Statistical Analyses of High Dimensional MicroRNA Data in Relation to Incidence and Survival After Cancer

Thor Schütt Svane Nielsen

Kongens Lyngby 2012
IMM-M.Sc.-2012-25

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

Abstract

Pancreatic cancer is globally the 4^{th} most common cause of cancer death and the overall 5-year survival rate among patients is less than 5%. Often the pancreatic cancer is already at advance stages when discovered, so the difficulties of an early diagnosis makes the life prognosis for these patients very dismal. Part of the problem with detecting this type of cancer in time, is that there are no typical symptoms. Incidence and prognosis prediction from high dimensional gene expression data have been subject to much research during recent years.

This thesis examines the relationship between microRNA expression profiles and their ability to predict correct diagnostics and expected survival from time of operation. This research area can hopefully reform future courses of treatment by providing patients with pancreatic cancer earlier diagnosis, and thus improve their prognosis.

This thesis deals with the statistical modelling of microRNA measurements from serum samples of both pancreatic patients and healthy controls. The analyses are divided into two parts. The incidence part focuses on the logistic model for predicting a binary outcome and the prognostic part considers Cox's proportional hazards model in order to handle censored survival times. However since parsimonious models are of clinical relevance, these models are used in combination with coefficient shrinkage techniques, where the shrinkage methods used here are univariate selection, backwards stepwise selection, Ridge regression, Lasso regression and naïve elastic net regression. These shrinkage methods require estimation of penalty parameters for which cross-validation have served as an excellent tool.

Results based on five different normalization methods indicate that models with only a few microRNAs are good predictors of cancer. The comparative study of the incidence analyses show no significant difference in prediction ability between the various shrinkage methods considered. The analyses of prognosis reveal no clear signal in the microRNAs in terms of predicting survival, which could be a result of scarce data. All in all, microRNA expression profiles are promising candidate biomarkers of pancreas cancer.

KEYWORDS: microRNA, pancreas cancer, normalization methods, incidence, generalized linear models, logistic regression, prognosis, survival analysis, Cox proportional hazards model, shrinkage methods.

Resumé

Kræft i bugspytkirtlen er globalt set den fjerde mest almindelige kræftrelateret død og overlevelsesprocenten på 5-års plan for disse patienter er mindre end 5%. Oftest er pancreaskræft allerede på fremskredne stadier når den bliver opdaget, så grundet vanskeligheder forbundet med en tidlig diagnose, er livsprognosen for disse patienter meget trist. En del af problemet med at opdage denne type af kræft i tide, er at der ikke er nogen typiske symptomer. Forudsigelse af incidens og prognose fra høj dimensionelle gen-profil data har været forsket i meget de seneste par år.

Dette kandidatspeciale undersøger sammenhængen mellem mikroRNA-profiler og deres evne til at forudsige den korrekte diagnose, samt den forventede overlevelsestid fra operationsdato. Dette forskningsområde kan forhåbentlig forbedre fremtidige behandlingsforløb og give patienter med kræft i bugspytkirtlen en tidligere diagnose, og dermed øge deres overlevelseschancer.

Dette kandidatspeciale omhandler statistisk modellering af mikroRNA-målinger fra serumprøver af både patienter med pancreaskræft og raske kontroller. Analyserne er inddelt i to dele. Incidensdelen fokuserer på den logistiske model, brugt til at forudsige et binært udfald, mens den prognostiske del anvender Coxs proportional hazards model der kan håndtere censurerede overlevelsestider. Men siden modeller med begrænset variable er kliniske relevante, er de nævnte modeller brugt i kombination med teknikker der kan indskrumpe koefficienterne, hvor metoderne brugt her er univariat selektion, baglæns trinvist selektion, Ridge regression, Lasso regression og naiv elastisk net regression. Disse shrinkagemetoder indebærer estimering af strafparametre, hvor krydsvalidering fungerede som et fremragende værktøj til dette formål.

Resultaterne baseret på fem forskellige normaliseringsmetoder, indikerer at modeller med kun få mikroRNA viser sig at være gode til at forudsige tilfælde med kræft. Det komparative studie af incidensanalyserne viser at der ikke er nogen signifikant forskel i evnen til at forudsige kræft, for de respektive shrinkagemetoder. Analyserne af prognose detekterer ikke noget klart signal i mikroRNAerne med hensyn til evnen til at forudsige overlevelse, hvilket kan være et resultat af et begrænset antal prøver. Alt i alt, mikroRNA-profiler er lovende biomarkører af kræft i bugspytkirtlen.

NØGLEORD: mikroRNA, kræft i bugspytkirtlen, normaliseringsmetoder, incidens, generaliserede linære modeller, logistisk regression, prognose, overlevelsesanalyse, Cox proportional hazards model, shrinkagemetoder.

Preface

This master thesis was prepared at the Department of Informatics and Mathematical Modelling (IMM) at the Technical University of Denmark (DTU) in cooperation with Danish Cancer Society Research Center and Herlev Hospital. It represents a partial fulfillment of the requirements for acquiring the Master of Science degree (M.Sc) in Engineering, cand.polyt. This final report concludes the two-year programme of Mathematical Modelling and Computation and was prepared over a six months period, corresponding to a workload of 30 ECTS points.

First and foremost, I would like to thank my supervisors Klaus K. Andersen and Christian Dehlendorff from Danish Cancer Society. For being supportive and motivating from the very beginning of the project, and a continuing source of priceless information, guidance and new project ideas when needed, which was very much appreciated. Secondly, my supervisor Per B. Brockhoff from DTU deserves thanks for his valuable comments and insights to certain parts of the report. Furthermore, I thank my collaborators from Herlev Hospital, especially professor Julia S. Johansen and surgeon Nicolai A. Schultz, for both supplying the data and providing aid in understanding relevant biological aspects. Last but not least, a special big thanks to my family, friends and coworkers for their gentle encouragements, patience and understanding throughout this entire project.

Kgs. Lyngby, 20$^{\text{th}}$ March 2012

Thor Schütt Svane Nielsen

Nomenclature

α	Tuning parameter in general / for naïve elastic net
β	Effect parameters (coefficients)
δ	Censorship
exp	Exponential function
$\hat{\Lambda}(t)$	Nelson-Aalen estimator of the cumulative hazard
$\hat{S}(t)$	Kaplan-Meier estimator of the survival function
$\Lambda(t)$	Cumulative hazard
λ_0	Tuning parameter for univariate method
λ_1	Tuning parameter for Lasso
λ_2	Tuning parameter for Ridge
log	Natural logarithm
$\mathbb{E}$	Expected value
$\ell(\cdot)$	Log-likelihood function
$\mathcal{B}$	Binomial distribution
$\mathcal{H}$	Hypothesis
$\mathcal{L}(\cdot)$	Likelihood function
$\mathcal{N}$	Normal distribution
$\mathcal{U}$	Uniform distribution
$\mathbf{C}$	Internal control normalized matrix
$\mathbf{Q}$	Quantile normalized matrix
$\mathbf{U}$	Mean normalized matrix
$\mathbf{U}_{120}$	Mean-120 normalized matrix
ρ	Spearman's rank correlation coefficient
P	Probability
AIC	Akaike information criterion
AUC	Area under curve
BIC	Bayesian information criterion
C_t	Cycling threshold

CI	Confidence interval
CP	Chronic pancreatitis
CV	Cross-validation
d	Uncensored subjects (deaths)
DM	Deviance measure
DOE	Design of experiment
$F(\cdot)$	Cumulative distribution function
$f(\cdot)$	Probability distribution function
FN	False negative
FP	False positive
FPR	False positive rate
GLM	Generalized linear models
$h(t)$	Hazard rate
HR	Hazard ratio
HS	Healthy subject
IM	Informative missing
IQR	Interquartile range
$IRLS$	Iteratively reweighted least squares
$L(\cdot)$	Loss function
L^1	L^1-space
L^2	L^2-space
$Lasso$	Least absolute shrinkage and selection operator
LM	General linear models
MAR	Missing at random
$MCAR$	Missing completely at random
MLE	Maximum likelihood estimation
MLR	Multiple linear regression
n	Number of samples
N/A	Not available (missing)
OR	Odds ratio
p	Number of parameters / probability
PC	Pancreatic cancer
$PDAC$	Pancreatic ductal adenocarcinoma
PI	Prognostic index
PM	Performance measure
$qrt–PCR$	Quantitative real time polymerase chain reaction
r	Pearson's product-moment correlation coefficient
$r(t)$	Risk set
ROC	Receiver operating characteristics
RSS	Residual sum of squares
$S(t)$	Survival function
T	Survival time
t	Time
TN	True negative
TP	True positive
TPR	True positive rate

Contents

Chapter 1

Introduction

Pancreas cancer is potentially a lethal disease that in most cases evolves very rapidly. Usually at the time of diagnosis, patients already have locally advanced or metastatic pancreatic cancer, where surgical procedure with curative intent is only possible for a smaller proportion. Earlier diagnosis of these patients is therefore crucial for their prognosis.

Prediction of pancreatic cancer patients and their expected survival based on gene expression profiles is thus an important application of genome-wide expression data. This thesis deals with microRNA expression profiles and tries to uncover the relationship between these profiles and both diagnostics, but also the time from operation to death. It is the hope that these results can help and be a part of a larger objective to archive more accurate incidence and prognoses determination, hence improving the treatment strategies for these patients.

The thesis deals with statistical modelling of data from pancreatic cancer patients provided by Herlev Hospital and Rigshospitalet. The main objective is to determine a subset of microRNAs which can be considered as good predictors of the incidence of pancreatic cancer, as well as a subset that gives information concerning the expected survival. At the time of writing there is no standard-

ized way of analyzing microRNA data in relation to incidence and prognosis. Substantial statistical challenges are connected with this topic, especially the fact that the number of microRNA variables are considerably larger than the samples available.

The main focus in this thesis consists of how data should be normalized and the methods for which data should be analyzed, such that the final results derived can be used in a clinical perspective. The latter involves methodology from logistic regression, Cox proportional hazards model and the use of shrinkage methods. The organization of the report can be described as follows.

Chapter 2: Clinical relevance. Provides a basic introduction to the biological concepts of microRNA and pancreas cancer which are the fundamental biological topics in this thesis.

Chapter 3: Data. Explains the underlying idea behind microRNA measurements and gives a thorough description of the data set provided, which is the foundation for all the analyses in this thesis.

Chapter 4: Methodology. Describes the theory behind the methods applied in the analyses. Overall the analyses can be subdivided into an incidence and prognosis part, with main focus on normalization methods, logistic regression, Cox proportional hazards model and shrinkage methods.

Chapter 5: Simulation study. This is a small theoretical study that seeks to understand how one certain normalization method cope with different types of noise typically encountered in this type of application.

Chapter 6: Results. Presents the results from the various analyses. This includes a comparative study and analysis using different normalization methods, for both the incidence and prognostic part.

Chapter 7: Discussion. Here the obtained results are discussed and put into a clinical perspective. Furthermore, the validity of the results is evaluated and other analysis approaches are considered.

Chapter 8: Conclusion. Summarizes the most important results along with a reflection on the work process and future research within this area.

Appendix. Consists of two supplemental parts to the thesis; some additional results and bibliography.

The analyses performed in this report was made using `R` version 2.14.1 and furthermore `Sweave` was used as a tool to embed relevant `R` code in LaTeX documents, where tables were generated with the `R` package `xtable` by Dahl [2009]. This ensured that the resulting output could be updated automatically if data or analysis changed, which was very helpful. All the `R` programming is enclosed on a CD.

All the actual microRNA names in the data have been coded due to confidentiality reasons, instead aliases created from a algorithm was used throughout the thesis.

Chapter 2

Clinical relevance

The human genome is organized in the famous double helix structure with high complexity. It is known that less than 2% of the total DNA, corresponding to about 23-25.000 genes, encodes for the production of protein, which is important for the body in relation to structure and reparation of bones, muscles, immune system, connective tissue etc.

Up until recently it was of scientific perception that the rest of our about 98% human genome, could be classified as so-called "junk DNA". More explicitly it consists of noncoding DNA (ncDNA), noncoding RNA (ncRNA), introns and so on. However this human material was considered waste, because there was no knowledge of it having any biological function, and the general belief was that it was just some immaterial leftovers from the human evolution over time. New analysis methods the past five years, have made it possible to conclude that this is not how the human biology works, far from it. This part of our DNA actually contains a lot of information systems, that do not encode for protein, but serve other biological purposes. Exactly how many distinct systems there exist is yet to be discovered, but one system is shown to be of great importance concerning cells regulation mechanisms; *microRNA* [Larsen 2011].

It is already well documented that microRNA (miRNA) plays an important role in cancer pathogenesis, apoptosis and cell growth, which is why this system of regulators have received such massive interest the past decade. Ideally these relatively new biomarkers, functioning as tumor suppressors or oncogenes, can help the health sector in the long run by earlier detection of various cancer types or other diseases, just by looking at an individual's miRNA profile. It is a well known fact that early diagnosis of cancer is crucial for the prognosis [Zhang et al. 2009].

So a lot indicates that miRNA will have tremendous impact on future medical routines. In the next section the miRNA will be explained from a more biotechnical perspective, defining more explicitly what a miRNA is.

2.1 MicroRNA

The first miRNA was actually discovered almost two decades ago, more specifically found by Lee et al. [1993] in the worm *Caenorhabditis elegans*. But it was not until the early 2000s that it was recognized as a distinct class in the bio community. In the last five years new miRNA discoveries have reached a seemingly exponential growth, which is related to the previously mentioned exploding interest within this area. This is illustrated in Figure 2.1.

There are at the time of writing 1527 known human miRNA sequences and this number is increasing (miRBase, last accessed November 2011), however there is a large variation in the knowledge of each individual miRNA. Some have widely known biological properties and are highly characterized, but a large part is still new to science, and hence a good basis for further research.

Concurrently with the rising number of new miRNA discoveries, a rigid, uniform system for miRNA nomenclature was to a great extent needed. One of the key problems was to distinguish miRNAs from e.g. siRNAs (small interfering RNAs), which is a class of double-stranded RNA molecules similar in terms of their functions and biological compound. Hence, the first thing done to ensure that only true miRNAs enter the miRBase Registry, was to demand that a certain combination of expression and biogenesis criteria was satisfied [Ambros et al. 2003].

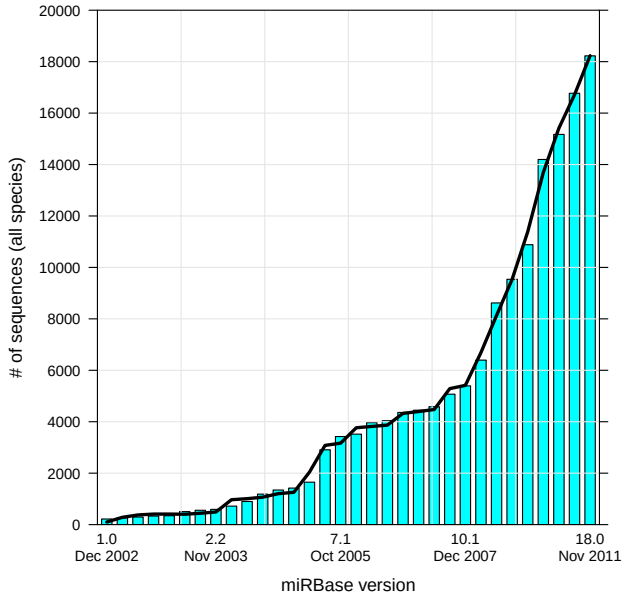

Figure 2.1: *Hairpin precursor miRNA entries, data found at miRBase.*

When novel entries fulfill the requirements that characterize them as miRNAs, a consistent naming scheme is applied. The miRNAs are assigned by sequential numerical identifiers according to experimentally confirmed miRNAs, before publication of their discovery. The number is connected with two prefixes, the first one consists of an abbreviation of 3-4 letters used to designate the species, e.g. *hsa* is used when the miRNA is found in a human gene (*Homo SApiens*)[1]. Second prefix specifies if the miRNA is a mature sequence (labeled *miR*), or precursor hairpins (labeled *mir*), related to the processing of miRNAs, these terms will be elaborated later. An example of a miRNA could be hsa-miR-101, which is most likely discovered before hsa-miR-136. Sequences whose mature miRNAs differ only at one or two nucleotides are given lettered suffixes, e.g. hsa-miR-10a and hsa-miR-10b, because they are very closely related. In a similar way, distinct hairpin loci that give rise to identical mature miRNAs, but are located in different regions of the genome, are given numbered suffixes, e.g. hsa-mir-219-1 and hsa-mir-219-2. Furthermore when two mature miRNAs originate from opposite arms of the same hairpin precursor, they are denoted with a *-3p* or *-5p* suffix. These suffixes refers to the three, respectively five prime

[1] *let-7* (*LEThal-7*) is one of the first discovered miRNAs and is special in the way that it is evolutionarily conserved from fly to human. The let-7 family comprises of twelve human genes encoding for nine distinct miRNAs (let-7a to let-7i).

untranslated regions (usually denoted 3'UTR and 5'UTR), which are particular coding regions of the messenger RNA (mRNA) [Griffiths-Jones et al. 2006].

MiRNA is a molecular group of short non-coding single-stranded RNAs with an average of 22 nucleotides. These very small RNA molecules are first being transcribed from the genome to primary miRNA (pri-miRNA) in the nucleus. The pri-miRNA is a long RNA precursor that contains a stem-loop structure of about 80 bases (also called hairpin structure because of its shape). The pri-miRNA is then cleaved into precursor miRNA (pre-miRNA) by the RNase III enzyme Drosha and Pasha protein. This pre-miRNA is likely to obtain the same characteristic hairpin structure, which basically is the specific miRNA sequence from the pri-miRNA. Next the pre-miRNA is transported from the nucleus to the cell's cytoplasm by a transport molecule called Exportin-5. Here the Dicer enzym processes the pre-miRNA into its mature form, which binds to a multiprotein complex, called RNA-Induced-Silencing-Complex (RISC). This multiprotein complex regulates gene expression posttranscriptionally by binding of a specific mRNA. The processing of miRNAs and their biological impact are only roughly described here, in reality there is more detailed knowledge of the process, however it was found beyond the scope of this thesis to describe this. Figure 2.2 gives an illustration of the described procedure [AppliedBiosystems 2006, Schultz et al. 2011].

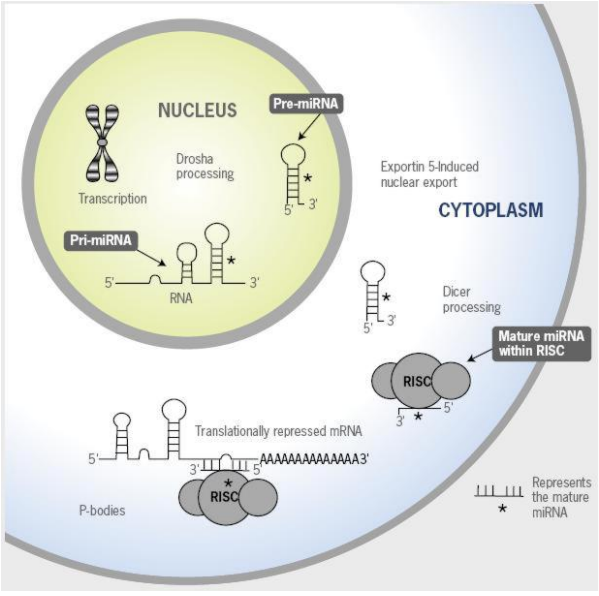

Figure 2.2: *Processing pathway of miRNA, provided by AppliedBiosystems [2006].*

So what used to be referred to as the biological equivalent of dark matter, miRNAs are now identified as key regulators of development, cell proliferation, differentiation, and the cell cycle. They are also known to have highly tissue-specific expression patterns, which makes them valuable biomarkers in separation of healthy and malignant tissue. Thus substantiating their role in transformation to malignant tissue and progression of malignant disease. This thesis' focus will be on miRNA profiling for pancreatic cancer patients, and this type of cancer is introduced in Section 2.2.

2.2 Pancreatic cancer

The pancreas is an essential organ for the functioning of the human body. The gland has dual functions in the human homeostasis. The exocrine part produce digestion enzymes and secrete them to the duodenum. The exocrine islands produce insulin and a hormone with the opposite functions called glucagon. It produces about $1.5l$ digestion liquid a day and this fluid neutralizes the stomach acid, along with splitting of proteins, fat and carbohydrates. The hormone insulin regulates the carbohydrate and fat metabolism in the body, and secretion of insulin is stimulated by consumption of meals. When the production of insulin is either to little or nonexisting, the usual diagnosis is diabetes [Patienthåndbogen 2008].

The pancreas typically weighs 100 to $150g$ and is between 12 and $15cm$ long. It is located deep down in the abdominal cavity, behind the stomach, where it is almost completely wrapped by the duodenum. The pancreas can be sectioned into three parts; the head (*caput*), body (*corpus*) and tail (*cauda*) [Patienthåndbogen 2008].

Jemal et al. [2010] states that pancreatic cancer (PC) is the 4^{th} most common cause of cancer death in United States, and the same was predicted for Europe in 2011 in the publication from Malvezzi et al. [2010]. Cancer in the pancreas is a highly lethal condition with an intimidating low survival rate, it has been reported that the overall 5-year survival rate among patients on a global plane, is less than 5% [Hidalgo 2010, Jemal et al. 2010].

Alone in Denmark, the average incidence of new pancreatic cancer patients per year from 2005-2009, were 445 men and 460 women. Getting the disease before reaching 50 years of age is a rare event, but it happens, however it is most likely to appear around the age of 65. The relative 1-year survival is 15% for men and women, when diagnosed in the period 1999-2003, and when looking

at the 5-year survival, the numbers are supporting the global percentage (3% for men, respectively 4% for women) [NORDCAN 2011].

Often the pancreatic cancer is already at advance stages when discovered, so the difficulties of an early diagnosis makes the life prognosis for these patients very dismal. Part of the problem with detecting this type of cancer in time, is that there is no typical symptoms, it includes e.g. weight loss, nausea, stomach pain and diarrhoea, which are all common symptoms. When patients present with jaundice, they often have advanced disease.

So far the only curable treatment is to surgically remove all cancer, however this is a complex procedure due to the fact that the tumor is not easily accessible, since it is placed behind other vital organs. The most common surgical treatment (*Whippel procedure*) for cancers involving the head of the pancreas, is to remove the pancreatic head, the duodenum and part of the common bile duct together (*pancreato-duodenectomy*). However it can only be performed if the patient is likely to survive major surgery and if the cancer is localized without invading local structures or metastasizing. Figure 2.3 is a simplified illustration of the stomach region and how the organs are connected before an eventual operation, and Figure 2.4 is after the surgical bypass is performed.

Figure 2.3: *Pancreas before operation, provided by Nicolai Schultz.*

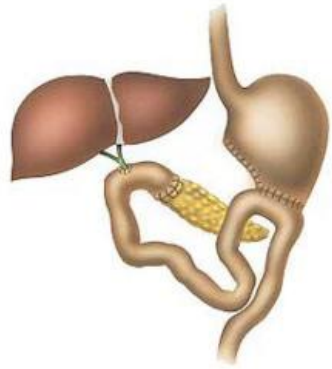

Figure 2.4: *Pancreas after operation, provided by Nicolai Schultz.*

Pancreatic cancer can be classified into a number of different histological types, but for all practical purposes this term refers to *pancreatic ductal adenocarcinoma* (PDAC), which is the most frequent and accounts for over 90%. About two-thirds of these tumors are located in the caput pancreatis, the rest can be diffuse or allocated between corpus and cauda pancreatis. There also exists very rare types, e.g. *neuroendocrine tumors*, that have a very different and atypical course of disease.

There is a certain clinical interest in malignant tumors located in the so-called papillary area, usually referred to as *periampullary tumors*. Besides caput pancreatitis, this group of carcinomas consists of *ampullary*, *duodenal* and *distal common bile duct cancers*. All of them resembles each other clinically and when scanned, the tumor type is determined by the location. Often it takes a histopathological examination to get the correct diagnosis. Figure 2.5 gives an overview of the papillary area.

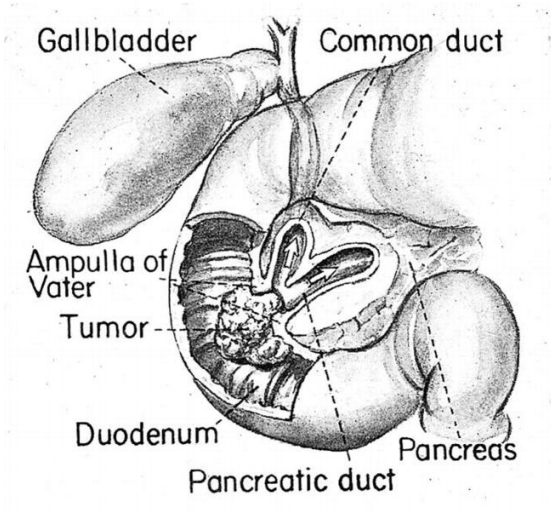

Figure 2.5: *An sketch of the papillary area, with a carcinoma of the ampulla of Vater, provided by Nicolai Schultz.*

Ampullary cancer is located in ampulla of Vater, which is an area formed by the union of the pancreatic duct and the common bile duct. It looks a lot like the common pancreatic cancer and is often noted as that, but has a better prognosis, mostly because of the critical localization which makes jaundice an early symptom. Duodenal cancer, as the name suggest, is placed in the duodenum, while common bile duct cancers are close to the gall bladder. All of these periampullary cancers usually express themselves with jaundice, because the tumor usually blocks the common bile duct, and hence accumulates gall matter. A fine example of real tissue infected with malignant carcinoma can be seen in Figure 2.6.

Chronic pancreatitis (CP) is a long-standing chronic inflammation of the pancreas which cause fibrosis and alters its normal structure and functions. This condition have no invasive potential, but the symptoms of pain, weight loss and sometimes jaundice mimics pancreatic cancer and it often cause diagnostic troubles. Not rarely are patients operated with a Whipple procedure for something

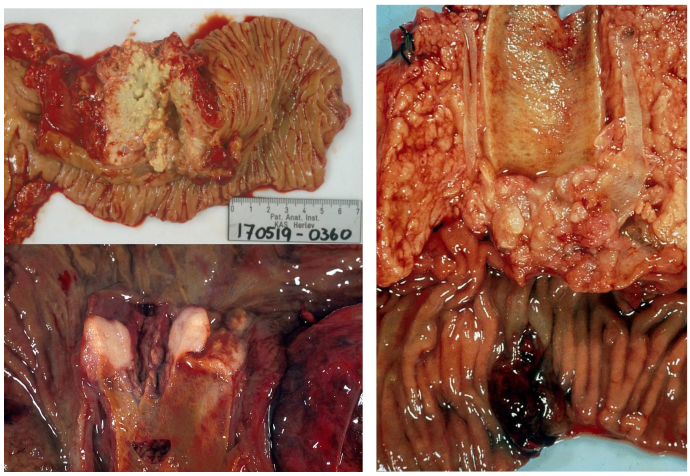

Figure 2.6: *Real life pancreatic (upper left and right) and ampullary (lower left) cancer tissue, provided by Nicolai Schultz.*

that turns out to be chronic pancreatitis. Persons with chronic pancreatitis regardless of its aetiology, is proved to have a higher probability of developing pancreatic cancer, however they are still two biologically different conditions.

In conclusion, people with pancreatic cancer are not in a very encouraging state, in light of their poor prognosis. Early detection is crucial for the possibility of operation and in general the chances of survival. Furthermore, the differentiation between various periampullary cancer types is troublesome due to clinical, radiological and histological similarities. Chronic pancreatitis mimics pancreatic cancer and is a daily clinical challenge for a pancreatic surgeon. Ideally these miRNA tissue-specific expression patterns, can help separate the pancreas cancer cases from those with chronic pancreatitis and healthy subjects (HS). Additionally reveal which miRNAs are significant regulators in relation to incidence and prognosis. This is the thesis' main focus. In Chapter 3, the descriptive and explorative analysis of the data set is presented.

Glossary

aetiology The cause of a disease.

apoptosis A process of programmed cell death by which cells undergo an ordered sequence of events which lead to death of the cell.

DNA Abbreviation for DeoxyriboNucleic Acid, which is an important substance responsible for the functioning of human bodies. DNA basically has its function to store information about your body. DNA has a capability to replicate itself and it is also responsible for production of RNA. Consists of two long chains of nucleotides twisted into a double helix and joined by hydrogen bonds between the complementary bases adenine (A) and thymine (T) or cytosine (C) and guanine (G). The sequence of nucleotides determines individual hereditary characteristics.

exocrine Gland that secretes outwardly through ducts.

gene The basic biological unit of heredity, i.e. genetic transmission from parent to child.

genome The total complement of genes in an organism or cell. For a human it is encoded in DNA and is divided into discrete units called genes.

histological The microscopic structure of tissue.

homeostasis The ability or tendency of an organism or cell to maintain internal equilibrium by adjusting its physiological processes.

intron Any nucleotide sequence within a gene that is removed by RNA splicing to generate the final mature RNA product of a gene.

junk DNA Noncoding regions of DNA that have no apparent biological function.

mRNA The messenger RNA contains a copy of the DNA strand, sort of chemical "blueprint", used for the protein synthesis.

nucleotide Generally a nucleotide is composed of a nucleobase (nitrogenous base), a five-carbon sugar (either ribose or 2'-deoxyribose), and one phosphate group. It is these molecules that, when joined together, make up the structural units of RNA and DNA. In DNA the nucleotides are adenine (A), thymine (T), cytosine (C) and guanine (G), but RNA uses uracil (U) in place of thymine.

oncogene A gene that has the potential to cause cancer.

pathogenesis The origin of a disease and the chain of events leading to that disease.

RNA Abbreviation for RiboNucleic Acid, which like DNA is also essential for life. Has the same structure as DNA, but one big difference is that the nucleotide thymine (T) is replaced by uracil (U). The sequence of nucleotides allows RNA to encode genetic information. All cellular organisms use messenger RNA (mRNA) to carry the genetic information that directs the synthesis of proteins.

siRNA Abbreviation for small interfering RNA, which is a class of double-stranded RNA molecules with 20-25 nucleotides in length.

tumor suppressor A tumor suppressor gene, or anti-oncogene, is a gene that protects a cell from one step on the path to cancer.

Chapter 3

Data

This chapter deals with the data set used throughout the thesis. The data is produced by a company named AROS Applied Biotechnology which specializes in miRNA extraction from most biological tissues and cells. It has been made accessible by Nicolai Schultz, who works as a surgeon at the Department of Surgical Gastroenterology and Transplantation Rigshospitalet, University of Copenhagen. Nicolai has also provided the clinical data associated with the patients involved.

The chapter consists of three sections. Section 3.1 deals with some background information concerning the experiment, and clarifies what the measured values for miRNA actually represent. Section 3.2 digs into the data, and looks at the variables available and how they distribute themselves, the so-called descriptive analysis. Section 3.3 digs even deeper and investigates the data further, in order to get an overview of how data behaves and reveal potential problems.

3.1 Background of miRNA measurements

The experimental process of miRNA extraction is not trivial, it includes a series of steps on both laboratorial and microbiological level, and in this section it will only be briefly described. Participants of this study have all submitted a blood sample and from this gotten the serum extracted. Serum is a fluid in the blood that is neither blood cells (white and red) or clotting factor (coagulation).

After a miRNA purification procedure that should ensure no proteins and other irrelevant molecular fragments remains, the samples are ready for being pipette onto so-called *TaqMan® array human microRNA A+B cards*. These cards are prefabricated from the company AppliedBiosystems™ and contain a total of 754 unique assays specific to human miRNAs. The A card focuses on the more highly characterized miRNAs, while the B card contains many of the more recently discovered miRNAs. These cards can be seen in Figure 3.1 [Applied-Biosystems 2010].

