DATA INTEGRATION IN PENALIZED REGRESSION MODELS

with application to genomics

LINN CECILIE BERGERSEN

THESIS FOR THE DEGREE OF

MASTER OF SCIENCE

Modelling and Data Analysis

Statistics Division, Department of Mathematics
Faculty of Mathematical and Natural Sciences
University of Oslo

June 2009

Acknowledgments

This thesis completes my Master's degree in statistics at the Department of Mathematics, University of Oslo. The thesis has been written in the period January 2008 to June 2009. Several persons have been helpful in this period, for which I am truly grateful and some of them I want to mention by name.

First and foremost, I would like to thank my supervisor Ingrid K. Glad, for introducing me to an exiting and challenging area of statistics and data analysis. She has always had time for discussions and provided me with invaluable guidance. Her contagious enthusiasm and interest in the project have inspired me through the whole process.

I would also like to thank Heidi Lyng who has been my co-supervisor, for the important assistance and knowledge of the biology part and for patiently discussing these issues.

Thanks to Marit Holden and Malin Lando for helpful discussions connected to related work, and to Arnoldo Frigessi for useful comments especially in the beginning of the process.

In addition, I would like to thank my fellow students at *Parameterrommet* (B800), for good company through long days and for all the great memories. A special thanks to Linn who helped pointing out typos.

I am also truly grateful to my dear parents, for their endless care and support. Thanks to my two four-legged girlfriends, for making it good to come home every day, and for always being able to make me laugh. And last but not least, to Andreas, for his love, encouragement and for all our joyful moments.

Linn Cecilie Bergersen
Oslo, June 2009

Contents

CHAPTER 1

Introduction

During the last few decades huge technological developments have taken place. Technological inventions have made an enormous impact on how we live our lives and completely changed our way of living. In scientific research the technological development during the information age has had a similar influence by enabling methods for collection and investigation of larger amounts of information. In most scientific disciplines the possibilities that arise when large amounts of data are available can be extremely important and may reveal new and valuable insight and discoveries.

The importance of using and developing proper methods to analyze these large amounts of data becomes even more crucial. In the analyses, it is most often a matter of finding patterns in data, which can be used to predict what will happen in the future. Throughout the ages we have been seeking patterns. An example may be a sailor seeking patterns in the weather before crossing an ocean, a store owner seeking patterns in his customers' preferences and demands, or a doctor seeking patterns in his patients' symptoms of disease.

The new challenges within statistical sciences arise with the explosive growth of information. While one in traditional statistical methodology assumed a few well chosen variables, today automatic methods for data collection leaves us with vast amounts of measurements without knowing which are relevant for the phenomenon under study (Donoho, 2000). Classical methods are not designed for these kinds of problems and may not be possible to use or may not behave as expected.

For instance in regression analysis, having a very large number of explanatory variables p when the sample size n is small, will not be in accordance with the assumptions in the usual regression model, where $p \leq n$. A lot of novel and effective strategies have been established to circumvent this problem, and shrinkage methods are one approach which is commonly used when doing regression with $p > n$, or even $p >> n$.

As the volume of existing data expands, an increased interest in data integration has also aroused. Methods combining information from different data sources could be of great relevance and importance in different scientific fields. One area where high-dimensional data frequently occur is within biology and medicine. Large high-dimensional data sets with thousands of covariates are a result of the great advances and new methods in biotechnology which are able to conduct high-throughput experiments of gene expression and other biological features of

interest. The underlying aim analyzing these data, is to search for novel biomarkers which can be used to predict outcome of a disease for future patients. Incorporating more than one type of such biological high-dimensional data in a single model may therefore be appropriate. By effectively taking advantage of known underlying biological processes, the idea of using more of the information available is just as beneficial from a biological point of view as from a statistical perspective.

The aim of this thesis is to propose a model for data integration of high-dimensional data in a regression setting where $p > n$. The suggested method will be a shrinkage method with L_1-penalties of the lasso type. By introducing penalty terms which could be uniquely defined for each covariate, the model may provide different amounts of shrinkage to the regression coefficients based on external information from additional data sources.

The model will be presented in a biological context and applied to a high-dimensional data set *The Radium Hospital Cervix Cancer Cohort Data*. The data set includes survival data and both gene expression measurements and aCGH data for patients diagnosed with cervical cancer at the Norwegian Radium Hospital in the period 2001-2004. The intent is to identify genes which are important for survival and to study the possibility of predicting the outcome for future patients. The aCGH data measures gains and losses in DNA copy number, which may cause changes in the expression of a gene. Abnormalities in a gene's expression level may disturb the primary function of the gene, which may cause highly aggressive disease and poor outcome. To take this process into account in a conventional regression model is unfeasible, but the aCGH data could be incorporated in the model by interpreting the data as prior information on each gene, giving genes within aberrated regions a larger chance of being selected in the final model.

The thesis is organized as follows: First the data at hand are introduced in Chapter 2. This also involves a description of cervix cancer and some of the underlying biological aspects related to cancer in general. In Chapter 3 and 4 the statistical background theory is presented by first reviewing the needed theory of survival analysis. Further it is focused on problems of high-dimensionality and some well known methods designed for handling such problems. Lasso regression analyses are carried out on both the gene expression data and aCGH data in Chapter 5. In this connection different methods for reducing the data set prior to the analysis are applied and advantages and disadvantages by reducing the data are pointed out. In Chapter 6 a method for data integration through genewise lasso penalization is introduced. The general idea is discussed and some asymptotic properties are derived. The computational aspects regarding the genewise lasso penalization procedure are discussed, as well as aspects regarding an additional tuning parameter q. To complete this chapter it is discussed how the methodology for integration of different data sources can be seen from a Bayesian perspective. The methodology introduced will in Chapter 7 be illustrated by an analysis of *The Radium Hospital Cervix Cancer Cohort*

Data, where the aCGH data are used in different penalization schemes to weight the penalties in a lasso regression analysis of the gene expression data. Both a biological validation of the selected genes, and a validation of the performance of the selected genes as biomarkers on a new independent data set are carried out. In the last chapter a summary and some concluding remarks are given to sum up the work and point out some possible topics for further research.

CHAPTER 2

The Radium Hospital Cervix Cancer Cohort Data

Cancer is one of the leading causes of human death in the western world and belongs to a complex group of diseases for which the clinical factors and causes vary a lot. The underlying biological course is in general similar for most cancers. Abnormal growth of cells develops to larger populations of cells which may invade tissue and metastasize and cause morbidity and if not treated, death of the individual (Ruddon, 2007, chap. 1). One way the cells can be disturbed is by genetic gains and losses, leading to abnormal copy numbers and again a change in expression level, which could be motive forces in tumor progression.

The technology of microarrays has been developed to be able to compare the genetic information in a tumor with a normal reference sample. The most recent microarrays are now able to contain information of ten thousands of genes. It can therefore be an enormously useful tool in the search for biological markers and contribute to possible new treatment strategies.

The data investigated in this thesis contain large scale microarray expression and gene copy number (aCGH) data for samples from cervix cancer patients. Survival data for the patients are also available. The aim is to collect the information in the data sets by combining methods dealing with problems connected to the high dimensions of these data. The data are supplied by Heidi Lyng at the Norwegian Radium Hospital and collected in a study approved by the regional committee of medical research ethics in southern Norway. The patients were all diagnosed in the period 2001-2004 at The Norwegian Radium Hospital having primary squamous cell carcinoma of the uterine cervix (Lyng et al., 2006).

Before describing the specific data at hand I will give a brief introduction to cervix cancer and some biological aspects related to cancer. A description of the technology of microarrays will also be given. It is natural to then consider the patients and the survival data, before the gene expression and aCGH data are described separately.

2.1. Cervix Cancer

Cervical cancer is one of the most common cancers among women. Yearly 470 000 incidences of the disease occur in the world, and 230 000 of these incidences end in death. 80% of the cases in

the world occurs in developing countries. This makes cervix cancer a very frequent disease and one of the most common causes of death for women in developing countries (Ruddon, 2007).

The causes for developing cervix cancer are related to an infection of a virus called Human Papilloma Virus (HPV). The virus is the most common sexual transmitting infection and belongs to a group of more than 100 viruses, where about 40 give rise to infections in the areas around the genitals. Only a smaller number of the different types of HPV is considered as high-risk type and associated with cervix cancer. HPV is a virus, which usually transmits through sexual contact and the chance of getting a HPV-infection is strongly connected to sexual habits (early sexual debut and multiple sex partners). It is estimated that about 20% of the population is infected at any time and about 70% of all women will be undergoing a HPV-infection during their life (Kreftregisteret, 2009, webpage). Not all infections are resulting in higher risk for cervical cancer and some are not even noticeable.

The virus is believed to be a necessary cause for developing cervix cancer since it may dispose cellular changes, but it is not sufficient (Hofvind et al., 2001). This means that not all women who get an infection will get cervix cancer, but the women who are diagnosed with cervical cancer have most likely had an infection. It is therefore believed that there may be other cofactors which are necessary for developing cervical cancer.

The disease often develops through stages and it is common to consider four different stages divided according to tumor size and how much the disease has spread. In the first stage the carcinoma is limited to the cervix, and in Norway this corresponds to 50% of all cases. Getting the diagnosis on an early stage gives larger possibilities for recovery and 9 out of 10 patients are still alive after 5 years. In the following stages, the tumor spreads to surrounding structures as upper and lower part of the vagina and after that to other parts of the body (Stage 4). When the disease has reached Stage 4, the situation is very serious and the recovery much worse. Only 1 out of 10 women are still alive after 5 years (Hofvind et al., 2001, Kreftregisteret, 2009, webpage). The differences observed in development countries compared to most of the western countries are due to organized screening programmes with routine PAP smears and gynecologic examinations in the latter countries, which detect most of the cases in an early stage and thereby reduce the incidence and mortality of cervix cancer. For example in Norway, all women in the age of 26-69 are invited to take a biopsy of the cervix every third year through an organized program.

2.2. Some Biological Aspects in Relation to Cancer

Changes in the expression of genes may cause the normal balance in the cell to be disturbed. This results in an imbalance between cell replication and cell death and thus a growth of a tumor cell

population. The gene expression changes can be due to different mechanisms including deletion and amplification of chromosomal segments (Ruddon, 2007, chap. 5).

Chromosomal DNA copy number corresponds to the number of copies of genomic DNA in the cell. Normally the copy number is 2 in each cell. Males have also one copy of each of the X and Y chromosomes, in contrast to women having two copies of the X chromosomes. As opposed to these normal cases, the copy number in cancer cells may vary substantially over the genome (Wieringen et al., 2007).

Genetic alterations by gains and losses may influence the gene expression levels to increase the ability of cells to reproduce and increase in number. An oncogene will promote cells to evolve to cancer cells when it is activated. A suppressor gene suppresses the cancer by controlling cells, such that they will not become cancer cells. If a suppressor gene is inactivated, increased profileration may occur and cells may become cancer cells. When the gene expression levels are influenced by genetic abnormalities, the proper function of the gene is disrupted. For example may suppressor genes be obstructed from functioning as a consequence of a copy number deletion. Oncogenes will be reinforced by an amplification (Bejjani et al., 2005, Ruddon, 2007, chap.1 and 5).

This indicates that copy number changes, which influence the function of different genes may lead to development and progression of cancers. It may, however, be challenging to identify the important genes. Not all genes which are overexpressed have to be in a region with increased copy number and not all regions with increased copy number will contain genes which are highly expressed (Bejjani et al., 2005). To only study genes which have an increased/decreased copy number, may therefore exclude relevant information. Thus both gene expression data, aCGH data and the understanding of these in combination are of interest in the search for biological markers in cancer research.

2.3. Microarray Technology

The technology of microarrays may be used both for measuring of gene expression and chromosomal copy number changes. The gene expression microarrays have been subject to much statistical research. Increasing interest has also been shown for the aCGH data, which may be used in the search for chromosomal regions for which the DNA is aberrated (van de Wiel & Wieringen, 2007). In the following, I will start by describing the practical methodology of gene expression microarrays. This will be similar for aCGH, apart from aCGH using chromosomal DNA instead of cDNA to hybridize to the array (Wieringen et al., 2007). Other differences will be considered in the end of the section.

Gene expression microarrays allow for comparison of mRNA levels in a tumor sample with a reference sample. The main advantage with the microarrays is the possibility to examine ten thousands of genes at the same time. mRNA is obtained from cancer tissue and a reference. The samples are labeled with different fluorescent dyes and hybridized together to a microarray. A normal use for the labeling is green fluorescent dye Cy3 and red fluorescent dye Cy5. The microarray is a glass-slide where samples of the mRNA have been spotted by an advanced printer technique. Each spot on the slide represent a gene and will be able to hybridize with the corresponding labeled mRNA derived. The position for each gene is known in advance. The mRNA which do not hybridize will be washed away before the array is scanned to detect the level of fluorescence in each spot. The level of fluorescence in a spot is considered proportional to the expression of a gene. That is, the amount of fluorescent in each spot reflects the amount of mRNA in the cell (Xiong, 2006, chap. 18). The more hybridization for a gene, the more intense will the signal be, indicating a higher level of expression for that particular gene. The red and green signals are combined in a color image representing the relative expression of a gene. Further image analysis techniques have to be applied to locate the spots and extract numerical data for the expression levels based on the pixel intensities (Quackenbush, 2006, Ruddon, 2007, chap.7).

In the extraction of the numerical data different preprocessing steps have to be performed before analysis, such as correction of saturated intensities, filtering of bad spots and lowess normalization are some examples, see Lando et al. (2009) for a description of the preprocessing performed on the data studied in this thesis. The data are typically represented in a matrix where one row represents a specific gene and each column represents a different biological sample (i.e. patients). Each entry in the matrix will correspond to a spot for one patient. The numerical value represents the relative expression level for a patient in a specific gene. The $\log_2$-ratio is convenient to use making the data more symmetric and more easy to compare both up- and down-regulated genes. The sign will indicate whether it is the red or the green channel that had the highest intensity.

The microarray technology used to extract the gene copy number data is similar to that described for gene expression, but DNA instead of mRNA is used in the hybridization. The DNA isolated from both the sample of interest and the reference will hybridize to a representation of the genome such that sequences may bind at different genomic locations (Pinkel & Albertson, 2005). As for the gene expression microarrays the hybridization intensity will be proportional to the relative copy number for the given sequences. The ratio for a sequence where no alteration in copy number is observed will be 0 on a logarithmic scale.

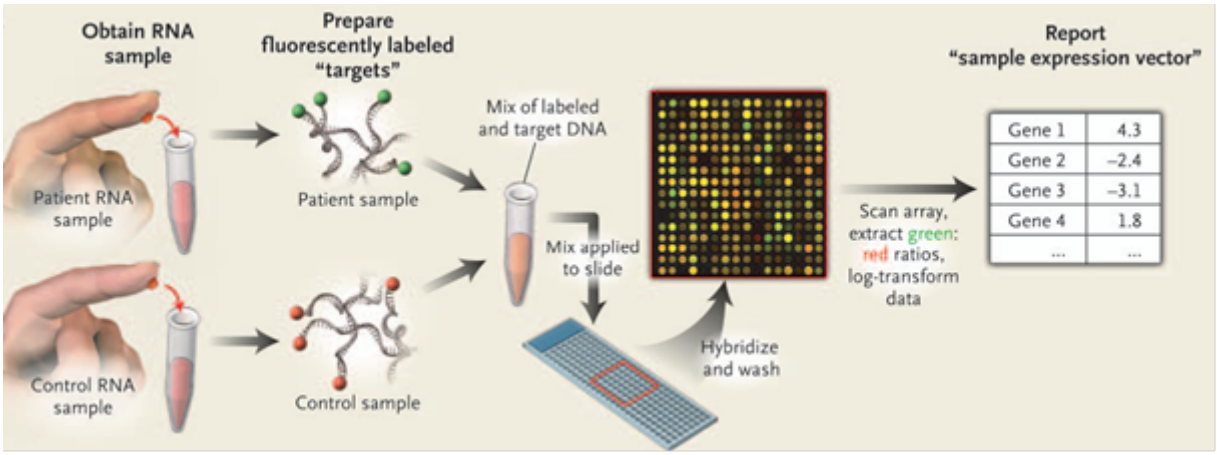


FIGURE 2.1. Illustration of the microarray method. Samples are extracted and labeled with different fluorescent dyes. The labeled samples are mixed together and allowed to hybridize to a microarray. The microarray is then scanned and the numerical data are extracted. The figure is from Quackenbush (2006)

2.4. Patients

The data set consists of 102 patients in total. All of them have been diagnosed with squamous cell carcinoma of the uterine cervix in the period 2001-2004 at the Norwegian Radium Hospital. All patients were from Stage 2 and 3 of the disease and received the same type of treatment; external irradiation and brachytherapy in combination. For a more detailed description of the therapy used, see Lando et al. (2009).

