation cannot be solved. In that case d_0 is set to be infinite and the estimate of s_0^2 simplifies to $\exp\{\bar{e}\}$.

The Moderated t-statistic

The posterior distribution of σ_i^{-2} in Expression (3.8) combined with the estimated hyperparameters is used to estimate a posterior variance, $\tilde{s}_i^{-2}$. By the properties of the gamma distribution, we have that

$$\tilde{s}_i^{-2} = \mathrm{E}\left(\sigma_i^{-2}|s_i^2\right) = \frac{d_0 + d_i}{d_0 s_0^2 + d_i s_i^2},$$

where $\tilde{s}_i^{-2}$ is the posterior mean of σ_i^{-2}. Taking the inverse of this expression yields the following expression

$$\tilde{s}_i^2 = \frac{1}{\mathrm{E}\left(\sigma_i^{-2}|s_i^2\right)} = \frac{d_0 s_0^2 + d_i s_i^2}{d_0 + d_i}.$$

Although this is, in fact, the inverse of the posterior mean of $\frac{1}{\sigma_i^2}$ given s_i^2, for simplicity we will refer to this as the posterior mean of σ_i^2. Inserting $\tilde{s}_i^2$ into the t-statistic yields the following expression

$$\tilde{t}_{i,k} = \frac{\hat{\beta}_{i,k}}{\tilde{s}_i\sqrt{v_{i,k}}}.$$

This expression reduces to the ordinary t-statistic if $d_0 = 0$. The $\beta_{i,k}$ coefficient is assumed to be nonzero with probability,

$$P\left(\beta_{i,k} \neq 0\right) = p_k.$$

When $\beta_{j,k} \neq 0$ a prior distribution is specified as

$$f(\beta_{i,k}|\sigma_i^2, \beta_{j,k} \neq 0) = N(0, v_{0,j}\sigma_j^2).$$

Furthermore, Smyth (2004) shows that the marginal distribution of $\tilde{t}_{i,k}$ for all miRNAs follow a mixture distribution of an ordinary t-distribution and a scaled t-distribution. Under the null hypothesis, $\mathrm{H}_0 : \beta_{i,k} = 0$, the moderated t-statistic follows an ordinary t-distribution with $d_0 + d_i$ degrees of freedom, that is

$$\tilde{t}_{i,k}|\beta_{i,k} = 0 \quad \sim t_{d_0+d_i}.$$

When $\beta_{j,k} \neq 0$, $\tilde{t}_{i,k}$ follows a scaled t-distribution with $d_0 + d_i$ degrees of freedom, that is

$$\tilde{t}_{i,k}|\beta_{i,k} \neq 0 \quad \sim (1 + v_{0l}/v_{i,k})^{1/2} t_{d_0+d_i}.$$

The respective mixing proportions of the distribution are p_k and $1 - p_k$.

The unconditional distributions of the moderated t-statistic $\tilde{t}_{i,k}$ and of s_i^2 are independent.

Improving Power for Differential Detection

limma improves power to detect differential expression by using the moderated t-statistic (Ritchie et al. 2015). We have established how the moderated t-statistic is derived and explained how the posterior estimate of the variance, $\tilde{s}_i^2$, is a compromise of the residual sample variance, s_i^2, and a pooled variance, s_0^2, derived from all miRNAs. If the number of samples is large and thus the degrees of freedom, d_i, of the sample variance distribution is high, then the posterior variance will be close to the sample variance. However, if the number of observations is small, then the pooled variance will have a greater impact on the posterior variance. This means that the posterior will be squeezed towards the pooled variance to hopefully obtain a better estimate.

A question to ask is how squeezing the variance of individual miRNAs towards a mean variance derived from all miRNAs improves the power. A possible way to motivate the use of a squeezed variance is by Stein's paradox. Stein's paradox states that when estimating three or more parameter means, there exist methods that combine the estimates that are more accurate than any method estimating each parameter separately (Efron and Morris 1977). The same concept is used in the *limma* method. It is believed that it is beneficial to include the average over all sample variances when estimating the variance of an individual miRNA. Generally, we believe that larger variances in expression data are overestimated and smaller variances are underestimated, and that we obtain better results by averaging them towards a common mean.

In addition to squeezing the sample variance towards an averaged variance, the degrees of freedom of the t-statistic under the null hypothesis increases. This means that using the posterior variance instead of the sample variance has two effects on the t-statistic. The variance estimate is shrunken or increased and the degrees of freedom increase. For miRNAs with small sample variance estimates, the variance is increased towards s_0^2, and the degrees of freedom is increased. When the degrees of freedom increase, the t-distribution becomes narrower with lighter tails. Furthermore, the estimated variance in the denominator is reduced, meaning that the t-statistic will be larger. This increases power for miRNAs with large variance. For miRNAs with small variance the power is also increased by the more degrees of freedom, but is reduced by increasing the estimated variance.

3.4.2 voom

When using *voom* a mean variance trend is estimated and used to estimate variances for each individual observation. In this way *voom* addresses a key problem when fitting normal based models to RNA-seq data; large counts have considerably larger standard deviations than smaller counts. Both *voom* and *limma-trend* are methods taking this into account. Using *limma-trend* a mean variance trend is also estimated, but it is used to estimate prior variances for each miRNA. These prior variances are then used instead of s_0^2 in the original *limma* method. As we believe variance modelling at the observation level to be beneficial we only consider *voom* in this thesis. *voom* is available in the R package *limma*. Other methods, such as those implemented in *edgeR* and *DESeq*, use count distributions to avoid having to fit normal based models to RNA-seq data (Robinson, McCarthy et al. 2010; Love et al. 2014). Law et al. (2014) states that this imposes serious limitations due to the reduced range of statistical tools associated with count distributions compared to the normal distribution.

Variance Modeling at the Observational Level

The standard deviations of read counts usually increase with increased number of counts. Applying the log-transformation to the counts counteracts this trend, but it overdoes it so that the trend appears descending. The *voom* method aims to model this trend without the need to specify the correct probability distribution.

Using *limma-trend* it is assumed that all observations of the same miRNA have equal variance. This assumption is violated if the sequencing depths differ between samples. The variance of an observation depend both on the mean of the miRNA and the sequencing depth of the sample. If the sample has a higher sequencing depth than the rest, then we would assume that observations from this sample have a higher variance than observations from other samples.

This information is, however, lost when accounting for the sequencing depth by converting the count to cpm values (cpm normalization is described in Section 4.2). To account for this, *voom* estimates variances at the observational level and operates on the raw read counts.

Precision Weights

To incorporate observation level variance into the *limma* pipeline, precision weight for each observation are estimated. The precision weights are defined as the inverse of the estimated variances.

Precision weights are estimated by fitting a linear model to the log cpm values $y_{i,j}$ for each miRNA i using ordinary least squares, as is done in *limma*. This results in a residual standard deviation for each miRNA, s_i, and fitted values, $\hat{\mu}_{i,j}$. The average log cpm values, $\bar{y}_i$, are then computed for each RNA sequence. This average is then converted back to a log count value by

$$\tilde{r}_i = \bar{y}_i + \log_2(\tilde{R}) - \log_2(10^6),$$

where $\tilde{R}$ is the mean of the library sizes R_j plus one. Next, a trend is fitted to the square root of the standard deviations s_i as a function of the average log counts $\tilde{r}_i$. The trend is fitted using a LOWESS curve, which is a form of local regression method (Strand 2019). We denote the trend fitted by the LOWESS curve as the function $lo()$ which takes a count value as input. Then all the fitted log cpm values, $\hat{\mu}_{i,j}$, from the linear model are also converted into fitted log counts by

$$\tilde{\lambda}_{i,j} = \hat{\mu}_{i,j} + \log_2(R_j + 1) - \log_2(10^6).$$

Consequently, by inserting the log counts into $lo()$ we can estimate the square root of the standard deviation for each observation taking the sequencing depth into account. Lastly the *voom* precision weights are the inverse of the variances obtained from the LOWESS curve, $w_{i,j} = \tilde{s}_{i,j}^{-2}$. The calculated weights paired with the log cpm values $y_{i,j}$ are input into the original *limma* pipeline.

3.5 The Elastic Net

The elastic net is a regularization method, first introduced by Zou and Hastie (2005). Regularization, sometimes called penalized regression, is a way of regularizing the estimation process of coefficients. These methods aim to shrink coefficient estimates and/or create sparse models with the goal of reducing overfitting. Lasso is a popular regularization method using an L_1 penalty to achieve sparse solutions (R. Tibshirani 1996). Another popular method is ridge regression. Ridge uses the L_2 penalty and does not perform variable selection, rather it only shrinks the coefficients estimates without letting any of them be exactly zero. The elastic net is a weighted combination of lasso and ridge and includes them both as special cases.

Consider n observation pairs $(\mathbf{x}_j, y_j)$, for $j = 1, 2, \dots, n$, where $\mathbf{x}_j = [x_{j,1}, \dots, x_{j,p}]^T$ is a vector of p explanatory variables. The elastic net minimizes the negative likelihood to estimate the coefficients, as is done for GLMs, but subject to some constraint. The general formulation of the elastic net minimization problem including an intercept, is

$$\min_{(\beta_0, \boldsymbol{\beta}) \in R^{p+1}} \left[-\ln L(\beta_0, \boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta}) \right], \tag{3.10}$$

where

$$
\begin{aligned}
P_\alpha(\boldsymbol{\beta}) &= (1-\alpha)\frac{1}{2}||\boldsymbol{\beta}||_{L_2}^2 + \alpha||\boldsymbol{\beta}||_{L_1} \\
&= \sum_{i=1}^{p}\left[\frac{1}{2}(1-\alpha)\boldsymbol{\beta}_i^2 + \alpha|\boldsymbol{\beta}_i|\right]
\end{aligned}
\tag{3.11}
$$

is the elastic net penalty (Hastie, R. Tibshirani and Wainwright 2016). The vector $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^T$ is the coefficient vector, and β_0 is the intercept. The shrinkage penalty in Expression (3.11) is a compromise between the L_1 lasso penalty and the L_2 ridge penalty, and the value α decides where the penalty lies on the scale between them. Choosing $\alpha = 1$ gives lasso regression and choosing $\alpha = 0$ gives ridge regression. Both α and λ are user specified values, though λ is often chosen through cross-validation. In this thesis we also choose α with the help of cross-validation, as further described in Section 5.3. The parameter $\lambda \geq 0$ is a tuning parameter and controls the complexity of the model. Increasing the λ value, will increase the impact of the penalty term and the coefficient estimates are shrunken towards zero. If we decrease λ, the penalty term is decreased, leading to less shrinkage and the model is able to fit more closely to the data. In particular, if $\lambda = 0$ the penalty term is zero and elastic net gives the same coefficient estimates as maximum likelihood.

3.5.1 The Variable Selection Property

Penalized regression methods are widely used with high dimensional datasets. Lasso can identify a small subset of relevant variables that provide good predictive accuracy. However, it tends to exclude highly correlated variables (Pavlou et al. 2016). If our goal is prediction and we do not seek insight into which and how the variables influence the response, this aspect of lasso is not problematic.

Ridge on the other hand, does not shrink any coefficients to zero and does not perform variable selection. In the extreme case of k identical covariates, ridge will give them all identical coefficients with $1/k$'th the size that any one of them would get if included in the model alone (Friedman et al. 2010). Thus, ridge has a way of treating correlated variables similarly, while lasso in theory only includes one of them. In these extreme cases with equal covariates the lasso problem breaks down. To explain this we fist consider the case of only including one of the identical covariates, x_j. For one given value of λ the estimated coefficient is $\hat{\beta}_j > 0$. If we then include a copy of this covariate as well, $x_{j'}$, then they can "share" this coefficient in infinitely many ways, $\tilde{\beta}_j + \tilde{\beta}_{j'} = \hat{\beta}_j$, while still resulting in the same loss (Hastie, R. Tibshirani and Wainwright 2016, Chapter 4). Thus, the coefficients are not well defined. This exact scenario is unlikely to happen in practice, but in expression analysis small groups of miRNAs or genes tend to be expressed together and are thus highly correlated. In addition, when the number of variables p exceeds the number of observations n, the lasso problem is no longer strictly convex, and it may not have a unique minimizer (R. J. Tibshirani 2013). However, this problem can be managed, but we do not explain this any further and refer interested readers to the paper by R. J. Tibshirani (2013).

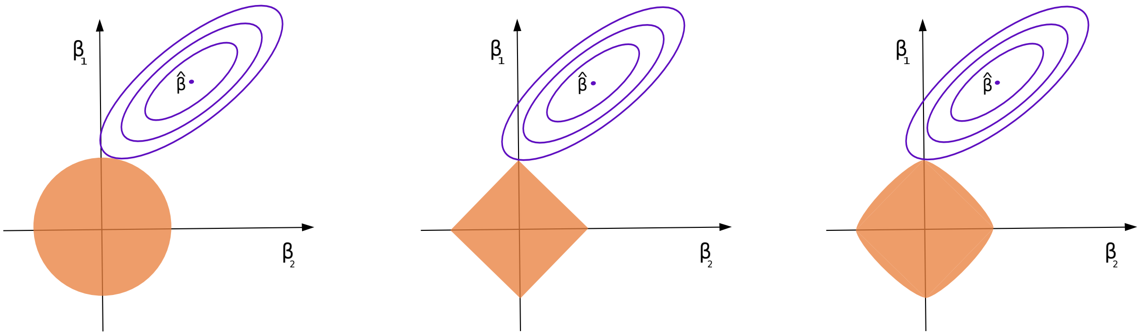


Figure 3.2: This figure is inspired by Figure 6.7 by James et al. (2013) on page 244. It represents a penalized regression case using a model with two variables. The ridge constraints are visualized to the left, the lasso constraints in the middle and the elastic net to the right. The point $\hat{\boldsymbol{\beta}}$ represents the maximum likelihood solution and the ellipses around $\hat{\boldsymbol{\beta}}$ represents paths were the value of the likelihood is the same.

The elastic net, being a compromise between ridge and lasso, inherits a way of including highly correlated variables while still performing variable selection. The method tends to select groups of correlated variables together or leave the whole group out. If we let α be close to, but less than 1, the elastic net behaves much like the lasso but will not break down in the case of equal variables. The elastic net problem is strictly convex for any $\alpha < 1$ and $\lambda > 0$, meaning a unique solution exists for any correlation between variables (Hastie, R. Tibshirani and Wainwright 2016, Chapter 4).

The constraints of the three regularization methods when using two variables are visualized in Figure 3.2. The maximum likelihood solution is represented by $\hat{\boldsymbol{\beta}}$ and the ellipses around are contours. All points on one of these contours have the same value of the likelihood, and as these ellipses expand away from $\hat{\boldsymbol{\beta}}$ the likelihood increases. The circle, diamond, and rounded diamond represents the ridge, lasso and elastic net constraints respectively, for one particular λ. If $\lambda = 0$, the maximum likelihood solution, $\hat{\boldsymbol{\beta}}$, lies within the constraint boundaries and the ridge, lasso, elastic net and maximum likelihood solution will be the same. If $\lambda \neq 0$ the coefficient estimates are given by the first point an ellipse intersects with the constraint region. Since the ridge constraint region is circular, this point will generally not be on one of the axes. Meaning the coefficients estimates will generally not be set to exactly zero. Lasso on the other hand, has sharp corners at the axes and the contours around the maximum likelihood solution will often intersect with the corners of the constraints first. This is what gives lasso a variable selection property. We notice that the elastic net constraints has corners like the lasso, but also rounded edges. The sharp corners encourage sparsity, while rounded edges encourage inclusion of correlated variables.

The variable selection property of elastic net and lasso have been demonstrated in multiple studies. Pavlou et al. (2016) reviewed frequentist and Bayesian shrinkage methods in a high dimensional setting and compared their predictive performance. They showed that both lasso and elastic net performed well in most scenarios. Elastic net was, however, found to select more variables than lasso and was superior when the data included correlated variables. They concluded that when variable selection is required and no high correlations are observed between variables lasso is a good option, while if there are high correlations elastic net is the preferred method. Benner et al. (2010) used simulated microarray data and found elastic net and lasso had the best predictive accuracy in moderately sparse scenarios.

3.5.2 Regularized Logistic Regression

For the regularized logistic regression case the minimization problem given in Expression (3.10) becomes

$$\min_{(\beta_0,\boldsymbol{\beta})\in R^{p+1}} \left[-\frac{1}{N}\sum_{j=1}^{n}\left[y_j \ln\left(\pi_j\right)+(1-y_j)\ln\left(1-\pi_j\right)\right]+\lambda P_\alpha(\boldsymbol{\beta}) \right]. \tag{3.12}$$

This optimization problem can not be solved analytically in the general case, thus we must employ numerical optimization procedures. There are multiple ways to solve this optimization problem, but we will present the solution proposed by Friedman et al. (2010) and implemented in the R package named *glmnet*. The method computes solutions for a sequence of λ values, in descending order, along a regularization path. This is an efficient way to compute solutions for a multitude of λ values, but also an efficient way even if only one are of interest (Friedman et al. 2010). For each λ value the algorithm starts with the solution from the previous λ value. Minimizing the log likelihood is difficult. The *glmnet* therefore, for each λ, minimizes the shrinkage term, $P_\alpha(\boldsymbol{\beta})$, along with a quadratic approximation of the likelihood, l_Q, multiple times until convergence. Coordinate decent is used to solve the optimization problem based on the penalized quadratic approximation. The algorithm can be summarized as follows

```
**Outer loop**: Decrement $\lambda$, start with $\{\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}\}$ for the previous $\lambda$
    **Middle loop**: Update the approximation of the likelihood with current estimate of $\{\beta_0, \boldsymbol{\beta}\}$
        **Inner loop**: Run a modified cyclical coordinate decent to solve the optimization problem for each penalized quadratic approximation to update $\{\beta_0, \boldsymbol{\beta}\}$
```

The Quadratic Approximation of the Likelihood

In a more explicit form the log likelihood in Expression (3.12) can be written as

$$l(\beta_0, \boldsymbol{\beta}|\mathbf{y}) = \sum_{j=1}^{n} y_j(\beta_0 + \mathbf{x}_j^T\boldsymbol{\beta}) - \ln(1+\exp(\beta_0 + \mathbf{x}_j^T\boldsymbol{\beta}))$$

(Friedman et al. 2010). This log likelihood is a convex function of β_0 and $\boldsymbol{\beta}$. For some current estimate, $[\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}]$, of the parameters a quadratic approximation to the log likelihood can be formed by Taylor expansion around the current estimates

$$\begin{aligned} l(\beta_0, \boldsymbol{\beta}|\mathbf{y}) &\approx l_Q(\beta_0, \boldsymbol{\beta}|\mathbf{y}) \\ &= -\frac{1}{2}\sum_{j=1}^{n} w_j(z_j - \beta_0 - \mathbf{x}_j^T\boldsymbol{\beta})^2 + C(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})^2, \end{aligned} \tag{3.13}$$

where

$$\begin{aligned} z_j &= \tilde{\beta}_0 + \mathbf{x}_j^T\tilde{\boldsymbol{\beta}} + \frac{y_j - \tilde{\pi}_j}{\tilde{\pi}_j(1-\tilde{\pi}_j)} \\ w_j &= \frac{1}{N}\tilde{\pi}_j(1-\tilde{\pi}_j) \end{aligned}$$

and

$$\tilde{\pi}_j = \frac{\exp(\tilde{\beta}_0 + \mathbf{x}^T\tilde{\boldsymbol{\beta}})}{1+\exp(\tilde{\beta}_0 + \mathbf{x}^T\tilde{\boldsymbol{\beta}})}.$$

Notice that the second term in Expression (3.13) is a constant, while the first term is on the form of a weighted mean square error for a linear regression model where z_j and w_j are the response and weight for observation j respectively. This quadratic approximation of the log likelihood can be solved by using the coordinate decent.

Coordinate Descent

The coordinate descent method, also known as the coordinate search method, is an optimization algorithm that iteratively searches along coordinate directions to find a point with lower function value (Nocedal and Wright 2006, Chapter 9). For each iteration all except one of the components are kept constant, while a search direction along a single component is chosen to minimize the objective function. For perhaps the simplest form of coordinate descent, cyclic coordinate descent, this is repeated for all components. After one cycle through all components, the whole process is repeated, starting again from the first component.

In the *glmnet* algorithm, coordinate descent is used to solve the penalized least-squares problem

$$\min_{(\beta_0,\boldsymbol{\beta})\in R^{p+1}} \left[-l_Q(\beta_0,\boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta})\right], \tag{3.14}$$

where we use the quadratic approximation to the log likelihood, l_Q, in place of the actual log likelihood, l, (Friedman et al. 2010). Inside the inner loop the likelihood function stays the same since $\tilde{\beta}_0$ and $\tilde{\boldsymbol{\beta}}$ are constant inside the inner loop. The approximation to the log likelihood is only updated in the middle loop. As initial solutions we set $\hat{\beta}_0 = \tilde{\beta}_0$ and $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$. We stress here, to avoid confusion, that $\tilde{\beta}_0$ and $\tilde{\boldsymbol{\beta}}$ is kept constant for each inner loop and instead we update $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ throughout the inner loop. Only once $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ (approximately) solve the optimization problem in Expression (3.14), then we update $\tilde{\beta}_0$ and $\tilde{\boldsymbol{\beta}}$ and thus also the quadratic approximation to the log likelihood, l_Q. For each step of the coordinate decent algorithm all variables except one $\hat{\beta}_k$ are kept constant. Thus, we have the minimization problem

$$\min_{\hat{\beta}_k\in R^1} \left[-l_Q(\hat{\beta}_0,\hat{\boldsymbol{\beta}}) + \lambda P_\alpha(\hat{\boldsymbol{\beta}})\right].$$

This minimization problem can be solved analytically. We can compute the partial derivative for β_k when $\beta_k \neq 0$ as

$$\begin{aligned}\frac{\partial}{\partial\beta_k}\left[-l_Q(\beta_0,\boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta})\right]|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}\,\beta_0=\hat{\beta}_0} &= -\sum_{j=1}^{N} w_j x_{j,k}(z_j - \hat{\beta}_0 - x_j^T\hat{\boldsymbol{\beta}}) + \lambda(1-\alpha)\hat{\beta}_k + \lambda\alpha\ \mathrm{sign}(\hat{\beta}_k)\\ &= -\sum_{j=1}^{N} w_j x_{j,k}(z_j - \hat{z}_j^{-k} - x_{j,k}\hat{\beta}_k) + \lambda(1-\alpha)\hat{\beta}_k + \lambda\alpha\ \mathrm{sign}(\hat{\beta}_k),\end{aligned}$$

where $\hat{z}_j^{-k} = \hat{\beta}_0 + \sum_{l\neq k} x_{j,l}\hat{\beta}_l$ (Friedman et al. 2010). We note that if $\lambda\alpha \geq \left|\sum_{j=1}^{N} x_{j,k}w_j(z_j - \hat{z}_j^{-k})\right|$ then the optimal solution for β_k given all other parameters is $\hat{\beta}_k = 0$. When this is not the case, the optimal value for $\hat{\beta}_k$ can be found analytically by finding where the gradient is zero for the two cases $\hat{\beta}_k < 0$ and $\hat{\beta}_k > 0$. This gives that the optimal solution for $\hat{\beta}_k$ is given as

$$\hat{\beta}_k = \frac{S\left(\sum_{j=1}^{N} x_{j,k}w_j(z_j - \hat{z}_j^{-k}), \lambda\alpha\right)}{\sum_{j=1}^{N} w_j x_{j,k}^2 + \lambda(1-\alpha)}$$

where $S(\kappa,\gamma)$ is the soft-thresholding operator defined as

$$S(\kappa,\gamma) = \text{sign}(\kappa)(|\kappa| - \gamma)+ = \begin{cases} \kappa - \gamma & \text{if } \kappa > 0 \text{ and } \gamma < |\kappa| \\ \kappa + \gamma & \text{if } \kappa < 0 \text{ and } \gamma < |\kappa| \\ 0 & \text{if } \gamma \geq |\kappa|. \end{cases}$$

The method first cycles through all variables once, as in cyclical coordinate descent. However, to exploit the sparseness in this setting it only iterates through the parameters in the active set until convergence. The active set consists of all parameters not estimated to be zero. Lastly, it cycles through the whole set again, and if this does not change the active set, it stops. If it does change the active set, the process is repeated.