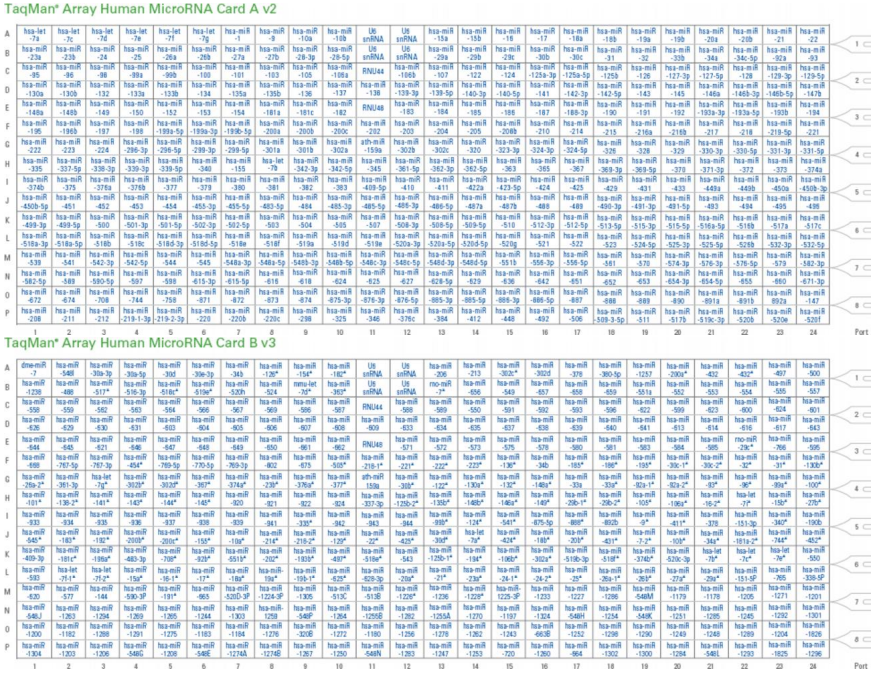

Figure 3.1: *TaqMan® array human microRNA A+B cards where each well contains individual miRNA reagent, provided by Applied-Biosystems [2010].*

As Figure 3.1 shows there are 16×24 wells per card and with two cards this gives a total of 768 wells. Now there were 754 different human miRNAs, so the remaining 14 spaces (7 pr. plate) are used for what is called endogenous controls. For every card of this particular kind, four candidate endogenous controls are selected. They are also called *housekeepers*, whose average can be used to normalize internal variation. One of these controls is quadrupled (sometimes referred to as the calibrator or reference sample), since it is essential when calculating the fold-change for relative expression analysis. The remaining three controls are replicated twice (one on each card). These concepts which will be elaborated in Section 4.1.3. Both A and B cards have been run for every single person in the population.

In order to determine the quantity of a specific miRNA in a certain sample, the previously mentioned TaqMan® system has been used. When the plates are put into the machine, a so-called *quantitative real time polymerase chain reaction* (qrt-PCR) is initiated. This process is used to amplify and simultaneously quantify a targeted DNA molecule, but in the context of targeting miRNA the qrt-PCR is combined with an initial reverse transcription. The real time refers to the possibility of observing the amplified gene material as the reaction progresses, i.e. for every cycle, opposed to the standard PCR, where the product of the reaction is only detected at the end.

The reactions are performed in a temperature block and in order to robustly detect gene expression from small amounts, such as miRNA, amplification of the gene transcript is necessary. Theoretically the targeted miRNA is doubled in each cycle, and to be able to measure this quantity, fluorescent light is added to the PCR mix. This fluorescent reporter is also amplified along with the transcript, and this can be seen as an amplification plot. An example is shown in Figure 3.2.

The amount of miRNA present in each well is determined by the number of cycles it takes to reach some threshold[2]. This quantity is called *cycle threshold* (C_t), and under the assumption of 100% amplification efficiency, the relationship between C_t values and PCR can simply be described as follows

1 PCR cycle = 1 increase in C_t value = twofold (2^1) of miRNA material
2 PCR cycles = 2 increase in C_t value = fourfold (2^2) of miRNA material
3 PCR cycles = 3 increase in C_t value = eightfold (2^3) of miRNA material
and so on.

[2]Strongly recommended to be decided by the manufacturer in order to ensure optimal threshold settings. This should (hopefully) result in 100% efficiency of amplification.

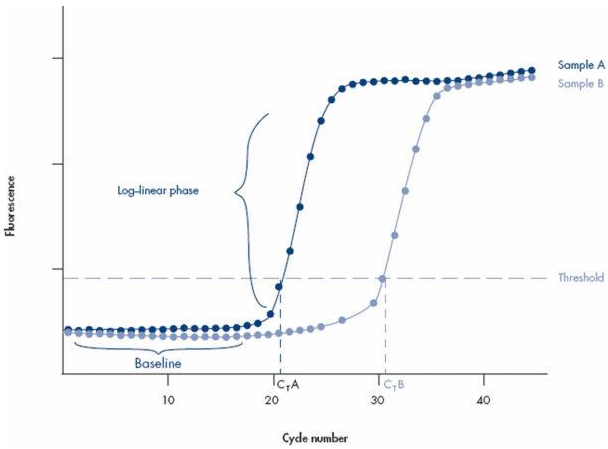

Figure 3.2: *Typical amplification plot for PCR, provided by QIAGEN.*

From the small example seen in Figure 3.2, it can be seen that sample A only needs about 22 cycles to reach its threshold, compared to the 31 sample B. This means that the concentration in sample A is much higher than B, about $2^(31-22) = 512$ times more, because lesser folds were needed before detection. It is possible that a concentration can be so low that the miRNA never reaches its threshold, which will usually express itself as "undetermined". Then the miRNA measurement is set as being missing, normally using a cut-off point of 40 cycles.

Keep in mind that C_t values themselves are not actual concentration quantities, but they are relative measurements. One makes the distinction between relative quantification and absolute quantification, which both are methods used to approximate the number of fold changes.

Absolute quantification gives the exact number of target molecules in a sample. Relative quantification compares the C_t values of ones target miRNA to another internal reference (such as an untreated control sample - the calibrator) or housekeeping genes from the same plate. This makes it possible to normalize for variation between different plates, hence making the plates comparable. A widely known comparative normalization procedure that tells you how many fold changes of amplicon occur, between cycles of the calibrator and target gene, can be calculated by the formula

$$\text{fold change} \quad = \quad 2^{-\Delta\Delta C_t} \tag{3.1}$$

where

$$\begin{aligned} \Delta\Delta C_t &= \Delta C_{t,\text{target}} - \Delta C_{t,\text{calibrator}} \\ \Delta C_{t,\text{target}} &= C_{t,\text{target}} - C_{t,\text{endogenous1}} \\ \Delta C_{t,\text{calibrator}} &= C_{t,\text{calibrator}} - C_{t,\text{endogenous2}}. \end{aligned}$$

It is important that the endogenous controls are picked, so they share similar properties such as stability and size as the target gene and calibrator. This is a good way of comparing C_t values of the cancer and control samples, however a drawback could be that it assumes that the target and reference amplification are equally efficient. This thesis will however not base the analyses on the fold change number and these types of normalization procedures, but instead use the raw C_t values as a starting point.

3.2 Description of clinical data

With the background of the clinical trial in place and a reasonable understanding of how the miRNAs are measured, it seems natural to move on to the data. The data consists of 226 persons whose serum was taken, and for every single one there was run an A and B card, resulting in 754 individual miRNA measurements (not including the endogenous controls). Furthermore a set of clinical information for each patient have been registered, e.g. sex, age, diagnosis etc. A summary of these variables is given in Table 3.1.

The cohort consists of patients treated at three medical departments; Rigshospitalet, Herlev Hospital and University of Heidelberg. Originally the clinical and miRNA data were found in two separate data sets, linked by the AROS number as the unique key. This number is important when following a certain subject, because it is given from the company's side, hence new measurements from the same person will also have this number. The patient number is more relevant to use in the context of the respective departments.

Variable name	Variable type	Variable explanation
AROS.A995.nr	integer	AROS number
Patient.nr	factor	Patient number
Age	integer	Age when included (yrs)
Sex	factor	Gender [1: male, 2: female]
Diagnosis	factor	Type of patient
Operation	factor	Operation
Cleansing.date	date	Date of purification
Operator	factor	Laboratory technician
Inclusion.date	date	Date of inclusion
Operation.date	date	Date of operation
Death.date	date	Date of death
Follow.up.date	date	Follow-up date
Date	date	Date of death (if occured), else follow-up date
Time	integer	Time from operation to event (days)
Status	factor	Event type [0: censored, 1: dead]
miR.uxc	numeric	miRNA1
⋮	⋮	⋮
miR.tae	numeric	miRNA754

Table 3.1: *Description of the variables in the serum data set.*

One of the most important clinical characteristic recorded, is the diagnosis of each subject. From this variable it is possible to see how many of the patients have an periampullary cancer and which are the healthy controls. The distribution of this variable can be viewed in Table 3.2.

Diagnosis	Name	n	%	Analysis grouping
0	Unknown	2	0.88	Not relevant
1	Pancreatic cancer	137	60.62	Pancreatic cancer
2	Ampullary cancer	4	1.77	Not relevant
3	Duodenal cancer	2	0.88	Not relevant
4	Common bile duct cancer	4	1.77	Not relevant
5	Serious cystadenoma	4	1.77	Not relevant
6	Solid tumor w.o. invasion	3	1.33	Not relevant
7	Chronic pancreatitis	20	8.85	Chronic pancreatitis
8	Neuroendocrine	1	0.44	Not relevant
444	Healthy control	49	21.68	Healthy subject
Analysis total		206	91.15	
Total		226	100.00	

Table 3.2: *Frequency table of the diagnosis variable. The types relevant for this thesis have been marked with a gray row color.*

Most of the samples in the cohort (n=137, 60.62%) have the most common form of pancreatic cancer (PDAC). The healthy controls represent the second largest group, however this group does not solely consists of the healthy subjects, it also includes the CP patients, since chronic pancreatitis per definition is not cancer. Together the CP+HS control group accounts for 30.53% of the

cohort. The remaining 8.85% samples, are unclassified and other periampullary cancer types, who are not relevant for the analysis in this thesis, so they are left out from this point on. This means that the original cohort is reduced to a total of 206 persons.

Besides the diagnosis variable, other useful information of the samples and their miRNA measurements are provided, among these are the gender, age (when included in the cohort) and operation status worth mentioning. The operator variable indicates which laboratory technician has purified which samples. It could be interesting in theory to see how variation between the different laboratory technicians influence the results, but it is inadequate. Only two names occur and a larger part of the samples have been purified by an unknown person, so it does not give much insight. The date of purification variable contains two unique dates, but it turns out that the cancer+chronic pancreatis patients have been purified on one day, and the healthy on another. This is problematic experimental planning, because it means that the effect of the purification is totally confound with the diagnosis. This issue will be discussed in much further detail in Section 3.3.1.

	Pancreatic cancer (n=137)	**Chronic pancreatitis (n=20)**	**Healthy subjects (n=49)**	**Total (n=206)**
Gender n(%)				
male	89 (65.0)	13 (65.0)	23 (46.9)	125 (60.7)
female	48 (35.0)	7 (35.0)	26 (53.1)	81 (39.3)
Age (yrs)				
mean	63.43	57.80	59.00	61.83
median	63.0	56.5	61.0	62.0
sd	9.74	10.04	7.12	9.45
range	31-86	42-85	41-66	31-86
Operation n(%)				
operated	96 (70.1)	0 (0.0)	0 (0.0)	96 (46.6)
inoperable	39 (28.5)	0 (0.0)	0 (0.0)	39 (18.9)
not relevant	2 (1.5)	20 (100.0)	49 (100.0)	71 (34.5)
Survival status n(%)				
death	60 (43.8)	0 (0.0)	0 (0.0)	60 (29.1)
censored	32 (23.4)	1 (5.0)	0 (0.0)	33 (16.0)
N/As	45 (32.8)	19 (95.0)	49 (100.0)	113 (54.9)
Survival time (days)				
mean				624.99
median				569.00
sd				419.35
range				6-1881

Table 3.3: *Summary of the variables gender, age, operation, status and time, divided into three analysis subgroups along with the total.*

In Table 3.3, appropriate descriptive statistics for selected variables are presented, stratified between the three main groups (PC,CP,HS) along with a total. Overall there are more men than women ($\approx 60/40$) in the cohort, only in the

healthy subjects group there is a slight overweight of women, which represents the general population fairly well. On average the patients are 61.83 years old, but in the PC group the mean age is closer to 65, which is in agreement with the literature. The patient's age vary from 31 years to 86 years. The operation status is only relevant for the patients with pancreatic cancer, since a surgical procedure is not relevant for CP cases and of course healthy subjects. As Table 3.3 shows almost 70% have been operated, but still a relatively large part was classified as inoperable, presumably because the tumor already was at an so advanced stage.

The status and time variables are useful in relation to prognosis after operation, and are constructed from the three date variables; operation, death and follow-up. The status indicates whether a patient has experienced an event, in this case death (status=1), or the end of follow-up, i.e. is *censored* (status=0). The censored patients will contribute equally to the analysis as the deceased patients, until censoring. The time variable is the time in days from operation to either death or end of follow-up. If the date of operation or both the death and follow-up date are missing, then the time will be defined as N/A. This is the case for about 32.8% of the 137 cancer patients, while 43.8% have died and 23.4% are censored. The mean survival time from operation is 624.99 days, or equivalently 1.71 years.

The basis of the analysis lies is miRNA measurements, and since this is still a new concept to science, there is no beforehand knowledge of how the statistical analysis should be performed. No one knows the truth to how the miRNA are correlated with each other, or if they can be regarded as independent etc. So it seems necessary to explore these miRNAs in greater depth to get a better understanding of how they are distributed before beginning the analysis. This is done in Section 3.3.

3.3 Description of miRNA data

The miRNAs are the center of this thesis, hence the key covariates. The hypothesis is that some of them are shown to be statistically significant regulators in relation to incidence and prognosis of pancreatic cancer. One of the problems, when working with miRNA data, is the classical challenge of having more parameters than observations ($p > n$, here the case is even $p >> n$). The reason is that the system of equations defining the regression model in classical regression analysis is underdetermined. Put in more mathematically terms, there will be more unknowns than equations available, making it impossible to find an

unique solution. So it is an obstacle that should be dealt with in some way, but since problem with high dimensional data is not unfamiliar, methods designed to cope with this problem exists.

For these miRNA data there seems to be many that have undetermined C_t values (or N/As), and the cause could be technical or simply that the concentration is just too low to detect. Hence, it could be interesting to look at how the N/A percentage of each miRNA are distributed. This is done in Figure 3.3.

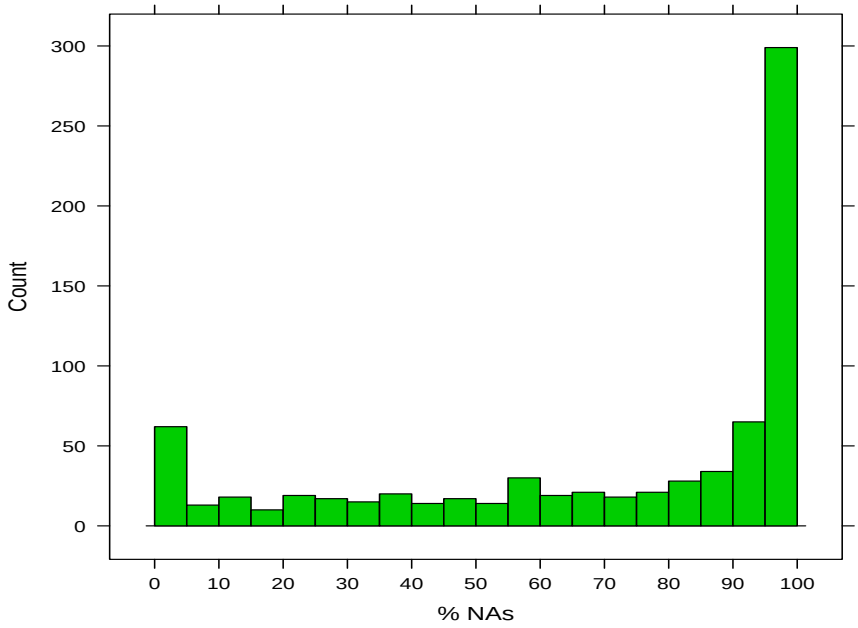

Figure 3.3: *Histogram showing the percentage of N/As for all miRNAs.*

This plot is very promising, because it indicates that many of the miRNAs have a high percentage missing, so they can most likely be discarded without losing to much information, and hence reduce dimensionality significantly. The question is however, where should the limit for exclusion be? No one knows the correct answer to this, so the choice have been based on statistical intuition and reasoning. Since almost 300 miRNAs have 100% missing measurements, it is fairly obvious that these should be removed from the analyses. This leaves ≈ 450 covariates, which still are too many parameters, so the criteria of exclusion was defined as miRNAs having more than 20 N/As, corresponding to around 10% measurements missing. The missing measurements must occur when miRNAs

have negligible presence, only a smaller proportion of the missing can be credited small machine fluctuations. This is the motivation behind the choice of limit, when a certain miRNA have more than 10% undetermined it seems reasonable to think that the concentration is sufficiently low to not being an influential covariate. For the serum data, 75 miRNAs satisfied this restriction, hence were eligible as candidate covariates in the analyses.

The assumption is now that the true miRNA predictors, are to be found in this subset, so the N/A limit is up for discussion. Even though this is a fairly strong assumption, it is a nice and easy way to work around the high dimension problem, and hopefully without losing to much valuable information. However, instead of getting rid of information that could be potentially valuable, adding information could be another way to address the N/A problem.

The concept of substituting missing values with some appropriate values, is called *imputation*. Several strategies exists on how to develop an appropriate imputation algorithm that fits the data, a large part lies in the assumption on why the data elements are missing. In this case the missing values would be classified as *informative missing* (IM) also called nonignorable nonresponse, because measurements are more likely to be missing when the concentration of miRNA is lower. The missingness pattern is systematical and the most difficult type of missing data to handle (see Section 4.1.2 for a wider definition of missing assumptions). In many cases there is no fix for IM data, approaches like *multiple imputation*[3] or *single conditional mean imputation*, which uses mean, median or another sensible value as a substitute of the missing values (*"best guesses"*). Both uses assumptions not quite meet here, although the latter could be applied with reasonable approximation, since the C_t scale for miRNA data is locked at $[0; 40]$. It is known that missing values present themselves when a high number of cycles still have not reached the threshold, so N/As could be replaced by some value between e.g. 35-40. On the other hand, this could result in distributions with heavy weight coming from high values, which is not desirable. Anyhow, these types of imputation methods are not the focus of this thesis, but still worth mentioning as an alternative approach [Harrell 2001, pp. 41-50].

In Figure 3.4 the average C_t level for every patient in the cohort is depicted. This is the mean of all the miRNA measurements w.r.t. every person, where the N/As have been excluded before calculation. Furthermore the patients have been ordered by their runorder, i.e. patient number one's plates have been run first and so on. In general there can be run ≈ 6 plates per day, so this trial has probably taken around a month to run.

[3] Uses random draws from the conditional distribution of the target variable given other variables, e.g. *bootstrapping* - a general purpose technique for obtaining statistical estimates by repeatedly simulating a sample of size n from some empirical distribution of the observed data, and then assessing how the computed statistic behaves over a number of repetitions.

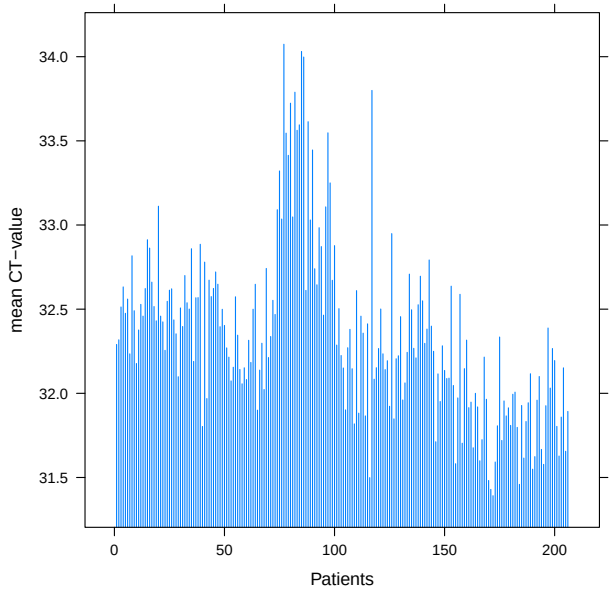

Figure 3.4: *Average C_t level for the 206 persons eligible for analyses, arranged by runorder.*

The general belief is that C_t values above 35 are not trustworthy, because after that many cycles it is most likely just some unspecific random fragments that is being recorded instead. Usually these unreliable miRNAs are removed from the analyses to strengthen the trust in results. The C_t grand average is 32.36, which implies that the measurements are generally in the high end - close to the cut-off point. This might not mean anything, but it is worth highlighting as an uncertainty factor.

Besides having a very high C_t level in general, there also seems to be a problem in the way the average is shifting. Theoretically the mean for all the patients would be expected to lie on some general level, with minor fluctuations due to measurement errors. But for these data there seems to exist several mean "jumps". The first about 75 patients clearly have a lower level than patients no. $\approx 76 - 100$, who have very high miRNA averages in comparison, but then the level drops again to a new low for the remaining patients. The reason for this trend is unknown, because as explained earlier an A+B card was run for every patient on different days, so the variation between plates, operator, runorder etc. is confounded, and hence for these data not possible to quantify.

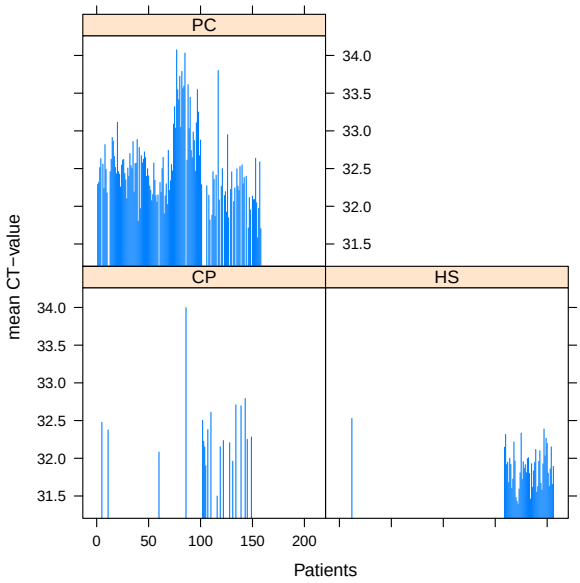

Figure 3.5: *Average C_t level divided into the three main groups; PC, CP and HS, arranged by runorder.*

To make matters even worse, in Figure 3.5 the miRNA average have been stratified between the three main groups; patients with pancreatic cancer, chronic pancreatitis and healthy subjects. This figure reveals that the PC patients generally are the ones having higher mean C_t compared to the healthy subjects. If this difference was caused by a true difference, it would be possible to distinguish between the two groups solely from this plot. This is however highly unlikely to be true, given past experience have shown us that nature often works in a far more complex way. Furthermore, if this were true, then chances that someone else already have discovered this are large. So a lot indicates that many of the irregularities present in the data are artificial variation, caused by e.g. technical errors. Unfortunately since this specific trial cannot be redone, the data needs to normalized, such that different patients are comparable in the analyses. However this is a statistical challenge for miRNA data and something still under discussion, because at present time there is still no evident normalization method that can be classified as the best and most robust. There are numerous normalization methods available though and this thesis will look into the theory behind some of them in Section 4.1, and how they perform individually on the data set in Chapter 6. Before moving on to the methodology applied, Section 3.3.1 goes into more detail with the issue of confounded factor sources,

and proposes how a more balanced design of experiment (DOE) can optimize these kind of data in the future.

3.3.1 Design of experiment

These types of miRNA trials have the potential to uncover some valuable knowledge regarding incidence and prognosis of cancer. The expenses of each trial however are very high and usually cost millions of kroner, making it crucial to obtain maximum information possible. Statistical methods in the field of design of experiments are therefore highly relevant. To get an idea of how to reach this goal, it is important to first define the three basic principles of experimental design; *randomization*, *replication* and *blocking*.

Randomization is according to Montgomery. [2008, pp. 12-13] the cornerstone underlying the use of statistical methods in experimental design. Most statistical methods use the assumption of errors being independently distributed random variables, and randomization usually makes this valid. Randomization refers to both the allocation of the experimental material and the randomly determined order in which the individual trials are to be run. By doing the randomization properly, unwanted extraneous effects present are averaged out, hence bias that has not been accounted for in the experimental design will be reduced.

Replications are independent repetitions of a certain factor combination and serve two important purposes. First, it allows the experimenter to obtain an estimate of the experimental error. Secondly, it supplies the experimenter a more precise estimate of some parameter which further strengthens the experiment's reliability and validity. Replicates are not to be confused with *repeated measurements*, where observations has been made on the same factor more than once, usually involving measurements taken at different time points.

Blocking is a design technique used to improve precision with which comparisons among factors of interest are made. The general idea is an arrangement of experimental units into groups (blocks), consisting of units that are similar to one another, reducing irrelevant sources of variation between units. This form of variability affecting the results, which blocking can systematically eliminate, is denoted *nuisance factor*. However, this requires that the nuisance source of variability is known and controllable.

These three concepts are to be kept in mind when planning an experiment, but it is no secret that for the data provided, they have been nowhere near the considerations. Even though the data given is the only available working

material it is still relevant, for future similar experiments to pinpoint certain improvements. Unfortunately for this type of experiments replication is not an existing dimension, because each subject will only get one A+B card on which only some controls are replicated.

Description	Levels	Associated with
Case control status	PC/CP/HS	Response
Survival time for cases	missing/below/above median	Response
Age for cases and controls	below/above median	Response
Gender for cases and controls	male/female	Response
Plate	$1 \dots n$	Reproducibility
Preparation/purification date	$1 \dots d_1$	Reproducibility
Analysis/running date	$1 \dots d_2$	Reproducibility

Table 3.4: *Factors suspected of influencing results.*

In Table 3.4 some of the factors assumed to be relevant for the experiment are listed. The table provides factor description, assumed number of levels for each factor as well as its role on the measurements. The case/control status, survival time, age and gender are called *fixed effects*, because the levels of these factors are of specific interest. Whereas the remaining factors are thought to be *random effects* influencing the precision or *reproducibility* of the experiment. Their factor levels are chosen at random from a larger population of possible levels, but where the objective is to draw conclusions about the entire population of levels [Montgomery. 2008, p. 505].

Reproducibility is defined as *precision under conditions where test results are obtained with the same method on identical test items by different operators using different equipment.* However since there are no replicates in this experiment and presumably only one operator, machine and laboratory at disposal, the variance components that reproducibility consists of cannot be estimated [Dehlendorff and Andersen 2011].

Despite each miRNA only has one measurement per patient, this still leaves randomization and blocking to be considered, in order to average nuisance factors out. The past experiment was done in a inappropriately manner, where the biggest issue was the purification order and the order in which the samples were analyzed. The major issue in general was that when they purified on the two occasions, the samples from cancer and chronic pancreatitis patients were done the first day and the healthy controls on another. Moreover the runorder was

as imbalanced as possible, because the overall sample order was PC→CP→HS. The result is clear from Figures 3.4 and 3.5, the sample mean C_t level for cases and controls is considerable different, which is most likely caused by the purification factor. However because of the poor designplan, this nuisance cannot be extracted from the data, the diagnosis of patients is confounded with both the purification and analysis date.

As mentioned before, blocking can aid in the prevention of nuisance factors having an effect. Now, for an experiment with n samples, each sample belongs to some combination of the four identified fixed effects. The plates are by construction totally confounded with patients and the plate-to-plate variation cannot be estimated free of patient-to-patient variation. This fact does not change unless replicates are made. The general idea is to treat the purification and analysis date variables as blocks and then distribute all samples out between these blocks, in order to remove unwanted variation as much as possible. This balancing out of the four factors could be performed in a *two-staged blocking procedure*, by following four steps.

1. Divide the purification days into d_1 blocks and assign each sample to a block. Allocation should be done in a balanced manner of the four factors, such that each block contains same proportion of each factor level, i.e. equally many males/females, cases/controls etc.
2. Randomize order of sample purification within each day.
3. Divide the analysis days into d_2 blocks and assign each sample to a block. Once again, the allocation should be balanced, but now treat the assigned purification day as one more factor to consider, beside the factors accounted for in step one.
4. Randomize analysis order for each day.

The above described design makes sure that the blocking effect can be taken out of the measurements. The importance of blocking cannot be stressed enough, but an equally important part of the design plan is the randomization of samples within each block. Both when treating purification as a blocking variable and the order in which the samples should be analyzed.

In closing, it has been possible to put the principles of DOE to the test in a new miRNA experiment. A similar A+B card experiment have been performed according to the two-staged blocking procedure just described, and in Figure 3.6 the results are visualized by stratifying on the diagnosis. The "JJ" refers to personal control samples taken from doctor Julia S. Johansen, who is in charge of the experiment[4].

[4]Unfortunately the data was delivered too close to the deadline of this thesis, so time did not allow any form of analyses on these data. Christian Dehlendorff should be credited for providing Figure 3.6.

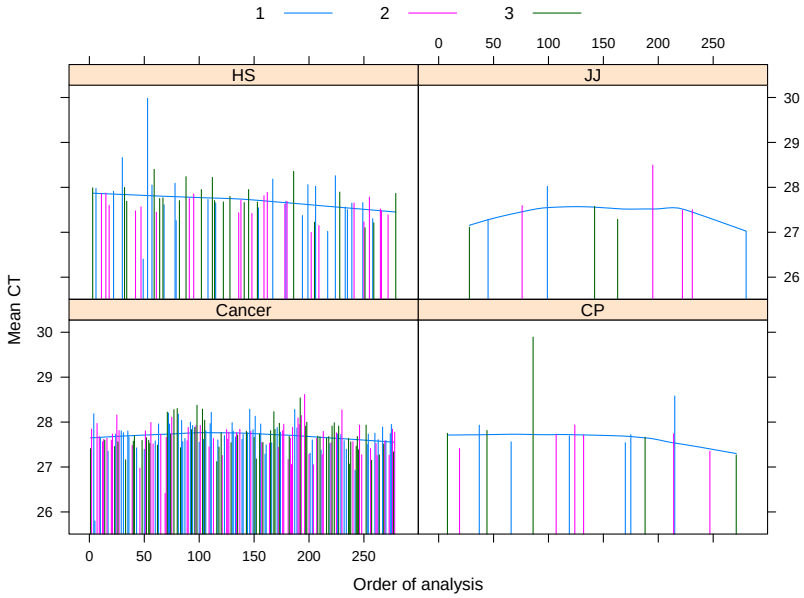

Figure 3.6: *Average C_t level arranged by order of analysis, stratified by diagnosis, for a new experiment using the described two-staged blocking procedure. A color denotes the designated purification date of a sample.*

The difference from the serum data that is without any form of design plan, is quite clear. The proper design plan ensures that the samples are analyzed in a much more balanced way concerning diagnosis and purification date, which readily can be derived from the width and color of the sample spectrum. First of all, the average C_t measurements in general are down to a more acceptable region, varying between 27 and 29 roughly (this is not a direct result of DOE though). Furthermore, the average C_t level of the samples in the various groups, are all on the same level, only with very few outliers (typically samples with many N/As). Of course there is minor variability, but there is no sudden unexplainable jumps in the mean, hence the samples are comparable. The results derived from this new experiment, further strengthens the claim of mean C_t level differences in Figure 3.5 between cancer and healthy subjects of the serum data, is artificially caused. In Figure 3.6 there is certainly an improvement of the data quality and if the miRNA measurements always looked like this, DOE would be redundant and the importance of normalization would be lesser. However, since the serum scenario is possible, the experiments should always strive to apply the principles of randomization, replication and blocking in a reasonable way.

Chapter 4

Methodology

This chapter deals with the statistical methodology applied throughout the thesis. However before describing things from a mathematical perspective, Figure 4.1 gives an overview of the main parts of the statistical analyses, and how these different parts are connected to each other.

Several important issues with the data was discovered in the previous chapter and it was highlighted how crucial it is to make well planned experimental designs. Since the reality for these data is something else, there exists a normalization issue. No standardized way of normalizing miRNA data has yet been establish, so five different methods have been used in this study. The five different normalization methods are described initially in Section 4.1. For one specific normalization method, called the *rank method*, a small simulation study has been performed. This is to validate the hypothesis that ranking of data produces more reliable results, and is more robust concerning these miRNA mean shifts, compared to working with the raw C_t values. The result of the simulation study is described later in Chapter 5.