A relapse of the disease is considered as the event of interest, and the time to the event is recorded for each patient. In the thesis both "time until a relapse" and "survival time" will be used for the time to the event. The observation times are thus defined as the time from the patients got their diagnosis and until the first event of loco regional or distant relapse and/or cancer related death. All patients who did not experience a relapse before end of study are censored. Some general comments on censoring will be given in Chapter 3.

The data are plotted for all of the patients in Figure 2.2 where the red bars correspond to the patients whose survival time is censored and the blue the observed survival times. All 102 survival times are sorted (from high to low) and plotted in the lower panel. In the top and middle panel the aCGH measurements and gene expression measurements for each patient are plotted. Note that we have both gene expression and aCGH data for only 95 of the patients. A summary of the survival data are also given in Table 2.1. From the table one may see that the observation times range from 2.95 to 71.11 months, and about 2/3 of the patient's survival times are censored. The number of patients in the gene expression data and aCGH data respectively are also given, together with the proportion of censored and observed cases in both data sets.

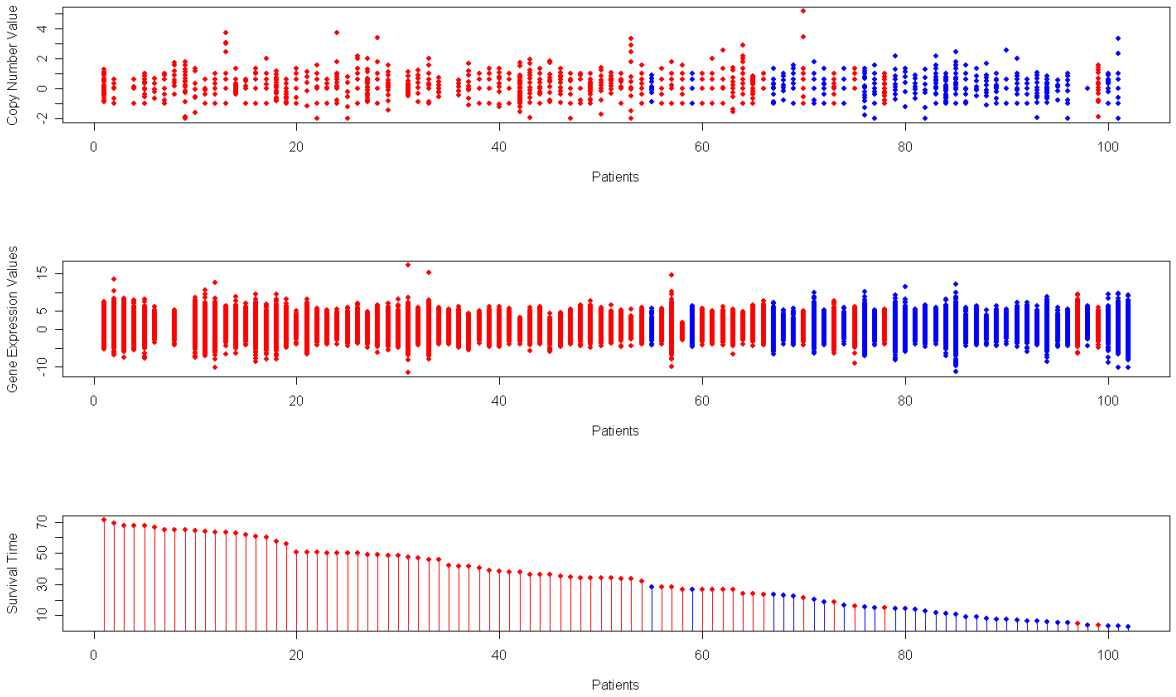


FIGURE 2.2. Plot of the survival times for each patient and their corresponding copy number and gene expression vectors. Red indicate data for the patients for which the survival times are censored and blue the observed survival times. The patients are ordered by their survival times.

The median survival times are also reported, and it seems obvious that the survival times for the observed cases are relatively smaller compared to the censored. For instance, for the observed patients in the gene expression data, the median of the survival times are 10.9 months whereas 41.6 months for the censored patients. This may not be unreasonable since most patients will either get a relapse in a short period of time or the patients will be cured.

There are a few patients standing out among the censored observation times. These patients are observed for less than five months which is short compared to the other censored observation times. This is because the patients died of another reason (not related to cancer), but they are still included in the study. There are six patients who are censored for this reason (Lando et al., 2009).

2.5. Gene Expression Data

The gene expression data are data extracted from microarray experiments as described in Section 2.3. The data are reported as the $\log_2$-ratio between the test and reference sample. A positive gene expression value corresponds to an up-regulation of the gene and a negative value corresponds to a down-regulation. Of the total 102 patients, gene expression measurements

Summary of Survival Data					
Data		Number of Patients	Median Survival Time	Minimum Survival Time	Maximum Survival Time
Gene Expression	All Patients	100	33.69	2.95	71.11
	Observed	32	10.94	2.95	28.20
	Censored	68	41.61	3.70	71.11
aCGH	All Patients	97	33.70	3.21	71.11
	Observed	31	11.25	3.21	28.20
	Censored	66	41.61	3.70	71.11

Table 2.1. Summary of the survival data showing the median, minimum and maximum survival time of all, observed and patients.

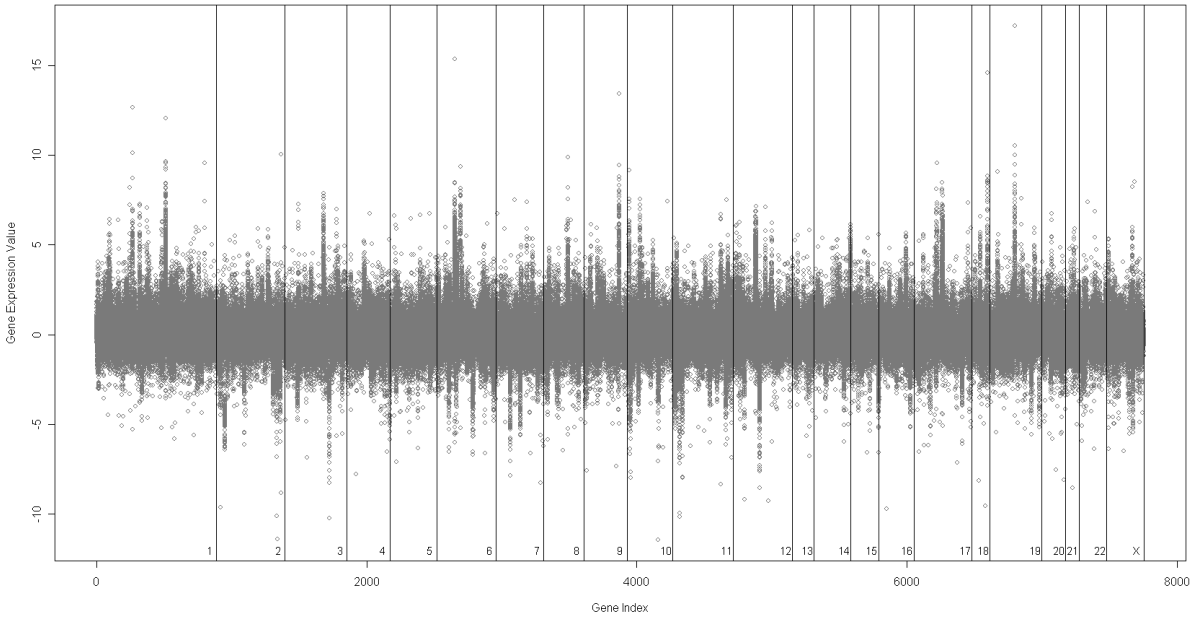
are available for 100 of them. The expression values are given for 12204 gene identifications and some of these identifications belong to the same gene. This is because some genes may be represented in more than one spot on the microarray, but the different spots are representing different parts of a gene.

The gene expression vectors were already normalized and imputed to take care of problems with missing values, according to standard methods and procedures. When a gene had more than 10% missing values, the gene was eliminated from the data. Since we in this thesis are interested in the genes for which we have both gene expression and aCGH data, we will concentrate on the genes for which this is the case. This involves elimination of genes where there are more than 10% missing values in the aCGH data and/or where aCGH measurements not exist. Genes for which we know the position, have both expression and aCGH data for, and which have more than 10% missing values in the gene expression data and more than 10% missing values in the aCGH data will constitute the data set. This corresponds to 7754 genes.

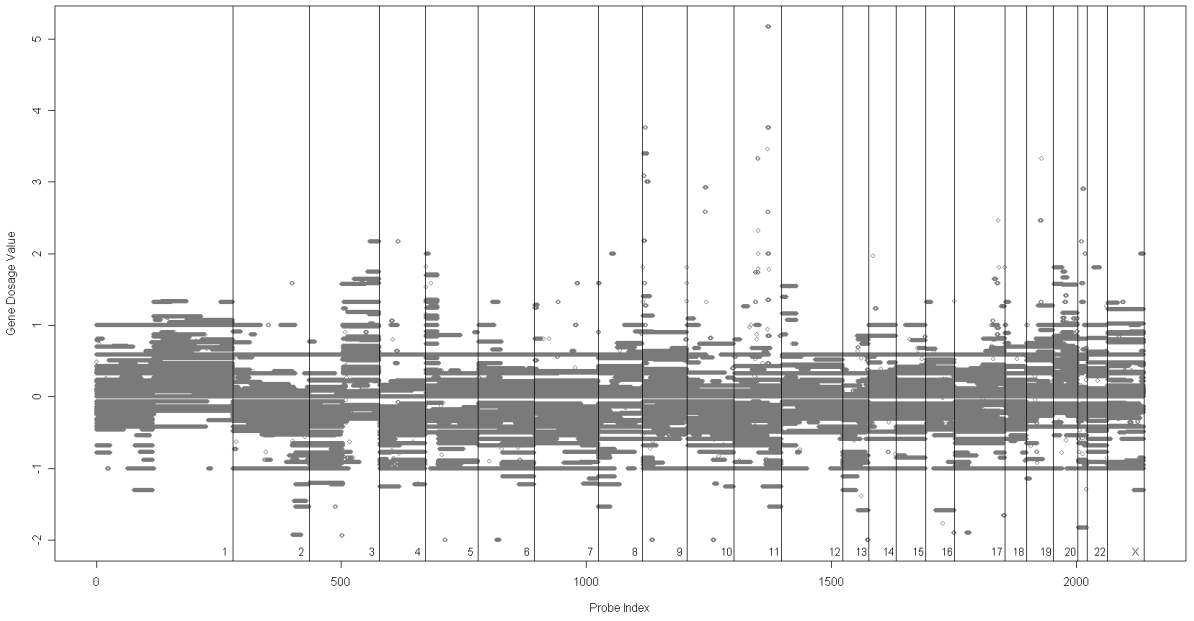
The gene expression data are plotted in Figure 2.3(a). The genes are located on chromosomes 1-X and in the plot the chromosomes are separated through vertical lines. The values of the $\log_2$-ratios in the data range from -11.4 to 17.2. The plot illustrates the behavior of the data, that is, some genes are up-regulated and some are down-regulated whereas others are more concentrated around zero.

2.6. Array Comparative Genomic Hybridization (aCGH) Data

We have copy number data for 97 patients. Each data vector represents a region of one or more genes which we call a "probe". When the probes with more than 10% missing values in the aCGH data and gene expression data are eliminated, 2138 probes are included. The probes contain information on the 7754 genes and are located on one of 23 chromosomes 1-X.



(a)



(b)

FIGURE 2.3.

(a) : Plot of the gene expression data with the position on the genome along the x-axis. The vertical lines separate the chromosomes.

(b): Plot of the aCGH data with the position on the genome along the x-axis. The vertical lines separate the chromosomes.

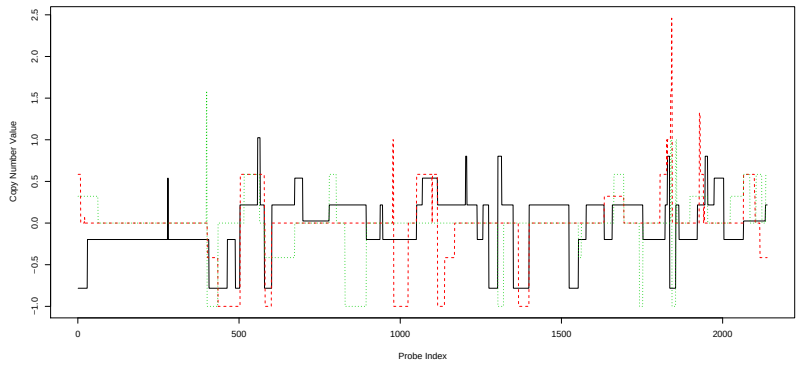


FIGURE 2.4. Plot of three aCGH data vectors for three patients. From the plot we may see the step form, indicating that many of the values for one patient are identical.

In a normal cell the copy number is 2, corresponding to 0 in our data which are the $\log_2$-ratios of the test and reference intensities. A loss in copy number (deletion) thus corresponds to a negative value. Amplifications which correspond to gains of DNA copy number are represented by positive values in the data. The copy number data are plotted in Figure 2.3(b) where the vertical lines separate each of the chromosomes. Note that -2 is the absolute minimum the values in the aCGH data can take, because it is not possible to loose more than two copies if the copy number in a normal cell is 2.

The aCGH data achieved by the procedures described in Section 2.3 are relative, which makes interpretation and comparison across experiments difficult (Lyng et al., 2008). Relative aCGH data are influenced by other features (total DNA content of the tumor cells, proportion of normal cell in the sample, experimental bias), than the DNA copy number. GeneCount is a method for genome-wide calculation of absolute copy numbers based on smoothed ratio levels, which account for these features. The absolute copy numbers for the aCGH data at hand have been extracted by GeneCount. The method is described in detail in Lyng et al. (2008). Note that in the following both "copy number" and "gene dosage" is used in the text to describe the aCGH data although the data are reported in terms of gene dosage, which is the output of GeneCount.

The nature of the aCGH data is somehow different from the expression data, since the vectors representing two different probes may be very correlated. Two or more probes may represent the copy number changes of neighboring regions on the chromosomes. This is often reflected by the vectors corresponding to neighboring regions as being very correlated. Some of them may also be 100% identical. If the data vector for one patient is plotted, the plot will remind of a step function. Copy number changes for three patients are plotted in Figure 2.4 to illustrate this. In addition to problems of collinearity when $p > n$, the aCGH data may be even more challenging to analyze in regression models because of the very high correlation among probe vectors.

CHAPTER 3

Survival Analysis

In this chapter some basic concepts of survival analysis will be introduced. The focus will be on the theory of survival analysis relevant for this thesis, which includes a brief introduction to some basic functions and aspects. The Kaplan-Meier estimator and the log-rank test will also be considered. Finally a more thorough presentation of the Cox proportional hazards model is given. The theory is mainly obtained from Aalen et al. (2008, chap. 1-4), which can be consulted for further explanation and references.