3.5.3 Inference on Elastic Net Coefficients

It is often of interest to determine the statistical strength of the included variables in the form of p-values or confidence intervals. However, using penalized regression makes this procedure difficult.

A naive approach is to use elastic net for variable selection and then include only the chosen covariates in a separate regression model. That way we can estimate the coefficients, without dealing with the penalization term, and compute p-values and confidence intervals. However, this way of performing inference will not produce correct p-values, and they will appear overly optimistic. After selecting the variables, we have already chosen the variables most related to the response and the computed confidence intervals and p-values does not take the variable selection into account. S. Zhao et al. (2021) present a defence of this naive approach, and discuss cases where it is possible to use this method for inference. However, it is only valid for some special cases, and they do not advocate applying this procedure to practical data analysis. They further conclude that in practice the approach will perform poorly when the sample size is small or moderate, and/or the assumptions are not met.

Lee et al. (2016) present a way of performing exact post-selection inference by conditioning on the on the selection event. This introduces a way to form valid confidence intervals for the selected coefficients and test whether all relevant variables have been included in the model. This is, however, still not developed for penalized logistic regression using elastic net.

Another way to perform inference is to use bootstrapping. However, since penalized estimation is a procedure that reduces the variance of estimators by introducing bias, the percentile method will not estimate the confidence interval of the true coefficient value. The accelerated bias-corrected method uses bootstrapping to compute confidence intervals of biased estimators. This method is further explained in Section 3.6.2.

3.6 Resampling Methods

In this section we describe two resampling methods; cross-validation and the bootstrap. We also describe the accelerated bias-corrected method for estimating confidence intervals of biased estimators using bootstrapping.

3.6.1 Cross-Validation

Cross-validation (CV) is a method often used for model assessment and model selection, especially when the aim is prediction. For example one usually uses CV to determine the shrinkage parameter λ in penalized regression. For both model assessment and selection we ideally want a way to test the model on unseen data. When evaluating a model we can consider a test error and a training error. The training error is computed based on how the model performs on the same data that was used to train the model, that is, to estimate the model parameters. The test error, however, is computed based on new data. The training error is often considerably different from the test error and the training error can dramatically underestimate the test error (James et al. 2013, Chapter 5). As we want to evaluate and compare models based on how they generalize to unseen data, we ideally want to use the test error for evaluation and assessment.

In cases where we need to perform model selection and model assessment, and have a considerable amount of data available, we can set a side both a validation and a test set. Then we can fit a model on the training set using a range of model parameters, then validate and determine the best model by comparing their performance on the validation set. After choosing the best model, we can compute the test error of the final model using the held out test set. However, in many cases we do not have enough data to be willing to reserve some of it for both validation and testing. We could use the same data for validation and testing, but if we use the same data to choose the best model and then evaluate its performance, we would again underestimate the test error. Double use of the data will create an overoptimistic estimate of the test error.

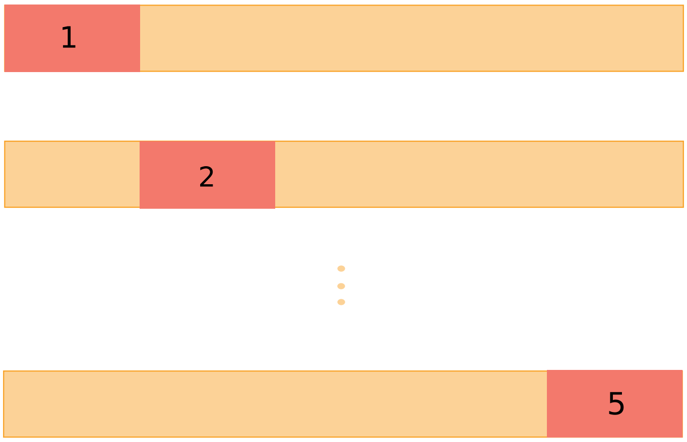


Figure 3.3: Visualization of data splitting of k-fold cross validation when using 5 folds. First, the part of the data numbered 1 is used for validation, while the rest is used for training. Next, the part numbered 2 is used for validation and the rest for training, and so on until all 5 folds are held out and used for validation once.

For model selection, the approach of randomly diving the data into two sets, setting one aside for validation, is called the validation set approach. However, as stated by James et al. (2013, Chapter 5) this method is highly variable depending on how the observations are randomized into the two sets. The validation set approach is a simple technique, but it tends to overestimate the error, as not all of the data are used for training. The method of k-fold CV reduces this variability by repeating the process. In k-fold CV the dataset is randomly split into k groups or folds of approximately equal size, n/k. Common values for k are $k = 3, k = 5$, and $k = 10$. First, the first fold is treated as the validation set and the model is fit using the remaining $k - 1$ folds. Then we can obtain an error estimate of the model performance on the validation set. For classification problems, different error estimates can be used such as misclassification error, area under the receiver operating characteristic curve (AUC) or binomial deviance. This procedure is then repeated k times, where each of the k folds is used as a validation set once. The data splitting of this procedure for $k = 5$ is visualized in Figure 3.3. We are then left with k estimates of the error, one for each validation set. The CV error is the sum of these errors averaged over the number of folds. When using the deviance as an error estimate, the CV error can be defined as

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}\text{Dev}(\mathbf{y},\hat{\mathbf{y}}), \tag{3.15}$$

where n_i is the number of observations in fold i.

Leave-one-out cross-validation (LOOCV) is a special case of k-fold CV where $k = n$. LOOCV has the advantage that there is no randomness included in the splits and we are able to use more of the data when fitting the model, thus reducing the bias. However, LOOCV can be computationally expensive as we need to fit and evaluate n models compared to k in k-fold CV.

Bias-Variance Trade off

A simple model, including only a few variables usually have a higher bias and lower variance than a more complex model. If we let the model include many variables then it is able to fit closer to the data, giving it a low bias but higher variance. This compromise is called the bias-variance trade-off, and can be used to justify the use of the penalty term in the elastic net model. We aim to find a model with the right balance, fitting the data as close as possible but without overfitting.

As mentioned the validation set approach can lead to overestimation of the test error, as a subset of the data are held out when fitting the model. The LOOCV on the other hand, will give almost unbiased estimates of the test error as we are able to fit the model to $n-1$ observations. This gives LOOCV a low bias, but also a high variance. When using LOOCV the models in the CV loop are fit using almost the same data, only one observation differs. This implies that the predictions of these models are highly correlated. The mean of highly correlated variables have a higher variance than less correlated variables.

The k-fold CV lies somewhere in between LOOCV and the validation set approach. As only a proportion $\frac{k-1}{k}$ of the data is used when fitting the models, the predictions of the models are less correlated than for LOOCV. This gives the method a higher bias, but also reduces the variance. Thus, when choosing k in CV we need to consider both the computational cost and training time, and the bias-variance trade-off.

The One Standard Deviation Rule

To avoid overfitting when using CV for model selection in general, we can use the one standard error rule. Consider a case where we want to determine a parameter deciding the complexity of the model, θ, where a larger value of θ gives a more complex model. Let $CV(\theta)$ be the CV error when using the complexity parameter θ. Then we can compute the sample standard deviation of the CV error for each θ. The one standard error rule is then to choose the simplest model (smallest θ) for which the error estimate lies within one standard error of the best model. The reasoning behind this rule is that when doing model selection, if the CV errors of two models are nearly equal then we want to choose the simpler model.

3.6.2 The Bootstrap

Bootstrapping can be used to estimate the uncertainty of methods or estimates. For models where we are unable to or it is difficult to calculate confidence intervals analytically, the bootstrap can be used to estimate the confidence intervals of the coefficients. There are multiple ways to perform bootstrapping, both parametric and nonparametric. We will focus on one nonparametric method for bootstrapping regression; the paired bootstrap. The theory in this section is mostly based on Chapter 9 in the book by Givens and Hoeting (2012).

The Bootstrapping Principle

Bootstrapping mimics the process of sampling from the distribution by sampling from the sample distribution. Consider n observations $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, and let them be realizations of the set of random variables $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ that are identically and independently distributed with density function f and cumulative distribution F. Furthermore, let the empirical distribution of the observed data be denoted as $\hat{F}$. Using the observed data we can compute an estimate of an unknown parameter θ, $\hat{\theta} = T(\hat{F})$. If we want to know the variance of this estimate, and we know the distribution F, we could sample n new observations from F and again compute $\hat{\theta}$. We can repeat this sampling many times to simulate the distribution of $T(\hat{F})$, and thus obtain an estimate of $\text{Var}(T(\hat{F}))$. The parameter θ could also be a function of both the data and the distribution, F, such that $\theta = \mathcal{R}(\mathcal{X}, F)$.

The concept of the bootstrap is to do this same sampling, but instead of sampling from the unknown distribution F, we sample from the empirical distribution $\hat{F}$. Consider sampling n observations B times from $\hat{F}$. In practice this means sampling n observations with replacement randomly from the original observations $\mathbf{x}$, B times. Denote a bootstrap pseudo-datasets as $\mathcal{X}^* = \{\mathbf{X}_1^*, \dots, \mathbf{X}_n^*\}$. The bootstrapping principle says that we may approximate the distribution of $\mathcal{R}(\mathcal{X}, F)$ by inserting the estimated distribution $\mathcal{R}(\mathcal{X}^*, \hat{F})$. This makes inference available without making any parametric assumptions.

The Paired Bootstrap

The paired bootstrap, also called bootstrapping the cases, can be used to do inference about the coefficients in a regression setting. Suppose that we have n observation pairs $\mathbf{z}_j = (y_j, \mathbf{x}_j)$ where $j = 1, \dots, n$. If the response and the covariates are measured from individuals selected at random, the pairs $\mathbf{z}_j$ can be seen as observations for i.i.d. random variables $\mathbf{Z}_j = (Y_j, \mathbf{X}_j)$ from a joint distribution. By fitting a regression model to the observed data we obtain a coefficient estimate $\hat{\boldsymbol{\beta}}$.

To obtain a bootstrap pseudo-dataset, sample $\mathbf{Z}_1^*, \dots, \mathbf{Z}_n^*$ at random with replacement from the observations pairs. Fit the regression model to the newly obtained pseudo-dataset to obtain a bootstrap coefficient estimate $\hat{\boldsymbol{\beta}}^*$. Repeat this procedure, sampling B bootstrap datasets and estimating B bootstrap coefficients to get an empirical distribution estimate of $\hat{\boldsymbol{\beta}}^*$. This empirical distribution is then used for inference about $\hat{\boldsymbol{\beta}}$.

Accelerated Bias-Corrected Percentile method

A simple method for doing inference is using the bootstrap estimates to construct a confidence interval. One way to do this is by using the percentile method. To estimate a 0.95% confidence interval using the percentile method, let $\alpha = 0.05$ and simply read off the $[(1-\alpha/2)100]$th and $[(\alpha/2)100]$th percentiles of the empirical distribution of the bootstrap estimates. However, the penalization term in the elastic net minimization problem leads to biased coefficient estimates. Therefore, the percentile method will not give a good estimation of the confidence intervals of the true coefficient.

The accelerated bias-corrected percentile method (BC_a) usually results in more accurate confidence intervals compared to the percentile method (Givens and Hoeting 2012, Chapter 9). The estimation of confidence intervals is improved by estimating and correcting for possible bias, and not taking the bias into account would create wrongly shifted confidence intervals. While the BC$_a$ also uses the percentiles of the bootstrap distribution to estimate the intervals, the method estimates new percentiles, ν_1 and ν_2 and uses them instead of the $1-\alpha/2$ and $\alpha/2$ percentiles.

The BC$_a$ interval is defined as

$$\left[\hat{\theta}^*_{(\nu_1)}, \hat{\theta}^*_{(\nu_2)}\right],$$

where the percentiles are defined as

$$\begin{aligned}\nu_1 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{(\alpha/2)}}{1-\hat{a}(\hat{z}_0 + z_{(\alpha/2)})}\right)\\ \nu_2 &= \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{(1-\alpha/2)}}{1-\hat{a}(\hat{z}_0 + z_{(1-\alpha/2)})}\right),\end{aligned}$$

and Φ is the standard normal cumulative distribution (Givens and Hoeting 2012, Chapter 9). These percentiles depend on the bias-correction factor, $\hat{z}_0$, the acceleration parameter, $\hat{a}$, and the user specified α quantiles of the normal distribution, $z_{(\alpha/2)}$ and $z_{(1-\alpha/2)}$. Givens and Hoeting (2012, Chapter 9) states that one of the simplest choices for estimating the bias correction factor is by letting $\hat{z}_0 = \Phi^{-1}\left(\hat{F}^*(\hat{\theta})\right)$, where $\hat{F}^*$ is the empirical distribution of the bootstrap pseudo data set and $\hat{\theta}$ is the estimate of θ obtained from the original model. In Chapter 4 in the book by Shao and Tu (2012) the bias correction factor is defined as

$$\hat{z}_0 = \Phi^{-1}\left(\#\{\hat{\theta}^*_b \leq \hat{\theta}\}/B\right),$$

where $\#\{\hat{\theta}^*_b \leq \hat{\theta}\}$is the number of bootstrap estimated coefficients less than or equal to $\hat{\theta}$. DiCiccio and Efron (1996) on the other hand, states that the BC_a algorithm, in its simplest from estimates z_0 by

$$\hat{z}_0 = \Phi^{-1}\left(\#\{\hat{\theta}^*_b < \hat{\theta}\}/B\right)$$

The only difference between these two estimates is that the latter only counts bootstrap estimates smaller than the original estimate, and not those that are equal.

The acceleration factor, $\hat{a}$, is defined as

$$\hat{a} = \frac{1}{6}\frac{\sum_{j=1}^{N}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(-j)})^3}{\left[\sum_{j=1}^{N}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(-j)})^2\right]^{3/2}},$$

where $\hat{\theta}_{(-j)}$ is the estimate of θ when observation i is left out of the training set, and $\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{j=1}^{N}\hat{\theta}_{(-j)}$. The acceleration factor is a measure of how fast the standard error changes on the normalized scale and allows for non-constant variance (DiCiccio and Efron 1996). If $\hat{z_0}$ and $\hat{a}$ are both zero, the interval becomes the same as for the percentile method.

Chapter 4

Data

In this chapter we present the dataset and explain how the pre-processing is performed. In addition, we include some exploratory analysis. The data were collected during the ProPACT study (Dotterud et al. 2010) which was a randomized, placebo controlled study, meaning the women were randomly split into a placebo and a treatment group. As mentioned in Chapter 1, the aim of the ProPACT study was to investigate the effect of a probiotic supplement, given to mothers, on the development of AD in their children. After two years the children were clinically examined and either diagnosed with AD or not. We are researching if the miRNA profile in the mothers breast milk is affected by the probiotic supplement and if this can explain the associated risk reduction in AD.

4.1 The MicroRNA Expression Data

A total of 415 women participated in the ProPACT study. Of these women, 211 were given the probiotic supplement while 204 were given the placebo. Several of the mother-child pairs were lost to clinical follow-up, or had missing values at the end of the study. For the probiotic and placebo group they were left with 61 and 63 complete biological samples respectively where the child attended the 2 year follow up.

	Probiotic	Placebo
Mother-child pairs	30	30
History of maternal atopy	9	17
AD at 2 years	10	20
Sensitized at 2 years	7	5*
Sex (female)	14	19
Presence of older siblings	15	15

Table 4.1: Table summarizing baseline characteristics of mother-child pairs. * One missing value of the characteristic: sensitized at 2 years, in the placebo group. Atopic sensitization was assessed by a positive skin prick test (SPT) or elevated specific immunoglobulin E (IgE).

From the mother-child pairs with a complete set of biological samples, 60 breast milk samples taken from the mothers 10 days postpartum, were semi-randomly selected based on some selection criteria. RNA-seq was used to quantify the expression values of 1396 miRNAs in these 60 samples and these data are analyzed in this thesis. We were not able to retrieve the exact information on how the 60 samples were selected or which selection criteria were used. However, we know the selection criteria was based on which supplement the mothers were given and whether or not the child had developed AD after 2 years. In the study by Simpson et al. (2015), which analyzed miRNAs expression in breast milk samples collected 3 months postpartum, 54 samples were also semi-randomly selected based on some selection criteria. The selection criteria they used are described in the article by Simpson et al. (2015).

The baseline characteristics of the mother-child pairs are summarized in Table 4.1. In the 60 samples there are 30 samples from the probiotic group and 30 from the placebo group. In the probiotic group, 10 children developed AD and 20 children did not. Thus, about 33% of the children in the probiotic group of this dataset developed AD, compared to 21% in probiotic group in the ProPACT study. In the placebo group, 20 children developed AD and 10 did not. Thus, about 67% of the children in the placebo group of this dataset developed AD, compared to 34% in placebo group in the ProPACT study. The dataset also contains some metadata relevant to the child developing AD. This includes information about the sex of the child, presence of older siblings and history of maternal atopy. There are 27 male children and 33 female children in the dataset. Parental atopy is known as a risk factor for developing AD (Ravn et al. 2020; Ruiz et al. 1992). There are 9 mothers with a history of AD in the probiotic group, compared to 17 in the placebo group.

4.2 Pre-processing of the microRNA data

Before the analyses are performed, the read counts are normalized and filtered. The preprocessing is performed in R and normalization is performed using the *edgeR* and *limma* packages in R (Robinson, McCarthy et al. 2010; Smyth 2004).

4.2.1 CPM Normalization

Normalization is an essential step in expression analysis and is used to ensure that technical bias has minimal impact on the results. Since higher sequencing depth lead to higher read counts, we scale the values to account for possibly varying sequencing depths. This is done by transforming the read counts to counts per million (cpm), as is suggested by Law et al. (2014). Let $r_{i,j}$ be the read count of miRNA i in breast milk sample j. The total number of mapped reads for sample j is

$$R_j = \sum_{i=1}^{M} r_{i,j},$$

where M is the number of miRNAs. This is also known as the library size of sample j. In addition, the $\log_2$ transformation is commonly used on the cpm values to deal with skewed data. We define the log counts per million for each count as

$$y_{i,j} = \log\left(\frac{r_{i,j} + 0.5}{R_j + 1.0} 10^6\right). \tag{4.1}$$

The counts are added a value of 0.5 to avoid taking log of zero. The library size is offset by one to make sure the denominator of the fraction in expression (4.1) is greater than zero, but also to make sure the fraction is strictly less than one.

4.2.2 TMM Normalization

Robinson and Oshlack (2010) introduce the trimmed mean of M-values (TMM) normalization and show that it dramatically improves result both in simulated and publicly available datasets. This is also an empirical strategy where the aim is to reduce technical bias introduced by the sequencing. However, it is often used in conjunction with cpm normalization as it corrects for another form of bias.

The number of reads of a miRNA depends not only on the quantity of that miRNA in the sample, but also on the quantity of other miRNAs. To see this, consider a hypothetical sample with only two miRNAs, *miR*-1 and *miR*-2, and assume they are both expressed equally with expression value 4. Consider another sample with the same two miRNAs plus two others, *miR*-3 and *miR*-4. Also in this sample all miRNAs have the same true expression value of 4. So, the true library size is 8 in the first sample and 16 in the second. If we sequence both samples to the same sequencing depth, so that both samples get a library size of 8, the read count of *miR*-1 and *miR*-2 would be 4 in the first sample. But in the second sample, we do not count all miRNAs present. The expected number of reads for all the miRNAs in the second sample is 2. We can scale by library size to cpm values, but this will still not show that there actually was an equal amount of *miR*-1 and *miR*-2 in both samples. A correct normalization would multiply the reads in the second sample by a factor of 2. TMM normalization tries to estimate this scaling factor.

To compute the TMM normalization factors, the algorithm first specifies a reference sample. We define two values $M^r_{i,j}$ and $A^r_{i,j}$ for each miRNA i in each sample j as

$$\begin{aligned} M^r_{i,j} &= \log_2\left(\frac{y_{i,j}/R_j}{y_{i,r}/R_r}\right) \\ A^r_{i,j} &= \frac{1}{2}\log_2\left(\frac{y_{i,j}}{R_j}\frac{y_{i,r}}{R_r}\right) \quad \text{for } y_i \neq 0, \end{aligned}$$

where r is the reference sample. For each sample, the M-values are trimmed by a total of 60%, 30% on each side. This leaves 40% of the M-values. The A-values are also trimmed, but only by 5%. This leaves a set, I^*, containing up to 40% of the miRNAs. This set does not need to be the same for all samples. Lastly, to compute the normalization factor for sample j, we take a weighted mean of the remaining M-values. This is defined as

$$TMM^r_j = \exp\left(\frac{\sum_{i\in I^*} w^r_{i,j} m^r_{i,j}}{\sum_{i\in I^*} w^r_{i,j}}\right)$$

where the weights are defined as

$$w^r_{i,j} = \frac{R_j - y_{i,j}}{R_j y_{i,j}} + \frac{R_r - y_{i,r}}{R_r y_{i,r}}.$$

Additionally, in the implementation of this function in in the R package *edgeR*, the normalization factors are scaled such that they multiply to one. This is, however, not mentioned by Robinson and Oshlack (2010). The TMM-value for each sample is then used to scale the library size, R_j, to compute the effective libraray sizes. These are then used instead of the original library sizes, when computing the cpm-values.

Using TMM normalization, we assume that the majority of miRNAs are not differentially expressed. Simulation studies done by Robinson and Oshlack (2010) indicate that the method is robust against deviations to this assumption until about 30% of miRNAs are differentially expressed.

4.2.3 Filtering

The dataset include many lowly expressed miRNAs, and some are not expressed at all in any samples. MiRNAs with overall low cpm values give us little to no information, so these are filtered out before starting the analysis. The reasoning behind filtering is based both on biological and statistical arguments (Chen et al. 2016). One reason is that a miRNA needs to be expressed at least at some level to be biologically interesting. Another reason is that we allow the mean variance relationship to be estimated more reliably by excluding low counts. Lastly, by filtering we reduce the number of tests in an already large multiple hypotheses problem. We filter based on the normalized cpm values in order to account for differences in library sizes but before calculation of the M-values. The reason being that we want to remove low-abundance miRNAs with unreliable M-values. In this thesis, we use a filtering procedure such that only miRNAs with a cpm value larger than 1 in at least 10 samples are kept in the dataset. After filtering, we are left with only 615 miRNAs.