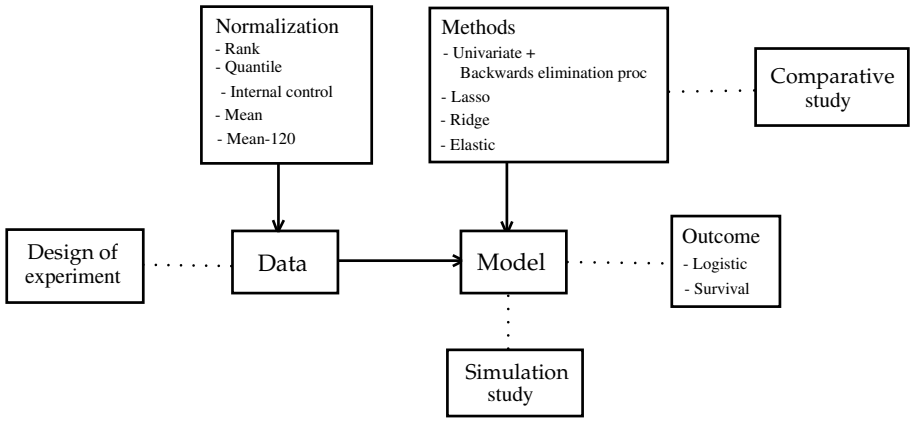

Figure 4.1: *Overview of this thesis' area of analyses, in form of a flow diagram.*

The statistical model depends on the type of response variable. *Incidence* deals with a binary response, i.e. whether a person is diagnosed with cancer or not. This is typically analyzed by *logistic regression*. In the study of incidence, the goal is to identify which miRNAs can be considered as predictors of cancer/healthy subjects, however a more formal introduction to incidence is provided in Section 4.2.

When dealing with *prognosis* another type of outcome is of interest; survival time. In this context survival time refers to the time from operation to death or end of follow-up. One of the analytical problems of survival analysis, is called censoring. This is essentially when we have some information about an individual's survival time, but do not know the exact survival time. A range of methods can be used to analyze survival data, but this thesis is restricted to only deal with the *Cox proportional hazards model*. It is probably the most used model within survival analysis. The fundamental concepts will be introduced in Section 4.3.

Furthermore, this thesis will also make use of the so-called *shrinkage methods*, methods that penalizes on the coefficients in various ways depending on the method used. These methods are very useful in $p > n$ situations, because they have the ability to shrink regressors with negligible influence (and sometimes eliminate, depending on the penalty term). Moreover, shrinkage methods also deals with *multicollinearity* among the regressors, which from a biological point of view makes sense to expect from miRNA data. These different modeling approaches have laid the ground for the comparative study, which tries to examine if either one have a better predictive performance. Moreover, it is very inter-

esting to see which miRNAs are identified by each method as good significant predictors, and which method - if possible - can be classified as most useful in a practical sense. The shrinkage methods can be applied regardless of the type of outcome variable, meaning that they can be used in relation to both incidence and prognosis. An introduction of the shrinkage methods considered in this thesis can be found in Section 4.4.

The methodology chapter ends with an introduction to *cross-validation* in Section 4.5, a useful method for finding optimal tuning parameters of the shrinkage methods.

4.1 Normalization methods

Normalization can be defined as *the process of isolating statistical error in repeated measured data, sometimes based on a property.* For the data available, the natural variation between patients is confounded with other sources of variation, such as the plates, machine, purification etc. So isolating and extracting the natural statistical noise solely, is simply not possible in this case. Normalization can however still be used to somehow even out the differences in mean C_t level for each patient, and by that making the patients more comparable in the analyses. The five different normalization methods considered are ranking of data, quantile normalization, use of an internal control, mean normalization and mean of the 120 most expressed miRNAs.

4.1.1 Rank normalization

The principle of rank normalization is to replace the actual C_t values with their ranks. This way, instead of working with the actual measurements, the ranks are used. Within each patient the C_t values for each miRNA are ordered ascendingly, and thus the first element is given rank 1, the second element rank 2 and so forth. This means that the miRNA with rank 1 is the most expressed, because it has the lowest cycling threshold or equivalently the highest concentration of miRNA material.

The method is fairly simple, but this is not the whole story, because there is also a need for handling ties. C_t measurement ties between two miRNA measurements is not a frequent event, but of course that does not change the fact that the situation needs to be addressed. When multiple measurements obtain

the same value, the median of the ranks involved is assigned to all of them. In the case of an even number of elements being alike, the mean of the two middle values is used. This is best illustrated with a small example, which is given in Table 4.1.

miRNA	**miR1**	**miR2**	**miR3**	**miR4**	**miR5**	**miR6**	**miR7**	**miR8**	**miR9**
C_t	21	N/A	32	34	N/A	23	N/A	25	32
Rank	1	8	4.5	6	8	2	8	3	4.5

Table 4.1: *Ranking example.*

In the example miR3 and miR9 both have $C_t = 32$, and since they are ranked 4 and 5, rank $\frac{4+5}{2} = 4.5$ is given to both. What also can be seen from this toy example, is how missing values are ranked. All miRNAs without C_t value are put at the end of the ordered list, and treated as having the same value. Three miRNAs are missing here (miR2, miR5, miR7), with the order number 7, 8 and 9, hence the median is 8 and assigned as the rank for all of them.

Missing values are an inevitable fact that needs to be handled for many types of data, miRNA data is no exception. There are several ways to get around the problem, each with its own pros and cons. The ranking algorithm implemented in this thesis, places all the N/A observations at the end and gives them a shared rank, just as the small example illustrates. However when a patient have an overweight of missing values, the disadvantage of this procedure reveals itself. The ranks will distribute themselves as an incrementing diagonal when measurements are present, but all the missing values in the end will lie in a single point. This clot will have some distance from the actual measurements, because the method chooses the median rank of all the N/As. So there is a lot of weight ascribed to a single rank, which could potentially influence the results. But since all miRNA with a high N/A percentage have been sorted out, the impact should be minimal here.

There is another disadvantage that can be added to the rank method. When the actual measurements are substituted with ranks, the effects are not easily interpreted, because now it is a pattern being analyzed instead of the true values. One increase in C_t value is not the same as one increase in rank, the latter is harder to explain the meaning of. However, using ranked miRNA data, each patient becomes its own control. This is the main motivation behind using this form of normalization, because the general C_t level for each patient is ignored. A simulation study in Chapter 5 will go more into detail concerning the benefits of using ranks compared to raw values.

4.1.2 Quantile normalization

The goal of quantile normalization is described by Bolstad et al. [2003] as *making the distribution of probe intensities for each array, in a set of arrays, the same.* Probe intensities in this context would be the miRNA C_t values, and the array refers to a sample/patient. The general idea comes from the quantile–quantile plot, which can be used for comparing the distribution of two data vectors. The plot will be a straight diagonal line if the distributions are the same. This concept have been extended to n dimensions, so that if all n data vectors have the same distribution, then plotting the quantiles in n dimensions gives a straight line along the line given by the unit vector $\left(\frac{1}{\sqrt{n}}, \cdots, \frac{1}{\sqrt{n}}\right)$. The method is described in more detail in the following.

Consider that the C_t values are contained in the matrix $\mathbf{X}$ given as

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix} \tag{4.1}$$

where p is the number of miRNAs and n is the number of samples. The original ordering of each column is given in the matrix $\mathbf{O}$. First step is then to sort each column such that $\tilde{x}_{1i} \leq \tilde{x}_{2i} \leq \cdots \leq \tilde{x}_{pi}$, where $\tilde{x}_{ji}$ is the j^{th} value in the sorted version of column i. Combining the n sorted columns gives $\tilde{\mathbf{X}}$ and the new index order within each column is given by matrix $\mathbf{D}$. Second step is to replace each row with its arithmetic mean, forming the matrix $\bar{\tilde{\mathbf{X}}}$. Finally the quantile normalized matrix $\mathbf{Q}$ is found by arranging $\bar{\tilde{\mathbf{X}}}$ back according to $\mathbf{O}$.

The quantile normalization algorithm can probably best be illustrated with a small example, consider a matrix with arbitrary C_t values

$$\mathbf{X} = \begin{bmatrix} 28 & 35 & 22 \\ 27 & 32 & 28 \\ 30 & 31 & 21 \end{bmatrix}. \tag{4.2}$$

First step is to sort within each column to obtain matrix $\tilde{\mathbf{X}}$ and $\mathbf{D}$

$$\tilde{\mathbf{X}} = \begin{bmatrix} 27 & 31 & 21 \\ 28 & 32 & 22 \\ 30 & 35 & 28 \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} 2 & 3 & 3 \\ 1 & 2 & 1 \\ 3 & 1 & 2 \end{bmatrix}. \tag{4.3}$$

Secondly the mean is taken over each row of $\tilde{\mathbf{X}}$

$$\bar{\tilde{\mathbf{X}}} = \begin{bmatrix} 79/3 & 79/3 & 79/3 \\ 82/3 & 82/3 & 82/3 \\ 93/3 & 93/3 & 93/3 \end{bmatrix}. \tag{4.4}$$

And finally by rearranging back within each column to the original order, the quantile normalized matrix is obtained

$$\mathbf{Q} = \begin{bmatrix} 82/3 & 93/3 & 82/3 \\ 79/3 & 82/3 & 93/3 \\ 93/3 & 79/3 & 79/3 \end{bmatrix}. \tag{4.5}$$

This procedure ensures that all the data vectors will lie on a straight diagonal line in an n dimensional quantile–quantile plot, i.e. the C_t distribution of all the miRNAs are made the same and by that comparability is archived. The method have been implemented in the `R` package `preprocessCore` from `Bioconducter` by Bolstad [2010], where the normalization can be done using the `normalize.quantiles` command. The algorithm ensures that tied values in a given column are also tied after normalization. So if a column contains some tied measurements, then for these ties an average of their values in the $\mathbf{Q}$ matrix are returned instead. This is similar to the rank method.

Probably the largest drawback of using this quantile normalization implementation, is the handling of missing values. From the documentation it can be seen that the function handle N/As using the assumption that the values are *missing at random* (MAR). Data are said to be MAR if *the probability of the observed missingness pattern, does not depend on unobserved data.*

To quote Diggle et al. [2002, p. 283], who explains this concept more generally.

> Let a complete set Y^* be partitioned into $Y^* = (Y^{(o)}, Y^{(m)})$, where $Y^{(o)}$ is a set containing the measurements actually obtained and $Y^{(m)}$ the set of measurements which would have been available, had they not been missing. R denotes a set of indicator random variables, indexing which elements fall into $Y^{(o)}$ and $Y^{(m)}$, respectively. Now, a probability model for the missing value mechanism defines the probability distribution of R conditional on $Y^* = (Y^{(o)}, Y^{(m)})$. The missing value mechanism is usually classified into
>
> - *completely random* (MCAR) if R is independent of both $Y^{(o)}$ and $Y^{(m)}$
> - *random* (MAR) if R is independent of $Y^{(m)}$
> - *informative* (IM) if R is dependent on $Y^{(m)}$.

The key point here is that a missing observation only depends on the observed data, which is a strong assumption. Unfortunately this does not quite hold for the data of interest, because theoretically a C_t value is missing if the concentration of miRNA is imperceptible small or absent (not detected after > 40 cycles). It would probably be more appropriate to assume the data being informative, so this is something that needs to be stressed in the analyses when applying the `preprocessCore` implementation of the quantile normalization algorithm.

To explain how the missing values are handled practically, a minor example is needed. Once again, assume we have a matrix of arbitrary C_t values, this time with a missing measurement

$$\mathbf{X} = \begin{bmatrix} 28 & 35 \\ 27 & 32 \\ 30 & 31 \\ 25 & \text{N/A} \\ 29 & 30 \end{bmatrix}. \tag{4.6}$$

Now what the function does when ordering the columns, it places the N/A observation first. The sorted matrix evaluates to

$$\tilde{\mathbf{X}} = \begin{bmatrix} 25 & \text{N/A} \\ 27 & 30 \\ 28 & 31 \\ 29 & 32 \\ 30 & 35 \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} 4 & 4 \\ 2 & 5 \\ 1 & 3 \\ 5 & 2 \\ 3 & 1 \end{bmatrix}. \tag{4.7}$$

The deviation from the regular procedure is that the mean for the first row cannot be calculated because of the missing value, but for the remaining rows the operation can be performed without problems. The second column (containing the missing value) can arrange the mean values according to the original order, but the first column is instead replaced with the quantiles of the mean distribution. The distribution is in this example (28.5,29.5,30.5,32.5) and since there are four elements, the five quantiles becomes $(Q_0, Q_{25}, Q_{50}, Q_{75}, Q_{100}) = (28.50, 29.25, 30.00, 31.00, 32.50)$. In conclusion the matrices $\bar{\bar{\mathbf{X}}}$ and $\mathbf{Q}$ becomes

$$\bar{\bar{\mathbf{X}}} = \begin{bmatrix} 28.50 & \text{N/A} \\ 29.25 & 28.50 \\ 30.00 & 29.50 \\ 31.00 & 30.50 \\ 32.50 & 32.50 \end{bmatrix} \quad \text{and} \quad \mathbf{Q} = \begin{bmatrix} 30.00 & 32.50 \\ 29.25 & 30.50 \\ 32.50 & 29.50 \\ 28.50 & \text{N/A} \\ 31.00 & 28.50 \end{bmatrix}. \tag{4.8}$$

Opposed to the rank method, the quantile normalization places the N/As first, which for miRNA data is not reasonable. Then the most expressed miRNAs will loose its leverage, which of course is bad. One way to manipulate the

implementation a little and get around this problem, is to negate all the C_t values before making the quantile normalization (and negate back afterwards). This trick results in a lack of means for the high C_t values instead, which is much more plausible from a practical perspective.

4.1.3 Internal control normalization

On miRNA A+B cards there are placed different endogenous controls, and the method of normalizing w.r.t. internal controls is straight forward. All patient samples have had the same controls measured, so the idea is to subtract some average of the controls from all the human miRNA measurements. Let $\mathbf{X}$ denote the matrix with C_t measurements unique p human miRNAs

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{4.9}$$

and let a column vector containing row-wise averages of some internal controls be denoted $\boldsymbol{c}$, given as

$$\boldsymbol{c} = \begin{bmatrix} \bar{c}_{1\cdot,ctrl} \\ \bar{c}_{2\cdot,ctrl} \\ \vdots \\ \bar{c}_{n\cdot,ctrl} \end{bmatrix}. \tag{4.10}$$

If there are some missing values present when these means are calculated, they will simply be removed from the calculations. The internal control normalization is found by

$$\mathbf{C}_{\cdot j} = \mathbf{X}_{\cdot j} - \boldsymbol{c}, \quad \text{where } j = 1, 2, \ldots, p. \tag{4.11}$$

The $\mathbf{C}$ is the internal control normalized matrix and the mathematics behind obtaining this result is fairly trivial. It can be an advantage to use this normalization, because the idea is that the plate variation is removed, assuming the internal controls are stabile. It is an essentials part of this method, to choose the internal controls correctly. Remember that all in all 14 of the 768 C_t measurements are from endogenous controls, and how they are divided onto the A and B cards can be seen in Table 4.2.

	Endogenous control	Alias	Rep
Card A			
	MammU6-4395470	calibrator/reference	4
	RNU44-4373384		1
	RNU48-4373383		1
	ath-miR159a-4373390	spike-in	1
Card B			
	U6-snRNA-001973	calibrator/reference	4
	RNU44-001094		1
	RNU48-001006		1
	ath-miR159a-000338	spike-in	1

Table 4.2: *Shows how the endogenous controls are divided out among the A+B card and the number of replicates.*

Under further examination of the various controls in the data, it seemed like a natural choice to base the mean vector $\boldsymbol{c}$ upon the U6 small non-coding RNA. There were several reasons for this

1. four replicates on each plate gives more reliable values
2. each replicate of RNU44 and RNU48 had an N/A percentage $> 70\%$
3. spike-in samples are traditionally not used for normalization.

Furthermore, no distinction between a control coming from an either A or B card is made when calculating the row-wise average, because they should be somewhat similar. A major downside of this method is the risk of the control measurements being contaminated in some way, i.e. unstable, because these are the foundation of the whole normalization. Either by not having comparable biological properties or simply some mistake during the qrt-PCR process that result in missing/erroneous values. The idea behind this method is to remove the plate variation on the basis of the same control placed on each plate, because the control measurement should be the same across plates. Unfortunately this internal control cannot tell anything about the variation coming from the purification process, which was previously identified as a nuisance factor. Thus, if the variation coming from the purification is significant, this becomes a problem.

4.1.4 Mean normalization

The fourth normalization method resembles the previous method, the difference is that instead of some internal controls, the mean of all miRNAs is used to

align the samples around zero. Again, consider the matrix containing the C_t measurements $\mathbf{X}$ defined as in Equation (4.9). Now define a vector with the means w.r.t. rows of $\mathbf{X}$ as

$$\bar{\boldsymbol{u}} = \begin{bmatrix} \bar{u}_{1\cdot} \\ \bar{u}_{2\cdot} \\ \vdots \\ \bar{u}_{n\cdot} \end{bmatrix}. \tag{4.12}$$

The mean normalized matrix $\mathbf{U}$ is then found by

$$\mathbf{U}_{\cdot j} = \mathbf{X}_{\cdot j} - \bar{\boldsymbol{u}}, \quad \text{where } j = 1, 2, \ldots, p. \tag{4.13}$$

The mean normalization is often a first choice, because it is a well-known way to make the samples comparable easily, and by that get an idea of the true effects in the data. It is assumed that all the variation removed is coming from the plates, operator, purification etc. which cannot be separated from each other, and they influence all miRNAs for the same person in the same way. The most important assumption, is that the variation is not coming from the type of sample (cancer/healthy).

4.1.5 Mean-120 normalization

The last method resembles the mean normalization a lot, only difference is in the way the $\bar{\boldsymbol{u}}$ vector is calculated. Instead of using all p miRNAs, only the 120 most expressed will be considered here. Remember that lower C_t value, corresponds to higher concentration and hence higher miRNA expression. Once again the matrix with the C_t measurements is defined according to Equation (4.9). First calculate the average C_t value for each miRNA (column means) and pick the 120 lowest. These miRNA will form the basis of the row-wise mean vector, such that

$$\bar{\boldsymbol{u}}_{120} = \begin{bmatrix} \bar{u}_{1\cdot,120} \\ \bar{u}_{2\cdot,120} \\ \vdots \\ \bar{u}_{n\cdot,120} \end{bmatrix}. \tag{4.14}$$

These values are then subtracted from the C_t values in the following way

$$\mathbf{U}_{\cdot j,120} = \mathbf{X}_{\cdot j} - \bar{\boldsymbol{u}}_{120}, \quad \text{where } j = 1, 2, \ldots, p. \tag{4.15}$$

Motivated by the hypothesis that it is only a handful of miRNA which are of significant importance, the mean-120 method does not consider all the variation

of the miRNAs, as opposed to the mean normalization. Since the truth is unknown, there is no telling of which method should be preferred. It could be that the 121^{th} miRNA had some effect of importance, but commonly these miRNA with high C_t values are regarded to be dominated by randomness, so it makes sense to exclude them from the normalization process.

This concludes the description of the various normalization methods used, and as the reader maybe have noticed there are individual pros and cons. The rank and quantile method are to a certain degree comparable, where the same goes for the internal control, mean and mean-120 method. The latter share the property that they all subtract some term from the C_t matrix, a term specific to each method. In order to get an overview, Table 4.3 gives a boiled down version of the motivation behind the methods, along with the assumptions associated with them.

Method	Idea/Purpose	Assumptions
Rank	Uses patterns instead of actual C_t values. Each patient becomes its own control.	Estimates based on ranking pattern.
Quantile	Distribution of each miRNA the same.	Assumes miRNAs have same quantiles. Assumes MAR. The way of weighing N/As is unreasonable in miRNA context.
Internal control + **Mean** + **Mean-120**	Subtract variation from each patient's C_t values based on some property.	Assumes all miRNAs on same plate are affected in the same way, i.e. by a constant shift.

Table 4.3: *Sums up the five normalization methods and highlight their idea and general purpose plus assumptions.*

The general advantage for all the methods is that they handle missing values, each in its own way. Though from miRNA perspective, the handling of missing values from the quantile implementation is criticizable, because it is not in accordance with the meaning of an undetermined measurement. Moreover, the assumption of measurements being missing at random is not met, so the use of quantile normalization is questionable in this situation. For the internal control, mean and mean-120 some variation is subtracted from each patient, under the assumption that all miRNAs are affected in the same way. With the internal control only the plate variation is removed, while the variation coming from the purification is not addressed. The idea with the mean normalization is that all variation is being accounted for, likewise for the mean-120, only difference is the number of miRNA considered. To conclude whether or not the assumption for these methods hold is difficult.

4.2 Incidence

It is a central area of this thesis to uncover miRNAs that are predictors of incidence of pancreatic cancer, where incidence in this context will be the measure of a person's probability of having pancreatic cancer versus being a healthy subject. The hypothesis is that some miRNAs indicate cancer by having either a high or low C_t expression, and it is this pattern that could be beneficial to identify concerning earlier detection and correct diagnosis. A closer look at the different statistical approaches available is needed, and in light of the tasks nature, the *generalized linear models* (GLM) is a good starting point.

4.2.1 Generalized linear models

Generalized linear models is best explained by first looking at its predecessor, the classical general linear model theory. The classical approach dates back about a century ago, and is the principle of fitting some dependent variable (usually called the response variable) on the basis of multiple independent, predictor variables (also referred to as the covariates).

The word *multiple linear regression* (MLR) is sometimes used interchangeable. It can also be described as formulating a linear model on the basis of multiple observed quantities, in order to predict the expected value of some outcome. The term "linear" does not refer to how the independent variables enter the model, but how the predictor is computed. The *general linear model* (LM) can be defined as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i(p-1)} + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{4.16}$$

where y_i is the response and ϵ_i are called residuals, the errors associated with the fitted model. In this classic general linear model these quantities are often assumed to be independently normally distributed with constant variance. The β's represents the p parameters to be estimated. Usually the matrix form is more preferable

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.17}$$

which corresponds to

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n(p-1)} \end{bmatrix}_{n\times p} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}_{p\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1} \tag{4.18}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. The general LM is very useful for data analysis, but limited in many ways. Probably the largest motivation behind the birth of the generalized linear model, is the assumption of the response variable $\boldsymbol{y}$ being quantitative and normally distributed. Many other type of response variables are met in practice, e.g. binary, count, categorical etc. to name a few of the most common. GLM provides a unified approach to model all types of responses by applying a so-called *link function*. To mathematically define this function, consider the general LM from Equation (4.17) and denote $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ as the linear predictor part of the data. Instead of modeling $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{y}]$ directly as a function of the linear predictor, now model some function $g(\boldsymbol{\mu})$ of $\boldsymbol{\mu}$, such that

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}. \tag{4.19}$$

The result of using this link function $g(\cdot)$, makes it possible to relax the assumption of $\boldsymbol{y}$ being independently normally distributed, and permit the distribution to be any that belongs to the exponential family of distributions. This general class includes well-known probability distributions such as Normal, Poisson, gamma and binomial, which all can be written on the form

$$f(y|\theta, \phi) \quad = \quad \exp\left[\frac{(y\theta - b(\theta))}{a(\phi)} + c(y, \phi)\right] \tag{4.20}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are some functions and θ is called the *canonical parameter*. Since the response variable of interest is binary (cancer/healthy), only the binomial distribution will be explained in further detail. This is also known as the logistic regression model [Olsson 2002, pp. 2-3, 36-37].

4.2.1.1 Logistic regression

The binomial distribution is the foundation of the logistic regression model, because it is used to predict the probability p of an event[5], here being a person's probability of having cancer, i.e. $p = \mathrm{P}[y = \text{cancer}] = \mathbb{E}[y]$. It is a special case of the exponential family distributions with $\theta = \log\left(\frac{p}{1-p}\right)$, $b(\theta) = n\log[1+\exp(\theta)]$, $c(y, \phi) = \log\binom{n}{y}$ and $a(\phi) = 1$. By inserting this into Equation (4.20), the probability distribution function becomes

$$f(y|p) \quad = \quad \exp\left[y\theta + n\log\left(\frac{1}{1+\exp(\theta)}\right) + \log\binom{n}{y}\right] \tag{4.21}$$

$$= \quad \exp\left[y\log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\binom{n}{y}\right] \tag{4.22}$$

$$= \quad \binom{n}{y} p^y (1-p)^{n-y}. \tag{4.23}$$

[5]Not to be confused with the earlier used denotation as number of parameters.

The canonical parameter $g(\mu) = \theta = \log\left(\frac{p}{1-p}\right)$ defines the *logit link* function, or logit transformation. This is a *canonical link* for the binomial distribution, because it naturally transforms the mean to a canonical location parameter of this distribution. The ratio $\frac{p}{1-p}$ is known as the odds in favor of the event and comparing whether the probability of a certain event is the same for two groups, can be computed as the odds ratio $\text{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$. The inverse of the logit link function is $p = \frac{\exp[g(\mu)]}{1+\exp[g(\mu)]}$ restricting the mean of the response on a $[0; 1]$ scale and by combining this with Equation (4.19) the probabilities can be expressed by the logistic function (hence the word logistic regression)

$$p_i = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i\boldsymbol{\beta})} = \frac{1}{1+\exp[-(\boldsymbol{x}_i\boldsymbol{\beta})]} \tag{4.24}$$

where $\boldsymbol{x}_i$ corresponds to the i^{th} row of $\mathbf{X}$ from Equation (4.18). The parameters $\boldsymbol{\beta}$ are usually found by *maximum likelihood estimation* (MLE), explained in Section 4.2.1.2 [Olsson 2002, pp. 37-42, 98].

4.2.1.2 Maximum likelihood

Pawitan [2001, p. 22, def. 2.1] defines the likelihood principle as *assuming a statistical model parameterized by a fixed and unknown θ, the likelihood $\mathcal{L}(\theta)$ is the probability of the observed data y considered as a function of θ*. Today the principle of likelihood plays a central role in statistical modelling and inference, because the *likelihood function* captures all information in the data about a certain parameter, including the uncertainty. For the binomial case of interest the likelihood function of a single experiment is given in Equation (4.23). Remember that the GLM models an n-vector of independent response variables, where each element is distributed binomially. The distribution $f_{\theta_i}(y_i)$ with the canonical parameter θ_i is determined by μ_i, and ultimately $\boldsymbol{x}_i\boldsymbol{\beta}$, seen from Equation (4.19). The likelihood function is thus the joint probability density for the actual observations considered as a function of $\boldsymbol{\beta}$

$$\mathcal{L}(\boldsymbol{\beta}) \quad = \quad \prod_{i=1}^{n} f_{\theta_i}(y_i). \tag{4.25}$$

However, there is a tradition to focus on the *log-likelihood function* instead, because it is computationally more convenient. One advantage of the log-likelihood is that the terms are additive, which is easier. Taking the logarithm of Equation (4.25) yields

$$\ell(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{n} \log[f_{\theta_i}(y_i)]. \tag{4.26}$$

The MLE is found by partially differentiating ℓ w.r.t. each element of $\boldsymbol{\beta}$, setting the resulting expressions to zero and solving for $\boldsymbol{\beta}$. In practice, numerical methods are designed to solve these equations, the in-built `R` function `glm` uses by default the *iteratively reweighted least squares* (IRLS) algorithm, where the objective is a quadratic approximation to the log-likelihood [Wood 2006, pp. 63-66].

This concludes the description of GLM with focus on the binary response, which is a central subject of this thesis. The number of explanatory variables have already been narrowed down to a much smaller proportion compared to the original set, but still the hypothesis is that only a handful miRNAs are of importance. A logistic regression performed on the raw values with 75 miRNA covariates, would result in a phenomenon known as *complete separation* of variables. It happens when the outcome variable separates a combination of predictor variables completely - perfect prediction. Mathematically it means that the log-likelihood function does not reach a maximum as the effects increases, making them infinitely large. This is due the level differences between cancer patients and healthy controls present in the data. So besides the normalization procedure, there is a need of methods that can penalize $\boldsymbol{\beta}$ such that infinitely large effects are avoided. These types of shrinkage methods are explained in Section 4.4, but first an introduction to basic survival analysis.

4.3 Prognosis

The focus is now being shifted towards the prognostic part of the thesis, where methods to analyze time to a single event are discussed. This collection of analyzing techniques all fall under the category survival analysis, because the most typical event is death, however the derived application of survival analysis is much broader today. In biomedical studies, analyses of independent variables influencing patients life prognosis have always been of great interest. In this context, it is relevant to examine if some miRNAs are significant indicators of a better or worse life duration from the time of operation. First, the basic concepts of survival analysis need to be introduced.

4.3.1 Basic notation and terminology

Survival analysis deals with survival time as the response, which essentially is a quantitative variable with a distinctive feature; it contains incomplete data. This lack of information is a result of only knowing an event has not occurred

in a given time period and not knowing if or when the event will happen afterwards, a key analytical challenge known as censoring. In other words, the survival time is not known exactly for some individuals, but should still be taking into account in the overall analysis. Censoring is generally caused by three reasons, either a person has not experienced an event before the study ends, lost to follow-up during the study period or withdraws from the study for some reason [Kleinbaum and Klein 2005, pp. 5-6].

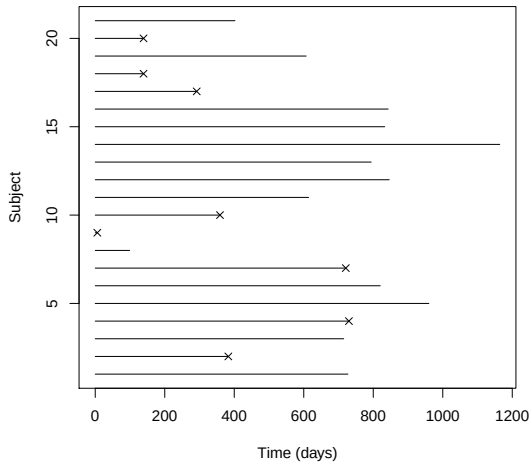

Figure 4.2: *Showing the survival times for a small segment of the data, where crosses indicate patients that have died and no crosses are censored patients.*

Figure 4.2 shows a small example of how survival data looks for a collection of subjects. The example is a small segment of the data used in this thesis. Each subject is represented by a line denoting the survival time in days, subjects with crosses as endpoints have experienced an event, i.e. died. In terms of notation, the random variable δ is used to denote censorship, such that $\delta = 0$ indicates a censored subject and $\delta = 1$ is a subject having experienced an event.

The key notation of a person's survival time is T, and any specific value of interest for the random nonnegative variable is denoted by t. The probability of an individual having an event of interest in the interval $[t, t + \Delta t[$ is given by

$$\mathrm{P}[t \leq T < t + \Delta t] \quad = \quad f(t)\Delta t \tag{4.27}$$

where $f(t)$ is the probability density function of a continuous random variable, describing the relative likelihood for this random variable to occur at a given point in the observation space. This is an approximation to the case where the difference in time is infinitely small, i.e. $\Delta t \to 0$. The *cumulative distribution function* is defined as

$$F(t) = P[T \leq t] = \int_0^t f(s)ds. \tag{4.28}$$

With the basic notation in place, it is now possible to define three fundamental concepts of survival analysis; the *survival function*, the *hazard rate* and the *cumulative hazard*.

4.3.1.1 Survival function

Also known as the *survivor function*, the survival function is a fundamental quantity that gives the probability of a person survives longer than some specified t. In other words, $S(t)$ is the probability that the random variable T taking values in $[0; \infty[$ exceeds the specified time t, which mathematically can be written as

$$S(t) = P[T > t] = \int_t^\infty f(s)ds = 1 - F(t). \tag{4.29}$$

This quantity has some theoretical quantities, which is readily seen from the hypothetical survival curve in Figure 4.3. Since it is a probability function it ranges between $0 \leq S(t) \leq 1$. The survival function evaluated at the starting point $S(0) = 1$, is the equivalent of being alive at the beginning of the study. Another extreme case is the theoretical $S(\infty) = 0$, which corresponds to increasing the study period without any limit, nobody would survive and the survival function must eventually fall to zero. This apply when the event considered is death, but there are however exceptions to this property, depending on the event of interest. An example could be studies where time to disease is measured for a group of patients, some patients will never get the disease and thus the survival function will not tend to 0 as $t \to \infty$, but instead towards some kind of positive value. The survival function is always a decreasing function though, since as time progresses more and more persons will eventually experience an event.