3.1. Basic Concepts of Survival Analysis

Survival analysis is a large field within statistics, which aims on studying the occurrence of events. An event could be any occurrence of scientific interest in a lifetime. Many scientific fields are interested in understanding the cause of events and to identify risk factors. In biology and medicine, survival analysis is central when the time until death or a certain development in a disease is studied. By understanding risk factors one may for example, investigate whether one should start with certain medical treatments or understand mechanisms of biological phenomena.

By introducing the concepts of censoring, analyses of survival data differ from analyses of other data in ordinary settings. A statistical framework especially suited for handling censored survival data is therefore needed in order to analyze the data properly.

We denote T as the time from an initiating event and to an event of interest (endpoint). T is thus called a survival time and is a nonnegative random variable. Although we use the term survival time, T could measure the time from a starting point to any event. It is worth noticing that the term survival time not necessarily relates to the study of death, it could measure the time until a relapse of a disease, a divorce or a failure of a technical system. In the analyses done in this thesis, survival time is solely used for the time until the patients experience a relapse of the disease or cancer related death.

3.1.1. Survival function. We let the random variable T denote the survival time with a cumulative distribution function $F(t)$. The survival function $S(t)$ is the probability of survival

beyond time t, and may be defined as

$$S(t) = P(T > t).$$

Thus $S(t)$ is the probability that the event has not yet happened at time t. Note that $S(t) = 1 - F(t)$, but it is customary to use the survival function in analyses of time to event data rather than the cumulative distribution. The relation between $S(t)$ and $F(t)$ makes it possible to obtain the density function as $f(t) = -S'(t)$.

In situations where the survival time is the time from a starting point and until death, the survival function will go to zero as t increases. When the event of interest not necessarily happens, that is, for instance if all patients will not experience a relapse of a disease, the survival function will decrease toward a positive value as t goes to infinity.

3.1.2. Hazard Rate. We now consider another central quantity in the theory of survival analysis. That is, the hazard rate $\alpha(t)$. While the survival function is specified as the unconditional probability that an event has not happened at time t, the hazard rate is defined with the help of a conditional probability. The conditional probability of experiencing an event in the next small time interval $[t, t + \Delta t]$, given that it has not yet happened at time t equals $\alpha(t)dt$. We may then define the hazard rate $\alpha(t)$ as

$$\alpha(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t). \tag{3.1}$$

The hazard rate could be any nonnegative function.

3.1.3. Censoring. As already indicated, censoring is one of the main things that has to be handled specifically when analyzing survival data. What we mean by "censoring", is nevertheless not yet defined. When a study is carried out and it is of interest to measure the time until a specific event for a group of individuals, one may experience at the end of study that not all individuals have experienced an event. For instance, studying cancer patients and the time from the patients get their diagnosis until they get a relapse, some patients will experience a relapse while others will not. These patients may experience the event later, but this will not be known when the data are analyzed. The data will thus contain both complete and incomplete observations of the event, and the incomplete observations are called censored survival times.

Throughout this thesis, we mean right-censoring when the term censoring is used, which is most relevant in this setting. There may be different reasons for a survival times being right-censored. The event may simply not occur before the study is ended, an individual may withdraw from the study, get lost in follow up or censored by other reasons. For example, may a cancer patient who experience death from another reason than cancer, be censored if death of cancer is the event of interest.

To express this more formally, define C to be the censoring time, that is the time from the initiating event to the individual gets censored. Let T be the complete observation time for the individual. The survival time may thus be expressed as $Z = \min(T, C)$ and it is common to introduce the censoring indicator $\delta = I(Z = T)$ which indicates whether the observed survival time are censored ($\delta = 0$) or not ($\delta = 1$). The survival data are then completely specified through the data pair (Z, δ).

3.1.4. The Kaplan-Meier Estimator. The Kaplan-Meier estimator can be used to estimate the survival function from a sample of censored survival data. Assume a sample of n individuals from a population, for which we have right-censored survival data and assume that there are no ties between the survival times. The Kaplan-Meier estimator can the be written as

$$\hat{S}(t) = \prod_{t_k^0 \leq t} \left\{ 1 - \frac{1}{Y(t_k^0)} \right\},$$

where $Y(t)$ is the number of individuals at risk "just before" time t, and $t_1^0 < t_2^0 < ...$ are the ordered times for which an event is observed. The estimated survival curve can thus be plotted in a Kaplan-Meier plot, by plotting $\hat{S}(t)$ versus t.

The Kaplan-Meier estimator may also be used to estimate the median (or other fractiles) survival time, and $100(1-\alpha)\%$ confidence intervals for $S(t)$ may be derived using that when evaluated at a given time t, $\hat{S}(t)$ is approximately normally distributed in large samples.

3.1.5. The Log-Rank Test. A highly relevant issue when studying survival data, is to compare the hazard rates for two or more populations. The log-rank test is one test which could be used for this purpose, performing a test on whether the hazard rates are equal. In the case of two samples this corresponds to

$$H_0 : \alpha_1(t) = \alpha_2(t) \text{ for all t} \in [0, t_0],$$

where $\alpha_1(t)$ is the hazard for Group 1 and $\alpha_2(t)$ for Group 2. A test statistic will then be

$$X^2(t_0) = \frac{Z_1(t_0)^2}{V_{11}(t_0)}, \tag{3.2}$$

where $Z_1(t_0)$ can be interpreted as the difference between the observed and the expected number of cases in the first sample and $V_{11}(t_0)$ is the variance of $Z_1(t_0)$. The test statistic $X^2(t_0)$ will be approximately chi-squared distributed with one degree of freedom under the null hypothesis.

The log-rank test can be extended to be applicable in situations with more than two groups, Aalen et al. (2008, chap. 3) may be consulted for an extensive description.

3.1.6. Counting Processes. Another important tool when survival data are to be analyzed is the theory of counting processes. In the following section we will make use of the formulation of counting processes and we therefore define some basic counting processes. A counting process $N(t)$ is defined as the number of events that has happened up to and including time t. Considering a small time interval $[t, t+dt)$ and assuming that only one event may occur in such an interval, the intensity process $\lambda(t)$ is the conditional probability that one observes an event given what is observed up to time t. For $n \geq 1$ individuals, one may define the aggregated counting process $N_{\bullet}(t) = \sum_{i=1}^{n} N_i(t)$, which counts how many of the individuals who has experienced an event, and the corresponding aggregated intensity process $\lambda_{\bullet}(t) = \sum_{i=1}^{n} \lambda_i(t)$.

3.2. Cox Regression

As often in statistics one turns to regression for the study of the effect of many covariates simultaneously. There are different approaches for doing regression when studying survival data, but one common approach is to model the effect the covariates have on the intensity process of a counting process (Aalen et al., 2008, chap. 4).

Suppose a set of data on the form $(Z_i, \delta_i, \boldsymbol{x}_i)$ for each individual i. Here Z_i is the observation time for the i_{th} individual, δ_i is a censoring indicator which is either 0 or 1 depending on whether the observed survival time Z_i is censored or not. Finally $\boldsymbol{x}_i$ is the vector $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ of covariate measurements for individual i which are assumed to be constant in time.

If we consider a counting process $N_i(t)$ for each individual i, $\lambda_i(t)$ is the intensity process, corresponding to the probability that an event occurs in the time interval $[t, t+dt)$ conditional on the past. Let $Y_i(t)$ be an indicator on whether individual i is at risk just before time t or not, and $\alpha(t|\boldsymbol{x_i})$ the hazard rate for individual i given covariates $\boldsymbol{x}_i$. We then have the relation

$$\lambda_i(t) = Y_i(t)\alpha(t|\boldsymbol{x}_i), \tag{3.3}$$

and want to identify differences in survival due to the set of covariates. It is possible to examine this if one is able to specify how $\alpha(t|\boldsymbol{x}_i)$ depends on the covariates $\boldsymbol{x}_i$. This dependency is usually described either by relative risk regression models or additive regression models. Relative risk regression is however most frequently used and Cox regression, which will be considered here is a relative risk regression model. In relative risk regression models one assumes a relationship of the form

$$\alpha(t|\boldsymbol{x}_i) = \alpha_0(t) r(\boldsymbol{\beta}, \boldsymbol{x}_i). \tag{3.4}$$

Here $\alpha_0(t)$ is called the baseline hazard at time t. The baseline hazard is assumed to be the same for all individuals, but is otherwise left unspecified. In the Cox regression setting, the baseline hazard may be considered as the hazard for an individual with all covariates equal to zero. The relative risk function $r(\boldsymbol{\beta}, \boldsymbol{x}_i)$ describes the effect of the covariates on the hazard. Because

the baseline hazard $\alpha_0(t)$ does not depend on the covariates, the model may be separated into one parametric and one nonparametric part. Hence, the model is usually referred to as being semiparametric. We therefore have to turn to the partial likelihood, since estimation of regression parameters $\boldsymbol{\beta}$ through ordinary likelihood methods is impossible.

By combining (3.3) and (3.4) we may express the intensity process for $N_i(t)$ as

$$\lambda_i(t) = Y_i(t)\alpha_0(t)r(\boldsymbol{\beta}, \boldsymbol{x}_i).$$

Then by introducing the aggregated counting process $N_\bullet(t)$, which register events among all individuals, the aggregated intensity process is given by

$$\lambda_\bullet(t) = \sum_{l=1}^{n} Y_l(t)\alpha_0(t)r(\boldsymbol{\beta}, \boldsymbol{x}_l).$$

The conditional probability that an event is observed for person i at time t, given the past and that there is an observed event at time t, is given as

$$\pi(i|t) = \frac{\lambda_i(t)}{\lambda_\bullet(t)} = \frac{Y_i(t)r(\boldsymbol{\beta}, \boldsymbol{x}_i)}{\sum_{l=1}^{n} Y_l(t)r(\boldsymbol{\beta}, \boldsymbol{x}_i)}. \tag{3.5}$$

The partial likelihood is thus obtained by multiplying the probabilities in (3.5) over all observed events times. We let $t_1^0 < t_2^0 < ...$ be the ordered survival times Z_i with $\delta_i = 1$, and assume that there are N events and that there are no tied events (tied events have to be handled specifically). Since each event contributes with one term as that in Expression (3.5), the partial likelihood is obtained as the product of the conditional probabilities over the observed event times $t_1^0 < t_2^0 < ... < t_N^0$. We therefore let (k) be the label for the individual experiencing an event at t_k^0 such that the covariate vectors corresponding to the N events are given as $\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}, ..., \boldsymbol{x}_{(N)}$.

$$L_{pl}(\boldsymbol{\beta}) = \prod_{k=1}^{N} \pi(k|t_k^0) = \prod_{k=1}^{N} \frac{Y_k(t_k^0)r(\boldsymbol{\beta}, \boldsymbol{x}_{(k)})}{\sum_{l=1}^{n} Y_l(t_k^0)r(\boldsymbol{\beta}, \boldsymbol{x}_l)} \tag{3.6}$$

We note that $Y_k(t_k^0)$ is always equal to one for individuals with an event (because they have to be at risk) and these may therefore be excluded in the numerator. Introducing the notation $\mathcal{R}_k = \{l : Y_l(t_k^0) = 1\}$ that is the risk set at time t_k^0 corresponding to the set of individuals who are still under study at a time just before t_k^0, we may write (3.6) as

$$L_{pl}(\boldsymbol{\beta}) = \prod_{k=1}^{N} \frac{r(\boldsymbol{\beta}, \boldsymbol{x}_{(k)})}{\sum_{l \in \mathcal{R}_k} r(\boldsymbol{\beta}, \boldsymbol{x}_l)}, \tag{3.7}$$

for a specified risk function. The censored survival times are not assumed to carry information on $\boldsymbol{\beta}$. The occurrence of censoring is however important for identifying the correct risk set for each t_k^0 and are thus included in the risk set $\mathcal{R}_k$ (Marubini & Valsecchi, 1995, chap. 6). For an individual with censored survival time c_i, the individual is considered to be at risk up to c_i and is thereafter excluded from the risk set.

For the Cox model the relative risk function is given by $r(\boldsymbol{\beta}, \boldsymbol{x}_i) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)$. Inserting this for $r(\boldsymbol{\beta}, \boldsymbol{x}_i)$ in (3.4), the Cox hazard function is given by

$$\alpha(t|\boldsymbol{x}_i) = \alpha_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i),$$

and the Cox partial likelihood;

$$L_{pl}(\boldsymbol{\beta}) = \prod_{k=1}^{N} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_{(k)})}{\sum_{l \in \mathcal{R}_k} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l)}. \tag{3.8}$$

We may then estimate the regression coefficients in the Cox model by maximizing the Cox log partial likelihood;

$$l_{pl}(\boldsymbol{\beta}) = \sum_{k=1}^{N} \{\boldsymbol{\beta}^T \boldsymbol{x}_{(k)} - \log \sum_{l \in \mathcal{R}_k} \exp(\boldsymbol{\beta}^T x_l)\}. \tag{3.9}$$

The size and sign of the estimated regression coefficients will indicate each covariate's influence on the survival. A positive coefficient $\boldsymbol{\beta}_j$ in front of a positive covariate will increase the hazard and contribute negatively on the survival and vice versa.

CHAPTER 4

Regression Analysis of $p > n$ Data

Analysis of high-dimensional data has developed into a large field in statistics, especially in connection with microarray data. One approach is regression analysis, where the aim is to sort out the most significant variables connected to the response and/or to predict outcome and examination of the prediction capability. For microarray data an aim will therefore be to sort out the most significant genes connected to survival or time to relapse and to predict these for new patients based on the gene expression/copy number profile. The regression coefficients can thus be estimated through Cox regression, but will often involve a need for special methods, because the number of variables (genes) p, in most cases are much larger than the number of samples n (e.g. in cancer research where the number of patients is limited). Different methods are designed especially for $p > n$ problems. These imply dimension reduction either by for example variable selection and/or shrinkage methods, which makes it possible to estimate the regression coefficients.

This chapter reflects some methods used in this regression perspective. The mathematical and computational need for such methods is described in Section 4.1, whereas the methods are described in Section 4.2 and Section 4.3 and later introduced to the Cox setting in Section 4.4. Finally in Section 4.5, K-fold cross-validation is described in the context of the regression methods discussed and some computational aspects are considered in 4.6. Unless otherwise is stated, the theory of this chapter is obtained from Hastie et al. (2001, chap. 2-3 and 7), which gives an exhaustive overview of the relevant statistical theory.

4.1. Challenges in Regression Analysis of $p > n$ Data

The method of least squares is a well known method for estimating regression coefficients and is maybe the most used method for this purpose. If we let $\mathbf{X}$ be the $n \times p$ matrix with n measurements of p explanatory variables, and $\mathbf{y}$ the vector of n responses, the method of least squares is concerned with minimizing the residual sum of squares given by

$$RSS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

The least squares estimates of the p regression coefficients will then be

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{4.1}$$

In some situations the least squares estimates are not possible to compute. If the columns of the matrix $\mathbf{X}$ are linearly dependent (i.e. the covariate measurements are correlated), the least squares estimates are inadequate. If the number of columns p is larger than n, two or more columns of $\mathbf{X}$ have to be collinear causing $\mathbf{X}^T\mathbf{X}$ to be singular. Since $\mathbf{X}^T\mathbf{X}$ has to be inverted in order to find the least squares estimates given in (4.1), and since a matrix has to be nonsingular to be inverted, the least squares method will not give satisfying results and specific statistical methods are required when $p > n$.

Since $\mathbf{X}$ in practice will not be exactly collinear, the matrix $\mathbf{X}^T\mathbf{X}$ will only be almost singular as opposed to exactly singular. It will be possible to invert the matrix, but this may lead to some very large entries along the diagonal of $(\mathbf{X}^T\mathbf{X})^{-1}$. This may be compared to taking the reciprocal of a very small number. Since the variance of each regression coefficient is given as the corresponding diagonal element in $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, this will cause least squares coefficients which may have a very large variance.