This filter criteria still include many lowly expressed miRNAs. But since we wish to analyze the effect of possible groups of miRNAs, they need not be as highly expressed to be biologically interesting as when examining their individual effects. In the *limma* analysis we do, however, analyze the effect of individual miRNAs, but as we are analyzing possible groups in the penalized regression analysis we use the same filtering for both analyses.

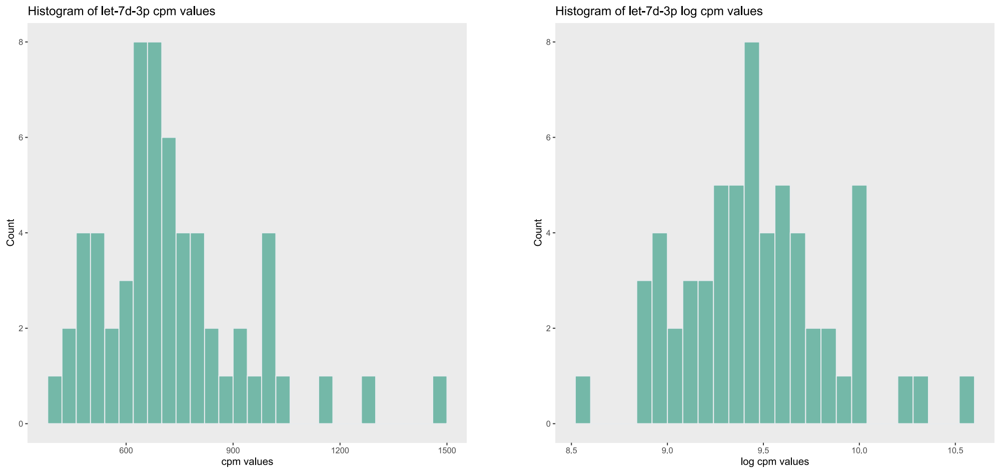
The dataset originally included some miRNAs with identical sequences but which have different names. These were given identical expression values during the RNA sequencing due to the fact that it was not possible to tell them apart. To avoid having many replicates of the same sequence, we join equal miRNAs sequences and give new names that reflect the merging. An overview of this procedure along with the new names are shown in Figure A.1 in Appendix A. This part of the pre-processing was also performed using R.

4.3 Data Exploration

In expression analysis it is common to use the $\log_2$ transformed cpm values. As mentioned, this is due to the fact that the distribution of cpm values is believed to be skewed, and a log transformation is used to shift the data. A histogram of the cpm values before and after a log transformation is shown for two random miRNAs in Figure 4.1 and 4.2. For the miRNA in Figure 4.1 the log transformation seem to transform the values of this miRNA to be more symmetrical, however, for the miRNA in Figure 4.2 the log transform seem to have shifted the values too far and the distribution now looks shifted the other way.

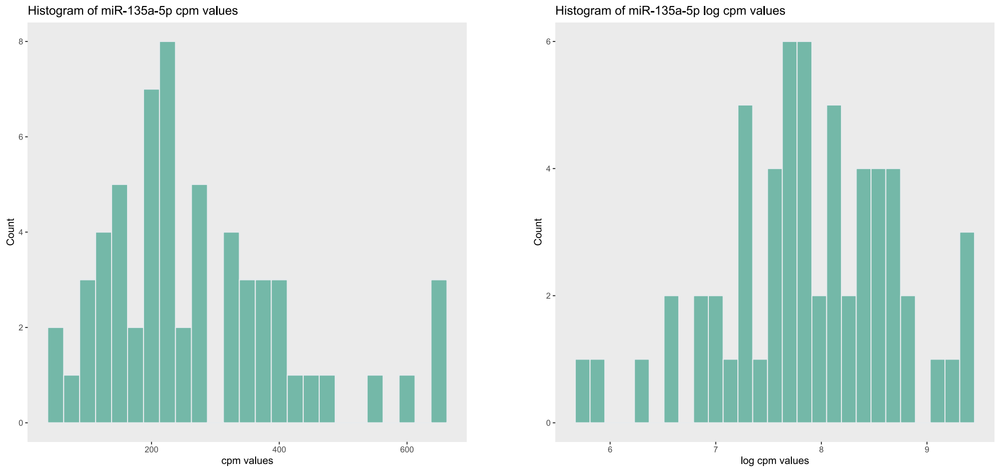
Histograms of sample means and sample variances of the miRNAs are shown in Figure 4.3. As seen in Figure 4.3 (a), most of the miRNAs in the dataset have a mean log cpm value below 5 and the mean over all miRNAs is 4.7. The histogram in Figure 4.3 (b) shows that most samples variances have a value between 0 and 2, and the mean of all sample variances is 0.69. However, there are some miRNAs with considerably higher variances than the rest, with *miR*-142-5*p* having the highest variance at 5.5.

The library sizes of each sample are shown in Figure 4.4. We see from this plot that the library sizes differ a lot between samples. This is likely due to differences in sampling depth when analysing the samples, but may also be due to differences between mothers. This visualizes the importance of accounting for sampling depth by scaling according to library sizes in order to get comparable values of the expression. Furthermore, as mentioned in Section 3.4.2 the variance depends on the sequencing depth, as higher sequencing depth likely leads to higher variance. The varying library sizes might suggest that estimating the variance for individual observations, as done in the *voom* method, will give better estimates than estimating



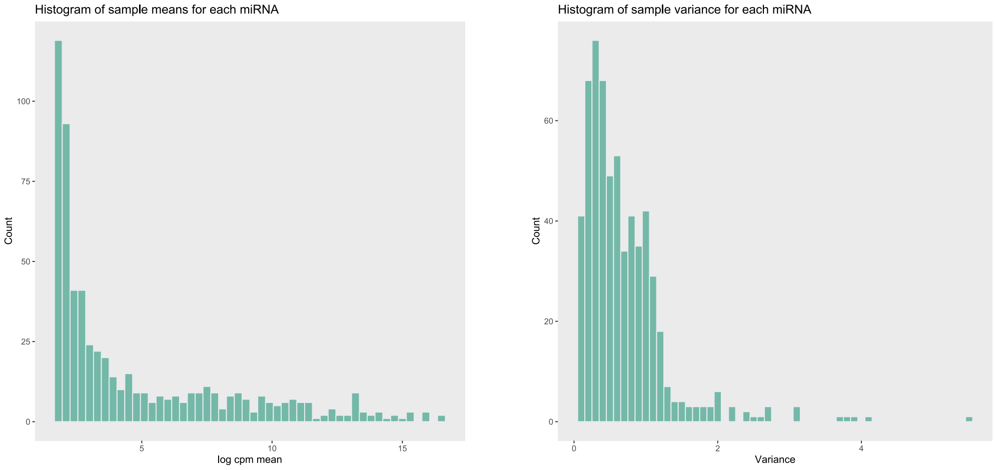
(a) **(b)**

Figure 4.1: Histograms showing the cpm values of *let-7d-3p* before and after a log transformation. The cpm values before a log transformation is shown in (a) and after a log transformation in (b).



(a) **(b)**

Figure 4.2: Histograms showing the cpm values of *miR-135a-5p* before and after a log transformation. The cpm values before a log transformation is shown in (a) and after a log transformation in (b).



(a) **(b)**

Figure 4.3: A histogram of sample means of each miRNA is shown in (a) and a histogram sample variances of each miRNA is shown in (b). The sample means and variances are computed after pre-processing, and given as log cpm values.

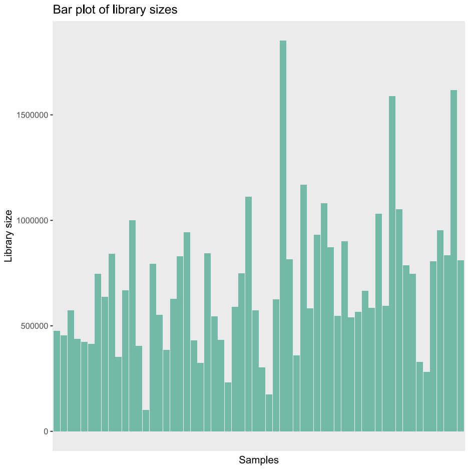
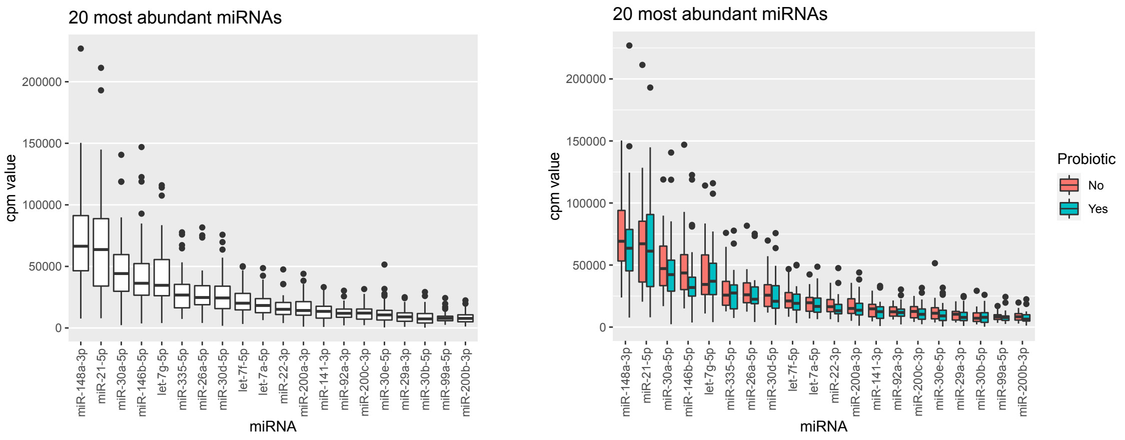


Figure 4.4: Library size of all samples in the dataset. There is one sample collected from each of the 60 mothers, 10 days postpartum. The library size is the total amount of miRNAs in a breast milk sample.



(a) The 20 most abundant miRNAs in the data set.

(b) The 20 most abundant miRNAs in the data set, separated by treatment group.

Figure 4.5: Boxplot of read count values of the 20 most abundant miRNAs in the dataset. The lower and upper hinges correspond to the 25th and 75th percentiles. The upper whisker extends from the hinge to the largest value no further than 1.5 times the inter-quartile range.

the variance only for each miRNA, as done in the *limma-trend* method. Thus, advocating for using *voom* instead of *limma-trend* to analyze these data.

Boxplots of the 20 most abundant miRNAs are shown in Figure 4.5. In Figure 4.5 the box of each miRNA are separated by probiotic and placebo supplement.

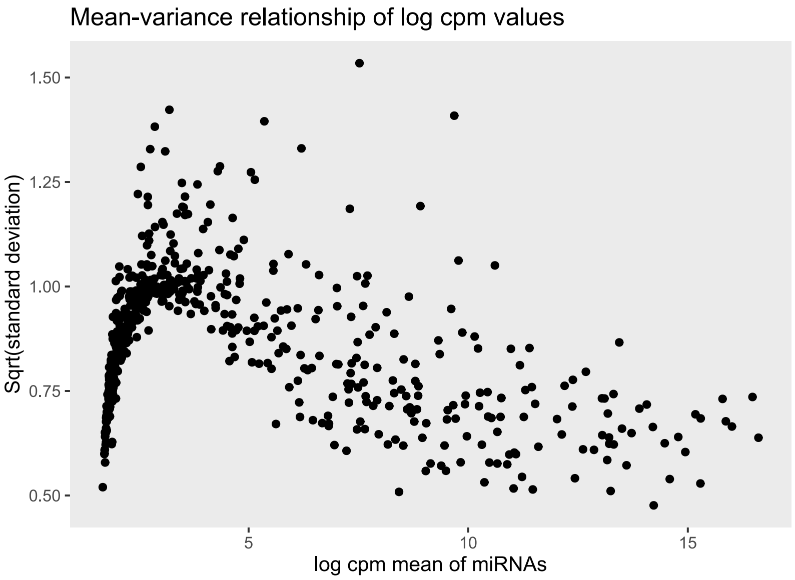


Figure 4.6: Standard deviation of the log cpm value of each miRNA plotted against the mean log cpm value.

The mean variance relationship of the log cpm values are shown in Figure 4.6. The standard deviation of each miRNA is plotted against its mean value. We notice that for mean log cpm values larger than 3 the trend appears descending as we would expect since we are using a log transformation. However, for miRNAs with a mean log cpm value smaller than 3 then trend does not behave as we would expect. This have been suggested to be a sign that we should have filtered out more of the lowly expressed miRNAs, because the variance can not be estimated well in these cases. This might be due to the fact that the expression level of these miRNAs are low inn all samples, and likely zero in many samples, thus it is hard to capture their variance. This might lead to the variance of individual of individual observations estimated by *voom*, being too small for low count values.

Chapter 5

Application of Statistical Methods to Data

In this chapter we present the data analysis plan used to approach the research question. As mentioned, previous papers have established a risk reduction in children developing AD when mothers were given a probiotic supplement before birth and while breastfeeding, compared to mothers given a placebo. We are interested in examining if miRNAs in breast milk at 10 days postpartum are possible contributors to this risk reduction. This research question is handled using two separate analyses. In the first analysis we are investigating if probiotics given to mothers before and after giving birth affect the miRNA profile in breast milk. We do this by examining whether any miRNAs are differentially expressed between mothers given probiotics compared to placebo. This analysis is performed using *voom*. In the second analysis, we want to examine if any miRNAs in the breast milk are associated with the children developing AD. This is done by fitting an elastic net model to the data, with AD as the response variable and miRNAs as the covariates. Before these two analyses, we perform an exploratory analysis of the data using hierarchical clustering.

5.1 Clustering of all MicroRNAs

We use hierarchical clustering to cluster both the breast milk samples from the mothers (observations) and the miRNAs (variables). Most miRNAs have low read counts and the distribution of the data is right skewed. Since both Euclidean and correlation based distances are sensitive to skewness and might not work well for this type of data (Datta and Nettleton 2014). As seen in Chapter 4, a $\log_2$ transformation somewhat reduces the skewness of the data and this will make it more applicable for these methods. For this reason, the clustering is performed on the log transformed cpm values. In addition, the log cpm values are centered to have a zero mean and scaled to have a standard deviation of one. This scaling is done only by miRNAs and not by samples. A detailed explanation of why scaling by miRNAs is necessary for Euclidean and correlation based distances is provided in Appendix A.

Three different clustering approaches are explored, all using the same link function but different dissimilarity measures. In all three approaches, we use the Ward linkage function as this linkage has been shown to perform well for gene clustering and significantly better than other linkages for clusters with nearly equal sample sizes (Freyhult et al. 2010; Y. Zhao et al. 2021).

We perform hierarchical clustering with two different correlation based distances, one using the absolute correlation and one using the correlation. These distances are defined as in Expression (3.2) and (3.1) respectively. When clustering using the absolute correlation based distance, miRNAs that are highly positively correlated and highly negatively correlated will be clustered together. We do not know how the probiotic possibly affects the miRNAs; we may think it works by down-regulating some while up-regulating others. Thus, the absolute correlation distance might be useful. However, using a correlation distance can also be advantageous, as it might be easier to group only positively correlated miRNAs. The reason this might be easier is that more noise might be introduced when using absolute correlation due to the fact that both positive and negative miRNAs are considered similar.

For both approaches, we create heatmaps and annotate the samples by probiotic supplement, history of maternal atopy and AD. This is done to obtain a visualization of clustering by any of these characteristics.

Additionally, we include a clustering approach using the Euclidean distance for completeness. The analysis is otherwise the same as the one described for the correlation based distances.

5.2 Analysis of the Probiotics Effect on MicroRNAs

In this section, we define the linear model to be used in the *voom* analysis. In this analysis we examine the effect of the probiotic supplement on miRNAs expression in breast milk.

5.2.1 The Multiple Linear Regression Model

We first define a linear model which is fitted to each miRNA. Since the samples in the dataset were not randomly selected, but based on some selection criteria as described in Chapter 4, we chose to adjust for AD. This is done by including AD as a covariate in the model. An interaction term between AD and probiotic supplement could also be included, but as seen in the previous analysis in the project thesis, this did not appear to be necessary (Omdal 2021). Maternal atopy or allergy have been shown to change the composition of factors in breast milk (Munblit et al. 2015). Thus, we choose to include maternal atopy as a binary covariate in the model. Furthermore, the gender of the child could also influence the expression of miRNAs in breast milk. The study by Xi et al. (2016) found that $miRNA$-30b and $miRNA$-378 were higher in the colostrum (breast milk the first 1-5 days after birth) received by girls than received by boys. Thus we also include gender as a covariate. We define the linear regression model as

$$\begin{aligned} y_{i,j} &= \mathbf{x}_j^T \boldsymbol{\beta}_i + \varepsilon_{i,j} \qquad (5.1)\\ &= \beta_{i,0} + x_{j,p}\beta_{i,1} + x_{j,ad}\beta_{i,2} + x_{j,matatopy}\beta_{i,3} + x_{j,gender}\beta_{i,4} + \varepsilon_{i,j}, \qquad (5.2)\end{aligned}$$

where $y_{i,j}$ is the log cpm value of miRNA i from sample j, $x_{j,p}$ is a binary covariate equal to one if sample j is from a mother given the probiotic supplement and zero if she was given the placebo supplement. The covariate $x_{j,ad}$ is binary and equal to one if the child from sample j had developed AD after 2 years and zero if not. The covariate $x_{j,matatopy}$ is binary and equal to one if the mother from sample j has a history of atopy and zero if not. The covariate $x_{j,gender}$ is binary and equal to one if the child from sample j is a boy and zero if it is a girl. This linear model is fitted to each miRNA using the *limma* package in *R*.

Additionally, we fit a linear model without the maternal atopy and gender covariates for comparison.

Because of the varying library sizes between samples in the dataset, we believe that the *voom* method has an advantage over limma-trend and use only *voom* to perform the analysis. Using the function *voom()* in the *limma* package, the mean-variance trend is estimated as explained in Section 3.4.2. This is done using the raw read counts and not the log cpm normalized values. Normalization is done internally in the function. Using the estimated trend, *voom* precision weights are calculated and used as input into the original *limma* pipeline along with the log cpm values. For each covariate, we obtain the log fold change between the two conditions and the corresponding p-value and adjusted p-value of that estimated coefficient. We are only interested in the differential expression between the probiotic and placebo group. To account for testing 605 hypotheses simultaneously, we consider the BH adjusted p-values to control the FDR. We set the threshold, or cut off value, of the adjusted p-values to be 0.1 as was done by Simpson et al. (2015) when analysing the samples collected 3 month postpartum.

5.2.2 Clustering the top 20 MicroRNAs

After performing the *voom* analysis we rank the miRNAs according to the p-value. We proceed with further analysis on the 20 top ranked miRNAs by creating heatmaps that include only these miRNAs. We replicate the clustering analysis described in 5.1, including all three clustering approaches.

5.3 Analysis of MicroRNAs Association with Development of AD

In this section, we describe the methods used in the second part of the analysis. This analysis is separate from the *voom* analysis. In this part of the analysis, we want to examine whether there are any miRNAs in the breast milk that are associated with the child developing AD. We hypothesize that miRNAs in breast milk cooperate, when consumed by the children, to reduce the risk of developing AD. Thus, we want to identify miRNAs associated with AD in a way such that that highly correlated miRNAs are either included or not together. To do this we use a logistic elastic net model with AD as the response variable and the miRNAs as covariates. The aim is to utilize the variable selection property of elastic net to select all miRNAs with an association to the child developing AD.

We present two cross-validation procedures for selecting the α and λ parameters in the elastic net model. We also present a way of estimating confidence intervals of the elastic net coefficients using bootstrapping.

5.3.1 Penalized Regression Model

In the penalized logistic regression model we let AD be the response variable. As in the first analysis, this a binary categorical covariate, defined as

$$y = \begin{cases} 1 & \text{if the child had developed AD after 2 years} \\ 0 & \text{if the child had not developed AD after 2 years.} \end{cases}$$

We let the cpm normalized expression values of the 605 miRNAs be included as covariates. We argue that including covariates for the mothers history of atopy and gender of the child might be needed in this analysis as well. As mentioned, the gender of the child and history of atopy of the mother might influence the miRNAs expression. Additionally, both of these factors affect the probability that a child develops AD. Thus, we can argue that this should be included in the model to reduce noise, but also to adjust for possible confounding. By including them using dummy variable encoding, one of the levels is included in the intercept. As we usually do not shrink the intercept, this would lead to only shrinking one of the levels. To allow shrinkage of both levels, we create an extended design matrix using one hot encoding. This would in the regular linear regression case be problematic as it would lead to an unidentifiable model. The covariate matrix would no longer have full rank and thus the solution (estimated coefficient vector) would no longer be unique. However, as we use penalized regression this is not a problem and we get a unique solution for the coefficient vector. We include one binary covariate for girl and one for boy. In the same way we include one binary covariate for the mother having a history of atopy and one for no history of atopy. The resulting model has a total of 609 covariates and one intercept. For the generalized linear model we let the linear predictor be defined as

$$\eta = \beta_0 + \mathbf{x}_{miRNAs}^T \boldsymbol{\beta}_{miRNAs} + \mathbf{x}_{basis}^T \boldsymbol{\beta}_{basis},$$

where

$$\mathbf{x}_{basis} = [x_{maternalAtopy}, x_{noMaternalAtopy}, x_{female}, x_{male}]^T$$

are the demographic and clinical covariates and $\mathbf{x}_{miRNAs}$ is a vector of covariates containing all 605 miRNAs. The coefficients $\boldsymbol{\beta}_{miRNA}$ and $\boldsymbol{\beta}_{basis}$ are the corresponding unknown coefficient vectors.

To connect the mean of the response variable to the linear predictor we use the logit link function such that

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \eta.$$

The elastic net minimization problem becomes

$$\min_{(\beta_0, \boldsymbol{\beta}) \in R^{610}} \left[-\ln L(\beta_0, \boldsymbol{\beta}) + \lambda P_\alpha(\boldsymbol{\beta})\right],$$

where

$$\boldsymbol{\beta} = [\boldsymbol{\beta}_{miRNAs}, \boldsymbol{\beta}_{basis}]^T$$

and

$$P_\alpha(\boldsymbol{\beta}) = (1-\alpha)\frac{1}{2}||\boldsymbol{\beta}||_{L_2}^2 + \alpha||\boldsymbol{\beta}||_{L_1}.$$

Prior to fitting the model and estimating the coefficients we standardize the covariates. Zwiener et al. (2014) suggest that standardizing, especially scaling the miRNAs to have equal variances, is an important step when analyzing RNA-seq data using regularized regression. The elastic net is implemented in the *glmnet* package in R, and we use this package to fit the model to the data.