Notice the hypothetical survival function in Figure 4.3 is a smooth curve. Unless some known distribution is assumed for the survival time, this is certainly not the case in practice when working with real data. When estimating the survival function a *step function* is obtained instead. A popular choice is called the *Kaplan-Meier estimator* which is also referred to as the *product limit method*,

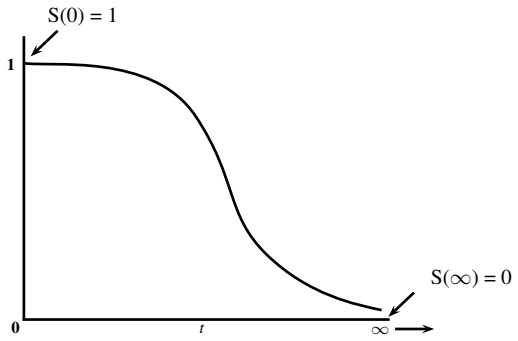

Figure 4.3: *Hypothetical survival curve.*

it estimates the survival function directly from the continuous survival times. Venables and Ripley [2002, p. 355] explain the Kaplan-Meier estimator in the following way.

Define the $r(t)$ to be the number of cases at risk just before time t, i.e. those that are in the trial and not yet dead (the risk set). Now, consider a set of intervals $I_i = [t_i; t_{i+1}[$ covering $[0; \infty[$, then the probability of surviving interval I_i can be estimated as $\frac{r(t_i)-d_i}{r(t_i)}$, where d_i is the number of deaths in interval I_i. Probability terms in the product will only appear for intervals in which a death occurs, so the limit becomes

$$\hat{S}(t) \quad = \quad \prod_{t_i \leq t} \frac{r(t_i) - d_i}{r(t_i)} \tag{4.30}$$

where the product is over time points at which deaths occur before t. The plot of a Kaplan-Meier estimator for the survival function, is a decreasing, piecewise constant function with jumps corresponding to the observed death times.

4.3.1.2 Hazard rate

Another fundamental term in survival analysis is the *hazard rate*. This is the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t. The hazard rate is defined by means of a conditional probability as

$$h(t) \quad = \quad \lim_{\Delta t \to 0} \frac{\mathrm{P}[t \leq T < t + \Delta t | T \geq t]}{\Delta t}. \tag{4.31}$$

Note that, the hazard rate and survival function are giving opposite information. While the survival function focuses on not experiencing an event, i.e. surviving, the hazard rate focuses on occurrence of event. Since the hazard is defined as a rate with a probability in the numerator, it is always nonnegative and has in principle no upper bound. There is a clear relationship between the survival function and hazard rate, given as

$$h(t) = \frac{f(t)}{S(t)} \tag{4.32}$$

that is useful for obtaining both functions knowing only one.

4.3.1.3 Cumulative hazard

The cumulative hazard is another way to represent the hazard rate, it can be described in words as an accumulation of the hazard over time, hence it is derived from integrating the hazard rate

$$\Lambda(t) = \int_0^t h(s)ds = -\log[S(t)]. \tag{4.33}$$

It can also be seen that by isolating in Equation (4.33), the survival function can be expressed in terms of the cumulative hazard by $S(t) = \exp[-\Lambda(t)]$. As for the survival function, the cumulative hazard also have non-parametric estimators, where the most common used is known as the *Nelson-Aalen estimator*. This estimator can be defined by applying similar reasoning as with the Kaplan-Meier, the hazard for each interval I_i is given by $\frac{d_i}{r(t_i)}$, so taking the sum of all times of death before t provides the Nelson-Aalen estimator [Venables and Ripley 2002, p. 356]

$$\hat{\Lambda}(t) = \sum_{t_i \leq t} \frac{d_i}{r(t_i)}. \tag{4.34}$$

This concludes basic survival analysis and equipped with this theory it seems natural to move on to the subject of how survival data is analyzed. In Section 4.3.2 a popular mathematical model for this purpose is introduced, the Cox proportional hazards model.

4.3.2 Cox proportional hazards model

It is a common objective in medical research to determine whether or not certain independent variables are correlated with the survival, i.e. if some covariates are

significant prognostic indicators. One way to satisfy this demand, is by applying one of the most popular methods for analyzing survival data today; the Cox proportional hazards model [Cox 1972].

First, let the matrix of covariates $\mathbf{X}$ and vector of coefficients $\boldsymbol{\beta}$ be defined as

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}. \tag{4.35}$$

Then the model consists of two parts, an underlying hazard function $h_0(t)$ which is the *baseline hazard*, i.e. the hazard for the respective individual when all independent variable values are set to zero. The second part is the effect parameters, the exponential to a linear sum of all the p explanatory (time-independent) variables. The Cox proportional hazards model for modeling the time it takes for an event to occur, is defined using the hazard rate for the i^{th} subject in the following way [Venables and Ripley 2002, p. 366]

$$h_i(t) \quad = \quad h_0(t) \exp(\boldsymbol{x}_i \boldsymbol{\beta}) \tag{4.36}$$

where $\boldsymbol{x}_i$ corresponds to the i^{th} row in $\mathbf{X}$ from Equation (4.35). It is seen that the covariates influence the hazard directly through a (log-)linear combination, so the simple interpretation of the effects is an attractive property. To illustrate this, consider some β_j being the effect of x_{ij} when corrected for the other covariates, it can be interpreted in terms of the *hazard ratio* (HR) when the x_{ij} is increased by one unit

$$\text{HR} = \frac{h_0(t) \exp[x_{i1}\beta_1 + \cdots + (x_{ij}+1)\beta_j + \cdots + x_{ip}\beta_p]}{h_0(t) \exp[x_{i1}\beta_1 + \cdots + x_{ij}\beta_j + \cdots + x_{ip}\beta_p]} = \exp(\beta_j) \tag{4.37}$$

from which it can be derived that when x_{ij} increases, the hazard increases when $\beta_j > 0$, and decreases when $\beta_j < 0$.

Equation (6.7) is called a *semi-parametric model*, because it consists of the baseline hazard which is non-parametric, and the parametric relative risk function. In contrast, a purely *parametric model* is one whose functional form is completely specified, except for the values of the unknown parameters, i.e. the survival time is assumed to follow a known distribution e.g. *Weibull*. The non-parametric element $h_0(t)$ is the main reason why the Cox proportional hazards model is such a popular choice, because it makes the model flexible since no specific distribution is assumed for the baseline group. This is very useful from a practical perspective, where the distribution is almost never known. Another way of saying this is that under the proportional hazards assumption the Cox

model is robust, in that the results from using the model will closely approximate the results from using the correct parametric model [Kleinbaum and Klein 2005, pp. 95-97].

The limitation of Cox proportional hazards model is the key assumption of the hazard rates for all subjects are proportional. Meaning that for any two subjects i and k with hazard rates $h_i(t)$ and $h_k(t)$, respectively, the relation

$$\frac{h_i(t)}{h_k(t)} = \frac{h_0(t)\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{h_0(t)\exp(\boldsymbol{x}_k\boldsymbol{\beta})} = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\exp(\boldsymbol{x}_k\boldsymbol{\beta})} = \exp[(\boldsymbol{x}_i - \boldsymbol{x}_k)\boldsymbol{\beta}] \tag{4.38}$$

is constant over time. The proportionality assumption is for example violated if the hazard rates cross, but even if this is not the case the proportional hazard assumption may still not be met [Kleinbaum and Klein 2005, p. 135]. There are several ways of verifying the use of Cox proportional hazards model, however this thesis is restricted to only consider *scaled Schoenfeld residuals*. The idea is to test for time trends in these residuals on the basis of a weighted mean [Venables and Ripley 2002, p. 371]. In `R` there exists a package named `survival` written by Therneau and Lumley [2011] that contains the function `cox.zph`, used for testing the assumption. In Section 4.3.2.1 the parameter estimation in the Cox proportional hazards model is introduced.

4.3.2.1 Maximum partial likelihood

Obtaining the maximum likelihood estimates of the parameters in the Cox proportional hazards model, is not as straight forward as in the GLM case, because the key feature of this model is the assumption of the outcome variable not following any specific distribution. Hence, in contrast to a parametric model a full likelihood based on the outcome distribution cannot be formulated. Thus, the likelihood function of the Cox proportional hazards model is based on the observed order of events rather than the joint distribution of events, a non-parametric method known as the *partial likelihood* [Kleinbaum and Klein 2005, p. 111].

Cox [1972] defined the partial likelihood function, and it can be explained by once again considering $r(t_i)$ as being the risk set containing the number of cases at risk of experiencing an event at time t_i. Let $i = 1, 2, \ldots, d$ denote the ordered uncensored subjects and $i = d+1, d+2, \ldots, n$ the censored subjects, then the partial likelihood function can be found by taking the product of conditional probabilities

$$\mathcal{L}(\boldsymbol{\beta}) \quad = \quad \prod_{i=1}^{d} \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{\sum_{j\in r(t_i)}\exp(\boldsymbol{x}_j\boldsymbol{\beta})} \tag{4.39}$$

with corresponding partial log-likelihood function

$$\ell(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{d} \left[\boldsymbol{x}_i \boldsymbol{\beta} - \log \left(\sum_{j \in r(t_i)} \exp(\boldsymbol{x}_j \boldsymbol{\beta}) \right) \right]. \tag{4.40}$$

Notice that no assumptions about the shape of the baseline hazard are needed. To get the parameter estimates that maximizes the partial likelihood, Equation (4.40) is differentiated w.r.t. to each element of $\boldsymbol{\beta}$ and set to zero. This system of equations is called *score equations* and it is solved computationally by iterative methods, most often using the *Newton-Raphson algorithm*. The `survival` package contains the function `coxph` used for fitting the Cox proportional hazards model to the data in this thesis.

4.4 Shrinkage methods

The main motivation behind these shrinkage methods is according to Zou and Hastie [2005] typically based on two aspects; accuracy of prediction and interpretation of the model. Especially the last one, scientist often prefers a simpler model because it puts more light on the relationship between the response and covariates. Parsimonious models are desirable when dealing with a large number of predictors, such as miRNAs. Three penalization techniques will be introduced here along with the *univariate method*, which is another general method to reduce the number of parameters, explained next in Section 4.4.1. The shrinkage methods will in general be introduced from a logistic point of view, but the methods apply in both the binomial and survival case.

4.4.1 Univariate method

The univariate method can be characterized as a selection method. It is a simple way to reduce dimensionality of parameters to a smaller subset, by univariate modelling. In this case, by performing $j = 1, 2, \ldots, p$ univariate logistic regression, i.e. for the j^{th} miRNA of interest the model becomes

$$g(\mu_{ij}) \quad = \quad \beta_{0j} + \beta_{1j} x_{ij}, \quad i = 1, 2, \ldots, n \tag{4.41}$$

where the hypothesis tested

$$\mathcal{H}_0 \quad : \quad \beta_{1j} = 0 \tag{4.42}$$

$$\mathcal{H}_1 \quad : \quad \beta_{1j} \neq 0. \tag{4.43}$$

The test of the j^{th} parameter is called a *Wald test*. If the large-sample conditions are valid, this test becomes $z = \frac{\hat{\beta}-\beta_0}{\hat{\sigma}_{\hat{\beta}}}$, where $\hat{\beta}$ is the MLE and β_0 the proposed value (here $\beta_0 = 0$) with the assumption that the difference between the two will be approximately normal [Olsson 2002, p. 47]. In the prognostic case, the univariate Cox proportional hazards model would be $h_{ij}(t) = h_{0j}(t)\exp(x_{ij}\beta_{1j})$, testing for the same hypothesis as in the logistic case.

The tests will generate j p-values, which are arranged in increasing order and the top ranked miRNAs are picked. Here λ_0 is a tuning parameter representing the p-value tolerance, so the miRNAs included in the model are the ones with a univariate p-value lower than this cutoff. There is no answer to what this value should be, it depends on the individual data and situation. However, cross-validation is a useful tool to get a statistical based indication, a method explained in Section 4.5. Even though univariate selection is a reasonable way to reduce variables considerably, it does not take correlation between covariates into account, which could be a problem in many situations. Despite this, the method was found applicable to determine which miRNAs to include in the starting model for the *backwards elimination procedure*.

4.4.2 Backwards elimination procedure

Backward stepwise selection is a strategy that starts with the large model and then sequentially deletes predictions under some criterion. It can only be used when $n > p$, but in combination with the univariate method, it constitutes a strong way to choose a parsimonious prediction model when $p > n$. Multiple criteria are available today, where the most popular are the *F–ratio*, *Akaike information criterion* (AIC) and *Bayesian information criterion* (BIC). The last two are closely related and applicable in settings where fitting is carried out by maximization of log-likelihood. Their generic forms are defined as

$$\text{AIC} = -2\ell + 2p \tag{4.44}$$

$$\text{BIC} = -2\ell + \log(n)p \tag{4.45}$$

where ℓ is the log-likelihood function, p represents the number of parameters in the fitted model and n the number of observations. The idea behind these criteria is a tradeoff between the deviance and number of parameters. This is important, because the deviance will decrease as the number of parameters in a model increases, so the parameter term compensates for this effect by favoring models with a smaller number of parameters. Lower values of the AIC/BIC index indicate the preferred model, that is the one with the fewest parameters that still provides an adequate fit to the data.

It can be seen that the BIC is proportional to AIC with the factor 2 replaced by $\log(n)$. Under the assumption $n > e^2 \approx 7.4$, BIC tends to penalize complex models more heavily, giving preference to simpler models in selection. Furthermore, BIC is asymptotically consistent as a selection criterion, i.e. given a family of models including the true model, the probability that BIC will select the correct one approaches one as the sample size becomes large. Because of this, the BIC has been preferred in this thesis [Hastie et al. 2001, pp. 56, 204-206].

4.4.3 Ridge

The curse of classic multiple linear regression is collinearity among the regressors, and it is from this not uncommon situation that *Ridge regression* originated from and was proposed as a method to circumvent the problem. The general idea is to shrink the regression coefficients, constraining $\boldsymbol{\beta}$, by imposing a penalty factor on their size. Ridge regression penalizes on the squared L^2–*norm* and when the response variable is continuous, the Ridge coefficients minimize a penalized *residual sum of squares* (RSS)[6] in the following way

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p-1} x_{ij}\beta_j \right)^2 + \lambda_2 \sum_{j=1}^{p-1} \beta_j^2 \right\} \tag{4.46}$$

where $\lambda_2 \geq 0$ is a tuning parameter that controls the amount of shrinkage. The larger the value of λ_2, the greater shrinkage [Hastie et al. 2001, p. 59]. Cross-validation can be used to obtain an estimate of the optimal parameter. An alternative and more general definition of Equation (4.46) can be written in terms of maximizing the log-likelihood function with the Ridge penalty term, which applies for GLM [Goeman 2010]

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \ell(\boldsymbol{\beta}) - \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\} \tag{4.47}$$

where $\|\cdot\|_2$ is the classic *Euclidean norm*. This can easily be expanded to apply for survival data, by replacing the log-likelihood with the partial log-likelihood. Ridge regression is good at dealing with high correlation between predictors, i.e. revealing grouping information, but one of the drawbacks is that it cannot produce a parsimonious model, because it keeps even small effects in the model due to the nature of the Ridge penalty. Another disadvantage of this method is that it does not handle the problem with missing values - an inherent problem

[6] Measure of the discrepancy between the data and an estimation model.

from the LM model, only complete cases can be considered, i.e. samples with all C_t measurements of the relevant miRNA present. The search of a regularization method that simultaneously performs variable selection and shrinkage, resulted in the *Lasso method*.

4.4.4 Lasso

This shrinkage method's name is an acronym for *least absolute shrinkage and selection operator*, which highlights the subtle but important differences from Ridge. Least absolute shrinkage refers to the L^1-penalty term now imposed on the regression coefficients, where the nature of this constraint will cause some of the coefficients to be exactly zero, hence selecting only a subset for the prediction model. The Lasso estimate is defined as [Hastie et al. 2001, p. 64]

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p-1} x_{ij}\beta_j \right)^2 + \lambda_1 \sum_{j=1}^{p-1} |\beta_j| \right\} \quad (4.48)$$

where the penalty factor $\lambda_1 \geq 0$ should be chosen to minimize an estimate of the expected prediction error. As mentioned in the Ridge case, this can be done by cross-validation. Equation (4.48) can also be written in terms of penalized log-likelihood optimization as

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \{\ell(\boldsymbol{\beta}) - \lambda_1 \|\boldsymbol{\beta}\|_1\} \quad (4.49)$$

where $\|\cdot\|_1$ is sometimes referred to as the *Manhattan norm*. Once again, Equation (4.49) can also be applied when using Cox regression by using the partial log-likelihood instead. The Lasso quickly became an appealing shrinkage method after its emerging, because of its sparse representation and the fact that it does both continuous shrinkage and automatic variable selection simultaneously. Despite showing success in many situations, Zou and Hastie [2005] have pointed out some limitations of the Lasso. In the $p > n$ case, the Lasso selects at most n variables of p candidates before it saturates, which is caused by the nature of the convex optimization problem. Moreover, if there is a group of variables among which the pairwise correlations are very high, then the Lasso has a tendency to select only one variable from the group, disregarding which one. In light of these problems, the Lasso is probably not an ideal method when working with miRNA data, but since many of the miRNAs with high N/A percentage have been discarded it can still give a decent result. However, just like the Ridge method, Lasso also have to remove cases with missing values. The last method applied is called a *naïve elastic net*, for which the ulterior motive was to combine the strengths of Ridge and Lasso.

4.4.5 Naïve elastic net

The thought of an even stronger regularization method lead to the idea of naïve elastic net[7], defined in the paper by Zou and Hastie [2005], describing it as *a stretchable fishing net that retains "all the big fish"*. This elastic net does automatic variable selection and continuous shrinkage simultaneously, and it can select groups of correlated variables. The basic principle is to exploit the advantages of Ridge regression and Lasso in some optimal combination. The penalized MLE using elastic net is found by

$$\hat{\boldsymbol{\beta}}^{\text{elastic}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \ell(\boldsymbol{\beta}) - \lambda_1 \|\boldsymbol{\beta}\|_1 - \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right\}. \tag{4.50}$$

Thus the elastic net combines the Ridge and Lasso penalties. However, the combination is more apparent when Equation (4.50) is written on a optimization problem form, thus let $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$, then

$$\hat{\boldsymbol{\beta}}^{\text{elastic}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p-1} x_{ij}\beta_j \right)^2 \tag{4.51}$$

$$\text{subject to } (1-\alpha) \sum_{j=1}^{p-1} |\beta_j| + \alpha \sum_{j=1}^{p-1} \beta_j^2 \leq t, \text{ for some } t. \tag{4.52}$$

The convex combination $(1-\alpha)\|\boldsymbol{\beta}\|_1 + \alpha\|\boldsymbol{\beta}\|_2^2$ is denoted the *elastic net penalty*. It is readily seen that the case $\alpha = 0$ corresponds to the Lasso, while $\alpha = 1$ is the Ridge regression. Furthermore, for all $0 < \alpha < 1$ the elastic net penalty function is strictly convex, thus having the characteristics of both Ridge and Lasso. Figure 4.4 (left) illustrates this argument, by showing the contour plot (first level) for the shrinkage methods when there are two parameters. The geometry of Ridge is a unit circle ($\beta_1^2 + \beta_2^2 \leq t$) while the Lasso is a diamond ($|\beta_1| + |\beta_2| \leq t$), and it is clear how varying α affects the constraint region for the elastic net. The RSS have elliptical contours centered at the full least squares estimate, and the methods find the first point where the elliptical contours hit the constraint region [Hastie et al. 2001, pp. 71-72].

The behavior of the discussed methods (Ridge, Lasso and elastic net) have been illustrated with a small example. Consider a simple linear regression model with the true parameters $\beta_0 = 2$, $\beta_1 = 3$ and $\beta_2 = 0.5$, the intercept is not important, but notice that β_1 is large and β_2 small. Figure 4.4 (right)

[7] Empirical evidence have shown that the naïve elastic net does not perform satisfactorily unless it is very close to either Ridge or Lasso, hence the word naïve. Improvements of the elastic net have been implemented, but they are beyond the scope of this thesis.

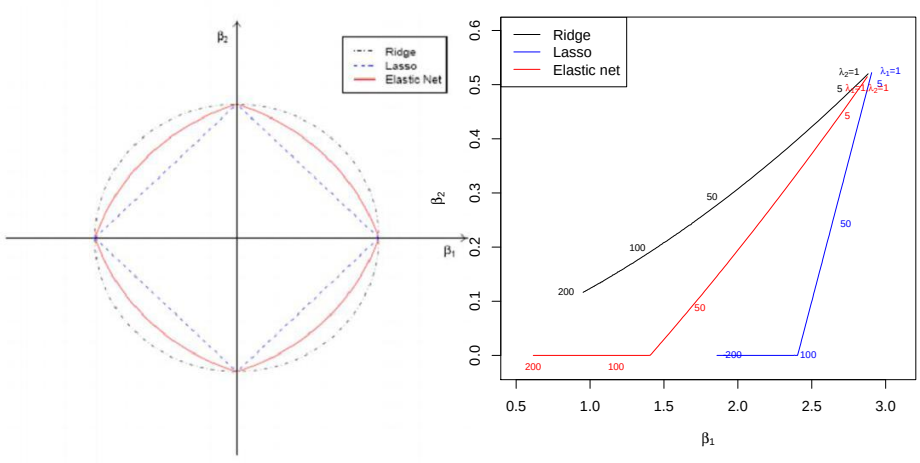

Figure 4.4: *(Left) Two-dimensional contour plot of the Lasso, Ridge and elastic net penalty with $\alpha = 0.5$, taken from Zou and Hastie [2005]. (Right) A small self-constructed example showing the effects' shrinkage path for the three methods, respectively.*

shows how the two coefficients are shrunk as the penalty factor is increased ($\{\lambda_1, \lambda_2\} \in \{1, 2, \ldots, 200\}$), for the Lasso, Ridge and elastic net with $\alpha = 0.5$. The Lasso is fairly quick to set the small $\beta_2 = 0$ and the method retain a close estimate of the true β_1, even after $\lambda_1 = 200$. The naïve elastic net have a tendency to overshrink in regression problems, and this phenomenon is also observed here. β_2 is set to zero as expected, but the elastic net shrinks β_1 much more than Lasso. The Ridge regression shrinks the parameters more simultaneously and does not set any of them equal zero, however if $\lambda_2 \to \infty$ the coefficients will tend to zero. In summary, all the regularization techniques performs as expected. Normalizing data to the same scale (i.e. same units) is important and should be stressed when applying these methods, otherwise the penalize factor works differently across the covariates.

In the above example the shrinkage methods was self-implemented, but when applied on the miRNA data, the `penalized` package written by Goeman [2011] was used. This `R` package has some useful properties, especially the fact that it supports both logistic regression and Cox proportional hazards model. Moreover, it allows optimization of the tuning parameters (λ_1 and λ_2) by cross-validation routines, a concept explained in Section 4.5.

4.5 Cross-validation

Hastie et al. [2001, p. 214] describes cross-validation as one of the most simple and widely used methods for estimating prediction error. It directly estimates the *extra-sample error*, defined as $\text{Err} = \mathbb{E}[L(Y, \hat{f}(X))]$, which is a generalization error when the method $\hat{f}(X)$ is applied to an independent test sample from the joint distribution of X and Y. This method is a useful tool when evaluating the model assessment for scarce data, which in real life usually is the case. Ideally, if enough data was present, a substantial part could be set aside to validate the model, but again this favorable situation almost never occurs.

The cross-validation algorithm is as follows, first the data is split into K roughly equally sized parts, called folds. For the k^{th} fold (test set), a model is fitted on the other $K-1$ folds (training set), and the prediction error for fold k using this fitted model is calculated. This is done for all K folds and gives an estimate of the prediction error.

Explained in more mathematical detail, let $\kappa : \{1, \ldots, n\} \mapsto \{1, \ldots, K\}$ be an indexing function that indicates the partition of the data. Denote by $\hat{f}^{-k}(x)$ the fitted function computed with the k^{th} fold left out, then the cross-validation estimate of the prediction error is given by

$$CV = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}^{-\kappa(i)}(x_i)). \tag{4.53}$$

When having a set of models $f(x, \alpha)$ indexed by some tuning parameter α, let $\hat{f}^{-k}(x, \alpha)$ denote the α^{th} model fit with the k^{th} fold of data left out. Then the cross-validation function is defined as

$$CV(\alpha) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}^{-\kappa(i)}(x_i, \alpha)) \tag{4.54}$$

which provides an estimate of the test error curve, where the tuning parameter $\hat{\alpha}$ that minimizes it, is the one of interest. The choice of K is not straight forward, as it depends on the data and objective. It can be classified as a tradeoff question between having low bias/high variance, or high bias/low variance. Consider one extreme case $K = n$ (also known as *leave-one-out* cross-validation), here the CV function is approximately unbiased for the true prediction error, but can have high variance because the n training sets are so similar to one another. Furthermore, the computational burden is often too immense to carry out in practice. On the other hand, in the case of e.g. $K = 5$, CV has a lower variance, but here bias could be a potential problem depending on how the performance of

the method varies with the size of the training set. As a general rule of thumb, five- and tenfold cross-validation are recommended as good compromises Hastie et al. [2001, pp. 215-216].

4.5.1 Receiver operating characteristics

Receiver operating characteristics (ROC) is a commonly used summary for assessing the tradeoff between *sensitivity* and *specificity*. This is appropriate for binary classifier systems where the discrimination thresholds are varied. Sensitivity is the *probability of predicting disease given the true state is disease*, while specificity is the *probability of predicting non-disease given true state is non-disease*. One of the objectives of this thesis is to identify the miRNAs making up the best model for predicting cancer vs. healthy, i.e. minimizing false positives and negatives as much as possible.

		Actual value	
		Positive	*Negative*
Predicted outcome	*Positive*	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)
	Total	P	N

Table 4.4: *Binary classification table.*

Consider the four outcomes of this binary classification experiment in a 2×2 *confusion matrix* in Table 4.4. The false positive outcome corresponds to a person being healthy, but predicted as having cancer. Sometimes referred to as a false alarm or in statistical terms a *type I error*, meaning that the test rejects a true null hypothesis[8]. A false negative is the equivalent of making a *type II error*, i.e. failing to reject a false null hypothesis. This is the case where a person has cancer, but is classified as healthy. The latter is generally considered to be a far more serious mistake, but overall the better model gives smaller prediction errors.

The *ROC curve* is a way of illustrating the predictive performance of a given model, using the sensitivity (true positive rate, TPR) and 1-specificity (false positive rate, FPR) as axes. It depicts the relative tradeoffs between true positive (benefits) and false positive (costs). Using the denotation from Table 4.4,

[8] $\mathcal{H}_0$: person is healthy
$\mathcal{H}_1$: person has cancer.

the definition of the axes are

$$\text{sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.55}$$

$$\text{1-specificity} = \text{FPR} = \frac{\text{FP}}{\text{N}} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{4.56}$$

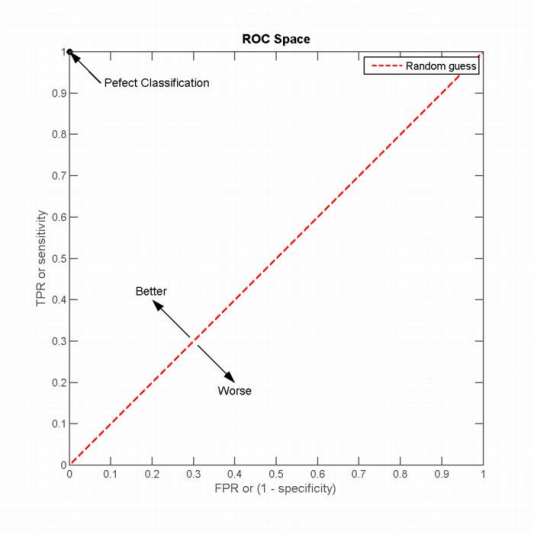

Figure 4.5: *Graphical representation of the ROC space, taken from Wikipedia.*

Figure 4.5 shows the ROC space and how TPR and FPR are used for the diagnostic test performance. The diagonal red line represents the completely random guess (like flipping a coin), models should lie above this line otherwise they are useless. Models with perfect classification lies in the upper left corner, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives).

In the search of an single quantity that measures a model's overall ability to discriminate between those individuals with disease and those without, the *area under curve* (AUC) was proposed. The curve is obtained by varying the probability cutoff point representing a particular decision threshold, deciding which persons to be classified as having the disease and being healthy. Fawcett [2006] highlights in his introduction to ROC analysis paper, an important statistical property of the AUC. It is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. An area of 0.5 indicates random predictions (no relevant classifier has an AUC lower than 0.5), while a value of 1 is perfect prediction. A

model having AUC greater than roughly 0.8 has some relevance predicting the responses of individual subjects, as rule of thumb. The AUC quantity in combination with cross-validation can serve as a measure for deriving some optimal tuning parameter, which gives the best possible prediction performance. The `R` package `ROCR` written by Sing et al. [2009], is used for calculating the AUC when needed.

Chapter 5

Simulation study

5.1 Objective

This theoretical chapter contains a preliminary performance analysis of the rank normalization method contra the raw C_t values. The hypothesis is that by rank normalizing data, more robust and reliable results are obtained, in the case of miRNA data with large mean C_t jumps. Since there is yet no statistical ground, a small simulation study have been created as way to validate this claim. The overall idea is to create a matrix of covariates that resembles true C_t measurements, i.e. share similar properties, and from a known linear predictor construct the response variable. This means that the truth is known, so the models fitted by using either the raw values or ranks can be evaluated under various noise conditions and compared according to some criterium. This is just the general outline of the study, a more comprehensive analysis plan will be given in Section 5.2.

5.2 Design of study

The matrix $\mathbf{X}$ containing the C_t values is simulated for each testcase from a normal distribution with $\mu = 25$ and $\sigma^2 = 1^2$, i.e. $\mathbf{X} \sim \mathcal{N}(25, 1^2)$, and has the dimension 250×768, the equivalent of having $n = 250$ patients and $p = 768$ miRNAs. This gives a covariate matrix that theoretically resembles noise-free miRNA data, and where the linear predictor used to define the response variable is arbitrarily chosen to be

$$\boldsymbol{\eta} \quad = \quad 0.5\boldsymbol{x}_3 - 1.1\boldsymbol{x}_{19} - 0.3\boldsymbol{x}_{119} + 1.1\boldsymbol{x}_{219} + \boldsymbol{x}_{300}. \tag{5.1}$$

The response variable is created by sampling from a binomial distribution with the probability of having cancer calculated from the logistic link function of the linear predictor, i.e. $\boldsymbol{y} \sim \mathcal{B}\left(n, \frac{1}{1+\exp[-(\boldsymbol{\eta}+\nu)]}\right)$, where ν is a scaling factor ensuring that there is a small overweight of cancer cases. For this study, appropriate proportions was obtained with setting $\nu = \sigma_\eta - \mu_\eta$.

With the constructed data set in place and knowledge about the true response, the purpose of the test is now to see how well the model is being fitted under various noise conditions, with and without rank normalization. These type of different noises have been identified from problems in the real data set and have been divided into four categories, i.e. four test dimensions that can be regulated which is explained next.

Sample frequency, ω Inspired by the mean C_t jumps that occur in the real data set, the sample frequency is a parameter that controls the per sample number where a new noise term should be added to the $\mathbf{X}$ matrix. In other words, it is the frequency for which samples should experience an alteration in the mean and standard deviation. In this study $\omega = 10, 30, 50, 75$ were tested, and when ω is high, fewer mean C_t jumps occur.

Sample mean, μ_ϵ For each set of samples determined by the frequency number, a small error term is added to the covariates with the distribution $\mathcal{N}(\mu_\epsilon, \sigma_\epsilon^2)$. The sample mean is defined as the product of some varied "loopingfactor" ρ and a random value between $[-1; 1]$, meaning that the average sample level can go up and down depending on the sign. The values tested are $\rho = 0, 0.1, 0.2, \ldots, 2, 3, 4, 5$.

Sample std. error, σ_ϵ The size of the standard deviation in the error term added for each sample frequency group, where $\sigma_\epsilon = 0, 0.1, 0.2, \ldots, 2, 3, 4, 5$.

No. of test cases, k This is the number of iterations to be run for each combination of noise conditions. Since there is a lot of possible combinations

$k = 10$ was chosen, so that the computation time would be kept on an acceptable level, but still provide a reasonable estimate of the truth.

Consider these as four levers controlling how often the mean C_t should jump, how high the jump should be, how big a variation should be associated with the jump and finally, how many times should it be replicated. Since each parameter had a number of levels it took quite some time to run the experiment. To be more specific, the total number of test runs can be calculated as the product of the number of different values each parameter assumes, resulting in $4 \cdot 24 \cdot 24 \cdot 10 = 23040$ test runs. One test run took approximately 30 seconds, so the whole simulation study took roughly $\frac{30 \cdot 23040}{60 \cdot 60 \cdot 24} = 8$ days to run.