The problems arising doing regression with many covariates are therefore both a statistical and a computational problem. Numerical inaccuracy will be a consequence of trying to invert an almost singular matrix, and statistically, the variance of the regression coefficients will be large and cause uncertain estimates and prediction results (Birkes & Dodge, 1993, chap. 8).

By the Gauss Markov theorem the least squares estimates will be BLUE (best linear unbiased estimator), which means that the least squares estimates are the estimates of the parameters having the smallest variance among all linear unbiased estimates (Hastie et al., 2001, chap. 3). In situations where $p > n$, these estimates with the restriction of being unbiased are not preferable if the variance becomes too large and may in prediction settings lead to a lower prediction accuracy. In this case it may exist estimates which are biased, but have a smaller variance and which will lead to better prediction results.

Generally there is a trade-off between bias and variance due to the model complexity. Typically, the more complex the model is, the variance will increase, but the bias will be lower. If the model complexity is decreased the lower variance and the higher bias. This is illustrated in Figure 4.1, where the prediction error curves for a "training" and a "test" set are given as a function of model complexity. This shows that one will be able to predict well for the training set (for which the model is constructed), when using a complex model. But for an independent "test set" the prediction error will increase if the model is too complex. The "right" model will take into consideration to find the perfect balance between bias and variance. Methods handling situations where $p > n$ and/or where the covariate vectors are linearly dependent, reduce the model complexity to trade a little bias for a reduction in the variance. Prediction accuracy may be improved by shrinking the regression coefficients or setting some regression

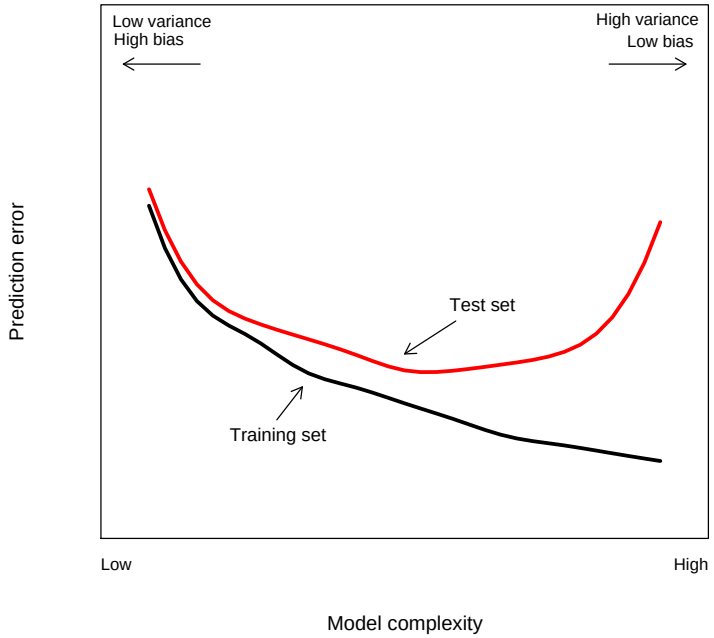


FIGURE 4.1. The figure illustrates the trade-off between bias and variance. The prediction error for a training and test set are drawn as functions of model complexity. From the figure it is obvious that in manner to get a low prediction error for an independent test set, one has to choose a model, which finds the perfect balance between bias and variance. Figure reproduced from Hastie et al. (2001, chap. 2)

coefficients equal to 0, as motivated above in terms of model complexity. Both setting some least squares coefficients equal to 0 and shrinking them will lead to biased estimates, but will reduce the variance (Hastie et al., 2001, chap. 3). Another reason for using a selection method to select a subset of covariates to use in the model is interpretation. By selecting a smaller number of variables, the results will often be easier to interpret compared to when a large number of predictors are used (Tibshirani, 1996). Subset selection and shrinkage methods are therefore widely used to improve interpretation and prediction accuracy when dealing with regression on $p > n$ data.

4.2. Subset Selection

There are several different approaches for doing subset selection. Selection of variables could be according to a number of different criteria, either specific for the data type at hand or by other mathematical criteria typically minimizing an estimate of the expected prediction error. When a criterion is chosen, the subset is produced by selecting all variables which satisfy the criterion. Variables which do not fulfill the criterion are eliminated from the model. In this way

only the variables in the subset are retained in the model. Best subset regression is one example, which for each $k \in \{0, 1, 2, ..., p\}$, finds the subset of size k, which minimizes the residual sum of squares. The subsets may vary in size and methods like forward and backward stepwise selection or combinations of these may be used to define the subsets.

4.3. Shrinkage Methods

Shrinkage methods shrink the regression coefficients toward 0 by minimizing the residual sum of squares subject to a constraint on the parameters. The minimization problem will therefore correspond to minimization of

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2, \qquad \text{subject to } \sum_{j=1}^{p}|\beta_j|^s,$$

for $s \geq 0$ and where $x_i = (x_{i1}, ..., x_{ip})^T$ is the covariate vector for person i. This is mathematically equivalent to

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|^s, \tag{4.2}$$

introducing $\lambda\sum_{j=1}^{p}|\beta_j|^s$ as a penalty term by the method of Lagrange multipliers and where λ controls the amount of shrinkage. In the case of a Gaussian likelihood this corresponds to subtracting the penalty term from the log-likelihood to be maximized. As λ increases, the regression coefficients are shrinked toward zero. Different values of s may be chosen and correspond to different constraint regions as illustrated in Figure 4.2. The constraint region has influence on whether the method is a selection or a shrinkage method depending on whether some regression coefficients are set to zero or not. The estimated solutions of the regression problem will be in the first point where the elliptical contours of the residual sum of squares hits the constraint region.

Shrinkage methods are improving the prediction accuracy by adding a penalty term as described above. Why this actually reduces the variance becomes more obvious by writing the problem in matrix form. The main reason is that an extra term is added to the diagonal elements of the matrix $\mathbf{X}^T\mathbf{X}$ to be inverted. This will cause not as high entries in the inverted matrix such that the regression coefficients become lower (shrinked), but also cause lower variance which is desired. More comments on this are given in the next section, considering ridge regression.

4.3.1. Ridge Regression. Ridge regression corresponds to minimizing (4.2) when $s = 2$, that is, minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2. \tag{4.3}$$

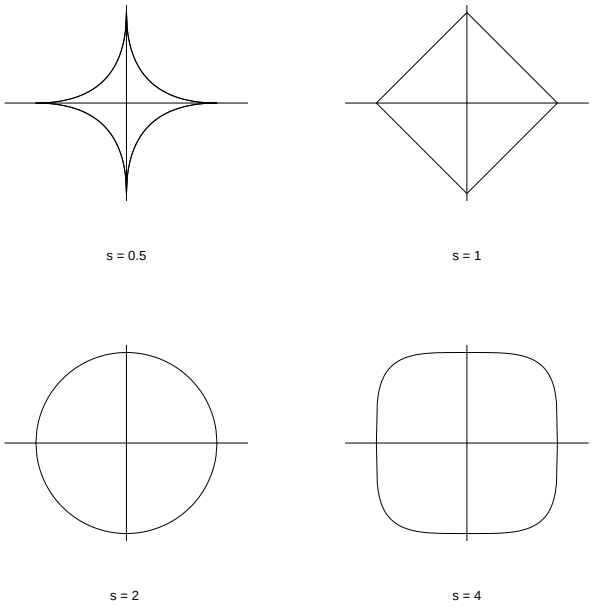


FIGURE 4.2. Different shapes of the constraint region depending on the value of s considering two input variables (Hastie et al., 2001, chap. 3). The constraint region when $s = 2$ corresponds to ridge regression and the region where $s = 1$ corresponds to the lasso.

In the case of two parameters, the constraint region corresponds to a circle and will rarely set regression coefficients equal to zero. Figure 4.3(a) gives a geometrical illustration of the ridge solution in the case of two covariates.

It is convenient to describe ridge regression also in matrix notation. The expression in (4.3) may be written as:

$$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}.$$

The ridge solutions will then be given by

$$\hat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\boldsymbol{I})^{-1}\mathbf{X}^T\mathbf{y}. \tag{4.4}$$

This shows that by adding the term λ to the diagonal of the matrix to be inverted, $\mathbf{X}^T\mathbf{X}$ is modified so that it is farther from singularity. From (4.4) it is obvious that this also makes the estimates smaller than the least squares estimates. It is also clear that since the least squares estimates are unbiased, the ridge estimates are not. But if we look at the variance, the penalty term in (4.3) should reduce the variance compared to the variance of the least squares estimates. The covariance matrix is given by (Gruber, 1998, chap. 3)

$$COV(\boldsymbol{\beta}_{Ridge}) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\sigma^2(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}, \tag{4.5}$$

which indicates that the variance given by the diagonal elements of the covariance matrix will be smaller than the diagonal elements in $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

4.3.2. The Lasso. As described in Tibshirani (1996) both subset selection and ridge regression have drawbacks. Subset selection gives a more interpretable model, but the performance may be unstable because the method is a discrete process where the variables are either included in the model or not. Ridge regression is a continuous process and is more stable, but may give results which are hard to interpret if the number of covariates is large. The least absolute shrinkage and selection operator (The lasso) was proposed by Tibshirani (1996) to keep some of the positive features of both subset selection and ridge regression. The lasso shrinks some coefficients and sets others equal to 0. The method differs from ridge regression simply by minimizing the residual sum of squares subject to another constraint region and will as a consequence produce some regression coefficients equal to zero.

The minimization problem in the lasso corresponds to (4.2) with $s = 1$, that is

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{4.6}$$

The constraint region when $s = 1$ will be a diamond in dimension 2 (see Figure 4.3). In higher dimensions it will, as opposed to when $s = 2$ in ridge regression, often be corners on the constraint region for the contours of the least square error function to hit. Since the solution of the regression problem will lie in the intersection of the constraint region and the contours of the residual sum of squares, the estimates will be equal to 0 much more often than for ridge regression, especially for higher dimensions. This will produce regression coefficients equal to 0, and classifies the lasso as a selection method in addition to being a shrinkage method.

In Tibshirani (1996) the covariance matrix is approximated by writing the penalty

$$\sum |\beta_j| \approx \sum \frac{\beta_j^2}{|\beta_j|}.$$

The lasso estimates may then be approximated with the solution in a ridge regression, that is

$$\boldsymbol{\beta}^*_{Ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-)^{-1}\mathbf{X}^T\mathbf{y}.$$

Here $\mathbf{W}$ is a diagonal matrix with diagonal elements $|\tilde{\beta}_j|$ and $\mathbf{W}^-$ is the generalized inverse of $\boldsymbol{W}$. Then the approximate covariance matrix of the estimates in the lasso will by using (4.5) be

$$COV(\boldsymbol{\beta}_{Lasso}) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-)^{-1}\sigma^2(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{W}^-)^{-1}, \tag{4.7}$$

where σ^2 can be estimated by an estimate of the error variance $\hat{\sigma}^2$.

4.4. Shrinkage Methods in the Cox Setting

The shrinkage methods may also be defined in terms of the likelihood instead of the residual sum of squares. In the case of a Gaussian likelihood, minimization of (4.2) corresponds to adding a penalty $-\lambda \sum_{j=1}^{p} |\beta_j|^s$ to the log-likelihood to be maximized.

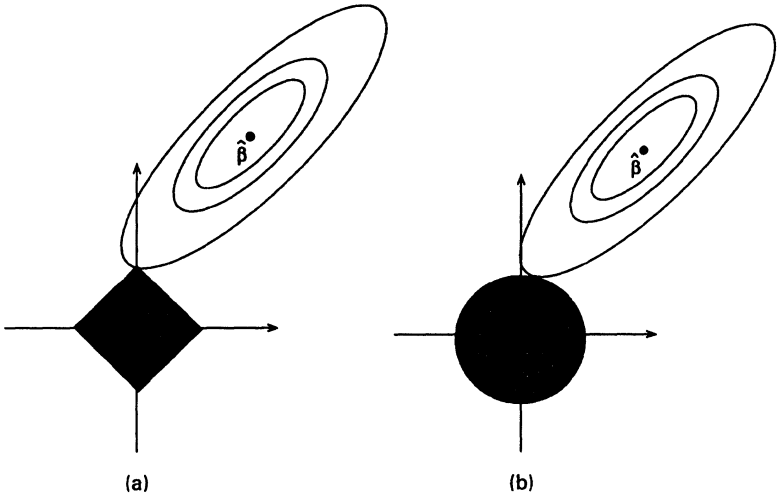


FIGURE 4.3. The two figures illustrate the ridge and the lasso solutions in two dimensions, which are in the intersection of the residual sum of squares error and the constraint regions. Similar illustrations may be given for the other constraint regions shown in Figure 4.2. The figure is from Tibshirani (1996).

Shrinkage methods in the Cox setting will correspond to maximizing a penalized version of the Cox log partial likelihood;

$$l(\boldsymbol{\beta}) = l_{pl}(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j|^s.$$

The lasso ($s = 1$) in the Cox setting was proposed by Tibshirani (1997). He assumes standardized input for the use of the lasso and this is also often assumed for ridge regression. This is to assure that the variables are treated in the same way by the penalization scheme, and is especially important if the variables are not measured at the same scale. This is mainly because one adds the same number λ to each diagonal entry in the matrix $\mathbf{X}^T\mathbf{X}$. In van Houwelingen et al. (2006) it is pointed out that in Cox regression analysis of microarray data there is no need for standardization because the covariates are already on the same scale. For instance, in the gene expression data all of the variables measure gene expression level and are therefore measured on the same scale. This is opposed to weight, age, height and other covariates of interest, which are measured in different units. Centering is neither necessary because a centering of the covariates will be compensated by the baseline hazard.

4.5. Cross-Validation

In the previous sections the models introduced all depend on a tuning parameter λ. In shrinkage methods the tuning parameter controls the amount of shrinkage whereas in subset selection the tuning parameter may be the number of variables to include in the model. In all of these methods the tuning parameter varies the complexity of the model.

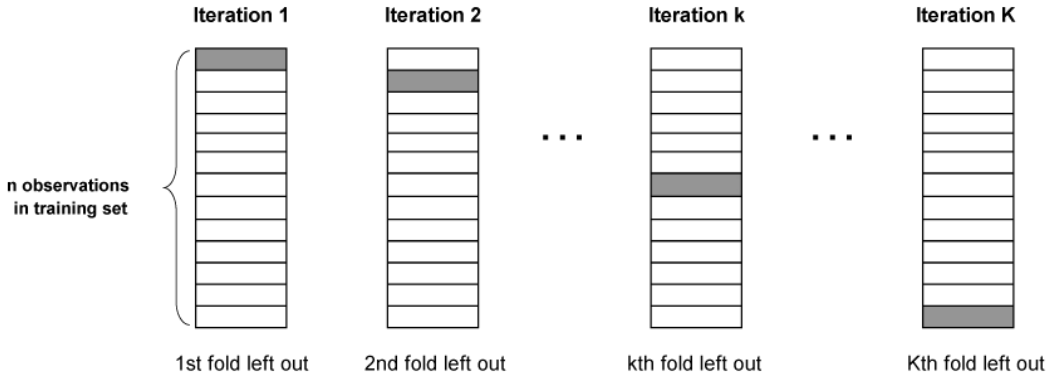


FIGURE 4.4. Illustration of K-fold cross-validation. We concentrate on the observations of the training set and these are divided into K folds. In each iteration the marked fold is kept outside the fitting of the model. By keeping one fold outside the estimation procedure we may consider the left out fold as a test set, and test the prediction capability for this left fold. By iterating and keeping one new fold outside the estimation in each iteration, we end up with K estimates of prediction capability. The K estimates may be combined to estimate the prediction capability of the overall model.