5.3.2 Nested Cross-Validation

When using an elastic net model we need to specify both the α and the λ parameter. As seen from the theory presented in Section 3.5, both parameters influence the level of penalization. The λ parameter is usually determined using CV, and this procedure is implemented in the *glmnet* package. However, there is no standard procedure for choosing α. One approach is to determine the α value based results from previous studies with a similar aim. Thus, just choosing a value without using the data. Another popular method is to perform CV of λ for a grid of α values, and choosing the (α, λ)-pair that results in smallest CV error. However, with this method, we cannot use the one-standard error rule. In addition, the results are dependent on the splitting of the data into folds and minor changes in these folds can lead to great variations in the CV error. Thus, the chosen parameters may vary a lot for different splits.

We propose a way to use CV to determine both α and λ and we will refer to this method as nested CV. Using this method we reduce the variation introduced by the splitting of the data.

Additionally, it is possible to use the one standard error rule. The proposed method is inspired by Wang et al. (2007), He et al. (2020) and Brownless (2020).

Choosing the α parameter can be seen as a form of model selection. This is because when choosing α we are choosing between lasso, ridge and everything in between. The nested CV algorithm has an outer loop for choosing α, and an inner loop finding the corresponding best λ. Thus, we can think of the nested CV as having an outer loop for model selection and an inner loop for parameter tuning. However, the nested CV is only used to find and optimal α value. Outside the nested CV, this α value is used in a new CV to determine the best λ for this α using the whole dataset. The reason we run a new CV is that when determining λ inside the nested CV we are not able to use the entire dataset, as is explained below. The outer and inner folds are determined as shown in Algorithm 2.

The nested CV algorithm is explained in detail in algorithm 3, but we give an overview of it here. For each α, we divide the observations into the same $k = 10$ folds. As in regular k-fold CV, each fold is used as a test set once. Consider the case where fold i is used as the test fold when evaluating $\alpha = a$. Then, as in regular CV the rest of the folds are used for training. We split this training set into 10 inner folds. Using $\alpha = a$ we perform regular CV using the function *cv.glmnet()* to choose the best λ for $\alpha = a$. After finding the best lambda, either by using λ corresponding to the smallest deviance or using the one standard error rule, this model is evaluated on the outer test set by calculating the binomial deviance. This way the model with $\alpha = a$ and the chosen λ can be evaluated on unseen data. Then, this is repeated, letting each of the outer folds be the test set once. After the outer CV, we can compute the mean CV-error for $\alpha = a$. This process is repeated for each α. We then compare the CV-error of all α values, and choose the smallest one. Using this chosen α and all the data available, we perform a new CV to determine the best λ. We do not set aside any data for a test set, as we are not really interested in prediction in this thesis, only variable selection, and want to use as much as possible of the limited data for this purpose.

Algorithm 2: *Create.CV.folds*

```
Input  : Number of observations $n$, number of folds $N$
$F \leftarrow [1, 2, \dots, N]^T$                 /* Create vector of numbers from 1 to N  */
$T \leftarrow \text{ceiling}(n/N)$                /* Find smallest integer $\geq n/N$ */
$F_{rep} \leftarrow F$ repeated $T$ times         /* Repeat $F$, $T$ times  */
$F_{rep} \leftarrow F_{rep}[1:n]$                 /* Get $n$ fold ID's */
$F_{id} \leftarrow$ shuffle $F_{rep}$             /* Get fold ID's in random order */
Output: $F_{id}$ vector of fold ID's
```

In this analysis we perform nested CV once where we chose the λ corresponding to the smallest deviance and once where we use the one standard error rule. We will refer to these two approaches as nested CV with λ_{min} and nested CV with λ_{1SE}. For both approaches, we use 10 inner and 10 outer folds. We define the sequence of α values as 19 values from 0.1 to 1 incremented by 0.05. Thus, including lasso but not ridge regression as we want to perform variable selection and are certain that not all covariates are relevant.

5.3.3 Repeated Cross-Validation

Repeated CV is the other method we use to choose α and λ. This is based on the common way of performing CV of an elastic net model, by performing CV over a grid of λ and α values. But instead of only doing this CV once, the CV and the splitting into folds is repeated 10 times.

We use the same sequence of α values as defined for the the nested CV. The λ sequence is

Algorithm 3: *Nested Cross-Validation*

```
Input  : Model matrix $\mathbf{X}$, Response variable $\mathbf{y}$, $\alpha$-sequence $\boldsymbol{\alpha}_{seq}$, number of folds in
         inner loop $k_{inner}$, number of folds in outer loop $k_{outer}$, lambda type $\lambda$-type
         (either choosing $\lambda$ with minimum deviance or choosing $\lambda$ using the one
         standard deviation rule)
$n_{obs} \leftarrow$ number of rows in $\mathbf{X}$                /* Get number of observations */
$\mathbf{F}_{id,outer} \leftarrow$ Create.CV.folds($n_{obs}, k_{outer}$)          /* Define outer folds */
Initialize $\mathbf{F}_{id,inner}$     /* Matrix to store inner fold id's for each outer fold
 */
for $k \leftarrow 1$ to $k_{outer}$ do     /* For each outer fold, define the inner folds */

    $n_{obs,k} \leftarrow$ number of rows in $\mathbf{X}[\mathbf{F}_{id,outer} = k, ]$
    $\mathbf{F}_{id,inner,k} \leftarrow$ Create.CV.folds($n_{obs,k}, k_{inner}$)      /* Define $k$'th inner folds */
    $\mathbf{F}_{id,inner}[, k] \leftarrow F_{id,inner,k}$            /* Store $k$'th inner folds in matrix */
end
CV$_{error} \leftarrow \emptyset$                  /* Vector to store CV errors for each $\alpha$ */
foreach $\alpha \in \boldsymbol{\alpha}_{seq}$ do
    Dev $\leftarrow 0$                               /* Used to compute CV error */
    for $k \leftarrow 1$ to $k_{outer}$ do
        $\mathbf{X}_{train,outer}, \mathbf{y}_{train,outer} \leftarrow \mathbf{X}[\mathbf{F}_{id,outer} \neq k, ], \mathbf{y}[\mathbf{F}_{id,outer} \neq k]$
        $\mathbf{X}_{test,outer}, \mathbf{y}_{test,outer} \leftarrow \mathbf{X}[\mathbf{F}_{id,outer} = k, ], \mathbf{y}[\mathbf{F}_{id,outer} = k]$
        cv-fit $\leftarrow$ cv.glmnet($\mathbf{X}_{train,outer}, \mathbf{y}_{train,outer,k}, \mathbf{F}_{id,inner}, \alpha, \lambda$-type )      /* Use
         cv.glmnet function to decide $\lambda$ */
        $\lambda_{best} \leftarrow$ best $\lambda$ from cv-fit                  /* Either $\lambda_{min}$ or $\lambda_{1SE}$ */
        model-fit$_k \leftarrow$ elastic net model with $\alpha$ and $\lambda_{best}$
        Pred$_k \leftarrow$ predict($\mathbf{X}_{test,outer}$, model-fit$_k$)  /* Predict on test set using the
         chosen model */
        Dev$_k \leftarrow$ compute-deviance(Pred$_k$, $y_{test,outer}$)
        Dev $\leftarrow$ Dev + Dev$_k / n_{obs,k}$   /* Divide by number of observations in test
         set */
    end
    CV$_{\text{error},\alpha} \leftarrow$ Dev$/k.outer$
    CV$_{\text{error}} \leftarrow$ CV$_{\text{error}} \cup$ CV$_{\text{error},\alpha}$
end
$\alpha_{best} \leftarrow \alpha$ with smallest CV$_{\text{error}}$
cv-fit $\leftarrow$ cv.glmnet($\mathbf{X}, \mathbf{y}, k_{outer}, \alpha_{best}, \lambda$-type )       /* Use cv.glmnet function to
 decide $\lambda$ */
$\lambda_{best} \leftarrow \lambda$ from cv-fit                          /* Either $\lambda_{min}$ or $\lambda_{1SE}$ */
model-fit $\leftarrow$ elastic net model with $\alpha_{best}$ and $\lambda_{best}$
Output: model-fit, $\alpha_{best}, \lambda_{best}$
```

defined as 200 values equally spread out from exp(1) to exp(−5). We let the number of folds in the k-fold CV be 10, and the number of repetitions also be 10.

Algorithm 4: Repeated Cross-Validation

```
Input Model matrix $\mathbf{X}$, Response variable $\mathbf{y}$, $\boldsymbol{\lambda}_{seq}$ sequence, $\alpha_{seq}$ sequence, number
  of repetitions $N$, number of folds $k$
$n_{obs} \leftarrow$ number of rows in $\mathbf{X}$
Initialize $\mathbf{F}_{id}$            /* Matrix to store fold id's for each repetition */
for $i \leftarrow 1$ to $N$ do
    $\mathbf{F}_{id,i} \leftarrow$ Create.CV.folds($n_{obs}, k$)      /* Define folds for i'th repetition */
    $\mathbf{F}_{id}[, i] \leftarrow \mathbf{F}_{id,i}$
end
for $i \leftarrow 1$ to $N$ do
    foreach $\alpha \in A$ do
        cv-fit $\leftarrow$ cv.glmnet($\mathbf{X}, \mathbf{y}, \mathbf{F}_{id}[, i], \alpha, \boldsymbol{\lambda}_{seq}$)   /* Use cv.glmnet() to calculate
          CV error of current $\alpha$ and all $\lambda \in \boldsymbol{\lambda}_{seq}$ */
        for $\lambda \in \lambda_{seq}$ do
            $CV_{error,i,\lambda,\alpha} \leftarrow CV_{error}$ from cv-fit      /* Store CV error from $\lambda, \alpha$ in
              repetition $i$ */
        end
    end
end
foreach $\lambda, \alpha$ pair do
    $mean.CV_{error,\lambda,\alpha} \leftarrow \sum_{i=1}^{10} CV_{error,i,\lambda,\alpha}$    /* Calculate mean CV error over the
      repetitions */
end
$\lambda_{best}, \alpha_{best} \leftarrow \lambda, \alpha$ pair with smallest $mean.CV_{error,\lambda,\alpha}$
model-fit $\leftarrow$ elastic net model with $\lambda_{best}, \alpha_{best}$
Output: model-fit, $\alpha_{best}, \lambda_{best}$
```

In the repeated CV method, a regular CV is performed for each (α, λ)-pair. In this way we can compute a CV error for all 200 × 19 possible combinations. This CV is then repeated 10 times. We are left with 10 CV errors for each (α, λ)-pair. By averaging over these 10 errors, we get a mean CV error for each pair. When comparing these mean CV errors, we choose the pair with the smallest error. This creates a more robust procedure compared to only performing CV over the grid of pairs once. In contrast to the proposed nested CV, this method treats α and λ equally. However, using repeated CV, we cannot use the one standard deviation rule. The repeated CV algorithm is presented in Algorithm 4.

We compare the results of the nested CV and the repeated CV. We will continue to compare the performance of the nested CV and the repeated CV when using bootstrapping. However, since the nested CV method has yet to be tested, we only include the results to compare this method to the repeated CV. For the analysis of the breast milk samples we consider the estimated coefficients of the model with model parameters chosen by the repeated CV.

5.3.4 Inference Using The Bootstrap

In order to estimate the uncertainty of the coefficient estimates, $\hat{\boldsymbol{\beta}}$, we compute confidence intervals. As we are not able to calculate confidence intervals analytically when using the elastic net, we aim to estimate them using the bootstrap.

We let the number of bootstrap samples be $B = 1000$. For each $b \in B$, we sample 60 observations at random with replacement from the mother-child pairs in the original dataset. This creates 1000 pseudo-dataset. For each pseudo-dataset, $\mathcal{X}^*$, we use the repeated CV and the nested CV algorithm to choose the value of the model parameters and fit the chosen elastic net model, as specified in 5.3.1. The model and its coefficient values are stored for all 1000 bootstrap datasets.

To summarize the result of the bootstrapping we create a barplot showing the proportion of times each covariate was included in the model (their coefficient was not set to zero). As we cannot include all 607 covariates in this plot, we only include those with a bootstrap coefficient estimate not equal to zero in at least 500 (50%) of the bootstrap samples. Additionally, we create a boxplot of the bootstrap estimated coefficients values for these same covariates. Bootstrapping is performed for both CV procedures for comparison of their selected parameters and how the parameters affect the estimated coefficients.

As the percentile method does not account for the coefficient estimates being biased, we aim to estimate confidence intervals of the coefficients using the BC_a method. When constructing the confidence intervals based on bootstrapping, we include the whole model selection procedure, thus including the CV to choose the parameters α and λ and the shrinking of the coefficients. Therefore, the confidence intervals reflect the variability of the entire variable selection procedure. The confidence intervals are constructed as described in Section 3.6.2. However, some alterations are made to the BC_a method to make it more applicable to the elastic net model. The calculation of the bias correction factor is modified. In Section 3.6.2 we present two very similar ways of calculating the bias correction factor. The only difference between the two is that one only counts bootstrap estimates smaller than the original estimate, while the other also counts those that are equal to the original estimate. In most cases, these two methods will create identical confidence intervals. However, we argue that since the elastic net shrinks the coefficients towards zero, these two methods may differ substantially. The reason is that for coefficients with $\hat{\beta} = 0$, many of the bootstrap estimates will likely also be zero. To illustrate this, imagine that the true, but unknown, parameter of interest is $\beta = 0$. Let the estimate from the original dataset be $\hat{\beta} = 0$ and assume that most of the bootstrap estimates also are zero, with the remaining being either positive or negative. Then, if we only count the number of bootstrap estimates smaller than zero, we would get that the bias correction estimate would be large and negative, when we intuitively would believe it should be close to zero. Similarly, if we count the number of bootstrap estimates smaller than or equal to zero, the bias correction estimate would be large and positive. However, with the alteration presented in this thesis, we would get that the bias is close to zero.

The alteration can be seen as an average of the two bias estimates presented in Section 3.6.2. We estimate the bias as

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\beta}_b^* < \hat{\beta}\} + \frac{\#\{\hat{\beta}_b^* = \hat{\beta}\}}{2}}{B}\right). \tag{5.3}$$

The bias correction factors presented in Section 3.6.2 and the bias correction factor in Expression (5.3) will likely be identical for cases where $\hat{\beta} \neq 0$, as the count $\#\{\hat{\beta}_b^* = \hat{\beta}\}$ then should be zero (due to the estimates being continuous and not likely to be repeated when they are different from zero).

In addition to the modification described above, we introduce another small alteration to avoid numerical computation issues. If the numerator in the inverse cumulative normal function of Expression (5.3) is zero, we let the numerator be equal to some small number, 0.001, to avoid $\hat{z}_0$ being set to negative infinity. However, this modification should rarely be necessary due to the first alteration fixing the issue of very negative bias estimates for cases with many or all repeated values.

We perform the bootstrapping and implement the accelerated bias-corrected method for estimating confidence intervals in R. We estimate 95% confidence intervals of all coefficients included in the chosen elastic net model. For ease in the presentation of the results we only include the estimates of coefficients included in at least 500 out of the 1000 bootstrap samples.

Chapter 6

Results

In this chapter we present the results of the two analyses described in the previous chapter; the probiotics effect on miRNAs and miRNAs association with AD. Before the results of these analyses are presented, we present the result of the exploratory clustering analysis.

6.1 Clustering All MicroRNAs

Here we present the result of the hierarchical clustering approaches described in Section 5.1.

Dendrograms and heatmap from hierarchical clustering using the Ward linkage function and the correlation based distance are shown in Figure 6.1. In this and all the following heatmaps, the sample dendrograms are shown at the top, and the miRNA dendrogram vertically on the left. The samples are annotated with information about each sample shown below the dendrogram. This is information about the supplement given to the mother, her history of dermatitis, and if the child developed atopic dermatitis. By inspecting the sample dendrogram, clustering into 5 groups seem reasonable. We notice that there is no clear grouping of the samples that can be easily explained by the annotated variables. When considering the miRNA dendrogram, the branches are very short at the bottom and longer at the top. There is one group with a particularly long branch compared to the others located at the top of the miRNA dendrogram. As we can see from the heatmap, all miRNAs in this group are highly expressed in the second sample group, counting from the left. There are a total of 62 miRNAs in the mentioned cluster, and 6 samples in the sample group where they have high expression values. Of these 6 samples, 5 children did not develop AD. In the one sample where the child developed AD, the mother had a history of AD and received the placebo.

Dendrograms and heatmap from hierarchical clustering using the Ward linkage function and the absolute correlation based distance are shown in Figure 6.2. The dendrogram of the samples shows a possible cluster containing 6 samples with a longer branch in the middle. These are the same samples that showed a similar pattern in the previous heatmap. The miRNAs show possible natural grouping into 4, 5 or 6 groups. For all three choices of the number of clusters, one cluster is much larger than the rest.

Dendrograms and heatmap from hierarchical clustering using the Ward linkage function and Euclidean distance are given in Figure A.3 in Appendix A. We will discuss the clusters and heatmaps in more detail after presenting the results from the two main analyses.

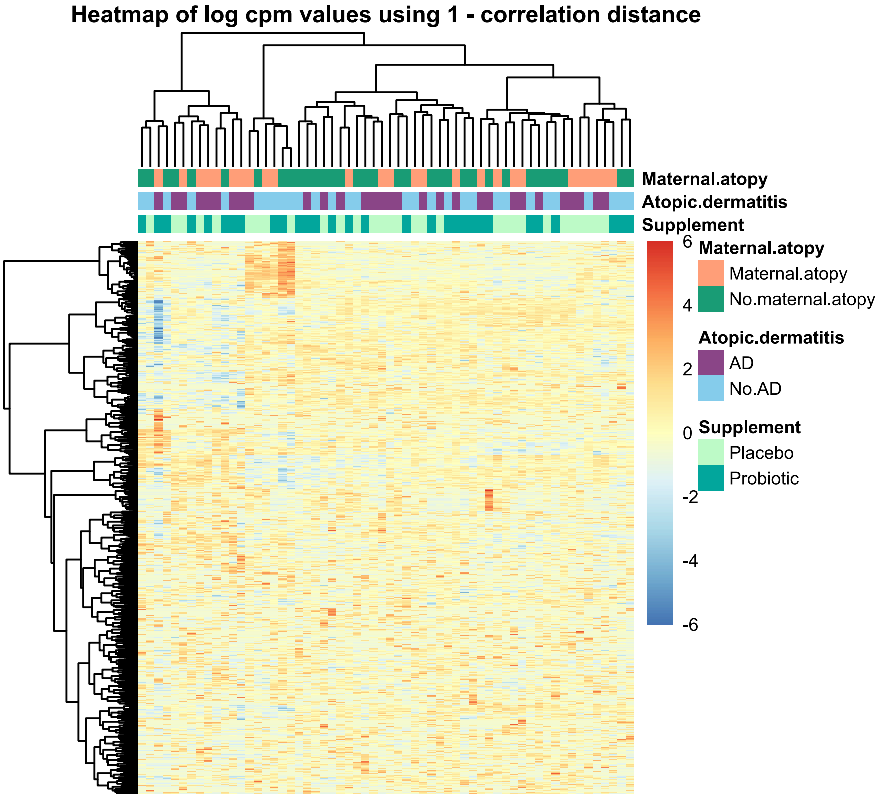


Figure 6.1: Heatmap with dendrograms of the hierarchical clustering of the log cpm values scaled by row. The samples are given by the columns and the miRNAs by the rows. Clustering is done using 1− correlation as dissimilarity measure and the Ward linkage function. Each sample is annotated by three variables; the supplement given to the mother, her history of dermatitis, and if the child developed dermatitis.

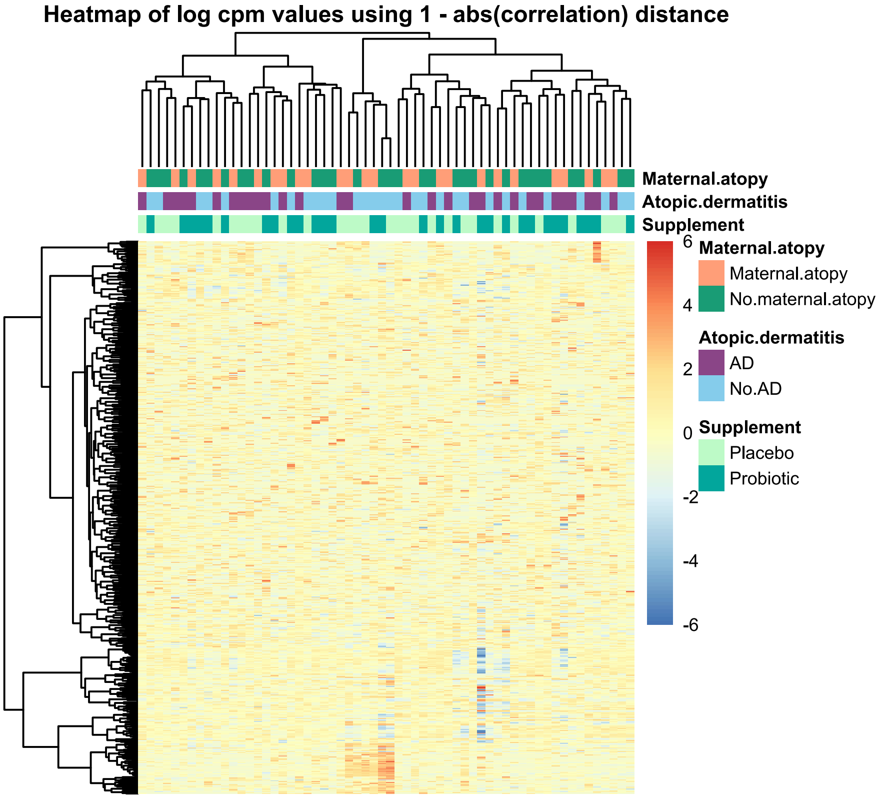


Figure 6.2: Heatmap with dendrograms of the hierarchical clustering of the log cpm values scaled by row. The samples are given by the columns and the miRNAs by the rows. Clustering is done using $1 - |\text{correlation}|$ as dissimilarity measure and the Ward linkage function. Each sample is annotated by three variables; the supplement given to the mother, her history of dermatitis, and if the child developed dermatitis.

6.2 Probiotics Effect on MicroRNAs

In this section we present the results of the analysis testing for differential expression of any miRNAs between the probiotic and placebo breast milk samples. We first present the results of the *voom* analysis and rank the miRNAs by *p*-value, then we present the results of the cluster analysis considering only these miRNAs.

6.2.1 Differentially Expressed MicroRNAs

The probiotic supplement is significantly associated with differential expression of only one miRNA, *miR*-577. The probiotic coefficient in the linear model for this miRNA have a raw *p*-value of 1.942 exp−5 and adjusted *p*-value of 1.174 exp−2. The log fold change between the two groups of this miRNA is −2.067. The miRNA *miR*-323*a*-3*p* also have a low adjusted *p*-value of 0.154. These results and an overview of the top 20 miRNAs, ranked by *p*-value, is shown in Table 6.1. An overview of the top 50 miRNAs is given in Table A.2 in Appendix A.