The analysis of the noisy data consisted of a preliminary univariate selection and the backwards elimination procedure, performed on both the raw values and on the basis of ranking. The tuning parameter λ_0 for the univariate selection was in this study originally fixed after the principle of *Bonferroni correction*, because determining λ_0 on the basis of cross-validation in each test run would simply prolong the computational time too heavily. The Bonferroni correction can be applied when testing p dependent or independent hypotheses on a set of data, performed simultaneously. Each individual hypothesis is then tested at a statistical significance level of $1/p$ times what it would be if only one hypothesis were tested. Usually a significance level of 0.05 is used, but unfortunately a Bonferroni correction of $0.05/768$ was shown to be too low, because non of the miRNAs tested was significant enough to be included in the starting model of the backwards step procedure, under this tolerance. After some try-and-error, the value as high as $\lambda_0 = 15.1/768 \approx 0.02$ gave reasonable result in the sense that separation issues with the covariates chosen in the model did not occur, which was the main priority. It might seem a little arbitrary to choose this value, but the same tolerance is used with and without rank normalization, hence the value of λ_0 is not expected to influence the comparability of the two, which is the only purpose in this context.

When the step procedure have obtained the final model based on raw values and ranks, respectively, then some kind of measure is needed to draw conclusions concerning the *goodness-of-fit* for each model. In this study, the final model deviance is used as the degree of fit measure, expressed as the percentage deviance "explained" by the model compared to the null model (model with intercept only). This quantity will be denoted deviance measure (DM) and defined as

$$\text{DM} \quad = \quad \frac{\epsilon_0 - \epsilon_m}{\epsilon_0} \cdot 100\%. \tag{5.2}$$

It is clear from Equation (5.2) that a low residual deviance ϵ_m implies a high DM, meaning that the model with the highest DM is the better model. Each

test run will result in two deviance measures, one for the analysis on the raw values and one for the ranks, however it is more practical with a single measure that easily can compare the two. Therefore let the substraction of the DM for each model be denoted *performance measure* (PM), defined as

$$\text{PM} \quad = \quad \text{DM}_{raw} - \text{DM}_{rank}. \tag{5.3}$$

This gives a basic measure of how well the two analysis procedures performs under various noise conditions. If PM < 0, then rank normalization is the best choice because it results in the best fitted model. Of course, PM > 0 then indicates that it is better to work with the raw values.

The results from the described simulation study are introduced in Section 5.3, but it is important to stress that this study could have been done in countless other ways and the results derived from it are not to be considered the sole truth. It is only meant to provide some idea of the best way to handle these troublesome miRNA data.

5.3 Results

Here some of the most important results of the study are presented graphically, but since it is hard to visualize more than three dimensions, multiple plots are created instead. Figure 5.1 only considers the behavior of the performance measure while varying the loopingfactor and sample frequency, disregarding the value of noise standard error. As a plotting tool *box plot* was chosen, since it is a nice way to depict multiple test cases under some condition, in order to get a clear overview.

It seems like the general trend is independent of how often a new noise term is added to the samples. Figure 5.1 shows that by increasing the size of the mean shift the PM becomes more negative, indicating that rank normalization of data provides the best fitted model under these conditions. However, in the beginning when the size of the mean jump is only moderate, there is very little difference between working with ranks and raw values (the latter would probably be preferred in this case). To conclude on the first scenario, when experiencing large mean C_t jumps between groups of samples, rank normalization could suggest more robust results. Next, Figure 5.2 now disregard the loopingfactor and concentrates solely on varying the amount of standard error of the noise term added.

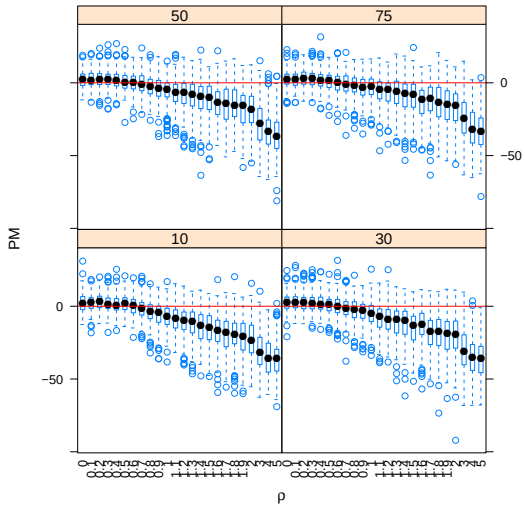

Figure 5.1: *The PM shown under different quantities of the loopingfactor, stratified by the sample frequency.*

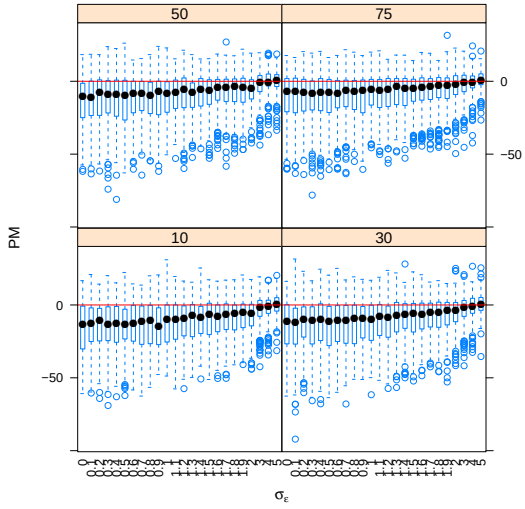

Figure 5.2: *The PM shown under different quantities of the noise standard error, stratified by sample frequency.*

It seems apparent that varying the size of the noise standard error, does not influence the performance of the two analysis methods significantly. There is slight evidence that the rank method provides best fit with smaller values of noise variation, but of course as the noising of data becomes large enough the prediction ability of both methods performs equally bad. Since the frequency of adding a new noise term only show signs of little effect, it is natural to keep this dimension fixed and see how the results distribute themselves when varying the loopingfactor and standard error at the same time. Figure 5.3 shows the results for $\omega = 30$.

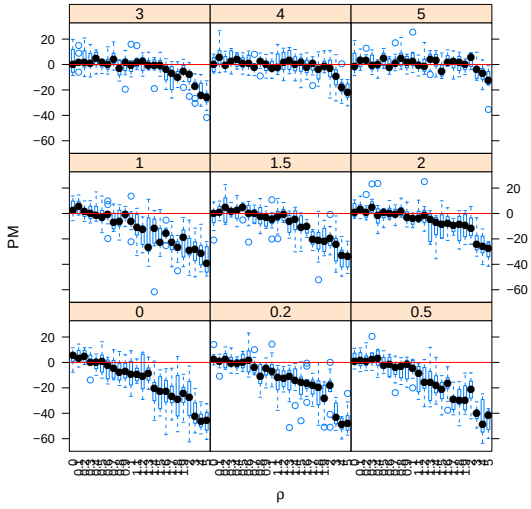

Figure 5.3: *The PM shown under different quantities of the loopingfactor for sample frequency $\omega = 30$, stratified by chosen values of the standard error.*

Beginning with one of the extreme cases where the loopingfactor is incrementally increased as the standard error is set to zero (lower left box plot in Figure 5.3), the same trend as in Figure 5.1 is observed. With a noise term close to being zero the raw values performs best which is also expected, but as the loopingfactor increases the rank normalization stands out as the most robust method to cope with these noise additions. As the standard error is also incrementally increased simultaneously the difference between the two methods smoothens out. Especially in the other extreme case where $\sigma_\epsilon = 5$ (top right box plot in Figure 5.3), the noise addition is so large and have become dominant, hence choosing to rank normalize data or not becomes insignificant. The same scenario have been tried with a higher sample frequency, i.e. $\omega = 75$ in Figure 5.4. This frequency implies only a couple of mean C_t shifts for a population of 250 patients, which

resembles the serum data even more.

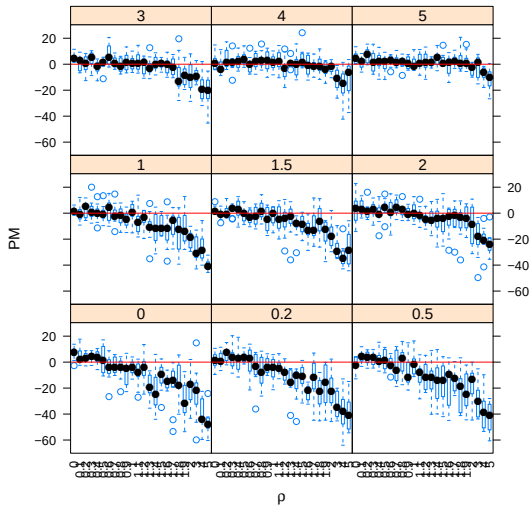

Figure 5.4: *The PM shown under different quantities of the loopingfactor for sample frequency $\omega = 75$, stratified by chosen values of the standard error.*

Not much changes when fewer jumps are simulated, the overall behavior of the results remains the same, leading to a conclusion of the simulation study. Rank normalizing data is a good idea in the presence of large mean differences between groups of samples, however in the case of very little noise the raw values are the better choice. Since the serum data provided for this thesis experience these large mean jumps, especially between PC patients and controls, the simulation study suggests rank normalized data will provide more reliable results.

It should however be kept in mind that there are many limitations to this simulation study and the results are to be taken with a "grain of salt". One limitation concerns the simulated covariate matrix $\mathbf{X}$ in which no missing values are generated. This is known not to be in agreement with real miRNA measurements, hence it is something to consider in order to improve the simulation study. Another limitation worth mentioning is the previous discussed tuning parameter λ_0 for the univariate selection method, here an overall value was fixed for both methods after the Bonferroni principle (to a certain degree). Strictly speaking, results closer to the truth would be obtained by applying cross-validation on the data for the respectively method, in each test run. Furthermore, if time allowed it all the normalization methods should tested and not just with the logistic

regression, but expanded to the prognostic case with Cox regression as well. However, despite these limitations, this small theoretical study still provides informative insight to the effect of normalizing miRNA data with properties as in the serum data. The different analyses on serum data in relation to incidence and prognosis, are found in Chapter 6.

Chapter 6

Results

This chapter presents the results from the analyses performed, by applying the methodology introduced in Chapter 4 on the data introduced in Chapter 3. The results have been divided into an incidence and a prognostic part. Section 6.1 includes the analyses based on logistic regression with the five different normalization methods, where the purpose is to identify a subset of promising miRNAs that function as indicators of pancreatic cancer. An additional comparative study is included here, where the shrinkage methods are evaluated in their ability to predict cancer/healthy subjects correctly. Section 6.2 includes the analyses based on Cox proportional hazards model with the five different normalization methods. The objective is here to identify the most significant miRNAs providing information of expected survival time from operation for PC patients. A similar comparative study is trying to highlight the ability of miRNAs to predict expected survival after operation.

6.1 Incidence

A number of analyses have been performed using logistic regression on the serum data, separated by small differences such as choice of normalization method and the number of folds used in cross-validation. Despite this, a general outline of the analysis procedure for incidence can still be given, which is described in the following.

As was mentioned in Section 3.2 there were 206 patients eligible for analyses, consisting of three groups; PC, CP and HS. Generally the CP patients will be regarded as healthy subjects, so the number of cases together with the controls is 69, while there were 137 cancer cases. Furthermore, there were 754 independent miRNAs from the beginning and those having > 20 N/As were excluded, leaving only 75 miRNAs that satisfied this criterion. This reduced data set can be characterized as the foundation for all the analyses regarding incidence, which can be summarized in a step-by-step procedure.

Step 1 Data is normalized by the chosen normalization method.

Step 2 The univariate selection method is applied to the normalized data set in order to find a subset of miRNAs for further analysis. The p-value tolerance λ_0 controls the number of miRNAs in this set, and is obtained by 20-fold cross-validation. The self-made cross-validation function used AUC as an evaluation measure to optimize after, hence the λ_0 chosen provides the best logistic model.

Step 3 The selected miRNAs are standardized by their *interquartile range* (IQR), i.e. divided by the distance between the 75^{th} percentile and the 25^{th} percentile. This provides more robust estimates of the coefficients, since outlier measurements are disregarded.

Step 4 Logistic regression is performed on the complete cases in a backwards elimination procedure, where the BIC is used as a model evaluator.

Step 5 Moreover, Lasso regression is performed with the miRNAs selected by the univariate method, as the starting model. The reason why only this subset of miRNAs is considered, and not the initial 75 miRNAs, is due to missing values. The generic problem inherited from the GLM restricts the shrinkage methods to only work with complete cases, and simply to many cases would be removed if all the 75 miRNAs were considered. The penalty parameter λ_1 is determined by 20-fold log-likelihood cross-validation, from the built-in function `optL1` of the `penalized` package.

Step 6 Furthermore, logistic regression with elastic net penalty is performed with the univariate selected miRNAs. The tuning parameters λ_1 and λ_2

are once again determined by 20-fold log-likelihood cross-validation, using the `cvl` function from the `penalized` package.

This is the general outline when data was analyzed with a binary outcome. There were minor discrepancies though, between normalizing with ranks and other methods. One thing is that ranking assigns a value to the missing values, meaning that all the 206 cases were maintained in the starting model for the backwards step procedure, Lasso and elastic net. Also, a re-ranking was needed after the univariate selection, otherwise the ranks corresponded to the old data set.

Ridge regression was left out as a candidate method for selecting significant predictors, since it does not create a parsimonious model, hence does not serve the purpose of selecting only a few candidate biomarkers. However, this certainly does not rule out Ridge regression as a useful method in other analyses situations, which is why it is included in the comparative study described in Section 6.1.1, where the prediction ability of each method is tested on the basis of AUC.

6.1.1 Comparative study

The main objective of this study is to compare the four different shrinkage methods' ability to fit a useful model in predicting cancer/no cancer. The shrinkage methods are not surprisingly the univariate method in combination with backwards stepwise selection, Ridge, Lasso and the naïve elastic net. The idea of a comparative study is derived from the paper by Bøvelstad et al. [2007], where seven methods were compared on their predicting survival from microarray data. The basic idea is to split the data set into a 2:1 training/test set, where each training/test set keep the same cancer-control ratio as the original data set, and then use the model fitted to the training set to predict the cases in the test set. To measure how well each method is predicting the test set, the AUC was used. The study is restricted to rank normalized data only, in order to avoid the trouble with missing values. As Bøvelstad et al. [2007] mentions in their article, a single split is not enough to establish a reliable evaluation of the prediction performance, therefore 50 iterations were run and relevant information were collected each time. The optimal tuning parameters are determined for each training/test set split by 10-fold log-likelihood cross-validation, except for λ_0 in the univariate selection; here AUC cross-validation was applied. The choice of 10-fold is considered to be a good tradeoff between computational time and reasonable estimates of the true optimal tuning parameters. This is also the number of folds Bøvelstad et al. [2007] used. A summary of the results can be seen in Table 6.1, where data from each test run can be seen in Appendix A.1.1.

k	λ_0	$\lambda_{1,l}$	$\lambda_{2,r}$	$\lambda_{1,e}$	$\lambda_{2,e}$	AUC_u	AUC_l	AUC_r	AUC_e
mean	5.649e-04	36.3188	1410.6050	36.1953	1410.6097	0.89314	0.93795	0.91987	0.92897
sd	6.429e-04	11.6041	568.3776	11.5109	568.3767	0.04189	0.02506	0.02873	0.02745
min	1.720e-05	8.3979	419.8538	8.4604	419.8541	0.79206	0.87814	0.84139	0.85783
max	2.688e-03	58.9925	3433.9514	57.9870	3433.9516	0.97164	0.98960	0.96503	0.98299

Table 6.1: *Comparison of the prediction performance for the four shrinkage methods, on the basis of AUC.*

Apparantly there is not much difference in the optimal penalty parameter found for the Lasso and Ridge, respectively, and the optimal parameters used in the elastic net penalty term. In terms of overall prediction ability, the univariate+backwards method seems to perform poorest of the four while Lasso in average performs best, i.e. have the highest AUC with the lowest deviance. Since Lasso is a shrinkage method that also functions as a screening tool, that is from its penalization nature also does subset variable selection by setting small coefficients to zero, it is reassuring to know that it seemingly also is the best method in terms of correct classification.

In general all the methods predict at an acceptable level ($\overline{\text{AUC}} > 0.89$) which is a very encouraging result, because it indicates that the miRNAs can serve as reliable predictors of patients with cancer and healthy subjects. However, the high average AUC could also have something to do with the presence of seperation of variables in the original data, due to bad experimental planning. Even though rank normalization tries to overcome this problem, there might still be some artificial seperation between PC and HS left, leading to better prediction results than actually true.

Since this comparative study is only based on ranks, there is no telling about how the shrinkage methods will perform using other normalization methods, so the conclusions drawn from this study are limited. The choice of using AUC as a model evaluation measure is only one way to go, other measures could also have been tested if time allowed it.

6.1.2 Rank

The clinical objective is to find a subset of miRNA predictors in relation to pancreatic cancer, hence this section provides the results derived from incidence analysis of the serum data, based on rank normalization. The analysis plan was given in Section 6.1. After rank normalizing data, the second step consisted of

finding the optimal p-tolerance from CV, which here was $\lambda_0 = 0.0001039$ resulting in 20 miRNAs constituting the starting model of the backwards elimination procedure.

The univariate method showed some inconsistency in the number of miRNAs selected, a small change in λ_0 could alter the miRNAs present in the resulting model to a certain degree. Further examination discovered that a large part of the p-values obtained were very small and close to each other. To determine whether the cumulative distribution function of these p-values $F_p(x)$ is uniform or systematically higher, i.e. the p-values are systematically lower, a one-sided and one-sample *Kolmogorov-Smirnov test* can be performed. The hypotheses tested are

$$\mathcal{H}_0 \quad : \quad F_p(x) = F_{\mathcal{U}(0,1)}(x) \tag{6.1}$$
$$\mathcal{H}_1 \quad : \quad F_p(x) > F_{\mathcal{U}(0,1)}(x). \tag{6.2}$$

The test statistic for the Kolmogorov-Smirnov test is defined as the largest vertical difference between the sample and theoretical cumulative distribution functions[9]. Here the statistic evaluates to $D_n^+ = 0.625$ with a p-value < 0.00001, so the null hypothesis is rejected suggesting that the p-values are systematically low. This result could also be derived visually from Figure 6.1, showing the empirical cumulative distribution function of the p-values.

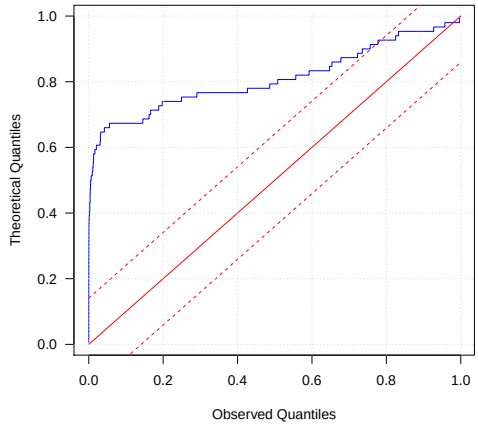

Figure 6.1: *One-sample Kolmogorov-Smirnov test of the univariate p-values for rank normalization.*

[9] $D_n = \sup_x |F_n(x) - F(x)|$.

The cumulative distribution of p-values is very steep in the beginning showing that there is many small values, so there is strong evidence of signal in the p-values. In Table 6.2 relevant information about the final model of the step procedure is provided.

miRNA	**OR**	$\mathbf{CI}_{0.95,low}$	$\mathbf{CI}_{0.95,high}$	**p-value**
miR.ecd	13.840	4.390	50.420	**<0.001**
miR.odd	0.190	0.060	0.530	**0.002**
miR.myc	5.410	1.950	18.730	**0.003**
miR.cgd	0.250	0.110	0.530	**<0.001**

Table 6.2: *The odds ratio, 95% confidence limits and p-values of the significant miRNAs, derived from the backwards step procedure for rank normalized data.*

In the end four miRNAs showed to be significant predictors of pancreatic cancer with rank normalized data. Starting with miR-ecd that has an odds ratio of 13.84, i.e. one IQR unit increase in ranks of this miRNA results in the odds of having pancreatic cancer increases by a factor 13.84. Clinically speaking an increase in rank means an increase in C_t and ultimately a decrease in miRNA material, from which it can be concluded that a decrease in the miR-ecd expansion indicates a patient having cancer. This fairly large odds ratio can be partly explained by the IQR standardization, where the ranks of each miRNA are divided by their range between Q_{25} and Q_{75}. A large range will result in smaller rank values and ultimately larger increase in odds ratio per unit rank increase. Furthermore, as explained in Section 4.1.1 the interpretation of a unit rank increase is not quite clear, as opposed to a unit C_t increase. Besides miR-ecd, miR-myc also seems to be a predictor of cancer with an odds ratio of 5.41, while an increase in miR-odd and miR-cgd strengthens the conclusion that a patient is healthy. The OR of the latter is probably easier interpreted with the inverse odds ratio. For miR-odd the $\text{OR}^{-1} = 1/0.19 = 5.26$, which can be interpreted as the odds of having cancer decreases by 5.26 for one unit IQR rank increase in miR-odd. For miR-cgd, the equivalent is true with an inverse odds ratio of 4.

It is hard to say how trustworthy these results are since the univariate method for these miRNA data is highly sensitive, which can also be seen from the wide range of the CIs. Therefore a comparison with miRNAs found by the Lasso and elastic net shrinkage method, respectively, is necessary in order to select the relevant miRNAs with more certainty. Table 6.3 provides a ranked list of the significant miRNAs found by the three methods, where Lasso used the penalty $\lambda_1 = 3.6$ and the elastic net $(\lambda_1, \lambda_2) = (3.6, 5.94)$.

	miRNA	Found by	Uni+step	Lasso	Elastic
1	miR.ecd	Uni+step, Lasso, Elastic	13.8385	2.8717	1.6669
2	miR.myc	Uni+step, Lasso, Elastic	5.4094	1.3659	1.2845
3	miR.cgd	Uni+step, Lasso, Elastic	0.2460	0.3693	0.5491
4	miR.odd	Uni+step, Lasso, Elastic	0.1911	0.4356	0.5852
5	miR.omd	Lasso, Elastic		1.5290	1.5590
6	miR.tyc	Lasso, Elastic		1.3722	1.2551
7	miR.wzc	Lasso, Elastic		0.9645	0.9416
8	miR.eud	Lasso, Elastic		0.8875	0.8751
9	miR.lld	Lasso, Elastic		0.7690	0.7493
10	miR.kmd	Lasso, Elastic		0.6081	0.6325
11	miR.pjd	Elastic			1.1204
12	miR.kyc	Elastic			1.1010
13	miR.gyc	Elastic			1.0435
14	miR.vyc	Elastic			1.0209
15	miR.zbd	Elastic			0.9706

Table 6.3: *Significant miRNAs found by the various methods on the basis of ranks, first ordered by number of times the given miRNA occurs for each method, and second by the magnitude of OR.*

The methods all find miR-ecd, miR-odd, miR-myc and miR-cgd to be significant (i.e. included in the final model), hence the conclusion is that these are the most prominent miRNAs for separating pancreas cancer patients from the controls. The OR of e.g. miR-ecd in the Lasso and elastic net method is much smaller than the OR obtained in the regular logistic fit, because of the penalty terms imposed. Lasso and elastic net also finds additional miRNAs not found by the backwards step which also have the potential of being biomarkers in relation to incidence, e.g. miR-lld and miR-eud which have an OR higher than 1 in both methods. In conclusion, there exists small differences in the number and type of miRNAs found by the various shrinkage methods, which was expected. Despite this, the ORs of miRNAs found by multiple methods are in agreement with each other, they "point in the same direction" which is important. It is interesting to see if the same applies across different normalization methods, which is analyzed in the forthcoming pages.

These were the results of analysis based on ranks which is regarded as a robust normalization method, i.e. a method that hopefully could remove nuisance factors such as day of purification and plate variation, but still maintain the true effects. Figure 6.2 shows the distribution of raw values and ranks for the four significant miRNAs in the logistic model. Generally speaking the distribution between the three groups PC, CP and HS is unchanged by the rank normal-

ization, however it makes separation between the cancer and healthy group for miR-odd even more apparent. This indicates that the rank method does have some positive influence on data, but everything comes with a price and in this case ranking of data complicates the interpretation of the odds ratios.

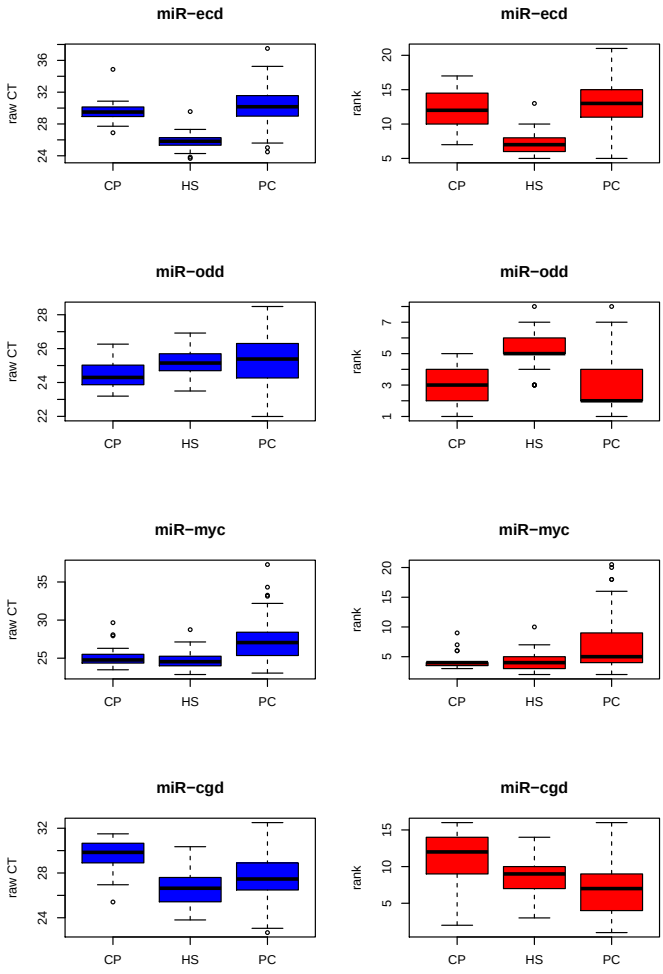

Figure 6.2: *Boxplot of miR-ecd, miR-odd, miR-myc and miR-cgd. (Left) Raw C_t values. (Right) Ranks, after the univariate selection.*

6.1.3 Quantile

The result of quantile normalization in terms of mean C_t level for each sample, can be seen in Figure 6.3. Each sample is now totally aligned disregarding ignorable discrepancies, making them more comparable for analysis.

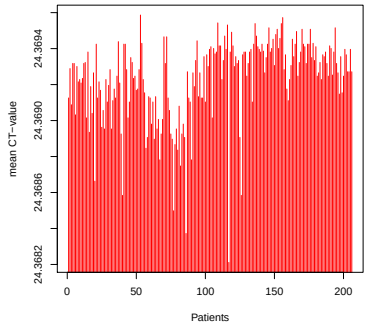

Figure 6.3: *Average C_t level for the 206 persons eligible for analyses after quantile normalization.*

Cross-validation with 20 folds determined the p-tolerance of the univariate method to be $\lambda_0 = 0.0011359$, and the one-side Kolmogorov-Smirnov test statistic $D_n^+ = 0.551$ has a p-value < 0.00001, rejecting the null hypothesis. Figure 6.4 shows that the empirical cumulative distribution of the p-values are significantly above the uniform.

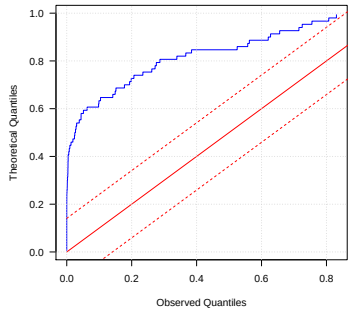

Figure 6.4: *One-sample Kolmogorov-Smirnov test of the univariate p-values for quantile normalization.*

The p-tolerance ensured that 20 miRNAs was included in the starting model of the backwards step procedure. The ending model is summarized in Table 6.4, where the first thing noticed is that miR-ecd and miR-cgd are once again present.

miRNA	**OR**	$\mathbf{CI}_{0.95,low}$	$\mathbf{CI}_{0.95,high}$	**p-value**
miR.ecd	18.990	6.600	65.700	**<0.001**
miR.pdd	9.250	3.450	27.990	**<0.001**
miR.tyc	3.710	1.370	11.540	**0.014**
miR.cgd	0.240	0.100	0.460	**<0.001**

Table 6.4: *The odds ratio, 95% confidence limits and p-values of the significant miRNAs, derived from the backwards step procedure for quantile normalized data.*

The ORs obtained on the basis of quantile normalization, is slightly more sensible in terms of interpretation, as opposed to those obtained from the rank normalization. This is because it is not a pattern being analyzed now, but instead the actual C_t values. The odds ratio of miR-ecd is 18.99 which is actually higher than in the ranking case. Apparantly one IQR C_t unit increase makes a large difference concerning the odds of having cancer. The OR of miR-cgd is almost the same as in the ranking case, where an increase in IQR C_t for this miRNA reduces the odds of having cancer by 4.17. Both miR-pdd and miR-tyc seems to raise the odds in favor of having cancer, when their miRNA expansion is reduced. The estimated ORs however, does not seem very reliable judging from their confidence intervals, which could be caused by the relatively small number of samples.

In the end the Lasso and elastic net regression were also run on quantile normalized data, this time with penalty term $\lambda_1 = 1.52$ for Lasso and $(\lambda_1, \lambda_2) = (0.829, 2.41)$ for the elastic net. The results are collected in Table 6.5, which gives an overview of which miRNAs the respective methods find along with their OR.

	miRNA	Found by	Uni+step	Lasso	Elastic
1	miR.ecd	Uni+step, Lasso, Elastic	18.9901	4.7506	2.5015
2	miR.pdd	Uni+step, Lasso, Elastic	9.2526	2.6063	1.9468
3	miR.tyc	Uni+step, Lasso, Elastic	3.7139	2.1279	1.7254
4	miR.cgd	Uni+step, Lasso, Elastic	0.2363	0.3074	0.3955
5	miR.egd	Lasso, Elastic		1.9763	1.8882
6	miR.wyc	Lasso, Elastic		1.3167	1.3089
7	miR.ged	Lasso, Elastic		0.8190	0.7467
8	miR.odd	Lasso, Elastic		0.5716	0.5238
9	miR.lld	Lasso, Elastic		0.5478	0.5741
10	miR.omd	Elastic			1.1788
11	miR.myc	Elastic			1.1715
12	miR.kyc	Elastic			1.0803
13	miR.gyc	Elastic			1.0802
14	miR.czc	Elastic			1.0557
15	miR.kmd	Elastic			0.9926
16	miR.vzc	Elastic			0.9349

Table 6.5: *Significant miRNAs found by the various methods on the basis of quantile normalization, first ordered by number of times the given miRNA occurs for each method, and second by the magnitude of OR.*

6.1.4 Internal control

The third analysis is based on normalization of data by the use of an internal control. As described in Section 4.1.3, U6 small non-coding RNA was chosen as the endogenous control, where it was possible to take the average of 8 measurements for each individual samples and subtract this value from the sample mean. The mean C_t level of the samples, after this normalization step was performed, can be seen in Figure 6.5. This normalization procedure does not seem to help in making the samples more comparable, since it does not solve the problem with these sample-to-sample mean shifts in C_t at all. This is only an attractive method when the choice of internal control is appropriate, which might not be the case here. This is probably why this is not a widely used normalization method.

The test of the empirical cumulative distribution of the univariate p-values following a uniform distribution, results in the test statistic $D_n^+ = 0.363$. The statistic is lower than those of the previous two analyses, however the null hypothesis is still rejected with a p-value < 0.00001.