The discussion opening this chapter points on the challenges in analyzing high dimensional data and the interplay of bias, variance and model complexity. This is illustrated in Figure 4.1. It was explained how the model which will produce a low prediction error for an independent test set, has to find the perfect balance between bias and variance. By estimating the performance of a group of models one may choose the model with the best estimated performance. In an ideal situation with *enough* data, this could be done by dividing the data into three parts: one which should be used to fit the models, a second part which should compare the performance of the suggested models, and a third part which should test the performance of the model on an independent test set. The comparison of the performance in the different models corresponds to estimating the performance of the model and evaluate them for different values of the tuning parameter λ, to finally select the (approximate) best model.

In most cases we are not in a situation with a lot of samples, and it would be insufficient to divide the data in three parts. In many situations, dividing into training and test set is costly enough. Setting aside too much data for testing may give a poor fit and having a too small test set, the estimated test error may not be as reliable. The evaluation step may, however be approximated. There are several methods for doing this including analytical methods as AIC, BIC and other related measures of test error, but the most widely used approach is cross-validation.

4.5.1. K-fold Cross-Validation. K-fold cross-validation uses one part of the training data to fit the model and a separate part to test the model. The general idea of K-fold cross-validation is to divide the data into K folds and leave one fold out to calculate the prediction error. This procedure is repeated for all K folds, leaving a new fold out at a time. The measures of prediction

error may be combined in an estimate of the prediction error;

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}^{-\kappa(i)}(\boldsymbol{x}_i, \lambda)). \tag{4.8}$$

Repeating this for a grid of λ values, one may minimize the cross-validation curve $CV(\lambda)$ to find the model, which minimize the prediction error. In expression 4.8 the cross-validation function is expressed in terms of the loss function $L(y_i, \hat{f}(\boldsymbol{x}_i, \lambda))$. The cross-validation function may just as well be written in terms of the log-likelihood. For the Cox regression setting the criterion of Verweij & van Houwelingen (1993) is one option. Verweij & van Houwelingen (1993) present leave-one-out cross-validation in survival analysis which is modified to yield K-fold cross-validation in Bøvelstad et al. (2007).

Assume n observations and denote the Cox log partial likelihood by $l(\boldsymbol{\beta})$. We let $l_{(-k)}(\boldsymbol{\beta})$ be the Cox log partial likelihood when the kth fold is left out and $k = 1, 2, ..., K$. The estimates of the regression coefficients $\boldsymbol{\beta}$, when the kth fold is left out and the tuning parameter λ is used, are denoted $\hat{\boldsymbol{\beta}}_{-k}(\lambda)$. The cross-validation function may then be written by

$$CV(\lambda) = \sum_{k=1}^{K}\{l(\hat{\boldsymbol{\beta}}_{(-k)}(\lambda)) - l_{(-k)}(\hat{\boldsymbol{\beta}}_{(-k)}(\lambda))\}.$$

The cross-validation criterion $CV(\lambda)$ gives a measure of the prediction capability of the models corresponding to different λ-values, and by maximizing $CV(\lambda)$ with respect to λ, we may optimize the prediction capability and find the optimal tuning parameter $\hat{\lambda}$.

To perform K-fold cross-validation we also have to define K. The maybe most commonly used are $K = 5$ and 10. $K = N$ corresponds to "leave-one-out" cross-validation which gives a nearly unbiased estimate for the true prediction error but may suffer from large variance (Hastie et al., 2001, chap. 7).

4.6. Computational Aspects

To carry out the ridge and lasso analyses in the Cox regression setting, programs developed by Bøvelstad et al. (2007) are applied. The main aim in Bøvelstad et al. (2007) was to compare the predictive performance of microarray data by using different dimension reduction methods. The programs are therefore especially designed to handle high dimensional microarrays in combination with survival data. The programs are coded in Matlab and R and take the data organized in a $n \times (p + 2)$ matrix as input. The output is the estimated cross-validation parameter λ and the estimated regression coefficients for the chosen method (i.e ridge, lasso).

The ridge procedure is coded in Matlab whereas the lasso is coded in R. Both programs apply the Cox regression model in combination with the shrinkage methods as proposed for the lasso

in Tibshirani (1997) and discussed in Verweij & Houwelingen (1994) for ridge regression. In the R program code for the lasso, the R package `glmpath` containing a lasso implementation of Cox regression is applied. By using a predictor-corrector method the entire path of coefficient estimates are determined as λ varies. Park & Hastie (2007) gives a comprehensive presentation of the implementation of the methods found in the `glmpath` package. See also `http://cran.r-project.org/web/packages/glmpath/glmpath.pdf`, which gives the instructions for use.

As described in the previous section, shrinkage methods require estimation of a tuning parameter. To estimate the cross-validation parameter λ, 10-fold cross-validation is used. The cross-validation criterion is, as described in the previous section, a more general form of the leave-one-out cross-validation criterion based on the Cox log partial likelihood (Verweij & van Houwelingen, 1993).

Some parameter values are to be determined in the program of Bøvelstad et al. (2007). In the analyses applied to the data we use their default values. All of them are evaluated by studying the manual of `glmpath` and is appropriate for our data set. We must have $\lambda > 0$ and the lower bound l in the grid of λ-values was set to be; $l = e^{-100}$. For further description of the programs of Bøvelstad et al. (2007), see `http://www.med.uio.no/imb/stat/bmms/software/microsurv/`.

CHAPTER 5

Lasso Regression Analyses of the Cervix Cancer Data

We have already addressed the statistical challenges involving a large number of covariates and a relatively small sample size related to regression analyses. Two different methods which enable regression in these settings, that is ridge regression and the lasso, were discussed in Chapter 4. Even if these methods make it possible to fit the regression models, it is not obvious that our concerns handling the large number of covariates are solved. Microarrays contain a large number of predictors, presumably only few are important for survival Park et al. (2007). Thus, keeping all available covariates in regression analyses may not be preferable since the data will contain a lot of noise (covariates not related to survival). Shrinkage methods are supposed to handle this by shrinking the coefficients of unimportant covariates, but including a huge number of covariates, which are not related to the response of interest, may interfere the analysis. No nonzero regression coefficients as a result in the lasso can be a consequence.

To overcome these issues it seems that reducing the data set prior to the analyses is convenient and necessary, even if shrinkage methods are applied. This also seems appropriate in the literature where reduction of the number of covariates is commonly used before the regression methods are applied. Fitting univariate Cox regression models for each gene and thereafter select significant genes based on the p-values is one approach, which is commonly used for this purpose, see for instance Park et al. (2007) among others. The problems, involved in selecting genes according to their significance in a univariate Cox regression model, are two-fold;

- Genes are selected according to their marginal influence on survival.
- The survival data are used for selection.

When applying selection methods, the aim is to select the genes that are related to the event of interest and exclude those that are not. Methods evaluating genes one by one may however, not fulfill this wish. Since genes may depend on each other in some unknown fashion (Bøvelstad et al., 2007), they may be found important in a joint analysis even if they are not explaining survival alone. If the aim is to exclude non significant genes, the fact is that some of the genes excluded by a univariate criterion may explain survival together with other genes. The other issue is related to prediction. A method which uses the survival data in different steps of the analysis may cause prediction results that are overoptimistic.

Park et al. (2007) also address the need of imposing methods to reduce the dimension of the data prior to running the lasso. They suggest to apply hierarchical clustering on the gene expression data to thereafter define "supergenes", which are the average of the genes in a cluster and use these "supergenes" as regressors in the lasso. A drawback with defining "supergenes", is that the actual gene expressions are not used in the analyses. Not defining "supergenes" as done in Park et al. (2007) may, however, reveal other challenges which are not desirable. For instance the selection of a representative may not be obvious and the determination of groups or clusters may not be sensible if there in no obvious grouping in the data.

Another approach discussed by van de Wiel & Wieringen (2007) is directed to aCGH data. The method aims on dimension reduction with minimal information loss and is concerned of defining a smaller number of regions, which can be used for further downstream analysis. Based on the high correlation between the vectors for the probes in the aCGH data, they determine regions consisting of vectors which are almost equal when coding for loss, normal and gain (-1,0,1) and which correspond to genes within the same chromosome. The regions are constructed by restricting the maximum distance between any two vectors in a region by a given threshold. A representative for each region could thus be used in further analyses. Also van de Wiel & Wieringen (2007) emphasize the benefits (interpretation, computational time) of dimension reduction in large data sets.

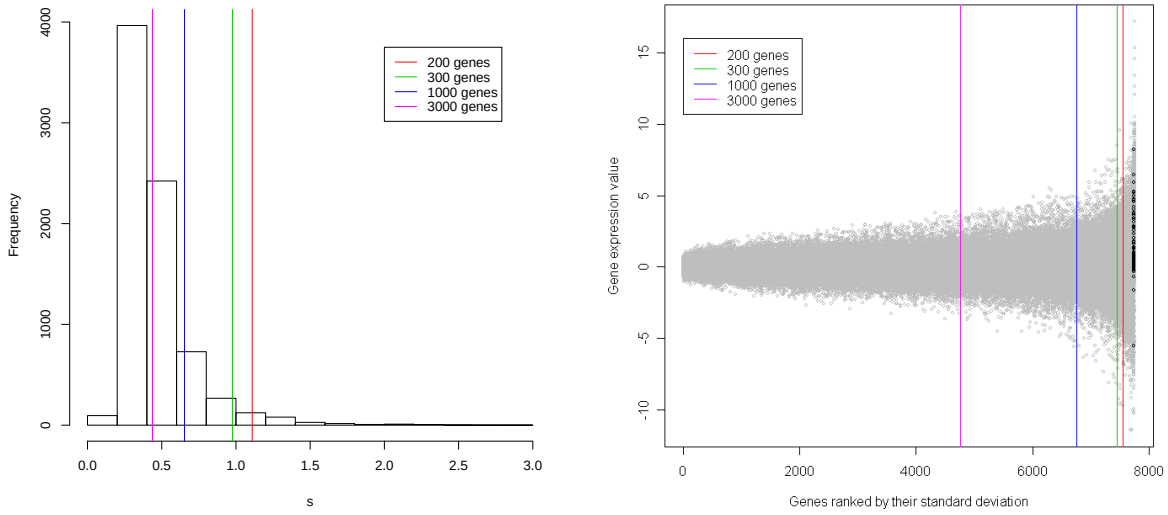
In the next sections we will study the gene expression data and the aCGH data separately by fitting Cox-lasso regression model. The aim of the analyses is to evaluate different methods, which can be used to reduce the data, and the analyses will illustrate some of the issues concerning data reduction. A couple of simple criteria are determined to define data sets of smaller size. All genes or regions that are selected in the analyses are listed in Appendix A.

5.1. Lasso Regression Analyses on the Gene Expression Data

The original gene expression data set were fitted in a Cox-lasso regression model. Only one gene was selected, that is gene (307660) which is located on chromosome 8 and corresponds to the copy number probe with probe-identification RP11-34M16. The gene symbol is FABP4. Two different approaches were used to reduce the number of covariates in the gene expression data;

- Variance
- Univariate Cox regression

5.1.1. Variance. The first example used variance as a criterion for selection. When searching for genes which cause differences in survival time, it is necessary that the gene expressions are different across the patients. The gene expression for a patient with poor survival should be



(a) (b)

FIGURE 5.1. (a): Histogram of the standard deviations for all of the genes and the thresholds indicating the four gene sets of size 3000,1000, 200 and 100 are marked as vertical lines.
(b): The gene expression data are plotted when the genes are sorted by their standard deviations. The genes to the right of the vertical lines, corresponding to the same thresholds as in (a), are those included in the gene sets.

different than the gene expression for a patient with no relapse of the disease, for the gene to explain these differences. Genes with higher variance in expression over the patients will therefore be crucial in the analysis and genes with less variance seem less important when explaining (predicting) survival.

To select genes according to variance, the empirical standard deviation of the gene expressions was calculated for each gene. The top ranked genes were picked out to be used in the analysis, whereas genes showing less variance were kept outside the regression. Since there are no obvious limit for which a standard deviation is high, the top ranked genes are used to determine the gene sets. Subsets of 3000, 1000, 300 and 200 top ranked genes according to their variance were defined and used as covariates in the lasso regression analyses.

A histogram of the calculated standard deviations are given in Figure 5.1(a), and the thresholds determining the sets of covariates are marked as vertical lines. In Figure 5.1(b) the data are plotted when sorted by their standard deviations in an increasing order from left to right. The vertical lines are corresponding with the thresholds in the histogram, and the genes to the right of the vertical lines are those included in the gene sets. From the plot it seems obvious that there are relatively large differences in variation from gene to gene. All four lasso regression analyses using variance as a criterion for data reduction prior to the regression, resulted in one

selected gene by the lasso. The gene is the same as the one selected when no preselection was used, and is marked black in Figure 5.1(b). In the plot we see that the gene has a large variance and was thus included in all of the four subsets.

Using variance as a selection criterion showed not to gain further insight, as the only gene selected is the same as for the total gene expression set of 7754 genes. The motivation for reducing the data according to variance was that a gene will not contribute to differences in survival if the expression of the gene is equal for all of them. It is, however, not necessarily the genes with the largest variance that are responsible for these contributions. When selecting genes according to their variance, one most likely include many genes which are subjects to noise or measurement error. Even if one may remove redundant information by excluding the genes with small variance, one is possibly still dealing with a high level of noise.

5.1.2. Cox Univariate Regression. A frequently used method for reducing the data set is to select genes, which are found to influence survival through univariate Cox regression. This is done by fitting a univariate Cox regression model for each gene and find which regression coefficients are significantly different from zero. Although reduction methods applying a univariate Cox regression to reduce the data is directly using the survival data, it is frequently used in the literature. This is not always a good method and when the (prediction) performance is to be evaluated, overoptimistic results may occur.

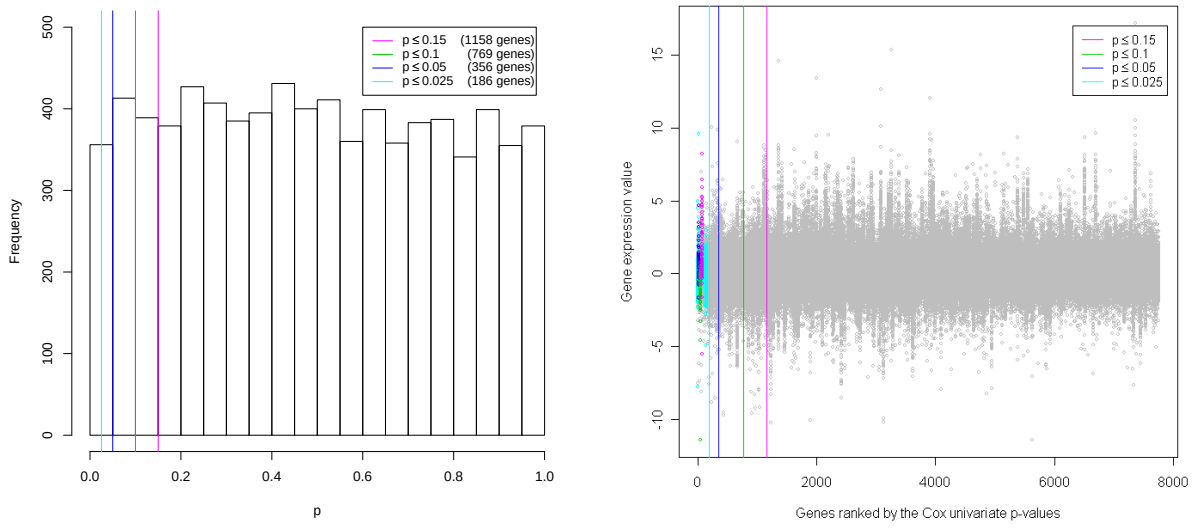
The univariate Cox regression model

$$h(t|\boldsymbol{x}_j) = h_0(t)\exp(\boldsymbol{x}_j\alpha),$$

was applied, where $\boldsymbol{x}_j$ is the gene expression vector for gene j. When the model is fitted, one may test the hypothesis $H_0 : \alpha = 0$, and small p-values indicate that gene j has an influence on survival and will be reasonable to include in the analysis.