A total of 47 miRNAs have a raw *p*-value below 0.05 but apart from the two previously mentioned they all have an adjusted *p*-value above 0.4. All of the 20 miRNAs have an average log expression value between 2 and 5. As seen in Chapter 4, this is the range where the mean of most miRNAs lies.

miRNA	$\log_2$ FC	Average log cpm value	*p*-value	adj. *p*-value
miR-577	-2.067	-4.606	0.000	0.012
miR-323a-3p	1.072	3.657	0.001	0.154
let-7e-3p	-0.990	-3.200	0.002	0.405
miR-2116-5p	0.887	3.053	0.003	0.405
miR-200c-5p	-0.922	-3.000	0.004	0.405
miR-517a/b-3p	-1.388	-2.980	0.004	0.405
miR-4668-5p	0.750	2.928	0.005	0.405
miR-33b-3p	1.037	2.816	0.006	0.486
miR-383-5p	0.858	2.766	0.007	0.495
miR-519a-5p	-1.238	-2.702	0.009	0.516
miR-199a/b-3p	-0.553	-2.607	0.011	0.516
miR-6516-5p	0.953	2.581	0.012	0.516
miR-582-3p	-1.105	-2.555	0.013	0.516
miR-378c	-0.336	-2.554	0.013	0.516
miR-34a-5p	-0.593	-2.516	0.014	0.516
miR-374b-3p	-0.526	-2.515	0.014	0.516
miR-3177-3p	0.775	2.466	0.016	0.516
miR-550a-3-5p	0.891	2.438	0.017	0.516
miR-3605-3p	0.855	2.425	0.018	0.516
miR-548ai	0.831	2.418	0.018	0.516

Table 6.1: Top 20 miRNAs ranked by *p*-value. The estimated $\log_2$ fold change between the probiotic and placebo group is given in the $\log_2$ FC column and average log cpm value is the mean log cpm value over all samples. The adjusted *p*-values are computed using the Benjamini and Hochberg method for controlling the FDR.

The estimated mean variance trend is shown in 6.3. In the same figure is the square root of the residual standard deviation for each miRNA after fitting the linear model to each miRNA is plotted against the converted mean $\log_2$ count. Note that this is not the same plot as in Figure 4.6 in Chapter 4 where we plot the square root of the standard deviation of the log cpm values

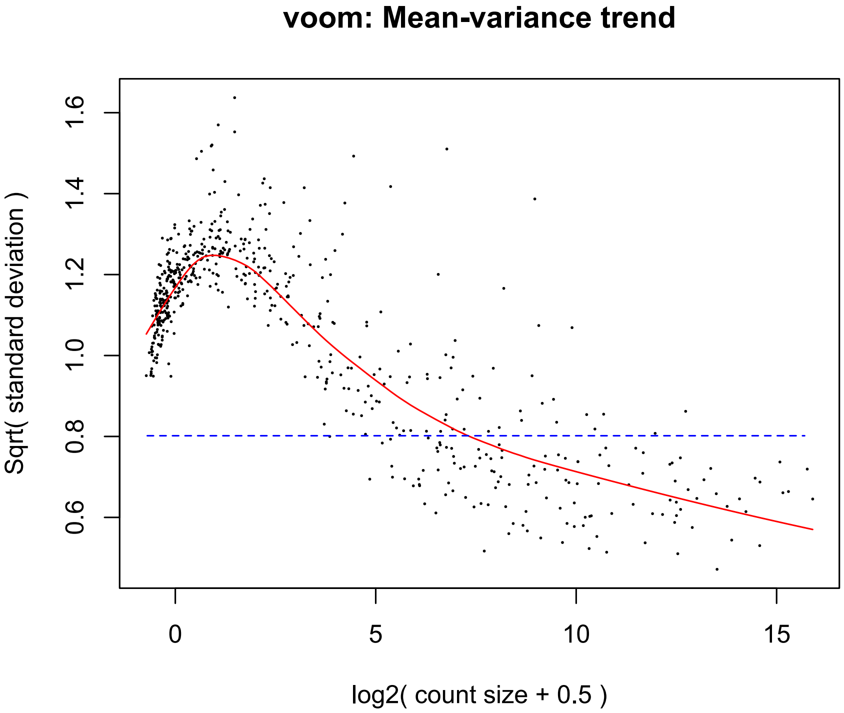


Figure 6.3: The square root of the residual standard deviation for each miRNA after fitting the linear model to each miRNA is plotted against the converted mean $\log_2$ count. Also shown as a red solid line is the estimated mean-variance trend using in the *voom* method. The dotted blue line is the 4'th root of the estimated prior by *limma*. As this is just a number, it is shown as a horizontal line.

of each miRNA against the mean log cpm value, but they show the same relationship between the mean and the variance of the expression values.

The results of the *voom* analysis when not including the maternal atopy and gender covariates in the model is given in Table A.1 in Appendix A. This smaller linear model rank some of the top 20 miRNAs in a different order, and include 6 different miRNAs in the top 20. However, *miR*-577 has a adjusted p-value below 0.1 when using both models. The three miRNAs; *miR*-577, *miR*-323a-3p and let-7e-3p are ranked highest by both models.

6.2.2 Clustering the top 20 MicroRNAs

The same clustering approach as the one performed on all 605 miRNAs is performed using only these 20 selected miRNAs. The dendrograms and heatmaps of hierarchical clustering using the correlation based and the absolute correlation based distances are shown in Figures 6.5 and 6.4. Since these heatmaps include considerably fewer miRNAs, the names of the miRNAs are also included in the plot. Using absolute correlation we do not see any apparent clustering of the samples that can be explained by the annotated characteristics. However, when performing clustering using the correlation based distance the samples seem to be clustered into two groups by supplement. The sample group on the left of the dendrogram in Figure 6.5 contains mainly samples where the mothers were given the probiotic supplement and the sample group on the right contains mainly samples where the mothers were given the placebo. Viewing both the heatmap and the dendrograms, it seems that there are also two miRNA groups. Most of

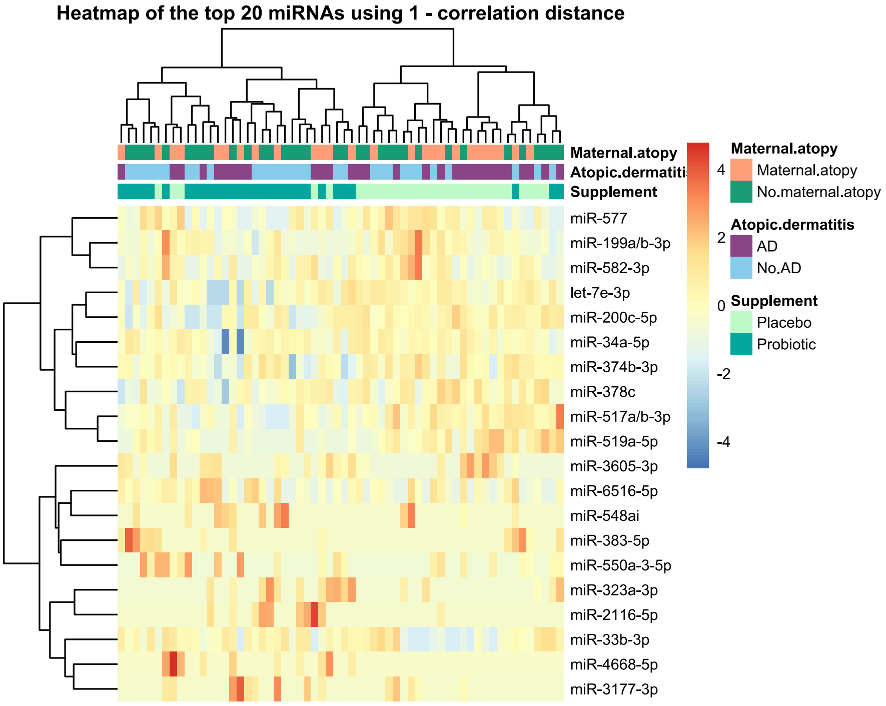


Figure 6.4: Heatmap with dendrograms of the hierarchical clustering of the log cpm values of the top 20 miRNAs from *voom* analysis. The samples are given by the columns and the miRNAs by the rows, and the data is scaled by row before clustering. Clustering is done using 1−correlation as dissimilarity measure and the Ward linkage function. Each sample is annotated by three variables; the supplement given to the mother, her history of dermatitis, and if the child developed dermatitis.

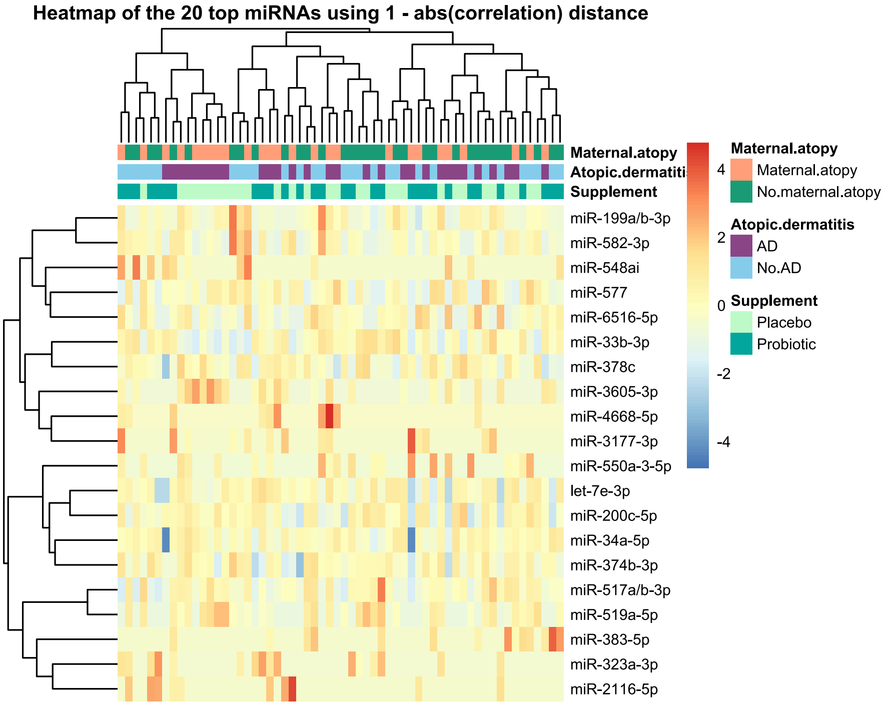


Figure 6.5: Heatmap with dendrograms of the hierarchical clustering of the log cpm values of the top 20 miRNAs from *voom* analysis. The samples are given by the columns and the miRNAs by the rows, and the data is scaled by row before clustering. Clustering is done using $1 - |\text{correlation}|$ as dissimilarity measure and the Ward linkage function. Each sample is annotated by three variables; the supplement given to the mother, her history of dermatitis, and if the child developed dermatitis.

the miRNAs in the upper group are down-regulated in the left sample group and up-regulated in the right sample group. For the miRNAs in the group at the bottom, some seem to be highly up-regulated for the left sample group and down-regulated in the right sample group.

The heatmap resulting from the clustering using the Euclidean distance is shown in Figure A.4 in Appendix A.

6.3 MicroRNAs Associated with Development of AD

To find miRNAs in the breast milk that are possibly associated with AD an elastic net model is used to perform variable selection. We first present the results of the two CV approaches for choosing α and λ in the elastic net model and the resulting models with the selected miRNAs. Then, we present the results of the bootstrapping, compare the two CV methods and present the estimated confidence intervals.

6.3.1 Nested Cross-Validation

The nested CV method with λ_{1SE} chooses $\alpha = 0.10$ and $\lambda = 1.61$ as the best model parameters. However, we note that the chosen α and λ vary noticeably for different seeds. The coefficient estimates of the resulting elastic net model are given in Table 6.2. For comparison, the nested CV method with λ_{min} chooses $\alpha = 0.15$ and $\lambda = 0.42$. The coefficient estimates of this model are given in Table A.3 in Appendix A.

	Repeated CV	Nested CV
Intercept	-3.760	-0.012
miR-1266-5p	0.126	0
miR-140-3p	-0.130	-0.008
miR-223-5p	-0.010	0
miR-33a-5p	-0.005	0
miR-342-3p	0.665	0.018
miR-3605-3p	0.288	0.011
miR-411-5p	0.006	0.007
miR-500a/b-5p	0.195	0
miR-524-5p	0.148	0.002
miR-570-5p	-0.148	0
miR-625-3p	-0.506	-0.034
miR-6515-5p	-0.727	-0.015
miR-671-5p	-0.050	0
matatopy	0.219	0.039
no Matatopy	-0.234	-0.039

Table 6.2: Table presenting estimated model coefficients when using repeated CV and nested CV for determining the model parameters α and λ of the elastic net model. The nested CV was performed using the one standard error rule. Covariates with estimated coefficients equal to zero in both models are not included in the table.

A unit increase in the log cpm value of any one of the miRNAs changes the odds of developing AD by a multiplicative factor of $\exp(\hat{\beta})$, where $\hat{\beta}$ is the estimated coefficient of that miRNA. We notice that that all coefficient estimates, when using the model parameters chosen by the nested CV with λ_{1SE}, are below 0.039 in absolute value and all of the miRNAs have coefficients estimates equal to or below 0.034 in absolute value. MiRNA miR-625-3p is the miRNA with the largest estimated coefficient in absolute value. This means that a unit increase in the log cpm value of any one of these covariates increases the odds by a multiplicative factor no greater than $\exp(0.018)$, which is approximately equal to 1.018, and decrease the odds by a multiplicative factor no less than $\exp(-0.034)$, which is approximately equal to 0.967. In other words, none of the miRNAs in this model are estimated to alone change the odds by more than 4% in any direction. We also notice that the two covariates for maternal atopy are included in the model and their estimated coefficient values is very similar, indicating that the model treats these two coefficients, which are 100% negatively correlated, almost equally. This is to be expected when using an elastic net model with $\alpha = 0.1$, meaning it is close to ridge regression. Ridge will favor treating them equally rather than letting one be zero and only including the other. Generally speaking will $\beta_1 - \beta_2 = C$ give the same results for all combinations of β_1 and β_2 but the penalty term in ridge will be minimized for $\beta_1 = -\beta_2$. Additionally, we notice that both gender covariates are set to zero.

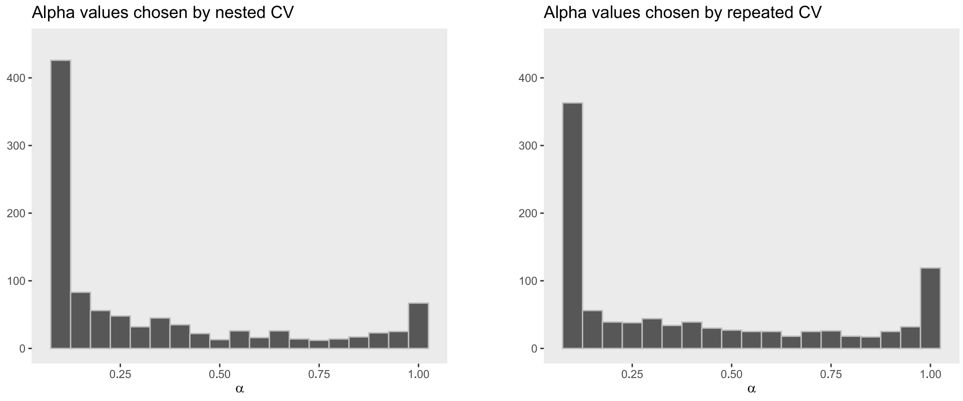
The nested CV method with λ_{1SE} and 10 inner and outer folds and the specified α sequence takes on average 37.28 seconds.

6.3.2 Repeated Cross-Validation

The repeated CV method, using 10 folds and 10 repetitions, chooses the model parameters $\alpha = 0.95$ and $\lambda = 0.10$. For this method as well, the model parameters vary a lot for different seeds. The coefficient estimates of the model using these parameters are given in Table 6.2. Only coefficients of variables included in the chosen model by the nested CV or by the repeated

CV are included in the table. We notice that the coefficient estimates of this model are generally larger in absolute value than in the model selected by nested CV. This is as expected as the λ parameter has a smaller value in this model. The miRNA *miR*-6515-5p has the largest coefficient value of -0.727, meaning a unit increase in the log cpm value of this miRNA is estimated to decrease the odds of developing AD by a multiplicative factor of 0.483. In other words, a unit increase in log cpm value of this miRNA is estimated to decrease the odds by 51.7%.

The repeated CV method using 10 folds, 10 repetitions, and the specified α and λ sequence takes on average 56.50 seconds. This is 19.22 seconds longer than the nested CV algorithm. One reason for the repeated CV algorithm being slower is that we specified the λ sequence to contain 200 values. In the nested CV we utilize the *cv.glmnet()* function in the *glmnet* package in R and this function chooses its own sequence of only 100 values of λ and adapts this sequence separately for each fold. Adapting the sequences leads to better convergence. This is, however, not possible for the repeated CV as we have to use the same λ sequence in all repetitions to be able to compute the average CV error.



(a)

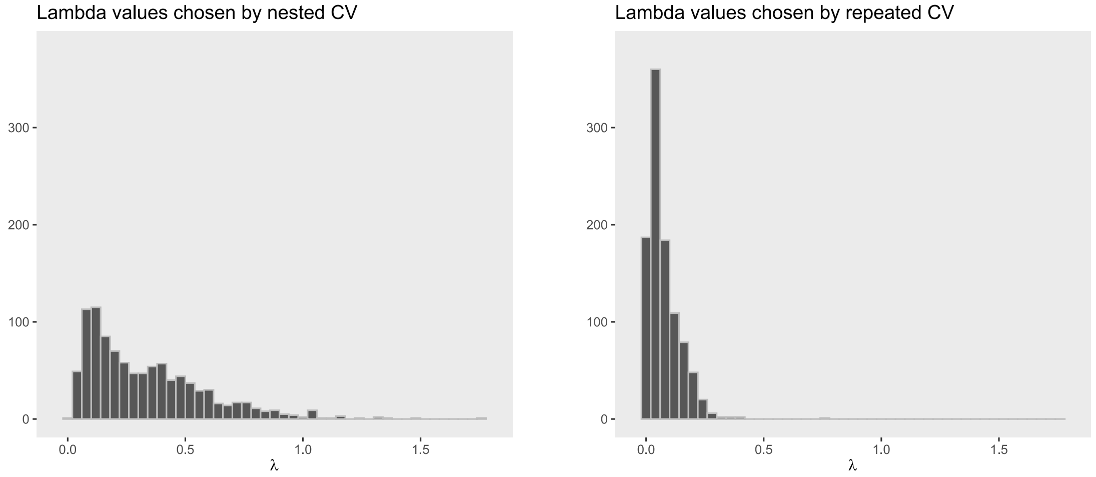
(b)

Figure 6.6: Histogram of the selected α values of the elastic net model for 1000 bootstraped datasets when using the nested CV in (a) and the repeated CV in (b).

6.3.3 Inference Using the Bootstrap

We perform paired bootstrapping and sample pseudo-datasets to perform inference of the coefficients. This is done for both the nested CV and repeated CV. For each method, this creates 1000 model parameter pairs and 1000 bootstrap estimates of each coefficient. Histograms of the chosen α values by the nested CV and the repeated CV, are shown in Figure 6.6 (a) and (b), respectively. Both CV methods choose $\alpha = 0.1$ most frequently, the nested CV more often than the repeated CV. The repeated CV chooses $\alpha = 1$, in other words a lasso model, more often than the nested CV. Histograms of the chosen λ values are shown in Figure 6.7. The range of chosen λ values is larger for the nested CV than for the repeated CV. For the repeated CV all λ values lie between 0 and 0.4, however, for the nested CV the λ values are more evenly spread out between 0 and 1 with even a few values above 1.5.

As we have not seen a nested CV used in this way to select model parameters of an elastic net model in any other papers, we only consider the elastic net model using the repeated CV for parameter selection in the rest of the analysis. That is, we end the comparison of the two CV procedures here. All of the following results from the bootstrapping were obtained using the repeated CV. However, the results of the bootstrapping using nested CV with λ_{1SE} are given in Appendix A.



(a)

(b)

Figure 6.7: Histogram of the selected λ values of the elastic net model for 1000 bootstraped datasets when using the nested CV in (a) and the repeated CV in (b).

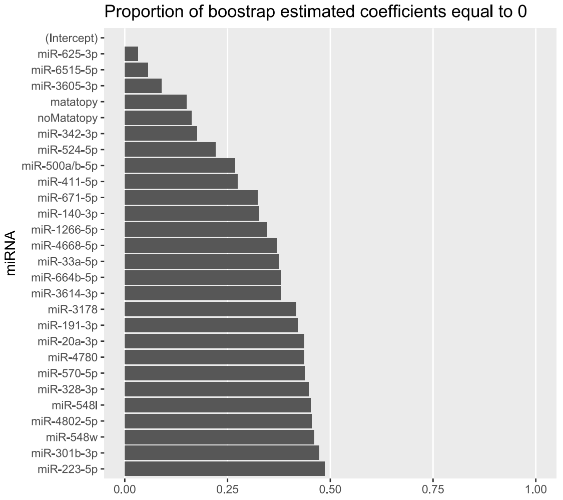


Figure 6.8: Barplot showing the proportion of times each coefficient was set to zero. The plot only includes coefficients not set to zero in at least 50% of the bootstrap models.

The barplot in Figure 6.8 shows the proportion of times each coefficient is zero in the bootstrap distribution. In this, and the next plots we only include variables selected (their coefficient is not set to zero) in at least 50% of the bootstrap models. Of the 605 miRNAs, only 25 are selected in at least 50% of the bootstrap samples. The two maternal atopy coefficients are also selected in at least 50% of the bootstrap estimated models. The intercept is always included as this is not penalized.

A boxplot of 1000 bootstrap realizations of these coefficients obtained by the paired bootstrap is shown in Figure 6.9. The intercept is not included in in this plot as its estimates are much greater than the rest of the coefficients, and by including it we would not be able to properly view the properties of the other coefficients.