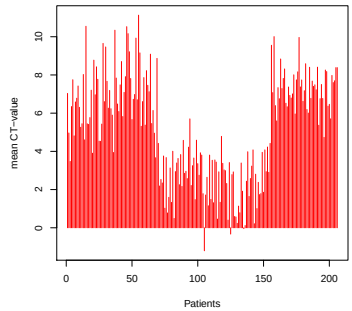

Figure 6.5: *Average C_t level for the 206 persons eligible for analyses after internal control normalization.*

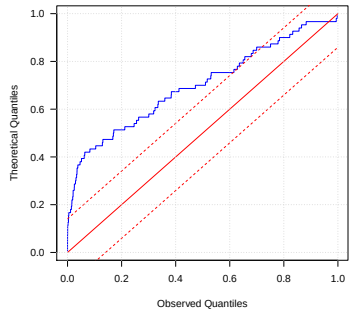

Figure 6.6: *One-sample Kolmogorov-Smirnov test of the univariate p-values for internal control normalization.*

It is clear from Figure 6.6 that the p-values are not within the 95% confidence band of the uniform distribution, visually verifying the result of the Kolmogorov-Smirnov test. The p-value cutoff was set to be $\lambda_0 = 0.0009134$, where 9 miRNAs had a p-value below this criteria. The resulting three miRNAs the step procedure decides to keep in the final model based upon lowest BIC, are summarized in Table 6.6. As seen before, miR-ecd and miR-cgd appears as significant predictors for a patient's probability of having cancer. The OR for miR-ecd is an extreme 44.72 which of course is a suspicious value, but apparently a raise in one IQR C_t unit of this miRNA raises the odds of having pancreatic cancer substantially. On the other hand, raising the miR-cgd by a single IQR unit decreases the odds of having cancer by 4.35, and for miR-lld the inverse odds ratio are even as high as 12.5.

miRNA	OR	$CI_{0.95,low}$	$CI_{0.95,high}$	p-value
miR.lld	0.080	0.020	0.220	**<0.001**
miR.ecd	44.720	13.850	197.480	**<0.001**
miR.cgd	0.230	0.060	0.740	**0.019**

Table 6.6: *The odds ratio, 95% confidence limits and p-values of the significant miRNAs, derived from the backwards step procedure for internal control normalized data.*

These results derived from the internal control normalized data should be taken with a grain of salt, because as Figure 6.5 indicated there were still large mean differences between samples present. This makes the results even more sensitive, hence more unreliable. Table 6.7 gives an overview of the miRNAs found by all the shrinkage methods using $\lambda_1 = 1.3$ for Lasso and $(\lambda_1, \lambda_2) = (1.31, 0.838)$. Not many miRNAs are found here compared to previous results, but notice that the ORs for miR-ecd in all the methods are unrealistically high.

	miRNA	Found by	Uni+step	Lasso	Elastic
1	miR.ecd	Uni+step, Lasso, Elastic	44.71500	21.5447	11.7827
2	miR.cgd	Uni+step, Lasso, Elastic	0.23436	0.4028	0.5202
3	miR.lld	Uni+step, Lasso, Elastic	0.08143	0.1381	0.1952
4	miR.ged	Lasso, Elastic		0.6752	0.6655
5	miR.eud	Elastic			0.9434

Table 6.7: *Significant miRNAs found by the various methods on the basis of internal control normalization, first ordered by number of times the given miRNA occurs for each method, and second by the magnitude of OR.*

6.1.5 Mean

The fourth incidence analysis is based on mean normalization. The basic idea is to subtract the mean from each individual sample, which forces the mean C_t for all samples to be aligned at zero. Figure 6.7 verifies that all the C_t averages is practically zero for each sample (the fluctuations should not confuse the reader, notice the magnitude of the y-scale).

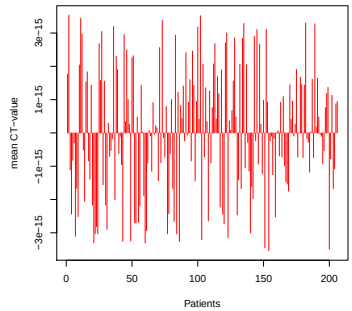

Figure 6.7: *Average C_t level for the 206 persons eligible for analyses after mean normalization.*

The next thing examined is if the p-values obtained from the univariate method are uniformly distributed, i.e. no signal is present in the covariates. Figure 6.8 depicts the empirical cumulative distribution of the p-values along with the cumulative uniform distribution and its 95% confidence bands. The one-sample Kolmogorov-Smirnov test statistic evaluates to $D_n^+ = 0.605$ with a p-value < 0.00001, so there is strong statistical evidence that the null hypothesis can be rejected.

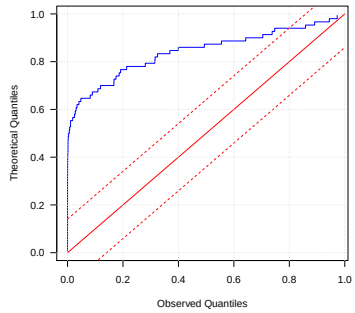

Figure 6.8: *One-sample Kolmogorov-Smirnov test of the univariate p-values for mean normalization.*

20-fold AUC cross-validation found the tuning parameter for the univariate selection to be $\lambda_0 = 0.0000088$ and 19 miRNAs passed this criterion. An IQR standardization was performed on these selected miRNAs, such that only the middle fifty is considered. The relevant miRNAs after the backwards step elimination of the logistic regression is completed, can be seen in Table 6.8.

miRNA	OR	$CI_{0.95,low}$	$CI_{0.95,high}$	p-value
miR.lld	0.220	0.090	0.470	**<0.001**
miR.tyc	5.970	2.680	14.530	**<0.001**
miR.egd	7.280	3.000	20.230	**<0.001**

Table 6.8: *The odds ratio, 95% confidence limits and p-values of the significant miRNAs, derived from the backwards step procedure for mean normalized data.*

The table shows that miR-lld reduces the odds of having cancer on one unit IQR C_t increase by $1/0.22 = 4.55$, where the miRNA also showed the same tendency with internal normalization. The miR-tyc was found by the quantile normalization before, but miR-egd have not previously been included in the final model after the step procedure. Here they both raise the odds of having cancer on a unit increase. All this indicates yet again that the univariate method is very sensitive, and that the normalization method have great influence on which miRNAs that are considered significant candidates in the final model.

The mean normalized data was also analyzed by the Lasso and elastic net regression. Lasso used a penalty parameter of $\lambda_1 = 1.96$ and the elastic net determined the optimal tuning parameters to be $(\lambda_1, \lambda_2) = (1.72, 2.95)$.

	miRNA	Found by	Uni+step	Lasso	Elastic
1	miR.egd	Uni+step, Lasso, Elastic	7.2846	3.5511	1.8489
2	miR.tyc	Uni+step, Lasso, Elastic	5.9710	2.7206	1.7213
3	miR.lld	Uni+step, Lasso, Elastic	0.2159	0.4166	0.5206
4	miR.gyc	Lasso, Elastic		2.2403	1.5239
5	miR.pdd	Lasso, Elastic		1.5111	1.6064
6	miR.czc	Lasso, Elastic		1.1238	1.2732
7	miR.ecd	Lasso, Elastic		1.0425	1.4277
8	miR.wyc	Lasso, Elastic		1.0036	1.0744
9	miR.ged	Lasso, Elastic		0.9054	0.8130
10	miR.omd	Elastic			1.2516
11	miR.wfd	Elastic			1.0736
12	miR.lyc	Elastic			1.0439
13	miR.xad	Elastic			1.0303

Table 6.9: *Significant miRNAs found by the various methods on the basis of mean normalization, first ordered by number of times the given miRNA occurs for each method, and second by the magnitude of OR.*

The top three miRNAs in Table 6.9 are found by all the shrinkage methods, raising the trust in them being the most prominent miRNAs for predicting cancer (for mean normalization). One curious observation is that miR-ecd, which showed to be extremely significant in other normalization methods, here has an OR barely over 1. This stresses the sensibility of these analyses performed, where one of the reasons is the low number of samples available.

6.1.6 Mean-120

The final normalization method resembles somewhat the previous; the mean-120 normalization. Only the mean of 120 most expressed miRNAs are now considered, i.e. the 120 miRNAs with the lowest average C_t value. The result of this normalization can be seen in Figure 6.9.

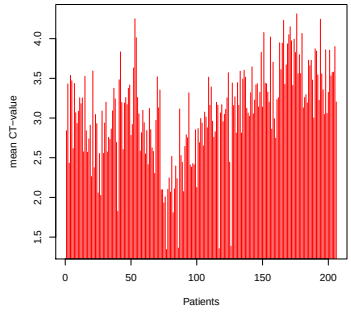

Figure 6.9: *Average C_t level for the 206 persons eligible for analyses after mean-120 normalization.*

It seems like the mean C_t of the samples have leveled out more compared to the original data, however there is still traces of differences present. This is not necessarily bad since the true picture of how things are supposed to look remains unknown, so it can not be concluded whether this normalization method have worsened the chances of finding the true effects of the miRNAs for these data or not. The cumulative distribution of the p-values from the univariate selection is provided in Figure 6.10.

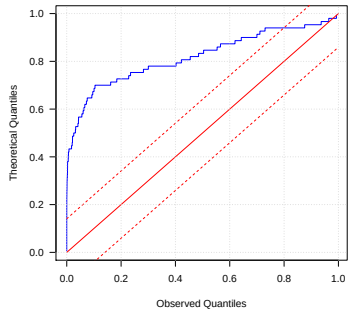

Figure 6.10: *One-sample Kolmogorov-Smirnov test of the univariate p-values for mean-120 normalization.*

Of course the Kolmogorov-Smirnov test statistic have been calculated as in the similar analyses, $D_n^+ = 0.603$ which concludes that the p-values are significantly not following a uniform distribution. The number of miRNAs used for further analysis was 12, based on the p-tolerance of $\lambda_0 = 0.0000046$. The results can be seen in table 6.10.

miRNA	**OR**	**CI**$_{0.95,low}$	**CI**$_{0.95,high}$	**p-value**
miR.ecd	10.530	2.760	47.210	**0.001**
miR.omd	3.830	1.250	13.330	**0.025**
miR.pjd	0.210	0.060	0.710	**0.016**
miR.gyc	6.330	2.410	19.960	**<0.001**
miR.tyc	4.830	1.780	13.970	**0.002**
miR.pdd	2.290	1.150	4.750	**0.021**

Table 6.10: *The odds ratio, 95% confidence limits and p-values of the significant miRNAs, derived from the backwards step procedure for mean-120 normalized data.*

The final logistic model that the backwards step procedure chose, contains a total of six important miRNAs. Worth noticing is the miR-ecd that has an OR of 10.53, this miRNA have been shown to have a high OR estimate before, by different normalization methods. Moreover, five of the six miRNAs have an OR above 1, so they all raise the odds of having pancreas cancer when one unit IQR C_t value is increased for the respective miRNAs.

Table 6.11 gives a summery of the miRNAs found by all shrinkage methods, on the basis of mean-120 normalization.

	miRNA	Found by	Uni+step	Lasso	Elastic
1	miR.ecd	Uni+step, Lasso, Elastic	10.5317	3.5681	3.1417
2	miR.gyc	Uni+step, Lasso, Elastic	6.3271	2.0220	1.9988
3	miR.tyc	Uni+step, Lasso, Elastic	4.8331	1.9011	1.9694
4	miR.omd	Uni+step, Lasso, Elastic	3.8277	1.4193	1.5294
5	miR.pdd	Uni+step, Lasso, Elastic	2.2885	1.5967	1.6204
6	miR.czc	Lasso, Elastic		1.5357	1.5364
7	miR.eud	Lasso, Elastic		0.9381	0.9091
8	miR.lld	Lasso, Elastic		0.6004	0.6034
9	miR.pjd	Uni+step	0.2139		
10	miR.vyc	Elastic			1.0211

Table 6.11: *Significant miRNAs found by the various methods on the basis of mean-120 normalization, first ordered by number of times the given miRNA occurs for each method, and second by the magnitude of OR.*

The penalization used for Lasso was $\lambda_1 = 2.44$, while elastic net applied $(\lambda_1, \lambda_2) = (1.71, 0.899)$ as the optimal amount of shrinkage determined by log-likelihood cross-validation. At this point it can be difficult to see the big picture concerning what miRNAs that should be flagged as good predictors, due to the fact that they have been selected on the basis of multiple normalization methods. Therefore, Section 6.1.7 provides a collection of all the results derived from the incidence analyses in tabular form, making the conclusion easier.

6.1.7 Conclusion

This section is dedicated to summarize on the results regarding incidence. Overall the comparative study revealed that the models derived from the four shrinkage methods, could classify a novel sample into cancer/healthy group with acceptable accuracy. The study was however only conducted for the rank normalization, thus lacking the power to conclude the same for all normalization methods. If this result is simply representing the truth in the miRNA data or if the separation between the cancer/control group is caused by artificial factors, remains unknown. In order to approach an answer, validation studies on other data needs to be performed.

The five normalization methods gave five different results, of course with some similarities, but it could be nice to get a measure of how different they actually perform. One attempt is to look at the p-values calculated from the different

univariate methods, and see how correlated they are across the different normalization methods. Two correlation coefficients seemed relevant in this context; *Pearson's product-moment correlation coefficient* and *Spearman's rank correlation coefficient*. The Pearson coefficient is a measure of the linear dependence of two variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in the interval $[-1; 1]$, where a coefficient of 0 indicates no dependence at all. Mathematically, the Pearson coefficient is calculated from a sample with n observations as

$$r = \frac{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)(x_{2,i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2}\sqrt{\sum_{i=1}^{n}(x_{2,i} - \bar{x}_2)^2}}. \tag{6.3}$$

The Spearman coefficient is another measure of statistical dependence between two variables, somewhat similar to Pearson but based on ranking the variables instead, hence non-parametric. By doing this, the coefficient measures how well the relationship between two variables can be described using a monotonic function. If the two variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are converted to ranked variables $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, the coefficient is defined as

$$\rho = \frac{\sum_i(u_{1,i} - \bar{u}_1)(u_{2,i} - \bar{u}_2)}{\sqrt{\sum_i(u_{1,i} - \bar{u}_1)^2 \sum_i(u_{2,i} - \bar{u}_2)^2}}. \tag{6.4}$$

The upper triangle in Figure 6.11 gives the Pearson and Spearman coefficient between the univariate p-values derived from the five normalization methods. The internal control normalization is definitely the method with the least correlation to the others, the highest Pearson and Spearman coefficient is reached between the ranking method. The lowest Spearman coefficient of 0.08 between internal control and quantile shows that there is almost no monotonic relationship at all. The linear relationship with mean normalization is a low 0.13, indicating almost no linear dependence present. All this shows that the internal control normalization seemingly is the least favorable choice of the five, regarding incidence analysis. After looking at Figure 6.5 again this conclusion makes sense, since this normalization clearly does not make the samples more comparable. Besides the internal control normalization, the methods have an acceptable relationship between their univariate p-values, with an exception of the mean and rank coefficients.

The lower triangle in Figure 6.11 is another way to visualize the results. Considering one of the small plots, the x- and y-axis are the p-values plotted on the logit scale for two methods, respectively. The red dotted lines separating into four quadrants represents a p-value of 0.001, i.e. logit(0.001)=-6.906755. A point in the 1$^{\text{st}}$ and 3$^{\text{rd}}$ quadrant means that the methods agree on the p-value being higher than 0.001 or lower, respectively. In the 2$^{\text{nd}}$ quadrant the method "on top" evaluates the p-value > 0.001 while the method "on the right" says < 0.001, and vice versa in the 4$^{\text{th}}$ quadrant. If all the p-values laid on the green line there would be perfect linear dependence between the methods, and the

Pearson coefficient would evaluate to 1. The solid red line is a monotonic fit, visualizing the Spearman coefficient.

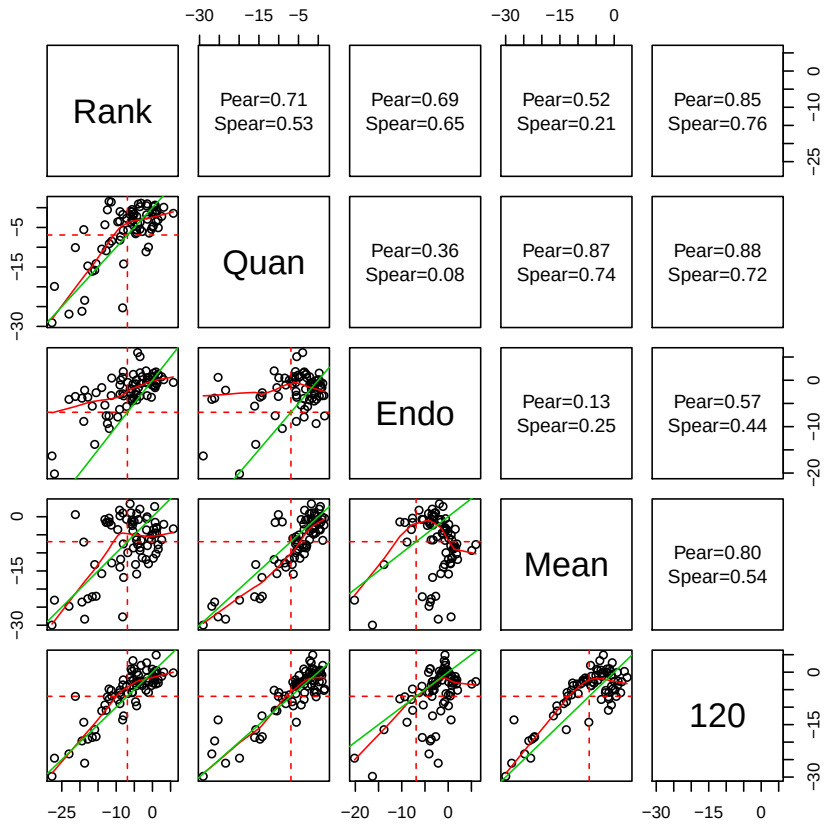

Figure 6.11: *Pairs plot of the p-values obtained from the univariate logistic regression, for the different normalization methods.*

Table 6.12 provides a summary of all the tuning parameters that has been used to obtain the final models. Generally the penalty parameters are higher for the rank method compared to others.

The most interesting thing is to find the subset of miRNAs that can be described as the most significant predictors. Since there are many candidate variables and not that many samples, the sensibility of the analyses is high (lack of power) and each normalization method proposes their individual set of relevant

Method	λ_0	$\lambda_{1,l}$	$\lambda_{1,e}$	$\lambda_{2,e}$
Rank	0.000104	3.596720	3.601383	5.936490
Quantile	0.001136	1.520203	0.829052	2.405713
Intctrl	0.000913	1.298118	1.308134	0.837915
Mean	0.000009	1.959673	1.723428	2.949099
Mean-120	0.000005	2.435876	1.708871	0.899139

Table 6.12: *Optimal tuning parameters used in the shrinkage methods for various normalization.*

miRNAs. The natural thing to do in this case is to trust the miRNAs chosen by most methods. Table 6.13 is motivated by this reasoning, it gives an overview of all the miRNAs selected by the backwards elimination procedure for each normalization method, where the miRNAs have been ordered by the number of methods that have them in common. The miRNAs miR-ecd, miR-tyc, miR-cgd, miR-pdd and miR-lld have all been found by two or more methods, where miR-ecd, miR-tyc and miR-pdd increases the odds of having pancreatic cancer on an IQR unit increase, while miR-cgd and miR-lld decreases the odds. Once again, it should be stressed that the ORs should be taken lightly, because they have been determined on the basis of relatively few observations which is also expressed by the large CIs.

Similar overview are provided for the Lasso regression in Table 6.14 and for the naïve elastic net in Table 6.15. These shrinkage methods are considering even more miRNAs in the final model, so only those having three or more in common are found interesting. The top-5 ranked miRNAs from the step procedure are all found in the top of the Lasso and elastic net as well, leading to the conclusion that these are in fact the most important miRNAs in relation to incidence. Other miRNAs such as miR-ged, miR-omd, miR-eud and miR-czc, show signs of being significant predictors, especially in the elastic net regression, however the incidence analyses top-5 miRNAs remain; miR-ecd, miR-tyc, miR-cgd, miR-pdd and miR-lld.

miRNA	OR_{rank} (95% CI)	p-value	OR_{quan} (95% CI)	p-value	OR_{endo} (95% CI)	p-value	OR_{mean} (95% CI)	p-value	OR_{120} (95% CI)	p-value	no
miR.ecd	13.84 (4.39-50.42)	**<0.001**	18.99 (6.6-65.7)	**<0.001**	44.72 (13.85-197.48)	**<0.001**			10.53 (2.76-47.21)	**0.001**	4
miR.tyc			3.71 (1.37-11.54)	**0.014**			5.97 (2.68-14.53)	**<0.001**	4.83 (1.78-13.97)	**0.002**	3
miR.cgd	0.25 (0.11-0.53)	**<0.001**	0.24 (0.1-0.46)	**<0.001**	0.23 (0.06-0.74)	**0.019**					3
miR.pdd			9.25 (3.45-27.99)	**<0.001**					2.29 (1.15-4.75)	**0.021**	2
miR.lld					0.08 (0.02-0.22)	**<0.001**	0.22 (0.09-0.47)	**<0.001**			2
miR.pjd									0.21 (0.06-0.71)	**0.016**	1
miR.omd									3.83 (1.25-13.33)	**0.025**	1
miR.odd	0.19 (0.06-0.53)	**0.002**									1
miR.myc	5.41 (1.95-18.73)	**0.003**									1
miR.gyc									6.33 (2.41-19.96)	**<0.001**	1
miR.egd							7.28 (3-20.23)	**<0.001**			1

Table 6.13: *Overview of the significant miRNAs found by the univariate selection + backwards elimination procedure for different normalization methods, ordered by the number in common.*

miRNA	OR_{rank}	OR_{quan}	OR_{endo}	OR_{mean}	OR_{120}	no
miR.lld	0.77	0.55	0.14	0.42	0.60	5
miR.ecd	2.87	4.75	21.54	1.04	3.57	5
miR.tyc	1.37	2.13		2.72	1.90	4
miR.pdd		2.61		1.51	1.60	3
miR.ged		0.82	0.68	0.91		3
miR.cgd	0.37	0.31	0.40			3
miR.wyc		1.32		1.00		2
miR.omd	1.53				1.42	2
miR.odd	0.44	0.57				2
miR.gyc				2.24	2.02	2
miR.eud	0.89				0.94	2
miR.egd		1.98		3.55		2
miR.czc				1.12	1.54	2
miR.wzc	0.96					1
miR.myc	1.37					1
miR.kmd	0.61					1

Table 6.14: *Overview of the significant miRNAs found by the univariate selection + Lasso for different normalization methods, ordered by the number in common.*

miRNA	OR_{rank}	OR_{quan}	OR_{endo}	OR_{mean}	OR_{120}	no
miR.lld	0.75	0.57	0.20	0.52	0.60	5
miR.ecd	1.67	2.50	11.78	1.43	3.14	5
miR.tyc	1.26	1.73		1.72	1.97	4
miR.omd	1.56	1.18		1.25	1.53	4
miR.gyc	1.04	1.08		1.52	2.00	4
miR.pdd		1.95		1.61	1.62	3
miR.ged		0.75	0.67	0.81		3
miR.eud	0.88		0.94		0.91	3
miR.czc		1.06		1.27	1.54	3
miR.cgd	0.55	0.40	0.52			3
miR.wyc		1.31		1.07		2
miR.vyc	1.02				1.02	2
miR.odd	0.59	0.52				2
miR.myc	1.28	1.17				2
miR.kyc	1.10	1.08				2
miR.kmd	0.63	0.99				2
miR.egd		1.89		1.85		2
miR.zbd	0.97					1
miR.xad				1.03		1
miR.wzc	0.94					1
miR.wfd				1.07		1
miR.vzc		0.93				1
miR.pjd	1.12					1
miR.lyc				1.04		1

Table 6.15: *Overview of the significant miRNAs found by the univariate selection + elastic net for different normalization methods, ordered by the number in common.*

6.2 Prognosis

The equivalent of the incidence analyses are now performed on survival data, where time from operation to death is the outcome of interest. This means that the number of samples available is reduced substantially, because only pancreatic cancer patients have experienced an operation (plus one patient with CP) as seen in Table 3.3. Unfortunately about one third of these cases additionally have to be removed from the analyses, which is a consequence of missing information about their death or follow-up date. Ultimately this gives a population of 93 patients with survival data, under half of what was available in the incidence analyses. The step-by-step analysis procedure for the prognosis part is summarized in the following.

Step 1 Data is normalized by the chosen normalization method.

Step 2 The univariate selection method for the Cox proportional hazards model is applied to the normalized data set in order to find a subset of miRNAs for further analysis. The p-value tolerance λ_0 controls the number of miRNAs in this set, and is obtained by K-fold partial log-likelihood cross-validation. The CV measure is calculated as in the Bøvelstad et al. [2007] paper, which is given by

$$CV(\lambda_0) = \sum_{i=1}^{K} \ell[\boldsymbol{\beta}_{(-i)}(\lambda_0)] - \ell_{(-i)}[\boldsymbol{\beta}_{(-i)}(\lambda_0)]. \tag{6.5}$$

The parameter λ_0 chosen is the one maximizing $CV(\lambda_0)$.

Step 3 The selected miRNAs are standardized by their *interquartile range* (IQR), i.e. divided by the distance between the 75^{th} percentile and the 25^{th} percentile. This provides more robust estimates of the coefficients, since outlier measurements are disregarded.

Step 4 Cox proportional hazards regression is performed on the complete cases in a backwards elimination procedure, where the BIC is used as a model evaluator. The proportional hazards assumption is tested for each of the miRNAs included in the final model, and the model as a whole.

Step 5 Moreover, Lasso regression is performed with the miRNAs selected by the univariate method, as the starting model. The penalty parameter λ_1 is determined by 20-fold partial log-likelihood cross-validation, from the built-in function `optL1` of the `penalized` package.

Step 6 Furthermore, Cox proportional hazards regression with elastic net penalty is performed with the univariate selected miRNAs. The tuning parameters λ_1 and λ_2 are once again determined by 20-fold partial log-likelihood cross-validation, using the `cvl` function from the `penalized` package.

6.2.1 Explorative analysis

It is always a good idea to explore the data properly before moving on to the analyses, in order to get an idea of how data behave and maybe discover irregular trends. In the context of survival analysis, one appropriate possibility is to look at the survival as a function of time. Figure 6.12 shows the Kaplan-Meier estimator of the overall survival for the whole population ($n = 93$), i.e. without adjusting for effects of age, sex, other cancers etc. The crosses indicate patients who have been censored.

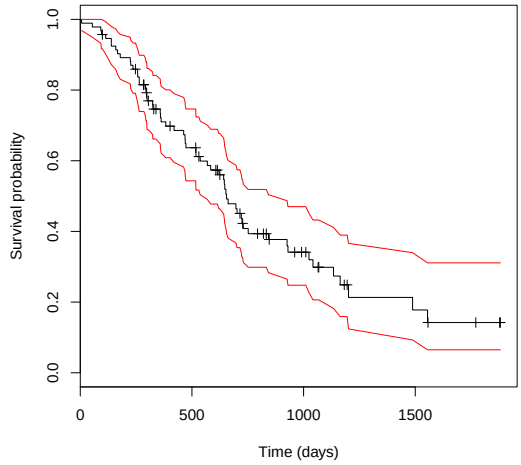

Figure 6.12: *Kaplan-Meier estimator of the survival function along with a 95% confidence band, for the serum data.*

After two years, more than half of the population have died or left the study, stressing the fact that pancreatic cancer is a lethal disease. The literature reports that the overall survival after 5 years is less than 5%, which seems to be in good agreement with this small sample population. It is hard to say precisely though, because of the sparse number of observations, which the wider spread of the 95% confidence band in the end supports. Besides survival probability, it is also interesting to look at the evolvement of the cumulative hazard over time, thus the Nelson-Aalen estimator is plotted in Figure 6.13.

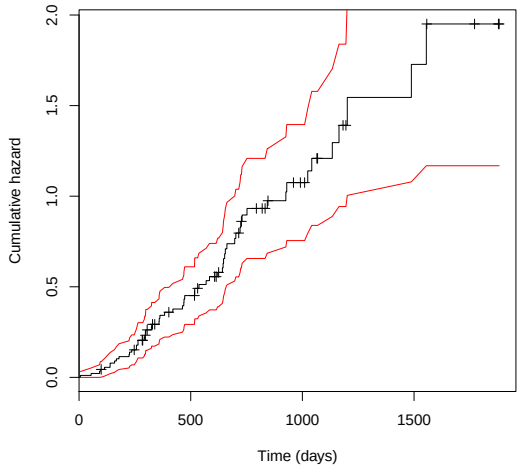

Figure 6.13: *Nelson-Aalen estimator of the cumulative hazard function along with a 95% confidence band, for the serum data.*

Here it seems like the hazard is close to being constant over time, since the shape of the cumulative function is neither concave (mortality rate decreases over time) or convex (mortality rate increases over time). Of course there is large uncertainty after approximately 4 years, due to the low number of samples. Next in Section 6.2.2, the comparative study for survival data is performed.

6.2.2 Comparative study

Analogously to the incidence case, a small comparative studies examines the shrinkage methods' ability to predict based on miRNAs, where prediction in this context is a cancer patient's survival. The same four shrinkage methods are being tested, this time by applying the Cox proportional hazards model. The paper from Bøvelstad et al. [2007] has once again laid the ground for inspiration of the comparative study design. The training/test set split was 2:1 for the incidence case, but because there is much fewer observations available for the prognostic analyses a 3:1 split was used instead, otherwise the lack of power in the analyses would become a problem. Bøvelstad et al. [2007] proposes three model evaluation criteria and the one chosen here is based on the *prognostic index* (PI). This is defined as the linear predictor for the patients in the test set

based on the coefficients estimated for the training set

$$\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{train}. \tag{6.6}$$

This PI is then used as a single continuous covariate in a Cox regression on the test set, i.e. the fitted model becomes

$$h_i(t) = h_0(t)\exp(\hat{\eta}_i\alpha) \tag{6.7}$$

where i is an index over the patients in the test set. To get an idea of the method's performance, the following hypothesis is tested

$$\mathcal{H}_0 : \alpha = 0 \tag{6.8}$$
$$\mathcal{H}_1 : \alpha \neq 0 \tag{6.9}$$

using likelihood ratio test, and the p-value obtained serves as a performance measure. Hopefully the p-value is as low as possible rejecting the null hypothesis, meaning that the linear predictor is significant for predicting expected survival. The optimal tuning parameters are determined for each training/test set split by 10-fold partial log-likelihood cross-validation. To get reliable results 50 iterations were run where a summary of these results are provided in Table 6.16.

k	λ_0	$\lambda_{1,l}$	$\lambda_{2,r}$	$\lambda_{1,e}$	$\lambda_{2,e}$	PI_u	PI_l	PI_r	PI_e
mean	0.0683674	29.4919	2826.9206	29.4705	2826.9224	0.49830	0.54278	0.39070	0.499781
sd	0.0516630	23.5411	3181.7506	23.6368	3181.7385	0.31876	0.34399	0.30225	0.350106
min	0.0008356	3.5667	9.4786	3.8830	9.8719	0.03981	0.03094	0.00650	0.009887
max	0.2471353	130.0868	10724.8107	131.0886	10724.8108	1.00000	1.00000	0.99275	1.000000

Table 6.16: *Comparison of the prediction performance for the four shrinkage methods, on the basis of likelihood ratio test.*

The results are very discouraging. The mean p-values for each method is close to 0.5, Ridge lying a bit under, all ranging from close to zero to one. This means that there is not enough statistical evidence to reject the null hypothesis on average. Either this is the result of miRNAs not saying anything about the expected survival for pancreatic cancer patients, or maybe there is simply to few samples to make any reasonable conclusion. To get an idea of the first presumption a simulation study should have been done to see if the true effects can be detected by miRNAs. Unfortunately there were not enough time to examine this further. The method that performs best is Ridge regression where the same conclusion was reached in Bøvelstad et al. [2007] as well, however in terms of prognostic prediction all methods fail for these data which should be kept in mind when performing the Cox regression analyses.

6.2.3 Rank

Even though the comparative study showed no sign of the miRNAs being capable of predicting survival after operation, the prognostic analyses still seemed relevant. First thing is to test for signal in the p-values derived from the univariate Cox model. Here the Kolmogorov-Smirnov test statistic becomes $D_n^+ = 0.577$ which rejects the null hypothesis with a p-value < 0.00001. Figure 6.14 verifies this results since the cumulative distribution for the p-values is higher than the one of the uniform distribution.