Four thresholds were defined to produce the four subsets; $p \leq 0.15, 0.1, 0.05, 0.025$. It is most common to use a significance level of 0.05 for this purpose, but the two higher levels are included in this example to include genes which are close to significant. For each of the four gene sets some genes were selected and are listed in Appendix A.

All of the genes selected by the lasso for the four different gene sets are included in the smallest gene set, that is, the set only containing genes which are significant on a 0.025 level. The number of genes selected increases when the significance level is low. This is sensible since we remove genes which do not show as strong significant influence on survival marginally. The genes with $p \leq 0.025$ are plotted in Figure 5.3, to be able to differentiate between the results for each of the four subsets. The analysis for the largest gene set including genes with a $p \leq 0.15$ only selected one gene (FABP4). In the set where $p \leq 0.1$, two genes were selected and three in the gene set



(a) (b)

FIGURE 5.2. (a): Histogram of the Cox univariate p-values for all of the genes. The thresholds indicates the four significant levels 0.15, 0.1, 0.05 and 0.025 and are marked as vertical lines.
(b): The gene expression data are plotted when the genes are sorted after their marginal ability of explaining the survival (p-values). Note that here it is the genes to the left of the vertical lines which are included in the analyses, for this to coincide with the histogram.

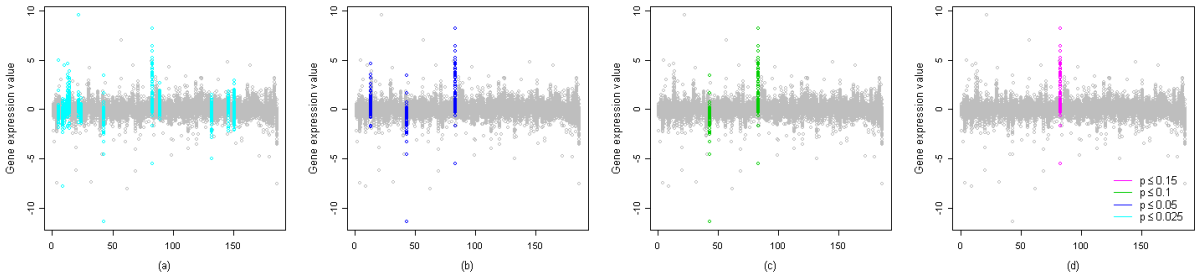


FIGURE 5.3. Genes selected by the lasso in the four gene sets with $p \leq 0.025, 0.05, 0.1$ and 0.15 respectively

containing genes with a univariate $p \leq 0.05$. For the 186 genes with $p \leq 0.025$, 13 genes were selected by the lasso.

Obviously an analysis where only genes which explains the survival one by one are included, will be more likely to select genes which are able to explain the survival together. Genes that are excluded because of a high univariate p-value, may however be important in a joint analysis. The excluded genes may therefore be just as relevant in the search of important biomarkers.

One should also stress that the survival data are not used when selection is performed. This is especially important when the prediction performance is to be assessed.

5.2. Lasso Regression Analyses on aCGH Data

Regression analysis with survival as an outcome and aCGH data as covariates has to our knowledge not been extensively studied. From a regression perspective, analyzing the aCGH data may be even more challenging than the gene expression data, due to the high correlation among aCGH probe vectors. It could also be less correlated to survival than expression data since a change in copy number does not effect survival unless it also influences the expression for corresponding gene(s). The aCGH data may, however, be of better quality than expression data and may reveal interesting regions related to survival.

For the aCGH data, zero probes were selected when the complete data set containing 2138 probes were analyzed in the lasso. It was thus convenient to reduce the number of covariates for these data as well. The four approaches studied to reduce the number of covariates in the aCGH data were

- Variance
- Univariate Cox regression
- Clustering

Here variance and univariate Cox regression are selection methods (the same as studied for the gene expression data) and the actual probes are used to fit the model. The last approach cluster the covariates into groups according to how similar they are and chooses a representative for each group to use in the regression.

One should note that it may be difficult to strictly compare the results of these analyses. This is due to the fact that the lasso tends to select one (Zou & Hastie, 2005) or some of the variables in blocks of correlated covariates. Because of the very correlated behavior of the aCGH data this may occur more frequently than for the gene expression data. It is thus a risk that the lasso selects different probes in different analyses, but which are very correlated. For a selected probe, the lasso could just as well has chosen another probe which are very correlated to the selected gene. This is important if the selected probes are to be compared for the different approaches and subsets.

5.2.1. Variance. Selecting probes by their variance can be motivated in the same way as for the gene expression data. A gene showing a large variation in copy number alteration across the patients is believed to influence the survival time for the patients. Probes which are equal

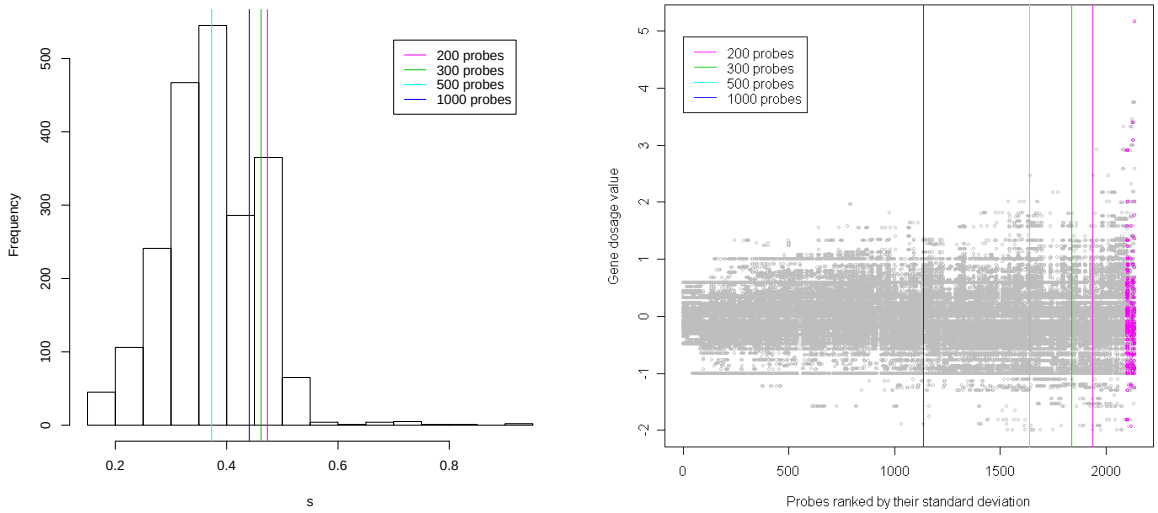
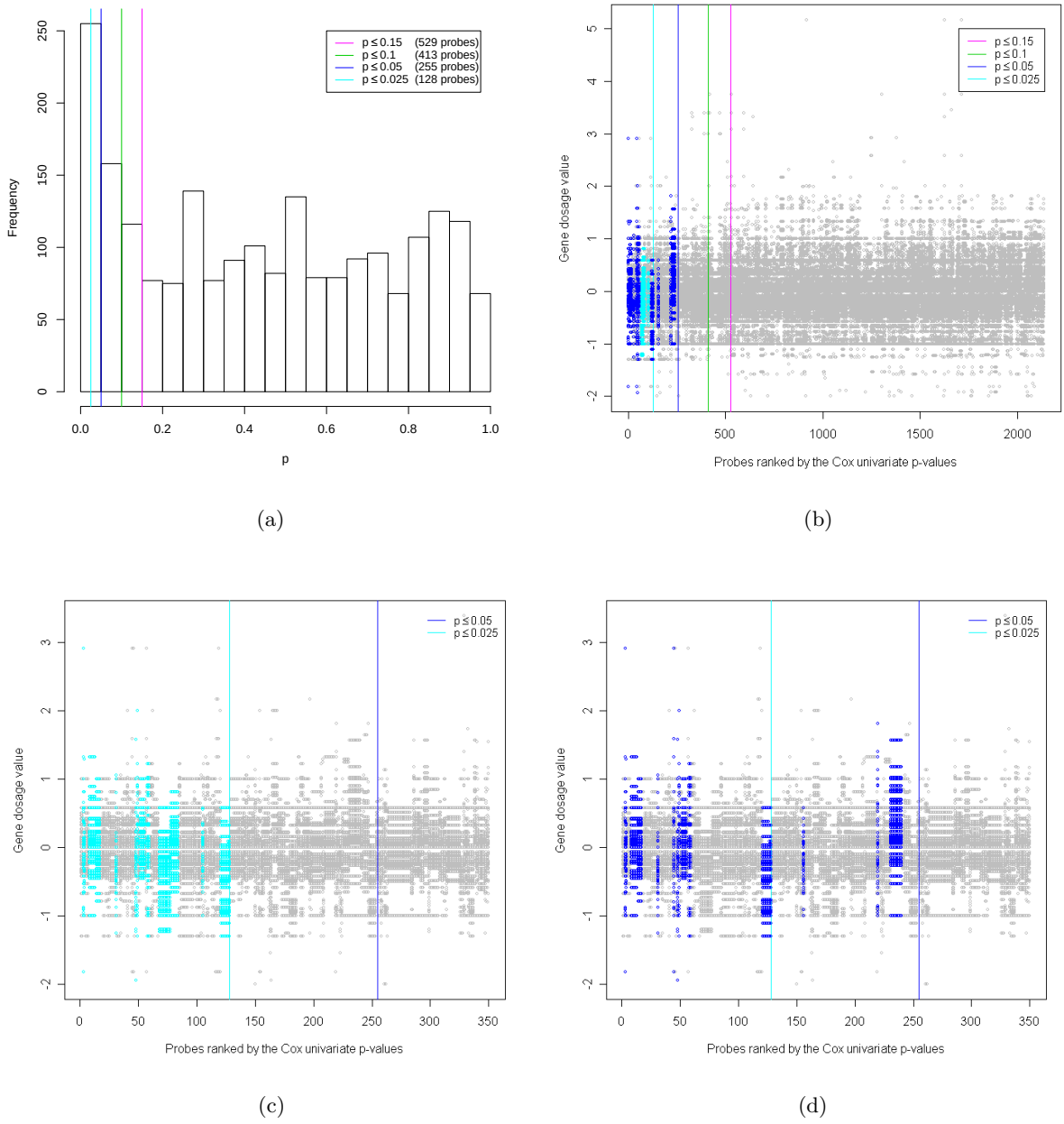


FIGURE 5.4. (a): Histogram of the empirical standard deviations for every aCGH probe. The vertical lines show the thresholds used to determine the gene sets of size $1000, 500, 300$ and 200 probes. (b) Probes plotted when sorted by their standard deviation and the colored lines correspond to those in the histogram in (a).

for all patients will not contribute to differences in the survival times, since they will all influence the survival in the same way.

As for the gene expression data the empirical standard deviation of each probe was calculated to reduce the aCGH data. The probes with the largest standard deviations are included in smaller subsets of size 1000, 500, 300 and 200. Figure 5.4(b)gives the histogram of the empirical standard deviations and a plot of the probes sorted by their variance. Six genes are selected for the subset of 200 probes. The probes are marked in Figure 5.4(b), and given in appendix A. For the large subsets, zero probes were selected by the lasso.

From the analysis where variance is used as a selection criterion for the gene expression data, it is discussed that including only genes with high variance may correspond to keeping a lot of noise in the analysis. To improve this method for the aCGH data one could for instance implement a sliding window, which checks that genes included in the subsets also have to be correlated with its neighbors, say a window of 5 or 10 neighbors. This could have excluded some of the noise, since the probes in the aCGH data are correlated in blocks and a probe is most likely not noise if it is correlated with its neighbors. Other problems may enroll using this approach connected to the many evaluations which have to be done; window size, the degree of dependence required, and the number of variables to constitute the subsets.



(a) (b)

(c) (d)

FIGURE 5.5. (a): Histogram of the Cox univariate p-values for each probe.
(b): Probes plotted when sorted by their p-values.
(c)-(d): The probes with the smallest p-values are plotted to get a clearer view of the selected probes.

5.2.2. Univariate Cox Regression. The second method studied for selection of probes in the aCGH data was univariate Cox regression. As for the same approach for gene expression, a univariate Cox regression model was fitted for each probe and a test was performed indicating the probes influence on survival. Small p-values in the test indicate that probe j has an influence on survival. The four thresholds $p \leq 0.15, 0.1, 0.05, 0.025$ define the subsets.

Including all probes with a p-value less than 0.15 and 0.1 did not lead to selection of any probes by the lasso. Reducing the data to only include probes with $p \leq 0.05$ and 0.025 corresponding to subsets of size 255 and 128 probes respectively, lead to selection of 47 and 42 probes. The aCGH data ordered by their univariate p-values are plotted in Figure 5.5(b). Some of the selected probes differ in the two analyses, as may be seen in Figure 5.5(c)-(d). This could be due to the fact that the lasso may not select the same probes in both settings if the probes are very correlated. This makes it somehow difficult to compare the results. It seems that when reducing the subsets from including probes with $p \leq 0.05$ to 0.025, some probes were excluded, but other probes with a p-value less than 0.025 were selected by the lasso.

As commented for the gene expression data, one should not rely completely on these results since the survival data are already used in the preselection. Since the probes included in the analysis are found to individually have a significant influence on survival, it is sensible that more probes are selected than for the other methods.

5.2.3. Clustering. The third approach is related to the approach described in Park et al. (2007). The method utilizes the special features of the aCGH data being very correlated. By dividing the probes into groups according to how similar they are, one may use only one representative for each cluster and thus reduce the number of covariates. This is reasonable considering the copy number probes which are very correlated and appear in blocks of nearly similar probes.

Hierarchical clustering was used to group the probes in a smaller number of clusters. Complete linkage and Spearman correlation as a distance measure were used. A representative for each cluster could then be defined. Both randomly chosen probes in a cluster and an average of the probes in a cluster were used as representatives. Using clustering as a method for reducing the number of covariates in the regression analysis did not lead to any nonzero regression coefficients.

There are a few decisions, which have to be taken when using this approach and other choices could have lead to different results. The number of clusters are chosen manually and the choice of representatives may not be the optimal to represent a cluster.

5.3. Integrated Data Reduction

Having both gene expression and copy number data available, it could be convenient to integrate the information in both data sources when reducing the data. If a gene's expression shows correlation with gene dosage it is considered as a possible driving force for cancer progression (Lando et al., 2009). Gene expression vectors which are correlated with the corresponding aCGH probe vector, are therefore believed to influence survival and are thus of specific interest when the gene expression data are analyzed. A similar approach may be used to reduce the aCGH

data as well. Probes which do not show a high correlation with gene expression are reasonable to exclude from the analysis since they can not influence the survival if none of the corresponding gene expression vectors are correlated with the probe.

To use correlation as a criterion, the Spearman rank correlation between each gene expression vector and its corresponding aCGH probe, was calculated after matching the vectors in the two data sources. The Spearman rank coefficient is equivalent to Pearson's product-moment correlation between the ranks of each gene's expression and dosage. Spearman correlation differs from the more commonly used Pearson correlation by not measuring the linear relationship, but rather any monotonic relationship between two data vectors. There is monotonic dependence between gene expression and gene dosage, if gene dosage increase when gene expression increase and vice versa. The coefficient may be interpreted the same way as Pearson correlation coefficient, where 1 indicates perfect positive correlation of the ranks indicating a monotonic increasing relationship, and -1 a monotonic decreasing relationship (Berrar et al., 2003, chap. 17).

5.3.1. Gene Expression Data. When the Spearman correlation coefficients was calculated, subsets with genes showing a relatively large correlation with gene dosage were defined. A histogram of the calculated Spearman correlations are given in Figure 5.6(a) where the thresholds used to define the gene sets are shown as vertical lines. The thresholds are decided to be $r > 0, 0.2, 0.3, 0.4$. Note also that the genes of interest are those showing a positive correlation whereas the change in gene expression showing a large negative correlation with gene dosage are not considered as a consequence of a change in copy number. The negative correlations are therefore of less interest in this setting.