Using the bootstrap estimates, we calculated the BC_a confidence interval of the coefficients

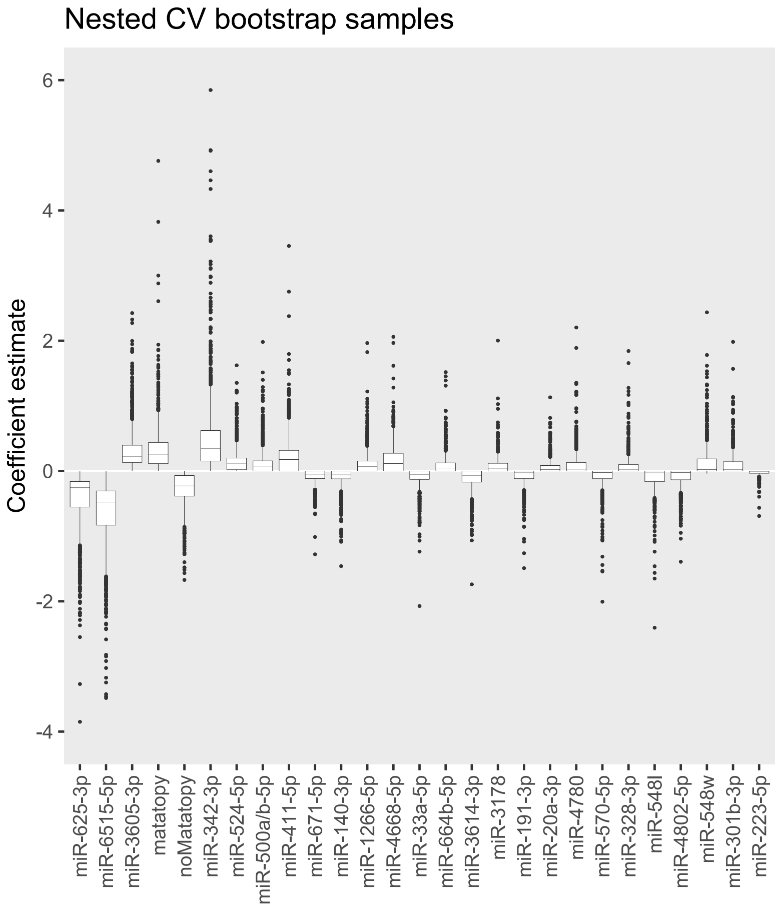


Figure 6.9: Boxplot of the estimated coefficient values of 1000 bootstrap datasets. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The whiskers extends from the hinge to the largest/smallest value at most 1.5 times the inter-quartile range from the hinge. The values beyond the end of these whiskers are plotted as individual points.

included in more than 50% of the bootstrap models. The resulting confidence intervals are shown in Figure 6.10.

The estimated confidence intervals, bias correction factor $\hat{z}_0$, acceleration factor $\hat{a}$ and the estimated percentiles for the selected variables and the intercept are given in Table 6.3. The estimated confidence intervals of the five miRNAs miR-342-3p, miR-3605-3p, miR-500a/b-5p, miR-625-3p and miR-6515-5p do not contain zero. The miRNA miR-3605-3p was also among the top 20 miRNAs ranked by the *voom* analysis. We also notice that many of the estimated confidence intervals have zero as their upper or lower limits. Only one of the coefficients included in at least 50% of the bootstrap samples have an estimated confidence interval which includes zero, but not as it upper or lower limit.

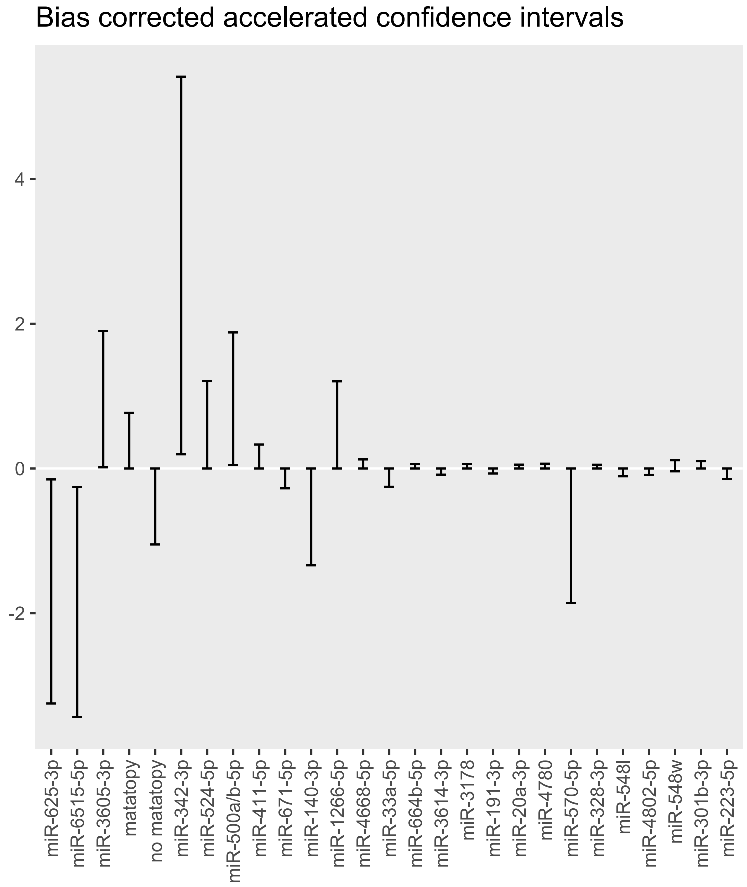


Figure 6.10: The estimated 95% confidence intervals of coefficients not set to zero in at least 50% of the bootstrap models. The confidence intervals are estimated using the modified bias-corrected accelerated method.

If we return to the heatmap in Figure 6.1, we can examine how these miRNAs cluster. We also want to investigate if the one differentially expressed miRNA from the analysis of the probiotics, miR-577, is clustered together with any of the miRNAs form the AD analysis. We let the clustering split the miRNAs into 11 groups, then miR-3605-3p, miR-6515-5p and miR-577 are grouped together, miR-342-3p and miR-625-3p are grouped together, and miR-500a/b-5p is an separate group. Even if we chose to split into only 4 groups, these miRNAs are split in the same way. None of these miRNAs are grouped into the cluster that we mentioned stands out in the heatmap.

	$\hat{\beta}$	Lower quantile	Upper quantile	z_0	a	ν_1	ν_2
Intercept	-3.760	-25.727	27.235	0.235	0.019	0.076	0.994
miR-625-3p	-0.506	-3.248	-0.151	-0.613	0.016	0.001	0.778
miR-6515-5p	-0.727	-3.435	-0.255	-0.527	0.015	0.002	0.825
miR-3605-3p	0.289	0.017	1.900	0.356	-0.014	0.099	0.995
matatopy	0.219	0.000	0.769	-0.176	-0.066	0.004	0.922
no matatopy	-0.234	-1.050	0.000	-0.030	-0.027	0.017	0.965
miR-342-3p	0.665	0.197	5.415	0.729	-0.016	0.299	1.000
miR-524-5p	0.148	0.000	1.207	0.361	-0.013	0.102	0.996
miR-500a/b-5p	0.195	0.049	1.881	0.852	-0.019	0.390	1.000
miR-411-5p	0.006	0.000	0.331	-0.589	-0.019	0.001	0.773
miR-671-5p	-0.051	-0.273	0.000	0.108	0.017	0.046	0.988
miR-140-3p	-0.130	-1.338	0.000	-0.745	0.006	0.000	0.684
miR-1266-5p	0.126	0.000	1.205	0.473	-0.009	0.151	0.998
miR-4668-5p	0	0.000	0.126	-0.896	-0.124	0.000	0.517
miR-33a-5p	-0.006	-0.253	0.000	0.300	0.036	0.103	0.997
miR-664b-5p	0	0.000	0.061	-0.880	-0.081	0.000	0.545
miR-3614-3p	0	-0.085	0.000	0.878	0.091	0.457	1.000
miR-3178	0	0.000	0.062	-0.812	-0.085	0.000	0.593
miR-191-3p	0	-0.069	0.000	0.805	0.046	0.385	1.000
miR-20a-3p	0	0.000	0.053	-0.779	-0.049	0.000	0.632
miR-4780	0	0.000	0.066	-0.779	-0.152	0.000	0.588
miR-570-5p	-0.148	-1.858	0.000	-0.863	0.038	0.000	0.611
miR-328-3p	0	0.000	0.052	-0.759	-0.118	0.000	0.615
miR-548l	0	-0.107	0.000	0.752	0.074	0.360	1.000
miR-4802-5p	0	-0.088	0.000	0.747	0.090	0.364	1.000
miR-548w	0	-0.039	0.115	-0.734	-0.109	0.000	0.636
miR-301b-3p	0	0.000	0.102	-0.718	-0.075	0.000	0.662
miR-223-5p	-0.010	-0.144	0.000	-0.151	0.052	0.020	0.968

Table 6.3: Results of the accelerated bias-corrected confidence interval computation. The values in the column $\hat{\beta}$ are the estimated coefficients of the elastic net model using $\alpha = 0.95$ and $\lambda = 0.10$. The factor z_0 is the bias correction factor, a is the acceleration factor, ν_1 and ν_2 are the corrected percentiles and the lower and upper quantile are the estimated confidence intervals. Only confidence intervals of coefficients not set to zero in at least 50% of the bootstrap models included. The coefficients are ordered by the number of times they where included in a bootstrap model in a decreasing manner.

Chapter 7

Discussion and conclusion

In this thesis, we have analyzed miRNAs as possible contributors to the risk reduction of AD associated with perinatal probiotic ingestion. We chose to split the analysis into two separate parts; one where we examined the probiotics effect on miRNAs and one where we searched for miRNAs associated with AD development. As we believe that miRNAs cooperate, we also performed a hierarchical clustering analysis to look for structures and groupings in the data.

We first discuss the analysis examining the effect of the probiotics and how the non-random selection process might have affected this analysis. Secondly, we discuss the analysis of miRNAs associated with AD and the validity of the confidence intervals. For both analyses, we discuss the methods we used, how they can be improved upon, which miRNAs we found and how these miRNAs were clustered.

7.1 Discussion of the Analysis of Probiotics Effect on MicroRNAs

In the first analysis, we chose a threshold for the adjusted p-values of 10%, as they also did in the previous study analyzing the breast milk samples collected 3 months postpartum (Simpson et al. 2015). Since this is more of an exploratory analysis, we have a higher tolerance for false positives than the commonly used threshold of 5%. The probiotic supplement was significantly associated with differential expression of only one miRNA, *miR*-577. The estimate of the $\log_2$ fold change corresponding to the effect of this miRNA was -2.067. In other words, the fold change of the probiotic supplement with respect to the placebo is 0.239. As the fold change is below 1, it means that for *miR*-577 the expression value is estimated to decrease when taking the probiotic supplement. As we have included other covariates as well, this fold change is the estimated change in expression value when taking a probiotic supplement compared to a placebo when all other covariates are constant. In addition to the probiotic covariate, we include AD, maternal atopy and gender as covariates. Thus, the fold change is the estimated change in expression value between mothers given a probiotic supplement and mothers given a placebo when the mothers have the same history of atopy, the gender of the child is the same and the outcome of the AD is the same. As we do not include interactions between probiotics and the other covariates, we implicitly assume that probiotics have the same effect when conditioned on all combinations of gender, history of atopy and AD.

The miRNA *miR*-323*a*-3*p* also has a small adjusted p-value of 0.154 and we note that the raw p-values of all the top 20 miRNAs are equal to or below 0.018. Although these were not differentially expressed at an FDR threshold of 10% in this analysis, we cannot conclude that they are not affected by the probiotic supplement. This is also the case for all the other miRNAs considered in this thesis that had an adjusted p-value above 0.1. We cannot conclude that they are not affected by the probiotics due to the fact that we are not proving the null hypothesis by not rejecting it. We thus chose to provide a ranking of the most important miRNAs in hope that this will be useful for further analyses.

In the study by Simpson et al. (2015), also using *limma* but analyzing the samples collected at 3 months, they found no conclusive evidence of differential expression between the probiotic and placebo groups in any of the miRNAs they considered. They found that four miRNAs had a raw p-value below 0.05, but none had an acceptable FDR. None of the miRNAs they reported are included in the top 20 miRNAs from this analysis. These results might indicate that probiotics affect miRNAs in the early stages of breastfeeding rather than in the later stages, and we suggest that further analysis focuses on this period. However, we used a less strict filtering procedure than Simpson et al. (2015), meaning we included more miRNAs with lower expression values. We also notice that *miR*-577 have an average cpm expression value of 0.041 and this miRNA was likely not included in their analysis.

The plot in Figure 6.3 show the estimated mean-variance trend used to compute the precision weights. For comparison, the estimated prior, s_0^2, calculated by the original *limma* method is shown. The prior, s_0^2, is the prior that would have been used for estimating the posterior variance, if we had used the original *limma* for the analysis instead of *voom*. The greatest difference between *voom* and *limma* is that *limma* use all miRNAs in the dataset to estimate only one prior variance, equal for all miRNAs, while *voom* allows for estimation of variance at the individual observation level. As seen in Figure 4.4 in Chapter 4, the library sizes in this dataset vary between samples. The different library sizes might be due to differences between the mothers or differences in sequencing depth of the samples. As seen in Figure 6.3, the variance depends on the count value. Thus, observations from mothers with a large library size, will likely have a larger variance, as the counts from these mothers are likely larger for all miRNAs. This highlights another advantage of *voom* over *limma*; *voom* estimates the precision weights (inverse of the variances) on the raw read counts and not after normalization as done in *limma*. As explained in Section 3.4.2, by estimating the variance after normalizing to cpm values, information about which observations had a high count value before normalization is lost and the estimated variance might not be as accurate.

The heatmap in Figure 6.5 appears to group the samples into a placebo and a probiotic cluster. The miRNAs in the top cluster appear to be up-regulated in the probiotic group compared to the placebo group. Furthermore, the miRNAs in the bottom cluster seem to be down-regulated in the probiotic group compared to the placebo group. Note that since the miRNAs are scaled, we cannot compare their values within one sample, i.e., we cannot state that the miRNAs in the top cluster are up-regulated compared to the miRNAs in the bottom cluster. The miRNA *miR*-577, which was found to be significantly differentially expressed between the placebo and probiotic group, belongs to the top cluster. If miRNAs contribute to reducing the risk of developing AD associated with a probiotic supplement, we believe that not only one but multiple miRNAs must contribute. MiRNAs are small molecules, and even though one miRNA can affect multiple genes, we believe they cooperate. Therefore, if the miRNAs are truly affected by the probiotics, it is not likely that only *miR*-577 is affected. In this study, we have a small sample size and an unfortunate selection procedure of samples. By providing a ranking and clustering of the miRNAs, further analysis may only focus on these miRNAs and thus reduce the number of tests.

The clustering with the absolute correlation based distance showed no signs of clustering by the annotated variables. As clustering is an unsupervised method, it might not cluster into the groups that are of interest. Also, we have no way of evaluating how well the clustering performs. However, as we have previously mentioned in Section 5.1, the absolute correlation might be more affected by noise in the data and this might be the reason why the correlation based clustering can cluster by which supplement the mother received and the absolute correlation clustering is not.

Both *edgeR* and *DESeq2* are methods often used to analyze RNA-seq data and could have been used instead of *voom* in this analysis. In contrast to *voom*, they are based on the negative binomial distribution (Robinson, McCarthy et al. 2010; Love et al. 2014). *DESeq2* uses a normalization method called regularized log transformation, which is recommended before cluster analysis. This method is an alternative to the cpm normalization used in *limma* and *edgeR*. The regularized log transformation stabilizes the variances such that they become approximately homoscedastic. In the paper by Love et al. (2014), the authors state that by using this transformation, we avoid spreading data for miRNAs with low count values such that random noise does not dominate the meaningful signal. Using this transformation instead of scaling the miRNAs to have equal variance might improve the cluster analysis. The reasoning of why scaling (or using some transformation to stabilize the variance) is necessary is explained in Figure A.2 in Appendix A.

In the paper by Soneson and Delorenzi (2013), they evaluated eleven methods for differential expression analysis of RNA-seq data using both simulated and real RNA-seq data. The authors concluded that methods combining a variance-stabilizing transformation with the *limma* method performed well under many different conditions and were relatively unaffected by outliers. Furthermore, they conclude that *DESeq* and *EdgeR* showed, overall similar accuracy with respect to gene ranking but that *DESeq* was often overly conservative, while *EdgeR* often was too liberal. However, which method is best is highly dependent on the dataset.

7.2 The Non-Random Selection Process of Samples in the Dataset

The non-random selection process of the dataset may complicate this first analysis. This might seem intuitive to some readers, but the reason is complex and difficult to describe in exact terms. We will, however, in a structured manner, try to relay the arguments why. Consider two cases; one where probiotics in fact impact miRNAs that affect the probability of the child developing AD, and a second where probiotics do not affect any miRNAs or only affect miRNAs not associated with AD.

In the first case, the selection process introduces some bias and makes it harder to detect differentially expressed miRNAs between mothers receiving probiotics and mothers receiving placebo. To illustrate this, consider a simplified example where a single gene's expression level solely determines the probability of developing AD. Let a high gene expression be equivalent to a high probability and let this gene be affected by a single miRNA in the mother's breast milk. A high miRNA value will down-regulate the gene expression. Let there be two groups; one in which the mothers ingest probiotics and one in which the mothers ingest a placebo. Let the probiotics have a 50% chance of increasing the miRNA expression value such that the probability of developing AD becomes zero and a 50% chance of not impacting the miRNA expression value. Assume next that for each of the two groups, we select semi-randomly, 15 mother-child pairs where the child developed AD and 15 pairs where the child did not develop AD. For the 15 breast milk samples where the mothers were given the probiotics and the child developed AD, we know that the probiotics have not affected the miRNA expression. We know this since it is impossible in this example to develop AD if the probiotic has affected the miRNA expression. Thus, none of these 15 samples will contribute to proving that the probiotics actually do affect the miRNA expression. Among the 15 samples without AD from the probiotics group, we expect that the probiotics contributed to lowering the miRNA expression value in more than half of the cases. The probiotics is expected to contribute in more than half of the cases because the probiotics is expected to affect the miRNA in half of all samples and we have deliberately selected samples from those that did not develop AD. Thus, on average more than half of the samples (more than 15/2) will provide information that probiotics affect the miRNA value. However, the information might be drowned out by the more than 15 other samples where the probiotic had no effect. Additional noise may also be introduced by the fact that the fraction of AD in each of the groups is not kept the same.

In this example, we have observed that even though the specific selection procedure potentially allows for discovery of differential expression, it makes it more challenging by oversampling cases where the probiotic had no effect. Of course, this was a simplified example, but the same intuition applies to a lot of other situations as well, such as when the probiotic supplement only increases the miRNAs expression by a bit instead of the stochastic and drastic change in the example. The selection procedure may also introduce other sources of noise, but having made our point we do not dwell on it further and accept that the selection process introduces noise.

For the second case, where probiotics do not affect any miRNAs or only affect miRNAs not associated with AD, the selection process should not introduce any additional bias. Thus, if we knew this was the case, the selection procedure would not be a problem for the analysis. However, the issue is that we have no way of knowing which case we are dealing with. Thus, even if the second case is the truth and the analysis shows no differentially expressed genes, we have, in addition to the uncertainty of any statistical experiment, the potential noise introduced by the sampling procedure if case one were in fact true.

In this thesis, we try to adjust for the non-random selection procedure by adding AD as a covariate in the linear model used in the analysis of the probiotics effect. This is, however, not a perfect fix. It might compensate for some of the issues with the non-random selection, but it also introduces some new issues. If probiotics do in fact affect the miRNAs that again affect the probability of the child developing AD, then in the linear model the AD coefficient might capture some of the information belonging to the probiotics coefficient, thus making it less likely for the probiotics coefficient to be seen as statistically significant in the analysis. Potentially, it might have been better to perform the analysis without adjusting for AD or if we had known which of the samples were non-randomly chosen, to only adjust for those. This information is, however, unfortunately not available to us.

In the previous paragraphs we have discussed the downsides of the non-random selection method. However, due to RNA-seq being expensive, the selection criteria were deemed necessary to ensure that they included enough samples from all four groups; AD and probiotic, AD and placebo, no AD and probiotic, no AD and placebo.

7.3 Discussion of the Analysis of MicroRNAs Association with AD

In the second analysis, of miRNAs association with AD, we used elastic net to perform variable selection. For this analysis we could have used the same *limma* analysis, and evaluate if AD was associated with differential expression with any of the miRNAs. However, by using a multivariable analysis we could take into account the correlation and cooperation between miRNAs. In the elastic net model, 13 miRNAs and the two maternal atopy covariates were selected. Maternal atopy and gender were included in the model to reduce noise and adjust for possible confounding, so we are mainly interested in the selected miRNAs. To ensure that both levels of the gender and maternal atopy covariates were penalized, an extended model matrix was used. However, this still does not ensure that the two levels are shrunken equally or that they are both included or excluded at the same time. The group lasso, introduced by Yuan and Lin (2006), is often used for cases where we have covariates that should be included or excluded together. However, we found no equivalent to this method for elastic net. As the goal was to select all miRNAs associated with AD, also those highly correlated, lasso was not appropriate and neither was group lasso. Ma et al. (2007) propose a supervised group lasso approach that takes into account the grouping structure of gene expression. They first divide the genes into clusters using a clustering algorithm, and identify important genes withing each cluster using lasso. Then, important clusters are selected using group lasso. They showed, using microarray datasets, that the approach is capable of identifying a small number of influential gene clusters and important genes within those clusters. We find this to be an interesting approach which could have been used in this analysis instead of the elastic net. Another interesting possibility would be to use a modified version of this supervised group lasso approach, where we would use elastic net instead of lasso to identify important miRNAs in each cluster.

We consider two CV methods for determining the α and λ parameters; nested CV and repeated CV. Both methods could also be performed using LOOCV, but as this is computationally more expensive this was not done in this analysis. Ideally we would perform a simulation study, evaluating the two CV methods on how well they perform variable selection, then proceed with the CV method best suited for this purpose. This was however not prioretized in this thesis.

Bootstrapping is used to evaluate how consistently each miRNA is selected by an elastic net model. The bootstrapping of the variable selection procedure using repeated CV resulted in 25 miRNAs being selected in more than 50% of the bootstrap models. However, when using the nested CV, only 12 miRNAs were selected in more than 50% bootstrap samples. When comparing barplots in Figure 6.8 and A.6 we notice that the bootstrap probability of 0 is slightly higher for all variables when using the nested CV. It is to be expected that nested CV results in a model with less and maybe more penalized variables compared to the repeated CV since we are in the nested CV using the one standard error rule to determine the best λ. From the histogram of α values in Figure 6.6 we notice that the nested CV choose a smaller α more frequently than the repeated CV. As the α parameter controls the weighting between the ridge and lasso penalty, this means that the nested CV favors a model closer to ridge for this dataset. As ridge does not perform variable selection, we would expect that only a few coefficients would be set to zero but this is the opposite of what we observe. However, this can be explained by nested CV choosing a wider range of λ values, as seen in Figure 6.7. We note that the two CV methods only determine the impact and form of the penalty term which only affect how many variables are selected and how much their coefficient estimates are shrunken. This implies that the same miRNAs are selected by both methods, it is just that a smaller value of λ or α allow additional miRNAs to be selected as well.

We used a slightly modified version of the accelerated bias-corrected method to compute confidence intervals. This results in five miRNAs with an estimated 95% confidence interval not contain zero. The five miRNAs were miR-342-3p, miR-3605-3p, miR-500a/b-5p, miR-625-3p and miR-6515-5p. In the study by (Simpson et al. 2015), where they analyzed the breast milk samples collected 3 months postpartum using *limma*, AD was found to be associated with differential expression of 13 miRNAs when using a raw p-value threshold of 0.05. However, none of these miRNAs had an acceptable FDR. None of the miRNAs they found are the same as the ones found in this analysis.

As seen from the results of the two analyses we have performed, none of the miRNAs found to be associated with AD was differentially expressed in the probiotic analysis. However, miR-3605-3p was included in the top 20 miRNAs with a raw p-value of 0.018, and an estimated confidence interval of $[0.017, 1.900]$ in the second analysis. It is also interesting to note that mir-3605-3p was clustered together with miR-6515-5p and miR-577 in the clustering using the correlation based distance. The miRNA miR-577 was the one miRNA that was differentially expressed between the probiotic and placebo group in the *voom* analysis.