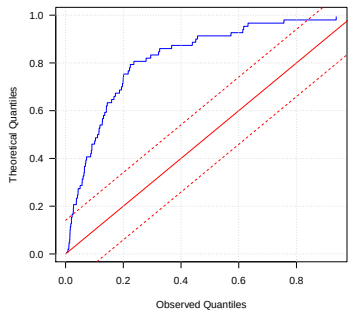

Figure 6.14: *One-sample Kolmogorov-Smirnov test of the univariate p-values for rank normalization.*

The p-value tolerance was obtained by partial log-likelihood CV and for the rank normalized data $\lambda_0 = 0.0431140$. There were 22 miRNAs selected as potential candidates and the result after running the backwards elimination procedure can be seen in Table 6.17.

miRNA	**HR**	$\mathbf{CI}_{0.95,low}$	$\mathbf{CI}_{0.95,high}$	**p-value**
miR.mmd	0.660	0.470	0.940	**0.020**
miR.wyc	1.840	1.200	2.810	**0.005**

Table 6.17: *The hazard ratio, 95% confidence limits and p-values of the significant miRNAs, derived from the backwards step procedure for rank normalized data.*

Only two miRNAs are found significant enough to stay in the final model. The miRNA miR-mmd has an hazard ratio of 0.66 indicating that on an IQR rank unit increase of this miRNA, the hazard of dying is decreased. For miR-wyz the hazard of dying is increased by a factor 1.84 on a unit increase, i.e. the expected

survival is worsened. The validity of the proportional hazards assumption for this final Cox model has been tested in Table 6.18.

	ρ	χ^2	p-value
miR.mmd	-0.038	0.113	0.737
miR.wyc	-0.108	0.767	0.381
GLOBAL		1.120	0.571

Table 6.18: *Test of the proportional hazards assumption for each significant miRNA in the model fit for rank normalization, along with a global test.*

First column is a correlation coefficient between the transformed survival time and scaled Schoenfeld residuals, second the χ^2 test statistic of the slope being zero and third the p-value of the test. Both the miRNAs have a low negative correlation and the p-value suggest that there is no statistical evidence of rejecting the null hypothesis, i.e. the proportional hazards assumption is not violated for any of the miRNAs. Furthermore, the global test of the model as a whole also comes to the conclusion that the proportional hazards assumption is satisfied. The more graphical approach to obtain these model diagnostics plots the scaled Schoenfeld residuals, this is done in Figure 6.15.

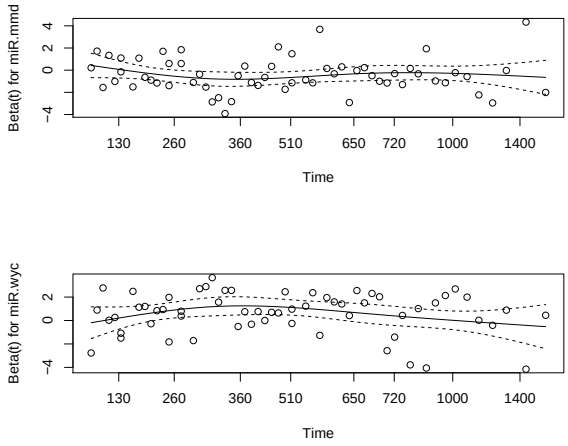

Figure 6.15: *Plot of the scaled Schoenfeld residuals against transformed time for each miRNA in the model fit for rank normalization, along with a four degree fitted natural spline (solid line) and its ±2 standard error confidence band (dashed lines).*

If violation of the proportional hazards assumption is to be perfectly rejected, then the smoothing function would be linear on the horizontal line of the log(HR). The lines would be at -0.416 for miR-mmd and 0.61 for miR-wyz, respectively. Despite minor fluctuations of the spline from these values there is clearly no danger of the assumption being violated.

Equivalently to the incidence case, Lasso regression and elastic net regression were also applied in the prognostic case. For rank normalized data the penalty factor $\lambda_1 = 3.15$ was the optimal choice found by 20-fold partial log-likelihood cross-validation. In the elastic net regression the two parameters were found to be $(\lambda_1, \lambda_2) = (3.21, 155)$, and the miRNAs found by all the regression analyses are summarized in Table 6.19 ordered by the miRNAs in common and the size of the HRs.

	miRNA	**Found by**	**Uni+step**	**Lasso**	**Elastic**
1	miR.wyc	Uni+step, Lasso, Elastic	1.8392	1.3992	1.0365
2	miR.mmd	Uni+step, Lasso, Elastic	0.6628	0.7845	0.9585
3	miR.dbd	Lasso, Elastic		1.0480	1.0364
4	miR.bkd	Lasso, Elastic		0.9107	0.9747
5	miR.fed	Lasso, Elastic		0.8940	0.9708
6	miR.tdd	Elastic			1.0221
7	miR.lyc	Elastic			1.0193
8	miR.wad	Elastic			1.0173
9	miR.dzc	Elastic			1.0130
10	miR.qzc	Elastic			1.0117
11	miR.uyc	Elastic			1.0091
12	miR.lzc	Elastic			1.0086
13	miR.kzc	Elastic			1.0030
14	miR.hyc	Elastic			1.0026
15	miR.zmd	Elastic			0.9929
16	miR.zvd	Elastic			0.9921
17	miR.nwd	Elastic			0.9873
18	miR.cae	Elastic			0.9827
19	miR.wtd	Elastic			0.9806
20	miR.vod	Elastic			0.9799

Table 6.19: *Significant miRNAs found by the various methods on the basis of ranks, first ordered by number of times the given miRNA occurs for each method, and second by the magnitude of HR.*

6.2.4 Quantile

This section deals with the results of quantile normalized data analyzed on the basis of Cox proportional hazards model. The test of p-values from the initial univariate selection method with $\lambda_0 = 0.0500952$ is $D_n^+ = 0.492$. The Kolmogorov-Smirnov test can be interpreted as the p-values derived are not drawn at random from a uniform distribution, which is also illustrated in Figure 6.16.

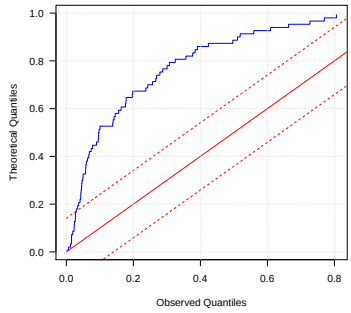

Figure 6.16: *One-sample Kolmogorov-Smirnov test of the univariate p-values for quantile normalization.*

There were 24 miRNAs selected for further analysis for which IQR standardization was performed, however only miR-qzc was included in the final model as seen in Table 6.20. When miR-qzc is increased one IQR C_t unit the risk of dying is 1.52. times greater.

miRNA	**HR**	$\mathbf{CI}_{0.95,low}$	$\mathbf{CI}_{0.95,high}$	**p-value**
miR.qzc	1.520	1.090	2.110	**0.013**

Table 6.20: *The hazard ratio, 95% confidence limits and p-values of the significant miRNAs, derived from the backwards step procedure for quantile normalized data.*

The assumption of the Cox model is also tested with the results provided in Table 6.21. The test indicate with statistical significance that the assumption of the hazards being proportional is not violated.

	ρ	χ^2	p-value
miR.qzc	-0.144	0.834	0.361

Table 6.21: *Test of the proportional hazards assumption for each significant miRNA in the model fit for quantile normalization, along with a global test.*

The Lasso regression was performed with $\lambda_1 = 6.74$, while the elastic net regression used $(\lambda_1, \lambda_2) = (6.62, 140)$. Not many miRNAs are found by the backwards elimination or the Lasso, it is clearly the elastic net that finds the majority. It should be noticed though that the HRs estimated by elastic net regression are fairly close to 1.

	miRNA	Found by	Uni+step	Lasso	Elastic
1	miR.qzc	Uni+step, Lasso, Elastic	1.521	1.2035	1.0599
2	miR.mmd	Lasso, Elastic		0.7917	0.9552
3	miR.wad	Elastic			1.0193
4	miR.kzc	Elastic			1.0114
5	miR.lyc	Elastic			1.0106
6	miR.dbd	Elastic			1.0091
7	miR.tdd	Elastic			1.0028
8	miR.fed	Elastic			0.9928

Table 6.22: *Significant miRNAs found by the various methods on the basis of quantile normalization, first ordered by number of times the given miRNA occurs for each method, and second by the magnitude of HR.*

6.2.5 Internal control

The third analysis is based on an internal control and it was this normalization method that showed most signs of being the odd one out in the incidence analyses. The result of the Kolmogorov-Smirnov test for this normalization method is seen in Figure 6.17, the null hypothesis is rejected with a p-value < 0.00001 derived from the test statistic $D_n^+ = 0.536$.

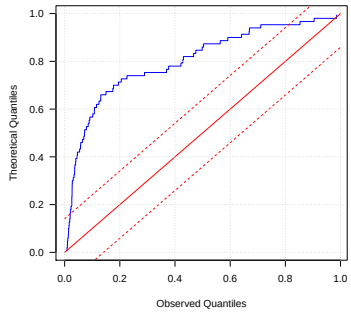

Figure 6.17: *One-sample Kolmogorov-Smirnov test of the univariate p-values for internal control normalization.*

Under the p-tolerance criteria $\lambda_0 = 0.0362152$ there were 25 miRNAs included in the starting model of the step procedure. The three miRNAs of the final model are summarized in Table 6.23.

miRNA	**HR**	$\mathbf{CI}_{0.95,low}$	$\mathbf{CI}_{0.95,high}$	**p-value**
miR.wad	8.450	1.550	46.080	**0.014**
miR.dzc	35.750	2.230	572.970	**0.012**
miR.vzc	0.010	0.000	0.350	**0.010**

Table 6.23: *The hazard ratio, 95% confidence limits and p-values of the significant miRNAs, derived from the backwards step procedure for internal control normalized data.*

Apparantly very extreme HRs are estimated, especially for miR-dzc and miR-vzc. If these results were to be trusted then one increase in miR-dzc would increase the hazard of dying by an 35.75 fold. On the other hand, by one unit IQR C_t raise in miR-vzc the hazard of dying decreases by $1/0.01 = 100$. From a practical perspective this seems unreasonable and the wide confidence intervals of these estimate stresses the large uncertainty connected with these results. The propotional hazards test of the model in Table 6.24 shows no sign of violation, either for the miRNAs individually or for the model as a whole. Table 6.25 lists all miRNAs found in the backwards regression procedure, Lasso regression with $\lambda_1 = 7.58$ and regression with the elastic net penalty $(\lambda_1, \lambda_2) = (6.57, 184)$, all based on internal control normalization.

	ρ	χ^2	p-value
miR.wad	0.098	0.319	0.572
miR.dzc	-0.124	0.692	0.405
miR.vzc	0.076	0.242	0.623
GLOBAL		2.145	0.543

Table 6.24: *Test of the proportional hazards assumption for each significant miRNA in the model fit for internal control normalization, along with a global test.*

	miRNA	Found by	Uni+step	Lasso	Elastic
1	miR.dzc	Uni+step, Elastic	35.74530		1.003
2	miR.wad	Uni+step, Elastic	8.45097		1.006
3	miR.vzc	Uni+step, Elastic	0.01318		1.001
4	miR.pjd	Lasso, Elastic		1.146	1.017
5	miR.wed	Lasso, Elastic		1.056	1.013
6	miR.lyc	Elastic			1.012
7	miR.kyc	Elastic			1.011
8	miR.uyc	Elastic			1.009
9	miR.lmd	Elastic			1.008
10	miR.tdd	Elastic			1.007
11	miR.ddd	Elastic			1.006
12	miR.kbd	Elastic			1.006
13	miR.dad	Elastic			1.005
14	miR.eed	Elastic			1.005
15	miR.dbd	Elastic			1.004
16	miR.gcd	Elastic			1.003
17	miR.gyc	Elastic			1.003
18	miR.hed	Elastic			1.003
19	miR.kzc	Elastic			1.003
20	miR.myc	Elastic			1.002
21	miR.fud	Elastic			1.001
22	miR.hyc	Elastic			1.001
23	miR.vyc	Elastic			1.000

Table 6.25: *Significant miRNAs found by the various methods on the basis of internal control normalization, first ordered by number of times the given miRNA occurs for each method, and second by the magnitude of HR.*

6.2.6 Mean

The analysis is repeated for the next normalization method which is the mean normalization and as always the p-values are tested with a one-sided one-sample Kolmogorov-Smirnov test. Here $D_n^+ = 0.469$ and the null hypothesis that the cumulative distribution of the p-values is uniform is rejected with p-value < 0.00001.

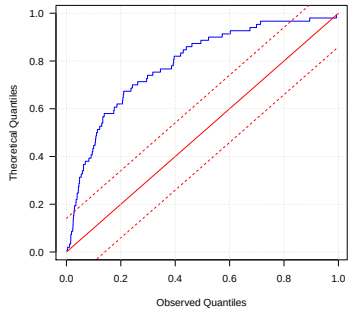

Figure 6.18: *One-sample Kolmogorov-Smirnov test of the univariate p-values for mean normalization.*

25 miRNAs passed a tolerance of $\lambda_0 = 0.0586029$, where the significant miRNAs in the final model is found in Table 6.26. As opposed to the previous prognostic analyses with other normalization methods, the number of miRNAs in the final Cox proportional hazards model is noticable larger. The miR-vhd increases the hazard of dying the most on one unit IQR C_t increase, i.e reduced expansion, but also miR-ddd and miR-lmd raises the hazard, strangely enough with the precise same quantity. The remaining miR-mmd, miR-myc and miR-zvd all contribute to a decrease in the hazard of dying when their amount of miRNA material is decreased. Actually miR-mmd was also chosen by the analysis based on ranks.

miRNA	HR	$CI_{0.95,low}$	$CI_{0.95,high}$	p-value
miR.mmd	0.280	0.120	0.650	**0.003**
miR.myc	0.180	0.050	0.620	**0.007**
miR.ddd	2.300	1.100	4.820	**0.027**
miR.zvd	0.240	0.080	0.700	**0.009**
miR.lmd	2.300	1.210	4.370	**0.011**
miR.vhd	11.260	2.790	45.390	**<0.001**

Table 6.26: *The hazard ratio, 95% confidence limits and p-values of the significant miRNAs, derived from the backwards step procedure for mean normalized data.*

The important test of proportional hazards assumption for the model is summarized in Table 6.27. Generally speaking everything looks fine because the global test rejects the alternative hypothesis, but miR-mmd is on a significance level of 5% violating the proportional hazards assumption.

	ρ	χ^2	p-value
miR.mmd	-0.259	4.766	**0.029**
miR.myc	-0.136	1.076	0.300
miR.ddd	0.131	0.788	0.375
miR.zvd	-0.011	0.006	0.936
miR.lmd	0.068	0.311	0.577
miR.vhd	0.126	0.926	0.336
GLOBAL		5.012	0.542

Table 6.27: *Test of the proportional hazards assumption for each significant miRNA in the model fit for mean normalization, along with a global test.*

The analysis ends with an overview in Table 6.28 of the miRNAs found by the various Cox regression techniques. The Lasso method penalized with $\lambda_1 = 6.04$ and the elastic net with $(\lambda_1, \lambda_2) = (5.79, 202)$.

	miRNA	Found by	Uni+step	Lasso	Elastic
1	miR.ddd	Uni+step, Lasso, Elastic	2.3008	1.0067	1.0141
2	miR.lmd	Uni+step, Lasso, Elastic	2.2998	1.0590	1.0168
3	miR.mmd	Uni+step, Lasso, Elastic	0.2824	0.6954	0.9623
4	miR.zvd	Uni+step, Lasso, Elastic	0.2406	0.9997	0.9748
5	miR.lyc	Lasso, Elastic		1.0386	1.0192
6	miR.dzc	Lasso, Elastic		1.0015	1.0105
7	miR.vhd	Uni+step	11.2624		
8	miR.myc	Uni+step	0.1813		
9	miR.wad	Elastic			1.0166
10	miR.fud	Elastic			1.0102
11	miR.kzc	Elastic			1.0074
12	miR.tdd	Elastic			1.0073
13	miR.dbd	Elastic			1.0052
14	miR.uyc	Elastic			1.0029
15	miR.fed	Elastic			0.9971
16	miR.lld	Elastic			0.9939
17	miR.nwd	Elastic			0.9898
18	miR.wtd	Elastic			0.9891
19	miR.vjd	Elastic			0.9850

Table 6.28: *Significant miRNAs found by the various methods on the basis of mean normalization, first ordered by number of times the given miRNA occurs for each method, and second by the magnitude of HR.*

6.2.7 Mean-120

The mean-120 is the fifth and last normalization method performed on the serum data regarding survival analysis of time from operation to death/end of follow-up. So far none of the p-values derived from the univariate selection with different normalization methods showed signs of being drawn at random from a uniform distribution. This is neither the case for mean-120 as seen in Figure 6.19, the statistic $D_n^+ = 0.518$ is with statistical certainty rejecting the null hypothesis with p-value < 0.00001 in a one-sample Kolmogorov-Smirnov test.

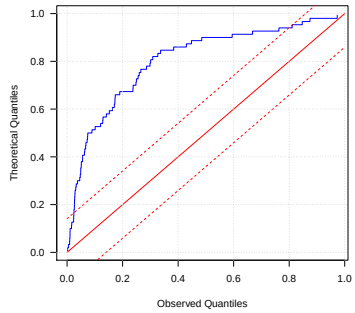

Figure 6.19: *One-sample Kolmogorov-Smirnov test of the univariate p-values for mean-120 normalization.*

IQR standardization was performed on the 15 eligible candidate miRNAs for the starting model in the backwards elimination procedure. These were chosen from having a p-value below the criteria $\lambda_0 = 0.0261484$. The resulting model from the backwards stepwise selection consisted of a single miRNA; miR-mmd which was found in previous analyses as well. Table 6.29 gives the HR, 95% confidence bands and the p-value for miR-mmd.

miRNA	**HR**	$\mathbf{CI}_{0.95,low}$	$\mathbf{CI}_{0.95,high}$	**p-value**
miR.mmd	0.580	0.410	0.820	**0.002**

Table 6.29: *The hazard ratio, 95% confidence limits and p-values of the significant miRNAs, derived from the backwards step procedure for mean-120 normalized data.*

The model indicates that if the IQR C_t level goes up one unit for this miRNA, i.e. the miRNA expansion decreases, then the hazard of dying is decreased by a $1/0.58 = 1.72$ fold. Furthermore, the assumption of proportional hazards seem to hold for this model, which can be seen from Table 6.30.

	$\boldsymbol{\rho}$	$\boldsymbol{\chi}^2$	**p-value**
miR.mmd	-0.037	0.048	0.826

Table 6.30: *Test of the proportional hazards assumption for each significant miRNA in the model fit for mean-120 normalization, along with a global test.*

Finally the miRNAs found by Lasso regression with the penalyty factor $\lambda_1 = 9.33$ and the elastic net regression with $(\lambda_1, \lambda_2) = (8.4, 82.5)$ are provided in Table 6.31. Here the list of significant miRNAs for predicting survival is quite short compared to those of other normalization methods.

	miRNA	Found by	Uni+step	Lasso	Elastic
1	miR.mmd	Uni+step, Lasso, Elastic	0.5831	0.7655	0.9212
2	miR.lyc	Elastic			1.0327
3	miR.tdd	Elastic			1.0241
4	miR.wad	Elastic			1.0123
5	miR.dzc	Elastic			1.0033

Table 6.31: *Significant miRNAs found by the various methods on the basis of mean-120 normalization, first ordered by number of times the given miRNA occurs for each method, and second by the magnitude of HR.*

This concludes the presentation of results derived from five different normalization methods. However, Section 6.2.8 provides a recapitulation of all the results in a more comparative manner, such that the greater overview is achieved.

6.2.8 Conclusion

This section is dedicated to summarize on the results regarding prognosis. The analyses clearly suffered under the scarce number of samples available and separation of variables still being an issue for some of the normalization methods. In the comparative study it was discouraging to see that all the shrinkage methods fail to predict survival. Whether the reason for this is due to miRNAs simply not being useful prognostic predictors or something else, is hard to say for certain. To get an idea of this, simulation studies and similar analyses on other data should be performed. The result of the comparative study did not prevent analyses of serum data though, because there could still be some effects in the miRNAs.

The pairs plot of the p-values derived from the univariate Cox model regression is seen in Figure 6.20. The internal normalization is clearly the method with the least correlation to the others, both in terms of Pearson and Spearman coefficients. Worst case being between internal control and rank, where the Pearson coefficient is practically zero. Between the four other normalization methods, correlation seems fine which is reassuring.

In the small plots, the x- and y-axis are once again the p-values plotted on the logit scale for two methods, respectively. The red dotted lines separating into four quadrants represents a p-value of 0.05, i.e. logit(0.05)=-2.944439. A point in the 1st and 3rd quadrant means that the methods agree on the p-value being higher than 0.05 or lower, respectively. In the 2nd quadrant the method "on top" evaluates the p-value > 0.05 while the method "on the right" says < 0.05, and vice versa in the 4th quadrant.

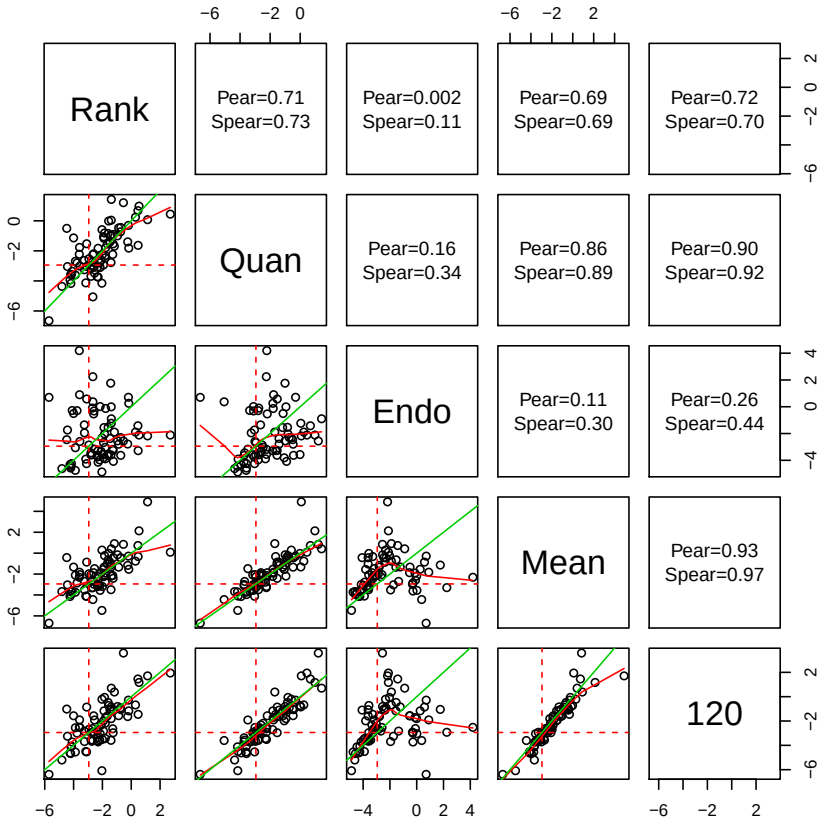

Figure 6.20: *Pairs plot of the p-values obtained from the univariate Cox model, for the different normalization methods.*

Method	λ_0	$\lambda_{1,l}$	$\lambda_{1,e}$	$\lambda_{2,e}$
Rank	0.043114	3.146331	3.208828	154.634503
Quantile	0.050095	6.740118	6.615210	139.665515
Intctrl	0.036215	7.581442	6.574792	183.606733
Mean	0.058603	6.040598	5.789969	201.522153
Mean-120	0.026148	9.326694	8.400988	82.536880

Table 6.32: *Optimal tuning parameters used in the shrinkage methods for various normalization.*

All the tuning parameters used for the various analyses are collected in Table 6.32. Overall the results derived from the five analyses of prognosis seemed somewhat unstable, which was also the conclusion of the comparative study, there is not much statistical evidence of miRNAs being able to predict the expected survival from pancreas operation. Even though the usefulness of miRNA predictors are questionable, the most significant subset can still be pointed out. Motivated by this, Table 6.33 have collected all the miRNAs found in the backwards elimination procedure based on the respective normalization methods. Only miR-mmd is worth mentioning because it was found three times, which seemingly on an IQR unit increase in C_t level reduces the hazard of dying. All other miRNAs are only found in one analysis which is to unreliable, this could be purely coincidence. The exact same conclusion is derived from the Lasso regression results, collected in Table 6.34, only miR-mmd shows repeatedly significance. Table 6.35 contains the miRNAs derived from using the elastic net penalty in Cox regression, here the list is longer and the number of miRNAs in common between different normalization methods are higher. The top-3 are miR-wad, miR-tdd and miR-lyc which are found by all five analyses, but in general not much weight is put on these because the HR estimates are very close to 1, i.e. the effects are close to nothing.

miRNA	HR_{rank} (95% CI)	p-value	HR_{quan} (95% CI)	p-value	HR_{endo} (95% CI)	p-value	HR_{mean} (95% CI)	p-value	HR_{120} (95% CI)	p-value	no
miR.mmd	0.66 (0.47-0.94)	**0.020**					0.28 (0.12-0.65)	**0.003**	0.58 (0.41-0.82)	**0.002**	3
miR.zvd	1.84 (1.2-2.81)	**0.005**					0.24 (0.08-0.7)	**0.009**			1
miR.wyc					8.45 (1.55-46.08)	**0.014**					1
miR.wad					0.01 (0-0.35)	**0.010**					1
miR.vzc							11.26 (2.79-45.39)	**<0.001**			1
miR.vhd			1.52 (1.09-2.11)	**0.013**							1
miR.qzc							0.18 (0.05-0.62)	**0.007**			1
miR.myc							2.3 (1.21-4.37)	**0.011**			1
miR.lnd					35.75 (2.23-572.97)	**0.012**					1
miR.dzc							2.3 (1.1-4.82)	**0.027**			1
miR.ddd											1

Table 6.33: *Overview of the significant miRNAs found by the univariate selection + backwards elimination procedure for different normalization methods, ordered by the number in common.*

miRNA	HR_{rank}	HR_{quan}	HR_{endo}	HR_{mean}	HR_{120}	no
miR.mmd	0.78	0.79		0.70	0.77	4
miR.zvd				1.00		1
miR.wyc	1.40					1
miR.wed			1.06			1
miR.qzc		1.20				1
miR.pjd			1.15			1
miR.lyc				1.04		1
miR.lmd				1.06		1
miR.fed	0.89					1
miR.dzc				1.00		1
miR.ddd				1.01		1
miR.dbd	1.05					1
miR.bkd	0.91					1

Table 6.34: *Overview of the significant miRNAs found by the univariate selection + Lasso for different normalization methods, ordered by the number in common.*

miRNA	HR_{rank}	HR_{quan}	HR_{endo}	HR_{mean}	HR_{120}	no
miR.wad	1.02	1.02	1.01	1.02	1.01	5
miR.tdd	1.02	1.00	1.01	1.01	1.02	5
miR.lyc	1.02	1.01	1.01	1.02	1.03	5
miR.mmd	0.96	0.96		0.96	0.92	4
miR.kzc	1.00	1.01	1.00	1.01		4
miR.dzc	1.01		1.00	1.01	1.00	4
miR.dbd	1.04	1.01	1.00	1.01		4
miR.uyc	1.01		1.01	1.00		3
miR.fed	0.97	0.99		1.00		3
miR.zvd	0.99			0.97		2
miR.wtd	0.98			0.99		2
miR.qzc	1.01	1.06				2
miR.nwd	0.99			0.99		2
miR.lmd			1.01	1.02		2
miR.hyc	1.00		1.00			2
miR.fud			1.00	1.01		2
miR.ddd			1.01	1.01		2
miR.zmd	0.99					1
miR.wyc	1.04					1
miR.wed			1.01			1
miR.vzc			1.00			1
miR.vyc			1.00			1
miR.vod	0.98					1
miR.vjd				0.98		1
miR.pjd			1.02			1
miR.myc			1.00			1
miR.lzc	1.01					1
miR.lld				0.99		1
miR.kyc			1.01			1
miR.kbd			1.01			1
miR.hed			1.00			1
miR.gyc			1.00			1
miR.gcd			1.00			1
miR.eed			1.00			1
miR.dad			1.00			1
miR.cae	0.98					1
miR.bkd	0.97					1

Table 6.35: *Overview of the significant miRNAs found by the univariate selection + elastic net for different normalization methods, ordered by the number in common.*

Chapter 7

Discussion

7.1 Summary of the results

The initial explorative analysis of the serum data discovered distinct average C_t levels between the pancreas cancer and healthy control samples, leading to artificial separation between these groups due to confounding. This was a result of poor DOE and remedies for avoiding confounding in miRNA experiments in the future were provided. The main principles to keep in mind when designing experiments are blocking, replication and randomization when reliable results are to be obtained. If the data at hand does not come from a well planned and executed experiment and nuisance factors tend to dominate the results, then normalization showed to play a crucial role. The conducted simulation study indicated that analysis based on ranks gave overall better prediction models compared to those based on the raw C_t values, where the measure for comparison was the deviance of the final model. However the study was limited in many ways, e.g. considering only a binary outcome (cancer/healthy) and one normalization method (out of five).

The main objective in this thesis was to find two sets of miRNA containing only a few miRNAs out the total 754. The first set consists of predictors of incidence i.e. those miRNAs constituting model that can discriminate samples into either the cancer or healthy group, with the number of misclassifications being as low as possible. The second set is the predictors relating to survival after pancreatic cancer operation, i.e. from looking at the increase or decrease of these miRNA expansions, information about a patient's hazard of dying is provided. The objective is motivated by the difficulties existing today with early and correct diagnostics of pancreatic cancer patients, and the lack of knowledge concerning the probabilistic life duration. The ideal situation where the true subset of miRNAs is found will hopefully lead to improves prognosis by for example earlier diagnosis and treatment.

The incidence analyses are based on the logistic regression which is a member of the GLM family, however due to fewer samples than covariates and the clinical objective in mind, shrinkage methods were applied to reduce dimensionality. The initial comparative study based on ranks examined the performance of the four shrinkage methods considered in this thesis; the univariate selection in combination with backwards stepwise regression, Lasso regression, Ridge regression and the naïve combination of the two called elastic net regression. The results showed that all these methods are able to separate the cancer and healthy patients from each other with high accuracy. Even though all regression techniques are performing at an acceptable level, Ridge regression was omitted in further analyses since it was not found relevant in this context because it does not reduce the number of miRNAs, which from a clinical perspective is needed.

The five normalization methods dealt with in this thesis are; ranks, quantile, internal control, mean and mean-120. There was strong evidence of the internal control normalization being different than the other four since the correlation with the other methods was fairly low. Analyses with different normalization methods resulted in different suggestions to the miRNA set of interest. There existed some overlaps and the miRNAs found in multiple analyses must be considered as those with the strongest signal value. Five miRNAs seemed to stand out and the subset of incidence predictors was found to be miR-ecd, miR-tyc, miR-cgd, miR-pdd and miR-lld.

Similar analyses were performed in the prognostic case, where the semi-parametric Cox proportional hazards model was used. The comparative study of the shrinkage methods did not present results as encouraging as the incidence case. The linear predictor in the Cox model was on average shown to be insignificant for all the shrinkage methods, leading to the conclusion that miRNAs are not good predictors of survival in these data. In the prognostic analyses with five different normalization methods, only miR-mmd seemed to be found more than once. It should be stressed that even though the signals in the miRNAs for these

data were not that strong, it does not mean that the idea of miRNA expression profiles functioning as prognostic indicators should be abandoned. Several limitations and assumptions concerning the data and methods are definitely to be taken into account, which is something that will be discussed in Section 7.2.

7.2 Validity of the results

It lies in the nature of a master thesis that certain restrictions must be made due to the fixed time span, and of course this thesis is no exception. One limitation is that the simulation and comparative studies are all based on rank normalization, thus the results derived may not apply to the other normalization methods. The ideal situation would be to see if all normalization methods provide better prediction models compared to working with the raw values and in the comparative studies to see if they are equally good/bad at predicting incidence and prognosis. In the light of the results obtained from the comparative study regarding prognosis, it would have been very informative with a small simulation study were the truth is known to see if the methods possesses the ability to predict survival from a subset of miRNAs.