In Figure 5.6 (b) the sorted gene expression values are plotted with the thresholds indicted as horizontal lines. The subsets of gene expressions correlated with copy number corresponds to those to the right of the colored lines. Genes selected by the lasso in the various analyses are also marked in colors. For the set of genes which are only restricted to be positively correlated with aCGH, two genes are selected. One of them is the gene (307660, chromosome 8) also selected when no preselection was used. This gene is marked black in Figure 5.6 (b). The other gene selected is located on chromosome 9 and is marked blue in the plot. Reducing the data further to only include those showing a higher correlation between expression and aCGH, that is $r > 0.2, 0.3$, involved removing the two selected genes from the first subset from the analysis. This results in that the lasso does not select any genes. Reducing the data even further, to only include genes where $r > 0.4$ gives a data set of 320 genes. The lasso is then selecting 5 genes and these are marked green in Figure 5.6(b). In this last analysis it seems that the reduction involves removing some noise and making the dimension of the problem smaller. This enables the lasso to select the 5 genes which are not found in the larger gene sets.

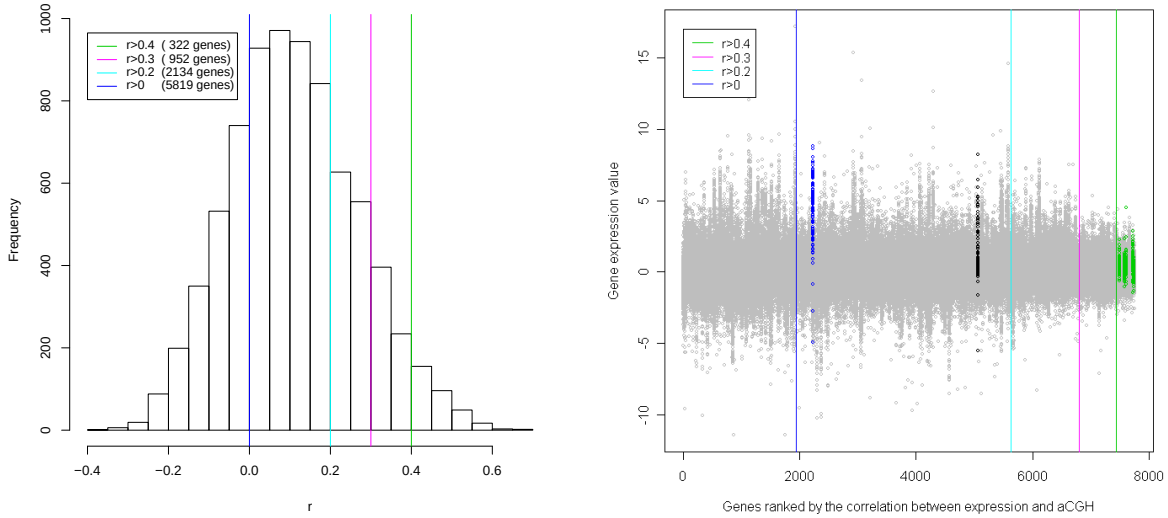


FIGURE 5.6. (a): Histogram of the spearman correlations for all of the genes, the thresholds defining the gene sets are marked as vertical lines.
(b): The gene expression data are plotted when the genes are sorted by the correlation with gene dosage. The genes to the right of the vertical lines, corresponding to the same thresholds as in (a), are those included in the gene sets.

These results confirm somehow that reducing the number of covariates may lead to selection of covariates which are not found when the number of covariates is large and where noise in the data may dominate the analysis. The thresholds here are arbitrary and one could imagine that other thresholds would yield a different number of selected genes. The fact that over 7000 covariates actually are completely removed in this last gene set should also be a concern. Having all of these data, it is preferable to include as much as possible in an analysis rather than deciding on arbitrary thresholds which either include the variables or completely remove them from the analysis.

5.3.2. aCGH Data. In the analyses of the aCGH data where correlation with gene expression was used as a criterion for reduction, the Spearman correlations was calculated and compared to find aCGH probes which were correlated with gene expression. The aCGH probes correspond to regions possibly involving more than one gene, and it is thus enough for a copy number probe to be highly correlated (exceeding a certain threshold) with one of the genes to be included in the smaller gene sets. For this approach, no probes were selected for any of the reduced data sets ($r \geq 0.4, 0.3, 0.2, 0$).

5.4. Discussion

Different strategies for reducing the data set prior to the lasso regression analysis were tried out in this chapter. We learn from some of these strategies that the resulting number of selected variables by the lasso strongly depends on the size of the data set. When selecting variables according to their univariate Cox p-value, more and more variables are found to be important by the lasso when the number of covariates decrease.

Reducing the number of covariates by only including genes with a high standard deviation did neither improve on the analysis on the complete data set. One reason for this may be that selecting genes with high variance also favors "noise-genes" to be included in the model. This does not make variance a good criterion for selection.

A good selection method should exclude most of the noise in the data. Selecting genes according to their ability of explaining survival one by one, is one approach where one may be be able to exclude most noise, because we know that the genes selected are related to survival one by one. It is, however, not a good approach to utilize the survival data in the selection, especially not when prediction is the aim.

Another drawback with most of the approaches discussed is that they utilize information on the genes one by one and not together. Since the Cox-lasso regression analysis is a joint analysis, taking into account how the gene expressions explain the survival together, selecting variables this way may not be optimal. Genes which may be able to explain the survival jointly can thus be completely excluded from the analysis.

When using reduction criteria as described in this chapter one has to determine strict thresholds for which the variables should be retained in the model or not. Making ad hoc choices for these thresholds as done here, may be too strict and does not include variables which should have been included or opposite, exclude variables which should have been included. This could be improved by evaluating different thresholds and choose the threshold/number of included variables by cross-validation.

For the reasons discussed above, to do selection prior to the regression analysis is not an optimal approach. Applying different methods for reducing the dimensions of the problem is, however, necessary in the context of microarrays. A reduction method should

- not include too much noise
- not utilize the survival data (the response)
- preferably take into account the gene's ability of explaining the response together with other genes

- not exclude too much of the data from the analysis

To fulfill all of these preferences is nevertheless difficult. Even if we want to reduce the dimension of the problem by selecting smaller subsets of genes, the last mark on the list is also important. The lasso selects only one probe to be significant in the analysis on the full dataset, from which one believes that reducing the data is necessary. This is also sensible considering the small sample size versus the large number of covariates. Preferably one would, however, rather use all of the data available in the analysis. Most of the methods in this chapter analyze the gene expression data and the aCGH data separately. Dividing the analysis in two parts also involves excluding a lot of information. That is, when the gene expressions are studied, we exclude the information available in the aCGH data.

The approach in Section 5.3 uses correlation between aCGH and gene expression, and is the only approach utilizing any information from both of the data sets and is hence a first step toward integration of the two data types.

CHAPTER 6

Data Integration by Genewise Lasso Penalization

Much attention has been directed to statistical regression analysis of gene expression data, for example through penalized regression. As described in Chapter 4, these methods may be used in regression settings when $p > n$. Even if these methods may reveal interesting results themselves, it is also of interest and importance to combine gene expression data with other types of data. By integrating other data types with gene expression data, one may be able to utilize more of the information available.

It is interesting to study the possibilities of combining information in different types of biological high dimensional data. By using data containing information of other biological features, one may be able to utilize underlying biological relationships in the analysis. One option may simply be to add the additional covariates to the model. This may nevertheless not be the optimal choice if two types of high dimensional data are to be analyzed, since the number of predictors in these data most often is very large. Adding even more covariates in a model may simply increase the difficulty of the problem.

Nygård et al. (2008) describe a method for inclusion of clinical variables in addition to gene expression data in a PLS regression model in the Cox regression setting. Since the number of clinical variables most often is small, they propose to include the variables in the model, but keeping them outside the PLS dimension reduction. In Ferkingstad et al. (2008) they present a method for multiple testing in an empirical Bayesian setting. The methodology allows for modulation of the posterior distribution of the null hypothesis based on external information. They manage to obtain a longer and differently ordered list of significant hypotheses by incorporating the external information on the covariates.

In this chapter a methodology for integration of gene expression data with other biological information is proposed. The aim in the integrated analyses is still to identify genes which are important for survival. Other types of biological data and the relationships between these and gene expression data are used in a suitable way to help identify differently expressed genes related to survival. The output will still be a list of genes which are assumed to influence the survival times, but the combined analysis may expose undiscovered genes, which may not be present when the gene expression data are studied alone.

The idea is to introduce genewise penalty terms λ_j instead of one common λ in the lasso analysis and to use other data sources, like for example aCGH data, to modify the penalty for each gene. The procedure involves two steps. The first step will be to define the genewise penalty modifications, and the second step will be to carry out the genewise penalized Cox regression analysis. The genewise penalties should reflect the importance of the present gene due to other biological knowledge than gene expression data and are supposed to make the penalty term in the regression smaller for genes which are believed to be of more importance due to external information. An individual penalty term will encourage the individual genes to be included in the model or not. Less penalty for a gene will indicate that the gene is more probable to have an influence on the survival times.

For *The Radium Hospital Cervix Cancer Cohort Data* different criteria may be considered when the genewise penalties are to be determined. Since both aCGH data and gene expression data are available it is convenient to use the aCGH data to define weights which should differentiate the amount of shrinkage for the covariates (genes). This is reasonable since we have the relation;

$$\begin{array}{c}\text{Genetic alteration}\\(\text{deletion/amplification})\end{array} \Rightarrow \begin{array}{c}\text{Differently}\\\text{expressed gene}\end{array} \Rightarrow \begin{array}{c}\text{Abnormal}\\\text{cell growth}\end{array} \Rightarrow \begin{array}{c}\text{Development of larger}\\\text{cell populations (tumor)}\end{array}$$

as described in Chapter 2.2. The relationship between aCGH data and gene expression data is that an increase/decrease in copy number might affect the gene to be differently expressed resulting in progression of tumors. This is what we want to utilize in the model. Exactly how this can be utilized in analyses of the Radiumhospitalet cervix cancer cohort data will be described in detail when the method is applied on the data in Chapter 7. The theory in this chapter will therefore not relate directly to the specific data at hand.

Next we will concentrate on the model in general and describe the effect of imposing individual penalty terms. Thereafter some asymptotic properties are considered, before a reparametrization to simplify the computations is described. Furthermore we discuss how to choose a new tuning parameter which will be a part of the model. To end this chapter a Bayesian perspective is described to illustrate the connections to a Bayesian setting.

6.1. Genewise Lasso Penalization

When the genewise lasso penalization is presented in this section, it is as for the general theory presented in Chapter 4, convenient to first describe the situation in a linear regression setting before presenting how the method may be applied in the Cox regression model, as will later be used in the analyses.

The lasso as it is described in Chapter 4 jointly shrinks the regression coefficients by adding a penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$ to the residual sum of squares to be minimized. By imposing a penalty term $\sum_{j=1}^{p} \lambda_j |\beta_j|$ where the amount of shrinkage λ_j is different for each gene, one may give some genes advantages such that they are more probable of being selected. For a large value of λ_j the regression coefficient for gene j is subject to a larger penalty and therefore less probable of being included in the model. A smaller value of λ_j will decrease the amount of shrinkage and will encourage the lasso to select gene j to be included in the model. This may be done by defining

$$\lambda_j = \lambda w_j.$$

Here w_j is a positive weight deciding the relative size of the individual penalty parameters λ_j, whereas λ is a global penalty common for all genes in the analysis.

In a linear regression setting a weighted lasso penalization analysis will correspond to minimizing the penalized residual sum of squares

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} w_j \left|\beta_j\right|, \tag{6.1}$$

where w_j is a predefined weight.

The lasso procedure in the Cox setting was described in Chapter 4.4 corresponds to maximizing a penalized version of the Cox log partial likelihood

$$l_{pl}(\boldsymbol{\beta}) = \sum_{k=1}^{N} \{\boldsymbol{\beta}^T \boldsymbol{x}_{(k)} - \log \sum_{l \in \mathcal{R}_k} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_l)\}. \tag{6.2}$$

In the situation where we want to do a weighted shrinkage, this corresponds to maximizing

$$l(\boldsymbol{\beta}) = l_{pl}(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} w_j |\beta_j|.$$

The estimated regression coefficients will indicate which genes have a copy number aberration and gene expression which together have an effect on survival. A summary of the general procedure for the weighted penalization is given in Table 6.1.

6.2. Asymptotic Properties

Fan & Li (2006) give a comprehensive overview of statistical challenges concerning high dimensionality and feature selection. The paper discuss how statistical challenges of this type arise in different disciplines as bioinformatics and health studies as well as financial econometrics. Further they define a unified approach to handle these challenges through penalized likelihood methods. Their notation and penalized likelihood method in the Cox regression setting was

Systematic Description of the Procedure in General

Assume the data matrix X is a $n \times p$ matrix consisting of p gene expression vectors for n patients.

(1) Determination of weights w_j.

(2) Fit a weighted lasso regression model with $X = (X_1, ..., X_p)$ as covariates. The penalty term will be weighted such that genes with a weight giving reason to believe the gene is related to survival through step 1, will have a larger probability of being selected by the lasso in Step 2.

The likelihood to be maximized will therefore be

$$l(\boldsymbol{\beta}) = l_{pl}(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} w_j |\beta_j|.$$

Here $l_{pl}(\beta)$ is the Cox log partial likelihood and λ is selected through K-fold cross-validation.

Output: p regression coefficients β_j indicating the influence gene j has on survival.

TABLE 6.1. The table describes the two steps in the procedure for a general weight.

introduced already in Fan & Li (2002) which consider variable selection for the Cox proportional hazards model and frailty model. Both Fan & Li (2002) and Fan & Li (2006) emphasize the possibility of letting the penalty function be defined individually for each variable, that is allowing for a variable specific λ_j and the option for incorporating prior information. To our knowledge this possibility has, however, not been pursued fully in applications to real data.

In the following mainly all theory, if not other is stated, is obtained from Fan & Li (2006) and Fan & Li (2002). For the rest of the section we assume that we are in a survival data setting as in Chapter 3, where we let T_i be the survival times, C_i the censoring times and $\boldsymbol{x}_i$ the covariates associated with the survival data for each patient i, where $i = 1, ..., n$. We define Z_i to be the observed survival time for patient i, that is $Z_i = \min(T_i, C_i)$ and δ_i the censoring indicator corresponding to $\delta_i = I(T_i \leq C_i)$. The ordered observed event times are given as $t_1^0, ..., t_k^0$ and (k) is the label for the event t_k^0 with corresponding covariate vector $x_{(k)}$.

In the following we let the partial log likelihood $l_{pl}(\beta)$ be as in (6.2) and assume that the regularity conditions A-D in Appendix B hold. Conditions A-D entail that the local asymptotic quadratic property for the partial likelihood is guaranteed. The local asymptotic quadratic property for the partial likelihood implies asymptotic normality of the maximum partial likelihood estimates. Defining the penalty function as $p_{\lambda_j}(\cdot)$ their penalized partial likelihood is given by

$$Q(\boldsymbol{\beta}) = \sum_{k=1}^{N} [\boldsymbol{x}_{(k)}^T \boldsymbol{\beta} - \log \{ \sum_{l \in R_l} exp(\boldsymbol{x}_l^T \boldsymbol{\beta}) \}] - n \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|).$$

By imposing some conditions on the penalty functions, the variable selection through penalized likelihood methods could be made more efficient. For instance it can be shown that under some conditions, the penalized partial likelihood estimator that converges at rate $O_p(n^{-1/2}+a_n)$ where a_n is defined below. Some results are stated in Fan & Li (2002) concerning the convergence rate of the penalized partial likelihood estimator.