The findings in this thesis show a slight possibility of miRNAs being part of the explanation behind the risk reduction. Further work might improve this analysis by performing a mediation analysis. As stated by Agler and De Boeck (2017), mediation analysis is a tool for testing possible causal relationships. That way we could, in a more structured manner, examine if miRNAs actually are mediators in the process of reducing the risk of developing AD. We would also suggest to use the results of this analysis as a way of filtering out miRNAs which are not as relevant.

7.4 Reliability of the Confidence Intervals

Elastic net has shown to perform well for variable selection in multiple studies (Cho et al. 2010; Giglio and Brown 2018; Zou and Hastie 2005). However, there is a debate about how to properly perform inference and whether or not the confidence intervals obtained by bootstrapping are even valid. In the paper by Zou and Hastie (2005) that originally introduced elastic net, the authors used bootstrapping to estimate standard deviations. Further, to adjust for the bias introduced by shrinking the coefficients, we used the accelerated bias-corrected method instead of the commonly used percentile method. However, the bias correction method can only adjust for bias which is possible to detect trough the bootstrapping procedure. Thus, we can question if we really are estimating the confidence interval of the true coefficient or the shrunken coefficient.

All but one of the estimated confidence intervals shown in Table 6.3 have zero as either their upper or lower limit. This may be a consequence of the true coefficients actually being zero or perhaps it is a consequence of using elastic net regression. Since we only have 60 samples compared to 609 covariates (miRNAs and basis covariates) it can be difficult to determine which miRNAs are actually associated with AD. The inclusion of the lasso penalty term will set many coefficients equal to zero, maybe even if they are associated with AD if there are other covariates that for this dataset appear to be a better predictor. With such few samples it can be a lot of randomness with respect to which predictor appear the most related to the response. We have tried to adjust for some of the bias introduced by the elastic net penalty using the slightly modified version of the computation of the bias correction factor. However, we still clearly see effects of the elastic net in the confidence intervals.

The *penalized* R package can also be used for regularization and allow for L_1 and L_2 penalties, and combinations of the two (elastic net). The main difference between this and the *glmnet* package is that they use another parameterization of the penalization term than the *glmnet* package, otherwise they both perform penalized regression using elastic net. In the vignette for the *penalized* package by J. Goeman et al. (2022) they mention that standard errors and confidence intervals can easily be calculated using the bootstrap but argue that they are not very meaningful for strongly biased estimators which arises in penalized regression models. Part of the reason is that the bias is such a major component of the mean squared error and the variance only contributes to a small part, and when using penalized regression it is hard to obtain a good estimate of the bias. We reason that this probably also applies to logistic regression as well. In the paper by Casella, Ghosh et al. (2010), the authors state that the bootstrap estimates of the standard error of the lasso estimator might be unstable and not perform well, and that they in fact are not consistent if the true $\beta = 0$. This also raises questions about the bootstrap estimated confidence intervals when using elastic net, and if they also are not valid. In the same paper by Casella, Ghosh et al. (2010), they develop a Bayesian version of the elastic net model using a Gibbs sampler which produce valid confidence intervals. This could have been an alternative to the frequentist elastic net used in this analysis, as it is possible that we would have been able to compute more accurate confidence intervals. However, the Bayesian elastic net was developed and tested only for prediction accuracy and they did not evaluate the model on how well it performs variable selection.

We have seen few other papers using the accelerated bias-corrected method to account for bias when estimating the confidence intervals, and as we do not know the true distribution of the coefficients it hard to evaluate them. Efron and Narasimhan (2020) have implemented a R package for computing the accelerated bias-corrected confidence intervals. In the paper introducing this package they use the BC_a method to create confidence intervals for coefficients from a *glmnet* model (they do not use any modification of the bias correction factor as we have done). They do, however, use a parametric bootstrap, unlike what we have done here where we use a nonparametric bootstrap. Still this serves a slight indication that they believe the bias corrected estimates can be used for penalized regression models.

Lee et al. (2016) explain how to perform exact post-selection inference by conditioning on the on the selection event, but the procedure for the combination of elastic net and logistic regression has not yet been implemented. If this is implemented in the future, it would be interesting to compare the confidence intervals produced by this procedure to the ones obtained in this thesis.

7.5 Conclusion

A perinatal probiotic supplement ingested by the mother before birth and during breastfeeding has been shown to reduce the risk of the child developing AD after 2 years (Dotterud et al. 2010). In this thesis we have analyzed miRNAs as possible contributors to this risk reduction. To investigate this we performed two analyses; one examining the effect of the perinatal probiotic supplement on the miRNAs expression values and one examining if any miRNAs in the breast milk are associated with AD in the children. We performed a univariate analysis of the probiotics affect on each of the 605 miRNAs using the *voom* method in *limma*. We found evidence that the probiotic supplement was associated with differential expression of one miRNA, *miR*-577, using a FDR threshold of 10%. We provided a ranking of the top 20 miRNAs according to p-value and a total of 47 miRNAs had a raw p-value below 0.05. In this study we have considered a high dimensional dataset with few samples and an unfortunate selection procedure. Both of these factors can possibly have reduced the power to detect more differentially expressed miRNAs. We provide ranking of the most important miRNAs which can be useful for further analysis by greatly reducing the number of tests.

In the second analysis, examining if any miRNAs in the breast milk are associated with AD, we performed a multivariable analysis. This was done to take into account the correlation and cooperation between miRNAs. Elastic net was used to perform variable selection as it has shown to perform well when variables are highly correlated. To perform inference of the selected miRNAs we used bootstrapping and a modified bias-corrected accelerated method to estimate confidence intervals of the coefficients. The estimated 95% confidence intervals of five miRNAs did not include zero. These five miRNAs were *miR*-342-3p, *miR*-3605-3p, *miR*-500a/b-5p, *miR*-625-3p and *miR*-6515-5p. The miRNA *miR*-3605-3p was also included in the top 20 miRNAs from the first analysis with a raw p-value of 0.018, but an adjusted p-value of 0.516. From the hierarchical clustering we also saw that this miRNA was clustered together with *miR*-6515-5p and *miR*-577 when using a correlation based distance. Thus, future analyses might consider focusing on these miRNAs.

In the study by Simpson et al. (2015), they also analyzed samples from the proPACT study to examine miRNAs as a contributor to the risk reduction. However, they analyzed samples collected at 3 months. They did not observe conclusive evidence of differential expression of any miRNAs between the probiotic and placebo groups, nor between the AD groups. The results of the analysis in this thesis might give an indication that probiotics affect miRNAs in the early stages of breast feeding rather than in the later stages. For further analysis we advocate the use of a multivariable analysis also of miRNAs affected by probiotics, as we believe the miRNAs are likely to cooperate if they are in fact contributors to the risk reduction.

Bibliography

Agler, Robert and Paul De Boeck (2017). 'On the Interpretation and Use of Mediation: Multiple Perspectives on Mediation Analysis'. In: *Frontiers in Psychology* 8. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.01984. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2017.01984.

Asher, M Innes, Stephen Montefort, Bengt Bjørkstén, Christopher K W Lai, David P Strachan, Stephan K Weiland, Hywel Williams and ISAAC Phase Three Study Group (2006). 'Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys'. In: *The Lancet* 368.9537, pp. 733–743.

Benjamini, Yoav and Yosef Hochberg (1995). 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing'. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1. ISSN: 00359246. URL: http://www.jstor.org/stable/2346101.

Benner, Axel, Manuela Zucknick, Thomas Hielscher, Carina Ittrich and Ulrich Mansmann (2010). 'High-dimensional Cox models: the choice of penalty as part of the model building process'. In: *Biometrical Journal* 52.1, pp. 50–69.

Brownless, J (2020). *Nested cross-validation for machine learning with python*. Online; accessed 2022-15-03. URL: https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/?fbclid=IwAR1Nl78lpB6CQ646h1hq9zJqRl3ntl42PJ5EIR_GGTJkiGJT352MSdzuX7M.

Bumgarner, Roger (2013). 'Overview of DNA microarrays: types, applications, and their future'. In: *Current protocols in molecular biology* 101.1, pp. 22–1.

Casella, George and Roger L Berger (2002). *Statistical inference*. Cengage Learning.

Casella, George, Malay Ghosh, Jeff Gill and Minjung Kyung (2010). 'Penalized regression, standard errors, and Bayesian lassos'. In: *Bayesian Analysis* 5.2. DOI: 10.1214/10-BA607. URL: https://doi.org/10.1214/10-BA607.

Chen, Yunshun, Aaron TL Lun and Gordon K Smyth (2016). 'From reads to genes to pathways: differential expression analysis of RNA-seq experiments using Rsubread and the edgeR quasi-likelihood pipeline'. In: *F1000Research* 5.

Cho, Seoae, Kyunga Kim, Young Jin Kim, Jong-Keuk Lee, Yoon Shin Cho, Jong-Young Lee, Bok-Ghee Han, Heebal Kim, Jurg Ott and Taesung Park (2010). 'Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis'. In: *Annals of human genetics* 74.5, pp. 416–428.

Datta, Somnath and Dan Nettleton (2014). *Statistical analysis of next generation sequencing data*. Springer.

DiCiccio, Thomas J and Bradley Efron (1996). 'Bootstrap confidence intervals'. In: *Statistical science* 11.3, pp. 189–228.

Dotterud, C K, O Storrø, R Johnsen and T Øien (2010). 'Probiotics in pregnant women to prevent allergic disease: a randomized, double-blind trial'. In: *British Journal of Dermatology* 163.3, pp. 616–623.

Efron, Bradley and Carl Morris (1977). 'Stein's paradox in statistics'. In: *Scientific American* 236.5, pp. 119–127.

Efron, Bradley and Balasubramanian Narasimhan (2020). 'The automatic construction of bootstrap confidence intervals'. In: *Journal of Computational and Graphical Statistics* 29.3, pp. 608–619.

Fahrmeir, Ludwig, Thomas Kneib, Stefan Lang and Brian Marx (2007). *Regression*. Springer.

Freyhult, Eva, Mattias Landfors, Jenny Önskog, Torgeir R Hvidsten and Patrik Rydén (2010). 'Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering'. In: *BMC bioinformatics* 11.1, pp. 1–14.

Friedman, Jerome, Trevor Hastie and Robert Tibshirani (2010). 'Regularization Paths for Generalized Linear Models via Coordinate Descent'. In: *Journal of Statistical Software* 33.1, pp. 1–22. DOI: 10.18637/jss.v033.i01. URL: https://www.jstatsoft.org/v33/i01/.

Giglio, Cannon and Steven D Brown (2018). 'Using elastic net regression to perform spectrally relevant variable selection'. In: *Journal of Chemometrics* 32.8.

Givens, Geof H and Jennifer A Hoeting (2012). *Computational statistics*. Vol. 703. John Wiley & Sons.

Goeman, Jelle, Rosa Meijer and Nimisha Chaturvedi (2022). *L1 and L2 penalized regression models*. URL: https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf.

Goeman, Jelle J and Aldo Solari (2014). 'Multiple hypothesis testing in genomics'. In: *Statistics in medicine* 33.11, pp. 1946–1978.

Halle, Kari K., Øyvind Bakke and Mette Langaas (2017). *Short note on multiple hypothesis testing TMA4267 Linear Statistical Models (V2017)*. URL: https://www.math.ntnu.no/emner/TMA4267/2017v/multtest.pdf.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Hastie, Trevor, Robert Tibshirani and Martin Wainwright (2016). *Statistical learning with sparsity*. Vol. 143. CRC Press.

He, Jianghua, Prabhakar Chalise and Yi Zhong (2020). 'Nested and repeated cross validation for classification model with high-dimensional data'. In: *Revista Colombiana de Estadística* 43.1, pp. 103–125.

Hochberg, Yosef (1988). 'A Sharper Bonferroni Procedure for Multiple Tests of Significance'. In: *Biometrika* 75.4, pp. 800–802. ISSN: 00063444. URL: http://www.jstor.org/stable/2336325 (visited on 22/05/2022).

Holm, Sture (1979). 'A Simple Sequentially Rejective Multiple Test Procedure'. In: *Scandinavian Journal of Statistics* 6.2, pp. 65–70. ISSN: 03036898, 14679469. URL: http://www.jstor.org/stable/4615733 (visited on 22/05/2022).

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Jiang, Yifan, Yao Jiang, Sheng Wang, Qin Zhang and Xiangdong Ding (2019). 'Optimal sequencing depth design for whole genome re-sequencing in pigs'. In: *BMC bioinformatics* 20.1.

Law, Charity W, Yunshun Chen, Wei Shi and Gordon K Smyth (2014). 'voom: Precision weights unlock linear model analysis tools for RNA-seq read counts'. In: *Genome biology* 15.2.

Lee, Jason D, Dennis L Sun, Yuekai Sun and Jonathan E Taylor (2016). 'Exact post-selection inference, with application to the lasso'. In: *The Annals of Statistics* 44.3, pp. 907–927.

Love, Michael I, Wolfgang Huber and Simon Anders (2014). 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2'. In: *Genome biology* 15.12, pp. 1–21. URL: https://doi.org/10.1186/s13059-014-0550-8.

Lyngra, Rose (2015). *Atopisk eksem*. Online; accessed 2021-02-24. URL: https://www.lhl.no/lhl-astma-og-allergi/eksem/atopisk-eksem/.

Ma, Shuangge, Xiao Song and Jian Huang (2007). 'Supervised group Lasso with applications to microarray data analysis'. In: *BMC bioinformatics* 8.

Munblit, Daniel, Robert J. Boyle and John O. Warner (2015). 'Factors affecting breast milk composition and potential consequences for development of the allergic phenotype'. In: *Clinical & Experimental Allergy* 45.3, pp. 583–601.

Nocedal, Jorge and Stephen J. Wright (2006). *Numerical optimization*. Springer.

Nuzzi, Giulia, Maria Cicco and Diego Peroni (Apr. 2021). 'Breastfeeding and Allergic Diseases: What's New?' In: *Children* 8. DOI: 10.3390/children8050330. URL: https://www.researchgate.net/publication/351220956_Breastfeeding_and_Allergic_Diseases_What's_New.

Omdal, Lene Tillerli (2021). *Analysing the Effect of a Probiotic Supplementation on miRNA Expression in Breastmilk*. Project thesis in the course TMA4500 at NTNU. Available upon request from the author.

Pavlou, Menelaos, Gareth Ambler, Shaun Seaman, Maria De Iorio and Rumana Z Omar (2016). 'Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events'. In: *Statistics in medicine* 35.7, pp. 1159–1177.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Ravn, Nina H, Anne-Sofie Halling, Aviva G Berkowitz, Maria R Rinnov, Jonathan I Silverberg, Alexander Egeberg and Jacob P Thyssen (2020). 'How does parental history of atopic disease predict the risk of atopic dermatitis in a child? A systematic review and meta-analysis'. In: *Journal of Allergy and Clinical Immunology* 145.4.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi and Gordon K Smyth (2015). 'limma powers differential expression analyses for RNA-sequencing and microarray studies'. In: *Nucleic Acids Research* 43.7, e47. DOI: `10.1093/nar/gkv007`.

Robinson, Mark D, Davis J McCarthy and Gordon K Smyth (2010). 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data'. In: *Bioinformatics* 26.1, pp. 139–140.

Robinson, Mark D and Alicia Oshlack (2010). 'A scaling normalization method for differential expression analysis of RNA-seq data'. In: *Genome biology* 11.3.

Rousseeuw, Peter J (1987). 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis'. In: *Journal of computational and applied mathematics* 20, pp. 53–65.

Ruiz, RGG, DM Kemeny and JF Price (1992). 'Higher risk of infantile atopic dermatitis from maternal atopy than from paternal atopy'. In: *Clinical & Experimental Allergy* 22.8.

Shao, Jun and Dongsheng Tu (2012). *The jackknife and bootstrap*. Springer Science and Business Media.

Simpson, Melanie Rae, Gaute Brede, Jostein Johansen, Roar Johnsen, Ola Storrø, Pål Sætrom and Torbjørn Øien (Dec. 2015). 'Human Breast Milk miRNA, Maternal Probiotic Supplementation and Atopic Dermatitis in Offspring'. In: *PLOS One* 10. DOI: `10.1371/journal.pone.0143496`. URL: `https://doi.org/10.1371/journal.pone.0143496`.

Sirevåg, Reidun (2021). *probiotika i Store medisinske leksikon på snl.no*. Online; accessed 2021-09-11. URL: `https://sml.snl.no/probiotika`.

Smyth, Gordon K (2004). 'Linear models and empirical Bayes methods for assessing differential expression in microarray experiments'. In: *Statistical applications in genetics and molecular biology* 3.1.

Soneson, Charlotte and Mauro Delorenzi (2013). 'A comparison of methods for differential expression analysis of RNA-seq data'. In: *BMC bioinformatics* 14.1.

Strand, Andreas (2019). *Module 7: Moving Beyond Linearity*. From module pages in TMA4268 Statistical learning for the spring semester of 2019 at NTNU. URL: `https://www.math.ntnu.no/emner/TMA4268/2019v/7BeyondLinear/7.html#loess`.

Thorndike, Robert L. (1953). 'Who belongs in the family?' In: *Psychometrika* 18, pp. 267–276.

Tibshirani, Robert (1996). 'Regression shrinkage and selection via the lasso'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tibshirani, Robert, Guenther Walther and Hastie Trevor (2001). 'Estimating the number of clusters in a data set via the gap statistic'. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 63, pp. 411–423.

Tibshirani, Ryan J (2013). 'The lasso problem and uniqueness'. In: *Electronic Journal of statistics* 7, pp. 1456–1490.

Wang, Lipo, Feng Chu and Wei Xie (2007). 'Accurate cancer classification using expressions of very few genes'. In: *IEEE/ACM Transactions on computational biology and bioinformatics* 4.1, pp. 40–53.

Ward, Joe H (1963). 'Hierarchical Grouping to Optimize an Objective Function'. In: *Journal of the American Statistical Association* 58.301, pp. 236–244.

Xi, Y, X Jiang, R Li, M Chen, W Song and X Li (2016). 'The levels of human milk microRNAs and their association with maternal weight characteristics'. In: *European journal of clinical nutrition* 70.4, pp. 445–449.

Yuan, Ming and Yi Lin (2006). 'Model selection and estimation in regression with grouped variables'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1, pp. 49–67.

Zhao, Sen, Daniela Witten and Ali Shojaie (2021). 'In defense of the indefensible: A very naive approach to high-dimensional inference'. In: *Statistical Science* 36.4, pp. 562–577.

Zhao, Yingdong, Ming-Chung Li, Mariam M Konaté, Li Chen, Biswajit Das, Chris Karlovich, P Mickey Williams, Yvonne A Evrard, James H Doroshow and Lisa M McShane (2021). 'TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository'. In: *Journal of translational medicine* 19.1, pp. 1–15.

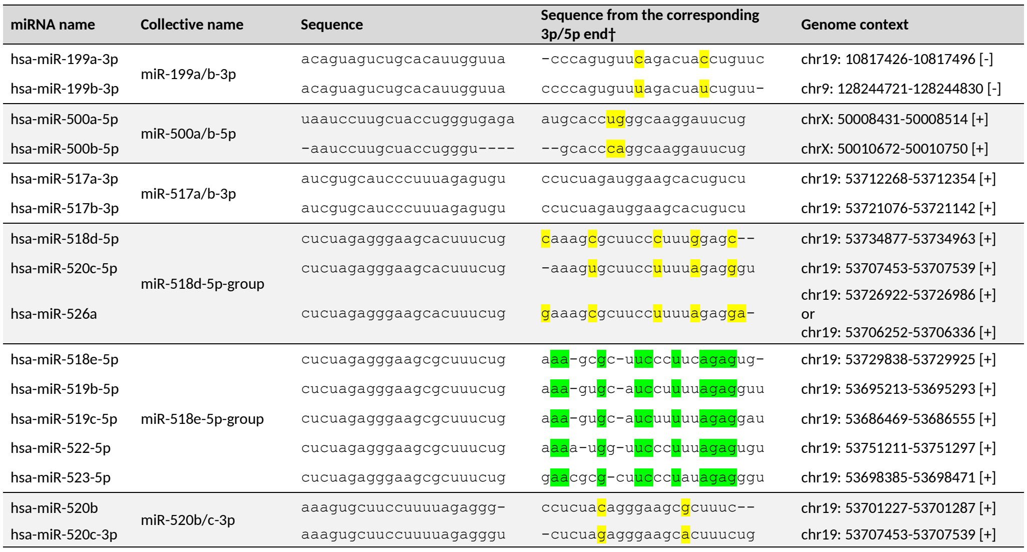
Zou, Hui and Trevor Hastie (2005). 'Regularization and variable selection via the elastic net'. In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2, pp. 301–320.

Zwiener, Isabella, Barbara Frisch and Harald Binder (2014). 'Transforming RNA-Seq data to improve the performance of prognostic gene signatures'. In: *PLOS One* 9.1, e85150.