The are two main reasons behind why the rank normalization was the method to be examined in greater depth. When the large differences in the mean C_t between the cancer and healthy samples was discovered, the idea of analyzing the pattern instead was a very attractive alternative. This way each patient becomes its own control and the general mean differences becomes less important to the results. Another advantage of this normalization method was the handling of missing values. The generic problem with missing values in GLM is easily overcome by converting the missing values to ranks, hence there is no need to exclude a proportion of the samples from an already scarce data set. The backside of doing this is that the missing values will lie as a clot in the tail of the rank distribution, which could have some influence on the results.

Other limitations are those given by the data and statistical methods. One large issue is the few observations available which definitely had an impact on the analyses and made the results in this thesis unstable. The miRNAs found in each analysis could be quite different and by changing some of the tuning parameters just a little showed how sensible the results were. This was also expressed by the width of the confidence intervals of the odds ratios and hazards ratios, where the majority had a wide range. Although, the latter is probably not only caused by the low number of samples, but also the artificial separation in mean C_t that existed between the cancer/control groups, which easied the

prediction of incidence. The normalization were supposed to prevent this from having an effect, but the results revealed that especially internal normalization did not level out all the mean C_t differences.

One of the main restrictions used throughout the thesis was the N/A fraction allowed for each miRNA. Here it was fixed at about 10% which seemed reasonable, but in principle this could be too low or high. It was however natural to exclude the miRNAs with near 100% or 100% missing because here there is strong evidence that the miRNA material is not present in the samples. The choice of excluding many miRNAs based on this criteria was also motivated by the $p >> n$ problem, the multivariate model alone can not handle more parameters than observations hence the dimensionality had to be reduced in some way and this seemed logical to do in miRNA context. However, the fact remains that even though only 10% missing measurements for the individual miRNA was allowed, some missing values still remained in the data. This was a problem in the analyses (except when using ranks) because GLM and Cox proportional hazards model can not handle missing values, hence only complete cases could be considered.

One way to handle missing values could have been to use imputation in the form of replacing the N/As with appropriate values. Many possible imputation strategies could be applied and since missing values are result of too high C_t values, a simple approach could be to replace missing values with a random number from some distribution e.g. the uniform between 35 and 40. By doing this the number of samples is maintained for all normalization methods, thus decreases the likelihood of overlooking important miRNAs. This is just a hypothesis, the effect of imputing miRNA data is not known since this have not been an area of examination. One other way to handle missing values is the use of indicator variables, i.e. including a binary variable for each miRNA that indicates presence/absence. This way it is possible to keep all the samples in the analyses and to both estimate the effect of miRNAs can be measured and the effect of one unit increase in C_t. To test whether the miRNA is significant regarding the response and what kind of effect that is significant, can then be seen from the p-values of the two mentioned effects.

The assumption of proportional hazards in the Cox model is important to the prognosis. Here the `cox.zph` function of the `survival` package served as a tool for testing this assumption using scaled Schoenfeld residuals. No signs of violation was detected in any of the prognostic analyses except for one case. To further strengthen these results, other procedures for testing the proportional hazards assumption could be performed. E.g. by generating time dependent covariates by creating interactions of the predictors and a function of survival time and then include these interactions in the model. If any of the time dependent covariates are significant then those predictors are not proportional.

From a biological point of view, correlation between the miRNAs should be expected and ideally models should account for this. The univariate selection method plays a central role of this thesis since it works as a bottleneck, i.e. is part of determining the set of miRNAs used in the other regression models. This method does not consider correlation between miRNAs and the extent of this problem is still unknown. Consider e.g. that one miRNA is strongly associated with the outcome and other miRNAs are strongly correlated with this specific miRNA. Then the univariate analysis might include some variables for further analysis that should have been excluded, because they would not have been showed significant had there been corrected for the strong miRNA variable. So it should be kept in mind that some miRNAs left out for other correlated miRNAs, potentially are strong predictors of incidence and prognosis.

Disregarding some of the mentioned limitations and taking into account the initial challenges connected with the original data, the results derived seem satisfying. It is very hard to say if the miRNA subsets obtained makes sense biologically or not and since this is still a relatively new research area the existing literature is sparse, making it difficult to check if the results are in accordance. One of the things uncovered by research though, is that serum samples are more unstable compared to e.g. tissue or whole blood samples in terms of storage, preparation and qrt-PCR, indicating that the basis for these results could be improved. One of most valuable lessons learned from this thesis is how to analyze these types of data, no one knows the best way to reach the objective so it has in many ways been a pioneer assignment.

To validate the results, studies should be performed where the most significant miRNAs found are measured on an independent sample. The validation would be cheaper because only a small number of miRNAs are considered and replicates can potentially be done to improve precision. If miRNAs are valid, it implies that the statistical evidence is very strong and miRNA expression profiles are one step closer to be accepted biomarkers applicable in the everyday clinical routines for pancreas cancer. Section 7.3 suggests alternative methods that could be applied in the analysis of miRNA data.

7.3 Alternative analyses approaches

The logistic regression model, Cox proportional hazards model and the four shrinkage methods applied in this thesis, are only a small part of the possibilities for analyzing high dimensional data. This section is dedicated to look at some of the alternative models, which could be considered in future research in

this area. Since the additional methods discussed here has not been examined in depth, they will only be briefly described.

Cluster analysis is the task of assigning a set of samples into groups that share similar properties, thus making it a candidate method in relation to incidence. There exists many different clustering algorithms based on distances, distributions etc. and they could be applied to the the task of separating cancer cases from the healthy subjects (maybe also looking at chronic pancreatitis cases as an individual group). However, ways to handle the missing values best for this analysis method should be considered [Wikipedia].

Another alternative method for dealing with high dimensional data, i.e. $p >> n$ situations, is called *sparse discriminant analysis*. It is a method which performs linear discriminant analysis with a sparseness criterion imposed such that classification, variable selection and dimension reduction is performed simultaneously. The fact that this method originates from the demand of a method providing easy interpretation of covariates and dimension reduction of biological and medical data, makes it an attractive alternative. As with the clustering, missing values should be handled somehow [Clemmensen et al. 2011].

Other existing companion shrinkage methods are e.g. *principal component analysis* which performs rotation of the coordinate axes. The orthogonal transformation is done such that the first principal component (corresponding to the first *eigenvector*) has the largest possible variance, that is accounts for as much of the variability in the data as possible, and the second principle component accounts for the second largest possible variance and so forth. Here the *eigenvalues* each represents a portion of the total variance and dimensionality of data can be reduced by only considering the largest eigenvalues. Another advantage is that the transformed variables becomes uncorrelated due to orthogonality, but the price is that the variables have to be scaled before transformation and the interpretation of the principle components is fuzzy. In some cases the first principal component is just the mean of all covariates, which is not very insightful. The largest drawback of using this approach is that it does not perform variable selection as such, i.e. does not give a set of miRNAs but a set of components instead, which is not clinically relevant.

An idea could be to perform regression on the principle components, this is known as *principal components regression* - a technique considered in the paper by Bøvelstad et al. [2007] with satisfying results in relation to survival. Bøvelstad et al. [2007] however mentions that a drawback of this method is that the first eigenvalues selected that accounts for as much variation in the gene expressions as possible, might not be associated with patient survival. An extension to overcome this problem is in the article denoted *supervised principal components regression* and it combines the univariate selection with principal

components regression, which results in a bivariate tuning parameter.

While the logistic regression model seemed like an obvious choice of underlying model for predicting incidence, the Cox proportional hazards model is not necessarily the best choice for the purpose of predicting survival of pancreatic cancer patient on the basis of miRNA expression profiles. The attraction of using the semi-parametric Cox model is that it does not assume any distribution for the underlying hazard, and keeping in mind that this biostatistical area of research is relatively new, the ordinary Cox model is an excellent starting point. However, had time allowed it, it could have been interesting to consider some parametric survival model, e.g. the exponential model, Weibull model or a log-logistic model. When an appropriate distribution have been specified, the idea is then to let the parameters of that distribution depend on then covariates. Whether this approach would have resulted in more interpretive results or not remains unknown for now. Also the *additive Aalen model* could be a possibility, where the coefficients influence the hazard additively instead of multiplicative. The Aalen model could potentially fit these kind of data better than the Cox model.

Chapter 8

Conclusion

In conclusion this thesis contains analysis of real data coming from pancreatic cancer patients and healthy controls, provided by Herlev Hospital. Subsets of those miRNAs showing significant prediction association with incidence and survival after operation respectively have been determined on the basis of five different normalization methods and application of four shrinkage methods.

Results showed that normalizing miRNA data was of great importance. All shrinkage methods could classify samples as cancer/healthy with few prediction errors. Prediction of survival from miRNA expression profiles for these data reveal no clear signal, where too few samples available could be one of the reasons. Section 8.1 gives short recommendations to persons working with and analyzing miRNA data in the future.

8.1 Recommendations

- If it is in any way possible to influence, then make sure that the company performing the sample analysis follow a comprehensive experimental design plan that accounts for as much nuisance as possible. One option could be to apply the two-staged blocking procedure suggested in Section 3.3.1, in order to avoid confounding.
- Normalize data to make samples more comparable in the analyses. Here the internal normalization is not an advisable method since it has displayed distinctive tendencies compared to other normalization methods.
- Base the relevant subset of miRNA predictors found in multiple normalization methods, i.e. choose only those found in more than one analysis.
- Look at the fraction of missing values in each miRNA and exclude those with nearly 100% or 100% (at least). This is a quick way to get rid of useless variables and reduce dimensionality, however the criteria for exclusion is not evidential so the fraction allowed is up for discussion. However, the use of imputation or indicator variables are alternatives worth mentioning.
- Use the univariate selection, Lasso or naïve elastic net method as initial screening methods to narrow down to a few relevant miRNAs. When suitable, supplement with the backwards stepwise elimination procedure.
- Estimate various tuning parameters by log-likelihood cross-validation.
- Validate on independent sample data set.

8.2 Future research

This concludes the work of this thesis and as already pointed out in the discussion, there are numerous ways to extent the analyses. The next logical step though must be to perform validation studies with a focus on those miRNAs derived on the basis of these analyses from incidence and prognosis, along with others selected from the literature. The experiments should be conducted in a reasonable manner such that the C_t measurements are not confounded with e.g. diagnose. The purpose would then be to analyze these selected miRNAs and see if some of the same effects are observed, which would indicate that there truly is a predictive signal in the given miRNAs.

More simulation studies could be performed to examine miRNA data in even greater depth, such that the understanding of normalization is further improved. Since missing values and lack of power seems to be a consistent problem in these kind of data, imputation seems like an important research topic to explore further.

Appendix A

Supplementary results

A.1 Comparative study

A.1.1 Incidence (Section 6.1.1)

k	λ_0	$\lambda_{1,l}$	$\lambda_{2,r}$	$\lambda_{1,e}$	$\lambda_{2,e}$	AUC_u	AUC_l	AUC_r	AUC_e
1	8.774e-04	53.5514	1642.3982	53.5827	1642.3936	0.8680	0.9545	0.9429	0.9342
2	2.755e-04	40.9662	654.1647	40.7199	654.1218	0.8549	0.9130	0.8946	0.9120
3	5.918e-05	38.0194	1820.0552	37.5188	1820.0538	0.8866	0.9688	0.9499	0.9679
4	1.720e-05	25.1196	1849.7227	25.2451	1849.7410	0.8781	0.9518	0.9083	0.9272
5	3.377e-05	50.7226	1425.1280	50.5976	1425.1261	0.8100	0.8781	0.8414	0.8578
6	3.275e-05	36.4051	1901.4472	36.3765	1901.4549	0.8592	0.9518	0.9367	0.9234
7	7.338e-05	23.2032	620.7596	23.2655	620.7544	0.9220	0.9631	0.9490	0.9641
8	6.103e-04	25.7415	1386.1711	25.7415	1386.1711	0.9008	0.9414	0.9017	0.9244
9	2.963e-04	24.1053	692.6669	24.1053	692.6668	0.9036	0.9064	0.9130	0.9187
10	3.600e-04	22.7110	1049.0769	21.7111	1049.0648	0.8771	0.9159	0.9026	0.9121
11	2.950e-04	40.3712	2160.9045	40.4054	2160.8991	0.9480	0.9631	0.9594	0.9490
12	1.028e-03	41.5700	1267.9067	40.5600	1267.8443	0.9716	0.9735	0.9527	0.9698
13	4.601e-04	43.1945	1270.5923	42.1670	1270.6798	0.9062	0.9188	0.9072	0.9110
14	4.291e-04	33.5807	1625.3417	33.5729	1625.3413	0.9623	0.9642	0.9487	0.9594
15	2.253e-03	22.7765	1937.6023	23.2896	1937.5829	0.9357	0.9102	0.9140	0.9149
16	3.039e-04	37.6742	1260.1835	36.6742	1260.1817	0.8980	0.9304	0.9246	0.9304
17	1.280e-04	41.0392	819.3059	40.9145	819.2967	0.8715	0.8904	0.8979	0.9026
18	2.991e-05	18.0378	1114.2971	17.9128	1114.2967	0.8521	0.9301	0.9263	0.9206
19	4.667e-04	45.9801	1219.3419	46.0425	1219.3445	0.8733	0.9139	0.9226	0.9323
20	1.645e-04	37.0367	815.2621	36.7868	815.2688	0.9083	0.8913	0.8516	0.8705
21	2.926e-04	37.0643	739.3318	37.3156	739.3237	0.8980	0.9159	0.9130	0.9333
22	3.933e-05	19.9186	1387.2900	19.9342	1387.2901	0.7921	0.9282	0.9045	0.9216
23	3.740e-05	25.4120	890.6915	25.1620	890.6892	0.8109	0.9400	0.8907	0.9081
24	2.997e-05	11.2751	1972.7970	11.2751	1972.7970	0.8641	0.9294	0.8859	0.8907
25	3.260e-05	58.9925	1722.8333	57.9870	1722.8243	0.8280	0.9159	0.8781	0.9017
26	2.880e-04	41.1541	969.2919	41.1541	969.2919	0.8965	0.9348	0.9026	0.9253
27	2.374e-05	49.4670	2213.1188	49.9671	2213.1214	0.8646	0.9188	0.8897	0.8772
28	2.959e-04	37.4956	3433.9514	37.5269	3433.9516	0.9509	0.9527	0.9461	0.9442
29	9.805e-04	27.5024	1805.2122	26.9936	1805.2363	0.9140	0.9716	0.9556	0.9546
30	3.138e-05	53.7316	1140.2854	53.9817	1140.3143	0.8182	0.9449	0.9101	0.9294
31	5.936e-04	34.0179	1342.9172	34.0179	1342.9172	0.8994	0.9371	0.9294	0.9294
32	2.688e-03	37.2116	2305.4401	37.3366	2305.4421	0.8979	0.9546	0.9461	0.9357
33	9.746e-05	31.3283	1522.6258	31.0787	1522.5970	0.8710	0.9036	0.8573	0.8677
34	2.257e-04	38.5652	1420.6371	37.5655	1420.6125	0.8956	0.9263	0.9197	0.9197
35	1.212e-03	29.2162	1318.4135	29.5934	1318.7363	0.9206	0.9263	0.9216	0.9253
36	1.834e-03	8.3979	419.8538	8.4604	419.8541	0.8690	0.9110	0.9081	0.9207
37	1.253e-04	35.4584	760.1849	35.9537	760.1167	0.9304	0.9565	0.9188	0.9574
38	6.686e-04	45.5329	1625.3881	45.4069	1625.3704	0.9518	0.9641	0.9471	0.9471
39	2.699e-04	49.3303	687.4877	48.8304	687.4789	0.8790	0.9253	0.9168	0.9301
40	1.415e-03	38.0217	1884.1752	37.4914	1884.1122	0.9521	0.9574	0.9439	0.9391
41	6.757e-04	27.6849	985.3906	28.6846	985.4140	0.9357	0.9272	0.9017	0.9225
42	2.124e-03	55.9660	969.6496	54.9780	969.7248	0.9386	0.9896	0.9556	0.9830
43	2.724e-04	31.8524	1003.6313	31.8524	1003.6313	0.8998	0.9575	0.9187	0.9527
44	1.165e-03	48.7040	2073.0089	48.7001	2073.0089	0.9691	0.9807	0.9642	0.9758
45	1.171e-03	48.9879	1460.1066	49.2375	1460.1212	0.8809	0.9499	0.9291	0.9527
46	1.741e-04	57.4852	2236.3928	57.4227	2236.3918	0.8738	0.9622	0.9650	0.9575
47	3.532e-04	31.8487	1140.7327	31.8331	1140.7324	0.9112	0.9698	0.9603	0.9641
48	1.500e-03	37.5002	971.7556	37.5588	971.7773	0.8526	0.9253	0.9036	0.9045
49	1.139e-03	26.7586	1840.4409	27.0086	1840.4369	0.9178	0.9487	0.9352	0.9429
50	2.968e-04	38.2522	1754.8856	38.2679	1754.8858	0.8866	0.9414	0.9301	0.9319

Table A.1: *Information about the prediction performance for the four shrinkage methods, on the basis of AUC, for each iterative step.*

A.1.2 Prognosis (Section 6.2.2)

k	λ_0	$\lambda_{1,l}$	$\lambda_{2,r}$	$\lambda_{1,e}$	$\lambda_{2,e}$	PI_u	PI_l	PI_r	PI_e
1	0.1286527	46.2511	10724.8107	46.2667	10724.8108	0.10333	0.11417	0.05068	0.087978
2	0.0833751	36.4178	3315.1176	36.2928	3315.1156	0.11526	0.11526	0.08885	0.105675
3	0.0611998	19.0303	8958.4754	18.0328	8958.3434	0.17811	0.10616	0.18235	0.173852
4	0.0921277	40.7142	1085.8926	40.6356	1085.8890	0.13572	0.34708	0.09530	0.347080
5	0.0698212	22.3364	1139.6223	22.0864	1139.6225	0.19091	0.43785	0.29037	0.320605
6	0.0452531	9.7688	818.8688	10.0187	818.8720	0.80653	0.63117	0.64888	0.634079
7	0.0324278	34.8309	2037.7888	35.3308	2037.7495	0.70996	0.92204	0.71423	0.903323
8	0.0362801	57.4236	7053.1278	57.6731	7053.0961	0.99021	0.66153	0.73275	0.661585
9	0.0541010	52.0258	98.3074	51.2408	97.6880	0.94595	1.00000	0.86456	1.000000
10	0.0403136	65.3131	5618.2196	65.8131	5618.2225	0.76146	0.58323	0.65794	0.886059
11	0.2140383	18.1995	853.9827	18.2151	853.9820	0.43520	0.55981	0.32337	0.149237
12	0.0768926	20.6095	4815.1456	20.8595	4815.1455	0.08730	0.41970	0.19761	0.195799
13	0.0411608	24.2508	1027.6267	24.2518	1027.6269	0.26537	0.29264	0.20940	0.225576
14	0.0819694	8.7142	847.7342	8.7454	847.7343	0.82824	0.91199	0.55181	0.583458
15	0.0096785	30.5610	530.5955	30.5688	530.5958	0.33546	0.07260	0.23784	0.089732
16	0.2471353	43.7350	10722.4002	43.6721	10722.3986	0.03981	0.08832	0.02328	0.067574
17	0.0729057	45.9730	8026.8754	45.7230	8026.8714	0.94207	0.94207	0.57884	0.618771
18	0.0008356	6.8788	63.2922	6.3851	63.3754	0.07518	0.09244	0.08731	0.092608
19	0.0093521	10.5353	2665.4744	10.5392	2665.4744	0.17888	0.12028	0.33220	0.341426
20	0.0193671	39.1050	865.8369	39.0425	865.8362	0.49133	0.75182	0.82511	0.815383
21	0.0556403	19.8991	104.6872	20.3987	104.6523	0.09167	1.00000	0.07353	1.000000
22	0.0580279	130.0868	8633.1012	131.0886	8633.1020	0.65041	1.00000	0.46336	1.000000
23	0.1379370	16.1542	3782.7775	16.1230	3782.7775	0.72047	0.72517	0.19982	0.180486
24	0.0486893	3.5667	9.4786	3.8830	9.8719	0.17568	0.17568	0.12198	0.139544
25	0.0430564	26.2513	1981.7145	26.3139	1981.7175	0.54584	0.76831	0.81018	0.705083
26	0.1074816	29.1224	7174.2840	29.6236	7174.2786	0.22961	0.54993	0.28186	0.220277
27	0.0787455	11.8242	448.2454	11.9492	448.2453	0.09258	0.03094	0.00650	0.009887
28	0.1013464	11.9084	335.9648	11.4073	335.9875	0.79463	0.79463	0.74899	0.920940
29	0.0425903	69.5083	1134.9101	69.5078	1134.9101	0.60552	1.00000	0.22484	1.000000
30	0.0881069	84.5844	4056.7054	84.6158	4056.7011	0.55861	1.00000	0.21261	1.000000
31	0.0751003	3.8268	37.8012	4.7654	38.1463	0.09058	0.10526	0.09165	0.104306
32	0.0238931	36.8177	2754.7916	36.9427	2754.7921	0.52544	0.89481	0.92993	0.928850
33	0.0320529	24.3265	184.6120	24.2664	184.6290	0.26683	0.45422	0.64203	0.393819
34	0.0282670	40.5684	1244.0578	39.5681	1243.9675	0.83991	0.89476	0.84224	0.846681
35	0.0388792	18.0435	256.5800	18.0747	256.5813	0.50089	0.50928	0.45830	0.495259
36	0.1283426	10.2376	96.3112	10.2981	96.3270	0.50443	0.77965	0.75728	0.726351
37	0.0932731	47.0167	4446.7400	47.1417	4446.7423	0.36413	0.19310	0.26990	0.224076
38	0.1313764	28.1610	10171.3871	28.0360	10171.3868	0.43902	0.08817	0.02636	0.014090
39	0.0157107	10.8859	470.4892	10.7609	470.4870	1.00000	1.00000	0.06392	1.000000
40	0.0561885	6.5597	258.0332	5.5627	258.1101	0.05694	0.05694	0.04414	0.039938
41	0.1796949	19.8159	255.0765	19.7211	255.1579	1.00000	1.00000	0.04314	1.000000
42	0.0290591	22.8143	3560.7551	22.7513	3560.7546	0.88498	0.97752	0.84234	0.830515
43	0.0642597	29.0997	1476.4374	28.0994	1476.4277	0.61177	0.56502	0.11716	0.348633
44	0.1012761	12.0760	4337.1762	12.0760	4337.1762	0.89897	0.92785	0.57032	0.546598
45	0.0206395	4.9463	439.1741	4.9307	439.1741	0.22941	0.42349	0.70895	0.702836
46	0.0443901	30.5258	2601.7212	30.5572	2601.7202	0.89342	0.39226	0.25507	0.283833
47	0.1161644	47.1324	6419.3298	47.3822	6419.3420	0.33883	0.06138	0.04523	0.020006
48	0.0376087	4.3258	88.9790	4.3258	88.9790	0.99527	0.26099	0.42395	0.495591
49	0.0087852	6.1625	72.2103	6.4117	72.2299	0.72960	0.70838	0.99275	0.921635
50	0.0149016	35.6738	3243.3028	35.5489	3243.2972	0.66325	0.63121	0.57420	0.589994

Table A.2: *Information about the prediction performance for the four shrinkage methods, on the basis of likelihood ratio test, for each iterative step.*

Bibliography

Ambros, V., Bartel, B., Bartel, D., Burge, C., Carrington, J., Chen, X., Dreyfuss, G., Eddy, S., Griffiths-Jones, S., Marshall, M., et al. (2003). A uniform system for microRNA annotation. *RNA*, 9(3):277.

Andersen, C., Jensen, J., and Ørntoft, T. (2004). Normalization of real-time quantitative reverse transcription-PCR data: A model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Research*, 64(15):5245.

AppliedBiosystems (2006). Illuminate the mystery of biological dark matter: miRNA expression analysis research tools. http://www.ambion.com/catalog/workflows/miRNA/ab_ambion_mirna_workflow.pdf.

AppliedBiosystems (2010). Taqman® array human microRNA cards. http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_054742.pdf.

Bloomston, M., Frankel, W., Petrocca, F., Volinia, S., Alder, H., Hagan, J., Liu, C., Bhatt, D., Taccioli, C., and Croce, C. (2007). MicroRNA expression patterns to differentiate pancreatic adenocarcinoma from normal pancreas and chronic pancreatitis. *JAMA: the Journal of the American Medical Association*, 297(17):1901.

Bolstad, B. (2010). *preprocessCore: A collection of pre-processing functions*. R package version 1.16-0.

Bolstad, B., Irizarry, R., Åstrand, M., and Speed, T. (2003). A comparison of

normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193.

Bøvelstad, H., Nygård, S., Størvold, H., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjærde, O. (2007). Predicting survival from microarray data - a comparative study. *Bioinformatics*, 23(16):2080.

Carstensen, B., Plummer, M., Laara, E., and Hills, M. (2011). *Epi: A Package for Statistical Analysis in Epidemiology*. R package version 1.1.33.

Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53(4):406–413.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 187–220.

Crawley, M. J. (2005). *Statistics: An Introduction using R*. Wiley. ISBN-10: 0-470-02297-3.

Dahl, D. B. (2009). *xtable: Export tables to LaTeX or HTML*. R package version 1.6-0.

Dehlendorff, C. and Andersen, K. (2011). Experimental plan for miRNA measurements. Confidential paper.

Diggle, P., Heagerty, P., Lian, K., and Zeger, S. (2002). *Analysis of Longitudinal Data*, volume 25. Oxford University Press, USA. ISBN-13: 978-0-198-52484-7.

Etheridge, A., Lee, I., Hood, L., Galas, D., and Wang, K. (2011). Extracellular microRNA: a new source of biomarkers. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Gilad, S., Meiri, E., Yogev, Y., Benjamin, S., Lebanony, D., Yerushalmi, N., Benjamin, H., Kushnir, M., Cholakh, H., Melamed, N., et al. (2008). Serum microRNAs are promising novel biomarkers. *PLoS One*, 3(9):e3148.

Goeman, J. J. (2010). L1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, 52(1):70–84.

Goeman, J. J. (2011). *penalized: L1 (lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model*. R package version 0.9-37.

Griffiths-Jones, S., Grocock, R., Van Dongen, S., Bateman, A., and Enright, A. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34:D140–D144.

Harrell, F. E. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Verlag. ISBN-10: 0-387-95232-2.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Prediction, Inference and Data Mining*. Springer Verlag. ISBN-13: 978-0-387-84857-0.

Hidalgo, M. (2010). Pancreatic cancer. *New England Journal of Medicine*, 362(17):1605–1617.

Jemal, A., Siegel, R., Xu, J., and Ward, E. (2010). Cancer statistics, 2010. *CA: a cancer journal for clinicians*.

Kleinbaum, D. G. and Klein, M. (2005). *Survival Analysis: A Self-Learning Text*. New York, Springer-Verlag, 2 edition. ISBN-13: 978-0-387-23918-7.

Kong, X., Du, Y., Wang, G., Gao, J., Gong, Y., Li, L., Zhang, Z., Zhu, J., Jing, Q., Qin, Y., et al. (2010). Detection of differentially expressed microRNAs in serum of pancreatic ductal adenocarcinoma patients: miR-196a could be a potential marker for poor prognosis. *Digestive Diseases and Sciences*, pages 1–8.

Larsen, H. (2011). Knapperne vælter frem på det genetiske kontrolpanel. *Politiken Viden*, (4):3.

Lee, E., Gusev, Y., Jiang, J., Nuovo, G., Lerner, M., Frankel, W., Morgan, D., Postier, R., Brackett, D., and Schmittgen, T. (2007). Expression profiling identifies microRNA signature in pancreatic cancer. *International Journal of Cancer*, 120(5):1046–1054.

Lee, R., Feinbaum, R., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854.

Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 - Proceedings in Computational Statistics*, 575–580.

Lægehåndbogen (2009). Pankreaskræft. http://laegehaandbogen.dk/mave-tarm/tilstande-og-sygdomme/bugspytkirtel/pankreaskreft-2316.html.

Liu, J., Gao, J., Du, Y., Li, Z., Ren, Y., Gu, J., Wang, X., Gong, Y., Wang, W., and Kong, X. (2011). Combination of plasma microRNAs with serum CA19-9 for early detection of pancreatic cancer. *International Journal of Cancer*.

Malvezzi, M., Arfé, A., Bertuccio, P. Levi, F., La Vecchia, C., and Negri, E. (2010). European cancer mortality predictions for the year 2011. *Annals of Oncology*, 22(4):947–956.

miRBase. Release 18, November 2011. http://www.mirbase.org/.

Montgomery., D. C. (2008). *Design and Analysis of Experiments*. John Wiley & Sons Inc., 7 edition. ISBN-13: 978-0-470-39882-1.

NORDCAN (2011). Kræftstatistik: Bugspytkirtel - Danmark. http://www-dep.iarc.fr/NORDCAN/DK/StatsFact.asp?cancer=130&country=208.

Olsson, U. (2002). *Generalized Linear Models: An Applied Approach*. Studentlitteratur. ISBN-13: 978-9-144-04155-1.

Patienthåndbogen (2008). Bugspytkirtelkræft. http://patienthaandbogen.dk/kreft/sygdomme/bugspytkirtelkreft-2187.html.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Clarendon Press. ISBN-13: 978-0-198-50765-9.

Plummer, M. and Carstensen, B. (2011). Lexis: An R class for epidemiological studies with long-lerm follow-up. *Journal of Statistical Software*, 38(5):1–12.

QIAGEN. Typical amplification plot. http://www.qiagen.com/resources/info/guidelines_rtpcr/dataanalysis_sybr.aspx.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN-10: 3-900051-07-0.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Springer, New York. ISBN-13: 978-0-387-75968-5.

Sarkar, D. (2011). *lattice: Lattice Graphics*. R package version 0.20-0.

Schultz, N., Werner, J., Willenbrock, H., Roslind, A., Giese, N., Horn, T., Wøjdemann, M., and Johansen, J. (2011). MicroRNA expression profiles associated with pancreatic adenocarcinoma and ampullary adenocarcinoma.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2009). *ROCR: Visualizing the performance of scoring classifiers*. R package version 1.0-4.

Springer, M. (2007). Identification of potential microRNA biomarkers for pancreatic cancer. http://www.biosciencetechnology.com/Application-Notes/2007/03/Identification-Of-Potential-MicroRNA-Biomarkers-For-Pancreatic-Cancer/.

Szafranska, A., Davison, T., John, J., Cannon, T., Sipos, B., Maghnouj, A., Labourier, E., and Hahn, S. (2007). MicroRNA expression alterations are linked to tumorigenesis and non-neoplastic processes in pancreatic ductal adenocarcinoma. *Oncogene*, 26(30):4442–4452.

Szafranska, A., Davison, T., Shingara, J., Doleshal, M., Riggenbach, J., Morrison, C., Jewell, S., and Labourier, E. (2008a). Accurate molecular characterization of formalin-fixed, paraffin-embedded tissues by microRNA expression profiling. *Journal of Molecular Diagnostics*, 10(5):415.

Szafranska, A., Doleshal, M., Edmunds, H., Gordon, S., Luttges, J., Munding, J., Barth Jr, R., Gutmann, E., Suriawinata, A., Marc Pipas, J., et al. (2008b). Analysis of microRNAs in pancreatic fine-needle aspirates can classify benign and malignant tissues. *Clinical Chemistry*, 54(10):1716.

Therneau, T. and Lumley, T. (2011). *survival: Survival analysis, including penalised likelihood*. R package version 2.36-10. Original Splus→R port by Thomas Lumley.

Therneau, T., Lumley, T., Halvorsen, K., and Hornik, K. (2011). *date: Functions for handling dates*. R package version 1.2-32. S original by Terry Therneau, R port by Thomas Lumley, Kjetil Halvorsen and Kurt Hornik.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer Verlag, 4 edition.

Wang, J., Chen, J., Chang, P., LeBlanc, A., Li, D., Abbruzzesse, J., Frazier, M., Killary, A., and Sen, S. (2009). MicroRNAs in plasma of pancreatic ductal adenocarcinoma patients as novel blood-based biomarkers of disease. *Cancer Prevention Research*, 2(9):807.

Wickham, H. (2011). *plyr: Tools for splitting, applying and combining data*. R package version 1.7-1.

Wikipedia. The free encyclopedia. http://en.wikipedia.org/wiki/Main_Page.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC. ISBN-13: 978-1-584-88474-3.

Zhang, Y., Li, M., Wang, H., Fisher, W., Lin, P., Yao, Q., and Chen, C. (2009). Profiling of 95 microRNAs in pancreatic cancer cell lines and surgical specimens by real-time PCR analysis. *World Journal of Surgery*, 33(4):698–709.

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