Appendix A

Additional Material

A.1 Merging of Sequences in the Dataset



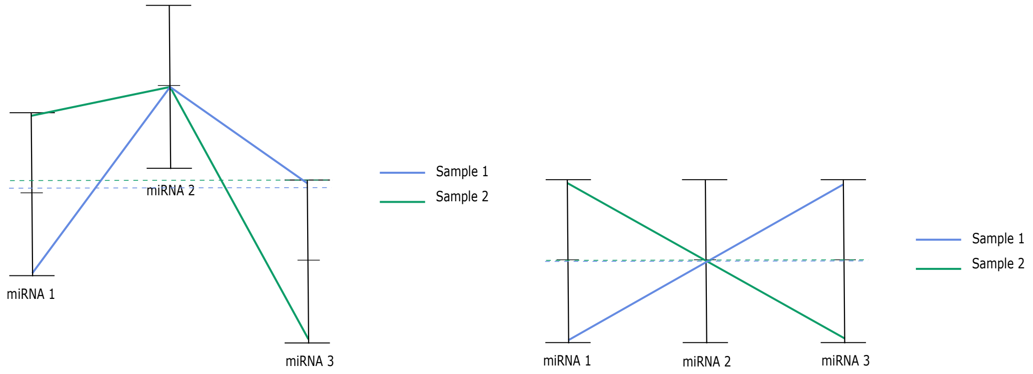
miRNA name	Collective name	Sequence	Sequence from the corresponding 3p/5p end†	Genome context
hsa-miR-199a-3p	miR-199a/b-3p	acaguagucugcacauugguua	-cccaguguucagacuaccuguuc	chr19: 10817426-10817496 [-]
hsa-miR-199b-3p		acaguagucugcacauugguua	ccccaguguuuagacuaucuguu-	chr9: 128244721-128244830 [-]
hsa-miR-500a-5p	miR-500a/b-5p	uaauccuugcuaccugggugaga	augcaccugggcaaggauucug	chrX: 50008431-50008514 [+]
hsa-miR-500b-5p		-aauccuugcuaccugggu----	--gcacccaggcaaggauucug	chrX: 50010672-50010750 [+]
hsa-miR-517a-3p	miR-517a/b-3p	aucgugcaucccuuuagagugu	ccucuagauggaagcacugucu	chr19: 53712268-53712354 [+]
hsa-miR-517b-3p		aucgugcaucccuuuagagugu	ccucuagauggaagcacugucu	chr19: 53721076-53721142 [+]
hsa-miR-518d-5p	miR-518d-5p-group	cucuagagggaagcacuuucug	caaagcgcuucccuuuggagc--	chr19: 53734877-53734963 [+]
hsa-miR-520c-5p		cucuagagggaagcacuuucug	-aaagugcuuccuuuuagagggu	chr19: 53707453-53707539 [+]
hsa-miR-526a		cucuagagggaagcacuuucug	gaaagcgcuuccuuuuagagga-	chr19: 53726922-53726986 [+] or chr19: 53706252-53706336 [+]
hsa-miR-518e-5p	miR-518e-5p-group	cucuagagggaagcgcuuucug	aaa-gcgc-uucccuucagagug-	chr19: 53729838-53729925 [+]
hsa-miR-519b-5p		cucuagagggaagcgcuuucug	aaa-gugc-aucgcuuuuagagguu	chr19: 53695213-53695293 [+]
hsa-miR-519c-5p		cucuagagggaagcgcuuucug	aaa-gugc-aucuuuuuagaggau	chr19: 53686469-53686555 [+]
hsa-miR-522-5p		cucuagagggaagcgcuuucug	aaaa-ugg-uucccuuuagagugu	chr19: 53751211-53751297 [+]
hsa-miR-523-5p		cucuagagggaagcgcuuucug	gaacgcg-cuucccuauagagggu	chr19: 53698385-53698471 [+]
hsa-miR-520b	miR-520b/c-3p	aaagugcuuccuuuuagaggg-	ccucuacagggaagcgcuuuc--	chr19: 53701227-53701287 [+]
hsa-miR-520c-3p		aaagugcuuccuuuuagagggu	-cucuagagggaagcacuuucug	chr19: 53707453-53707539 [+]

Figure A.1: Overview of the renaming of miRNAs with the same sequence. †Yellow highlights indicate the differences in the sequences of corresponding $3p/5p$ ends whilst green highlights indicate the similarities in these sequences. Note on $miR\text{-}517a/b\text{-}3p$: both the $3p$ and $5p$ end of these mature miRNAs have identical sequences, however there are differences in the loop part of the stemloop sequence.

A.2 Standardization of MicoRNAs Before Clustering

One can argue that standardization is not necessary when clustering samples by thinking that all miRNAs have equal mean or that different means do not matter, and that the variance does not influence the clustering. However, as can be seen from our dataset the assumption of equal means is not a plausible assumption. We explain why we believe centering and scaling is necessary for each individual approach below.

As mentioned, we standardize each miRNA such that they have mean zero and standard deviation equal to one before the clustering. When performing clustering of miRNAs using a correlation based distance, a standardization makes no difference as it done as part of the calculation of the correlation, and standardizing twice gives the same results. However when clustering the samples using a correlation based distance we argue it is important to standardize for each miRNA. Firstly, if we do not scale such that the miRNA have equal variance then miRNA with a small variance will have a low impact on the clustering compared to those with a high variance. One can argue that biologically miRNAs with a high variance are more interesting than those with a very small variance. However, we have found no biological evidence that this is always the case and neither found any reasoning for giving them more influence in the clustering. Additionally we have to account for the variance depending on the mean, thus we need to scale the miRNAs to have equal variance. Secondly, we argue that centering of the miRNAs is essential. There are many reasons for this, but we will present two that describe why it is important for both of the correlation based distances. Lets say for simplicity that we have scaled the miRNAs to have equal variance. Then lets consider two samples. The two samples have an extremely large expression value for one particular miRNA. Then say this miRNA have a mean value that is much larger than both the sample means. Consider now another miRNA with a mean almost equal to the sample means. The two samples have extremely large values for this miRNA as well. Then the first miRNA will have a much greater impact on the correlation between the two samples than the second miRNA, just because it has a larger mean. Additionally, if the two samples have an extremely low value for a miRNA with a high mean Then this will almost not impact the clustering at all, since the samples expression values for this miRNA is close to the sample mean. In addition to this, the correlation will in some cases be wrongly computed. To illustrate this we consider yet another example. Two samples and their expression values for three miRNAs are showed in Figure A.2. In both figures the miRNAs are scaled to have equal variance, and in Figure A.2 (b) they are also centered. As is clearly visible from Figure A.2 (b) the two samples are negatively correlated.



(a) **(b)**

Figure A.2: The figures shows an example of three miRNAs before and after centering. The three miRNAs are assumed to be scaled to have equal variance. The two lines between the miRNAs represents their values in two different samples. The two dotted lines are the mean of the two samples over the three miRNAs.

Before the centering the samples appear to be positively correlated. They will also be considered positively correlated when computing the correlation between them, even though they are actually negatively correlated.

For the Euclidean distance, if we cluster the miRNAs without standardizing, then we will only cluster miRNAs that have similar expression values overall. We argue this will not give an interesting clustering considering the research question in this thesis, as we will only cluster miRNAs that are highly expressed in all samples together and those that are lowly expressed in all samples together. This is not as interesting as we want to examine if some miRNAs are up-regulated and down-regulated together between the groups; placebo and probiotic, or AD and no AD.

A.3 Hierarchical Clustering with Euclidean Distance

The hierarchical clustering of all miRNAs using the Euclidean distance and the Ward linkage is shown in Figure A.3. The hierarchical clustering of the top 20 miRNAs using the Euclidean distance and the Ward linkage is shown in Figure A.4.

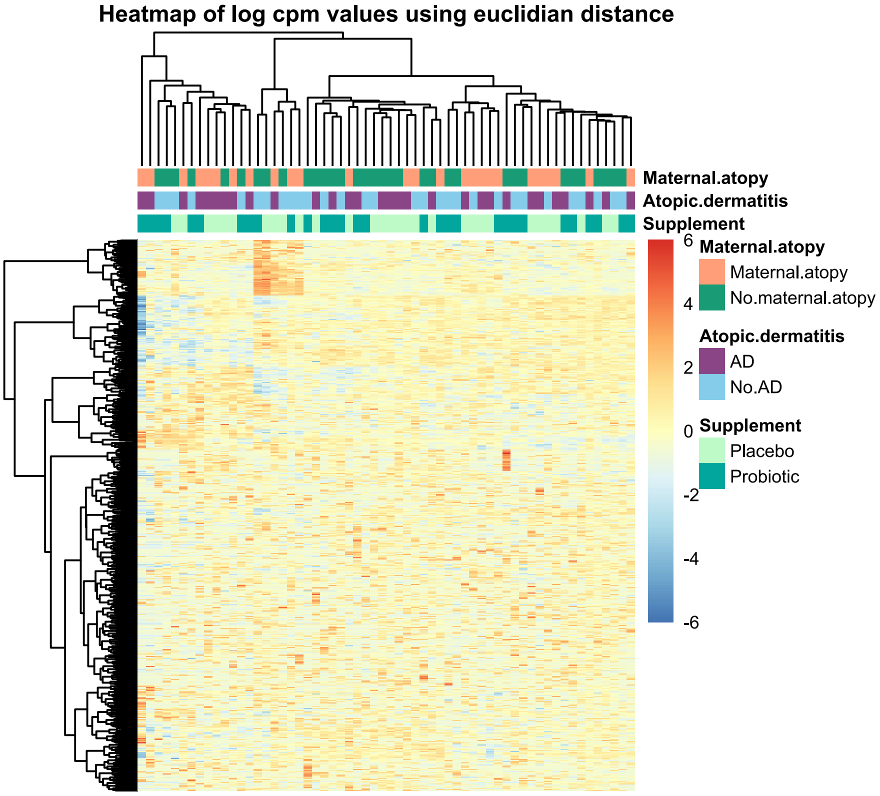


Figure A.3: Heatmap with dendrograms of the hierarchical clustering of the log cpm values scaled by row. The samples are given by the columns and the miRNAs by the rows. Clustering is done using the Euclidean distance and the Ward linkage function. Each sample is annotated by three variables; the supplement given to the mother, her history of dermatitis, and if the child developed dermatitis.

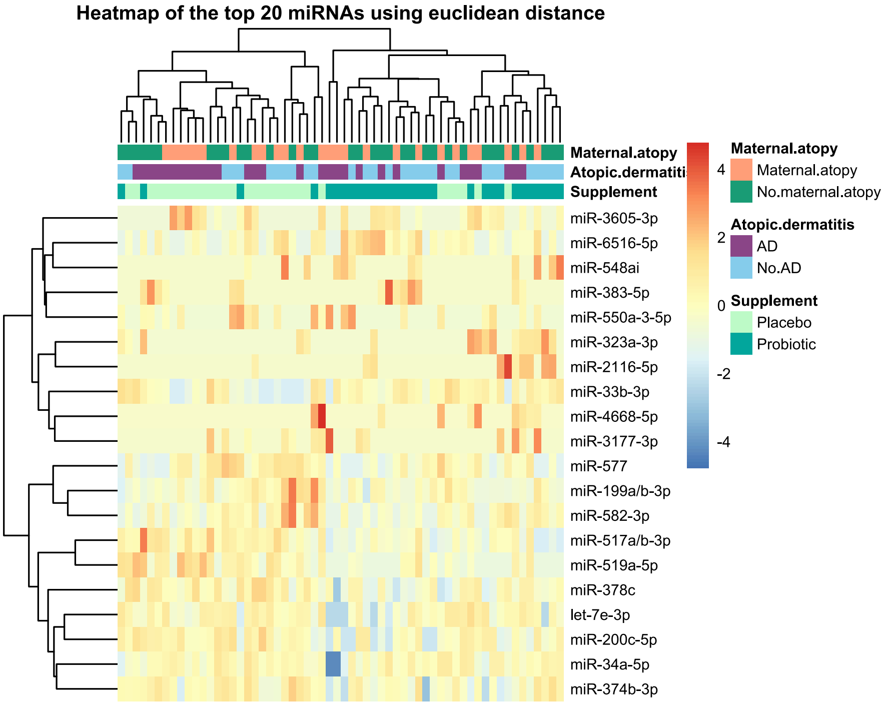


Figure A.4: Heatmap with dendrograms of the hierarchical clustering of the top 20 miRNA ranked by the voom analysis. The log cpm values scaled by row. The samples are given by the columns and the miRNAs by the rows. Clustering is done using the Euclidean distance and the Ward linkage function. Each sample is annotated by three variables; the supplement given to the mother, her history of dermatitis, and if the child developed dermatitis.

A.4 Analysis of Probiotics Effect without Gender and Maternal Atopy Covariates

The log fold change, average expression, *p*-value and adjusted p-value for the probiotic coefficient, when not including the gender and maternal atopy covariate in the linear model of the voom analysis, is shown in Table A.1.

miRNA	$\log_2$ FC	Average log cpm value	*p*-value	adj. *p*-value
miR-577	-2.112	2.970	0.000	0.011
miR-323a-3p	1.015	0.343	0.001	0.200
let-7e-3p	-0.943	3.855	0.002	0.455
miR-33b-3p	1.113	2.710	0.003	0.455
miR-2116-5p	0.824	0.126	0.005	0.486
miR-200c-5p	-0.856	3.026	0.006	0.486
miR-383-5p	0.870	0.315	0.006	0.486
miR-517a/b-3p	-1.285	4.066	0.006	0.486
miR-199a/b-3p	-0.557	7.465	0.008	0.496
miR-582-3p	-1.155	2.449	0.008	0.496
miR-519a-5p	-1.162	1.842	0.014	0.731
miR-190a-5p	-0.754	4.300	0.015	0.731
miR-3613-5p	-0.419	8.521	0.023	0.731
miR-374a-3p	-0.388	6.974	0.024	0.731
miR-378d	-0.401	6.581	0.026	0.731
miR-34a-5p	-0.526	6.271	0.027	0.731
miR-6516-5p	0.832	1.778	0.027	0.731
miR-4668-5p	0.588	0.098	0.027	0.731
miR-548l	0.653	0.382	0.028	0.731
miR-3605-5p	0.969	1.981	0.028	0.731

Table A.1: Top 20 miRNAs when using the linear model without including gender and mothers atopy as covariates. The miRNAs are ranked by *p*-value. The estimated $\log_2$ fold change between the probiotic and placebo group is given in the log FC column and average log cpm value is the mean log cpm value over all samples. The adjusted *p*-values are computed using the Benjamini and Hochberg method for controlling the FDR.

A.5 Top 50 MicroRNAs in the Analysis of Probiotics Effect on MicroRNAs

In Table *A*.2 shows the top 50 miRNAs, ranked by *p*-value, when performing the *voom* analysis using the linear model defined in 5.2.

miRNA	$\log_2$ FC	Average log cpm value	*p*-value	adj. *p*-value
miR-577	-2.067	-4.606	0.000	0.012
miR-323a-3p	1.072	3.657	0.001	0.154
let-7e-3p	-0.990	-3.200	0.002	0.405
miR-2116-5p	0.887	3.053	0.003	0.405
miR-200c-5p	-0.922	-3.000	0.004	0.405
miR-517a/b-3p	-1.388	-2.980	0.004	0.405

miR-4668-5p	0.750	2.928	0.005	0.405
miR-33b-3p	1.037	2.816	0.006	0.486
miR-383-5p	0.858	2.766	0.007	0.495
miR-519a-5p	-1.238	-2.702	0.009	0.516
miR-199a/b-3p	-0.553	-2.607	0.011	0.516
miR-6516-5p	0.953	2.581	0.012	0.516
miR-582-3p	-1.105	-2.555	0.013	0.516
miR-378c	-0.336	-2.554	0.013	0.516
miR-34a-5p	-0.593	-2.516	0.014	0.516
miR-374b-3p	-0.526	-2.515	0.014	0.516
miR-3177-3p	0.775	2.466	0.016	0.516
miR-550a-3-5p	0.891	2.438	0.017	0.516
miR-3605-3p	0.855	2.425	0.018	0.516
miR-548ai	0.831	2.418	0.018	0.516
miR-378d	-0.439	6.581	0.019	0.516
miR-190a-5p	-0.757	4.300	0.019	0.516
miR-548l	0.708	0.382	0.020	0.516
miR-7976	0.732	0.606	0.020	0.516
miR-760	0.860	0.373	0.022	0.527
miR-146b-5p	-0.273	15.865	0.025	0.539
miR-30e-5p	-0.291	13.896	0.025	0.539
miR-342-3p	0.224	9.824	0.025	0.539
miR-26a-1-3p	0.597	0.170	0.026	0.541
miR-3605-5p	1.005	1.981	0.027	0.543
miR-1260a	0.616	0.210	0.029	0.553
miR-25-5p	0.992	1.106	0.029	0.553
miR-516b-5p	-1.311	2.189	0.033	0.580
miR-3913-3p	0.799	0.425	0.034	0.580
miR-7706	0.748	3.340	0.034	0.580
miR-330-3p	-0.407	4.561	0.035	0.580
miR-561-5p	0.543	0.107	0.036	0.580
miR-570-5p	0.715	0.223	0.036	0.580
miR-374a-5p	-0.267	11.264	0.039	0.601
miR-374a-3p	-0.364	6.974	0.042	0.612
miR-146a-5p	-0.320	12.382	0.042	0.612
miR-2115-3p	0.912	1.370	0.043	0.612
miR-6777-5p	0.679	0.540	0.043	0.612
miR-1976	0.689	0.427	0.045	0.615
miR-6842-3p	0.720	0.558	0.046	0.615
miR-224-3p	-0.434	5.448	0.049	0.622
miR-4536-5p	0.578	0.186	0.049	0.622
miR-548v	0.647	0.370	0.050	0.622
miR-3613-5p	-0.365	8.521	0.052	0.622
miR-148b-3p	-0.218	12.129	0.053	0.622

Table A.2: Top 50 miRNAs when using the linear model including gender and mothers atopy as covariates. The miRNAs are ranked by p-value. The estimated $\log_2$ fold change between the probiotic and placebo group is given in the log FC column and average log cpm value is the mean log cpm value over all samples. The adjusted p-values are computed using the Benjamini and Hochberg method for controlling the FDR.

A.6 Nested CV Bootstrapping Results

Here we present the results of the bootstrapping when using the proposed nested CV. A boxplot of the coefficients estimates are shown in Figure A.5. We also include a barplot showing the proportion of times each coefficient was set to zero in the bootstrap samples, this is shown in Figure A.6.

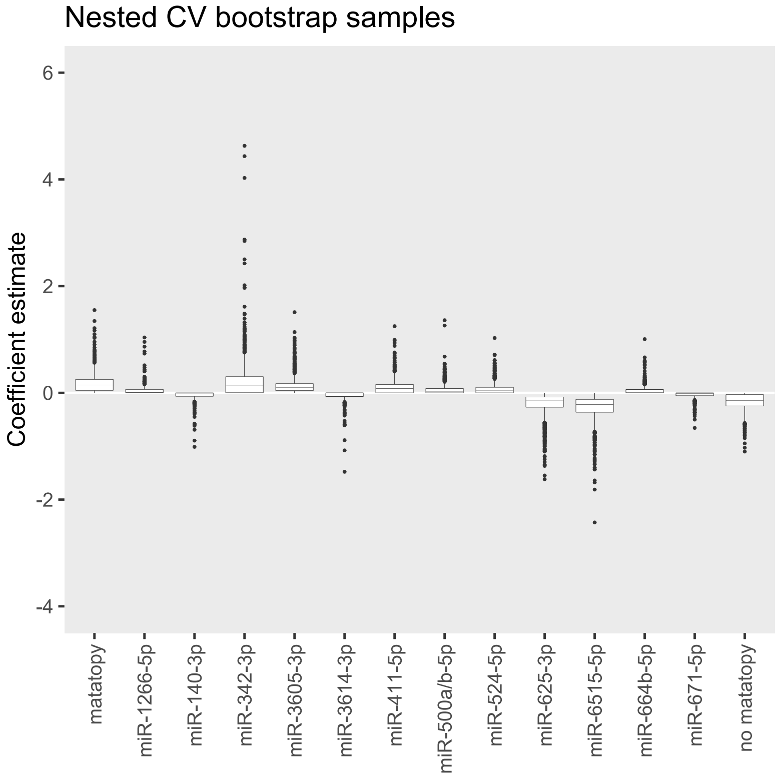


Figure A.5: Boxplot of the estimated coefficient values for 1000 bootstrap samples when using the nested CV to determine the model parameters of the elastic net model. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The whiskers extends from the hinge to the largest/smallest value at most 1.5 times the inter-quartile range from the hinge. The values beyond the end of these whiskers are plotted as individual points.

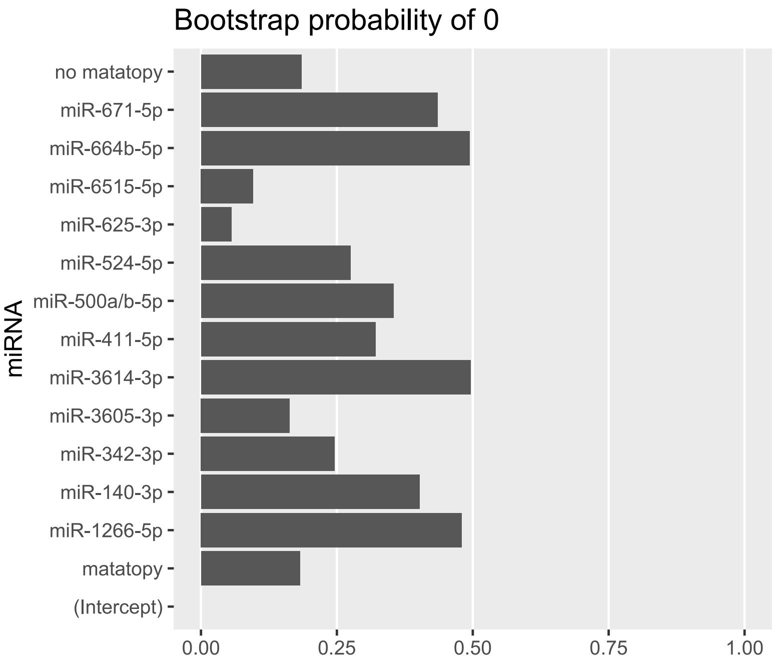


Figure A.6: Barplot showing the bootstrap probability of zero for coefficients not set to zero in at least 50% of the bootstrap models when using the nested CV for determining the model parameters. The bootstrap probability of zero is the proportion of times they where set to zero.

Using the nested CV algorithm with the minimum λ criterion to choose λ and α of the elastic net model, results in the model coefficients shown in Table A.3

	Nested CV using minimum λ
Intercept	-2.38
miR-103a-2-5p	-0.01
miR-126-5p	-0.04
miR-1266-5p	0.09
miR-1323	0.01
miR-140-3p	-0.09
miR-142-5p	-0.00
miR-181a-5p	0.05
miR-181b-5p	0.03
miR-191-3p	-0.07
miR-200c-5p	0.01
miR-203b-3p	0.03
miR-20a-3p	0.04
miR-21-5p	-0.06
miR-223-3p	-0.00
miR-223-5p	-0.03
miR-23a-5p	-0.01
miR-26a-1-3p	0.09
miR-301b-3p	0.07
miR-3178	0.05
miR-328-3p	0.04
miR-330-3p	-0.04
miR-33a-5p	-0.06
miR-342-3p	0.31
miR-3605-3p	0.17
miR-3614-3p	-0.08
miR-378c	0.06
miR-378d	0.01
miR-379-5p	0.01
miR-3912-3p	0.01
miR-411-5p	0.18
miR-421	-0.01
miR-450a-5p	-0.01
miR-4668-5p	0.10
miR-4772-3p	-0.01
miR-4775	-0.04
miR-4780	0.05
miR-4802-5p	-0.05
miR-494-3p	0.02
miR-500a/b-5p	0.08
miR-505-3p	-0.05
miR-517c-3p	0.01
miR-518c-3p	0.01
miR-520a-3p	0.01
miR-524-5p	0.10

miR-525-5p	0.05
miR-526b-5p	0.02
miR-532-3p	-0.01
miR-548ah-5p	0.01
miR-548e-5p	-0.02
miR-548l	-0.08
miR-548w	0.05
miR-570-5p	-0.06
miR-625-3p	-0.19
miR-629-3p	-0.01
miR-629-5p	-0.01
miR-6511a-3p	0.05
miR-6515-5p	-0.34
miR-664b-5p	0.07
miR-671-5p	-0.07
miR-744-5p	-0.01
miR-885-5p	0.01
miR-99b-3p	-0.03
matatopy1	0.21
noMatatopy1	-0.21

Table A.3: Table presenting estimated model coefficients when using nested CV with λ_{min}. The model parameters of this mode are $\alpha = 0.15$ and $\lambda = 0.42$. Covariates with estimated coefficients equal to zero are not included in the table.

A.7 Code

All the code used to perform the analyses in this thesis can be found in the following Github repository: `https://github.com/leneomdal/Master-thesis-miRNA-analysis`.

NTNU

Norwegian University of
Science and Technology