We consider an asymptotic set up with the penalty parameter $\lambda_{n,j}$ depending on the sample size n and let $\boldsymbol{\beta}_0 = (\beta_{10}, ..., \beta_{d0})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)$ be the true regression coefficients and where $\boldsymbol{\beta}_{10}$ are the s nonzero regression coefficients and $\boldsymbol{\beta}_{20}$ are the remaining regression coefficients which all are zero. Further we define the two sequences

$$\begin{aligned} a_n &= \max\{p'_{\lambda_{n,j}}(|\beta_{nj0}|) : \beta_{nj0} \neq 0\} \\ b_n &= \max\{|p''_{\lambda_{n,j}}(|\beta_{nj0}|)| : \beta_{nj0} \neq 0\}, \end{aligned} \tag{6.3}$$

which are needed in order to prove that there exists a local maximizer $\hat{\boldsymbol{\beta}}$ for the penalized partial likelihood function that converges at rate $O_p(n^{-1/2} + a_n)$. The following result can be found in Fan & Li (2002):

THEOREM 1. *Assume that* $(\boldsymbol{x}_1, T_1, C_1), \ldots, (\boldsymbol{x}_n, T_n, C_n)$ *are independent and identically distributed according to the population* $(\boldsymbol{x}, T, C)$*,* T *and* C *are conditionally independent given* $\boldsymbol{x}$*, and Conditions (A)-(D) hold. If* $b_n \to 0$*, then there exists a local maximizer* $\hat{\boldsymbol{\beta}}$ *of* $Q(\boldsymbol{\beta})$ *such that* $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_p(n^{-1/2} + a_n)$*, where* a_n *is given by (6.3).*

The proof of Theorem 1 can be found in Fan & Li (2002). In order to show that Theorem 1 hold for the method proposed for data integration in this chapter we define the two sequences a_n and b_n in (6.3) for our specific case. Note that w_j is decided externally prior to the analysis and is not necessarily dependent on the sample size n. The penalty function in our situation thus corresponds to

$$p_{\lambda_{n,j}}(|\beta_{j0}|) = \lambda_{n,j} \sum_{j=1}^{d} |\beta_{j0}|$$

and the first and second derivatives are given by

$$\begin{aligned} p'_{\lambda_n}(|\beta_{j0}|) &= \lambda_{n,j} \\ p''_{\lambda_n}(|\beta_{j0}|) &= 0. \end{aligned} \tag{6.4}$$

We then have

$$\begin{aligned} a_n &= \lambda_{n,j} \\ b_n &= 0. \end{aligned} \tag{6.5}$$

For the result of Theorem 1 to hold, we must have $b_n \to 0$ when $n \to \infty$. This is obviously satisfied. It also follows from Theorem 1 that if $\lambda_{n,j}$ is chosen properly and if $a_n = O(n^{-1/2})$, there

exists a $\sqrt{n}$-consistent penalized partial likelihood estimator. Since $a_n = \lambda_{n,j}$, $\sqrt{n}$-consistency requires that $\lambda_{n,j} = O_p(n^{-1/2})$.

Fan & Li (2002) also demonstrate the oracle property. For a selection method to enjoy the oracle properties it should show

- Consistency in variable selection (it identifies the right subset model).
- Asymptotic normality of the estimated regression coefficients.

The following theorem from Fan & Li (2002) is necessary in order to show that a method possess the oracle property. First we define

$$\boldsymbol{\Sigma} = diag\{p''_{\lambda_{n,j}}(|\beta_{10}|), ..., p''_{\lambda_{n,j}}(|\beta_{s0}|)\}$$

and

$$\boldsymbol{b} = \left(p'_{\lambda_{n,j}}(|\beta_{10}|)sgn(\beta_{10}), ..., p'_{\lambda_{n,j}}(|\beta_{s0}|)sgn(\beta_{s0})\right)^T,$$

here s is the number of components of β_{10}, that is the number of true regression coefficients that are different from zero.

THEOREM 2 (Oracle property). *Assume that the penalty function $p_{\lambda_{n,j}}(|\theta|)$ satisfies the condition*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_{n,j}}(\theta)/\lambda_{n,j} > 0.$$

If $\lambda_{n,j} \to 0, \sqrt{n}\lambda_{n,j} \to \infty$ and $a_n = O(n^{-1/2})$, then under the conditions of Theorem 1, with probability tending to 1, the $\sqrt{n}$ consistent local maximizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ in Theorem 1 must satisfy:

(i) *(Sparsity) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;*

(ii) *(Asymptotic normality)*

$$\sqrt{n}(\boldsymbol{I}_1(\boldsymbol{\beta}_{10}) + \boldsymbol{\Sigma}) \left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\boldsymbol{I}_1(\boldsymbol{\beta}_{10}) + \boldsymbol{\Sigma})^{-1}\boldsymbol{b}\right\} \to N\left\{\mathbf{0}, \boldsymbol{I}_1(\boldsymbol{\beta}_{10})\right\},$$

where $\boldsymbol{I}_1(\boldsymbol{\beta}_{10})$ is the first $s \times s$ submatrix of the Fisher information matrix $\boldsymbol{I}(\boldsymbol{\beta}_0)$ of the partial likelihood.

From Theorem 1 we found that for a $\sqrt{n}$-consistent penalized partial likelihood estimator to exist, $\lambda_{n,j} = O_p(n^{-1/2})$ must hold. Since both $\lambda_{n,j} = O_p(n^{-1/2})$ and $\sqrt{n}\lambda_{n,j} \to \infty$ can not be satisfied simultaneously, the oracle property does not hold for the genewise lasso penalization procedure. This is however the same as for the standard lasso, which neither enjoys the oracle property of the same reason that the two conditions cannot be satisfied at the same time. By these results, the genewise lasso penalization procedure has proven to hold the same asymptotic properties as

the lasso and that by introducing the individual penalty parameters λ_j the asymptotic behavior of the lasso is not changed.

The idea of imposing individual weights to the penalization in a lasso regression model has also been discussed in Zou (2006) where they define the adaptive lasso. The adaptive lasso is later introduced in relation with the Cox proportional hazards model by Zhang & Lu (2007) and studied in high dimensional settings in Huang et al. (2006). The computational scheme of the adaptive lasso is much similar to the one used in the integrated analyses done here. The difference lies in the purpose of the weights and how they are defined. The motivation for the adaptive lasso is that there exists some scenarios where the lasso is inconsistent for variable selection (Zou, 2006). The adaptive lasso is therefore suggested as a method enjoying the oracle property in Theorem 2, and is shown in Zou (2006).

The adaptive lasso defines the individual penalty parameters as $\lambda_j = \lambda w_j$ and the weights w_j as being the reciprocal of a consistent estimator for the regression coefficients β; which are to be estimated in the model. Any consistent estimator can be used and will make the adaptive lasso, a method having the oracle properties. One option is to define the vector of weights as

$$\mathbf{w} = \frac{1}{|\hat{\beta}_{OLS}|^q}, \tag{6.6}$$

where $q > 0$, and $\hat{\beta}_{OLS}$ is the vector of ordinary least squares estimates, if these exist, estimated from the same single dataset. The paper also shows that the nonnegative garotte, which is another popular variable selection method, may be considered as a special case of the adaptive lasso setting $q = 1$.

The vector of weights in (6.6) is suggested for the adaptive lasso when collinearity is not a concern. When the covariate vectors are very correlated, for instance if the number of explanatory variables is larger than the number of samples ($p > n$), the ordinary least squares estimates can not be computed and the weights have to be chosen in a different way. A practical solution would be to use the estimated regression coefficients from a ridge regression (Zou, 2006). Another option suggested by Huang et al. (2006) is to use the marginal regression estimators to obtain the initial estimators and weights. Under a partial orthogonality condition Huang et al. (2006) show that the adaptive lasso with these initial weights indeed has the oracle properties as well.

The main difference from the adaptive lasso and the integrating lasso procedure considered here, is thus that the weights in the integration procedure will be based on external data in addition to the gene expression data, hence representing real prior information. The determination of w_j should be carefully evaluated such that the individual penalty terms are sensible. If a gene's influence on survival is reflected through λ_j being small, the gene is believed to have an influence

on survival and vice versa. Apriori unimportant variables receive larger penalties if λ_j is large and are more likely to be discarded in a selection process. Some comments on how to determine the genewise weights w_j are considered separately in Section 6.4 for the integration procedure.

6.3. Reparametrisation and Computational Aspects

The separate analyses using ridge and lasso in a Cox regression may be done straightforwardly using the program code in Bøvelstad et al. (2007) which is described in Section 4.6. The genewise penalty terms, however, have to be handled with special care. To perform the weighted analysis we have introduced the individual penalty parameters

$$\lambda_j = \lambda w_j,$$

where w_j is a positive weight which is small if gene j is believed to influence the response.

The model we want to fit corresponds in a linear regression setting to minimization of

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |w_j \beta_j|. \tag{6.7}$$

To be able to use the standard program to fit the model with individual weights, a manipulation of this criterion is convenient. We may rewrite (6.7) as

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \frac{x_{ij}}{w_j} w_j\beta_j)^2 + \lambda \sum_{j=1}^{p} |w_j \beta_j|.$$

Defining $\alpha_j = w_j\beta_j$, the standard lasso procedure may be used to minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \frac{x_{ij}}{w_j} \alpha_j)^2 + \lambda \sum_{j=1}^{p} |\alpha_j|,$$

with respect to α_j, which corresponds to a standard lasso minimization criterion with each entry in the data matrix given as $\frac{x_{ij}}{w_j}$. Note that this transformation does not change the collinearity among the variables and will neither make the estimation more complicated nor easier.

A similar approach may be used in the Cox regression setting, by defining the data matrix to have entries $\frac{x_{ij}}{w_j}$, and maximizing

$$l(\boldsymbol{\alpha}) = l_{pl}(\boldsymbol{\alpha}) - \lambda \sum_{j=1}^{p} |\alpha_j|,$$

with respect to $\boldsymbol{\alpha}$. The estimated parameters $\hat{\boldsymbol{\alpha}}$ can in this way be obtained from the program code in Bøvelstad et al. (2007). When $\hat{\alpha}_j$ is found by using the algorithm, one may find the actual regression coefficients $\hat{\beta}_j$ from the weighted analysis as

$$\hat{\beta}_j = \frac{\hat{\alpha}}{w_j}.$$

Systematic Description of the Two-Dimensional Cross-Validation Procedure

(1) Define η_j for gene j.

(2) Set a grid of q-values.

(3) For a fixed q ;

- **a** Define weights $w_j = \frac{1}{|\eta_j|^q}$.
- **b** Define the input data $\frac{x_{ij}}{w_j}$ for the lasso procedure as described in Section 6.3.
- **c** Compute $CV(q, \lambda)$.

(4) Repeat all parts of step 3 by iterating through all q-values in the grid.

(5) Find the pair of (q, λ) which maximizes $CV(q, \lambda)$.

(6) Fit the final model with the cross-validated parameters $(\hat{q}, \hat{\lambda})$.

TABLE 6.2. Overview of the two-dimensional cross-validation procedure.

The estimated regression coefficients $\hat{\beta}_j$ then reflect the importance of covariate j in the model. The variables that are selected, that is, those with a corresponding regression coefficient $\hat{\beta}_j \neq 0$, represent the variables which are most important for the response and which are found to be important when the information from the two data sources is combined.

6.4. Two-Dimensional Cross-Validation

The individual amount of shrinkage depends on individual weights determined through prior knowledge connected to the covariates. To weight the penalty term in the regression analysis, it is important to carefully evaluate the possible alternatives to use as weights. The weights should be positive and reflect the covariates' importance for the response due to external information.

In the following we will define η_j to represent any quantity that will be used to weight the penalty term. The values of η_j should increase with the prior believed importance of the covariate. The weights may then have the form

$$w_j = \frac{1}{|\eta_j|^q}, \tag{6.8}$$

such that a small value of η_j, indicating a less importance, produces a large weight w_j which will cause a large penalty for covariate j. To have one example of such a η_j in mind, one could think to η_j as the rank correlation between copy number and gene expression for gene j. Various possibilities for η_j will be discussed in connection to the data at hand in Chapter 7.

By introducing the weights as in expression (6.8) we also introduce a second tuning parameter q in the model. While the tuning parameter λ controls the amount of shrinkage imposed on the

coefficients simultaneously, the tuning parameter q controls the form on the weight function and its ability to distinguish between variables which will be given a relatively high or low weight. Depending on the size of $|\eta_j|$, different values of q will be necessary to form reasonable and effective weights. The value of q could be any positive value and is not restricted to integers. By setting the parameter q manually without any assessment, one risks to overfit the model. It will therefore be sensible to cross-validate q as well as λ. The genewise lasso penalization thus involves cross-validation of two parameters q and λ. The values for q and λ chosen through cross-validation should be the pair of q and λ that optimize the prediction capability. This involves a two-dimensional cross-validation procedure with maximization on a two-dimensional grid.

To implement such a cross-validation procedure, one should first define a cross-validation criterion. We want to choose parameter values for λ and q that maximize the prediction capability. We can then use a criterion on the same form as discussed for the one-dimensional cross-validation of λ in Chapter 4.5, but where the criterion varies both with q and λ;

$$CV(q,\lambda) = \sum_{k=1}^{K} \{l(\hat{\boldsymbol{\beta}}_{(-k)}(q,\lambda)) - l_{(-k)}(\hat{\boldsymbol{\beta}}_{(-k)}(q,\lambda))\}.$$

A grid of q-values is decided, and for a fixed q, we may search for the optimal λ value through K-fold cross-validation in the same way as for the ordinary lasso model. By iterating through all values of q, defining the cross-validated log-likelihood for each grid value, we may find the best pair $\hat{q}$ and $\hat{\lambda}$ as the pair maximizing $CV(q,\lambda)$. The estimates $\hat{q}$ and $\hat{\lambda}$ are the two parameters optimizing the prediction capability. A systematic description of the two-dimensional procedure is given in Table 6.2 and the R-script performing two-dimensional cross-validation for the genewise lasso procedure is given in Appendix C.

6.5. Bayesian Interpretation

The method introduced for data integration in this chapter is presented in a classical Cox regression setting. Individual penalty terms are added to the partial log-likelihood. The method may just as well be considered in a Bayesian framework which may illustrate and reveal different aspects of the method. Bayesian regression methodology is concerned with utilizing prior information, that is, external information and/or expert knowledge on the covariates in the analysis. This makes it natural to give a Bayesian interpretation of the introduced method which is presented especially with the aim to combine the gene expression variables with the external information in the aCGH data.

6.5.1. Bayesian Regression. To set up a Bayesian regression model one has to specify a conditional distribution for the response data $p(\boldsymbol{y}|\boldsymbol{\theta})$ i.e. the likelihood, and a prior distribution

$p(\boldsymbol{\theta})$ for the parameters $\boldsymbol{\theta}$. By combining these, one may compute the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ which represents our updated knowledge about the parameters when the data are introduced (Hastie et al., 2001, chap. 8).

In the classical regression approach the only information that is used explicitly is the data. The Bayesian approach for regression differs from the classical approach by introducing a prior distribution for the parameters $\boldsymbol{\theta}$, which reflects beliefs one has about the parameters prior to the analysis. To specify this distribution one should utilize what one knows or believes about the parameters and which values are likely for them, and translate this prior knowledge into the form of a probability distribution $p(\boldsymbol{\theta})$. When the prior distribution is specified, it should be combined with the conditional distribution of $\boldsymbol{y}|\boldsymbol{\theta}$ to obtain the posterior distribution of $\boldsymbol{\theta}|\boldsymbol{y}$ through Bayes's formula

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta}). \tag{6.9}$$

From the posterior distribution one is able to make inference about the parameters. The parameters may be estimated by the posterior mode $\hat{\boldsymbol{\theta}} = \arg\max_\theta p(\boldsymbol{\theta}|\boldsymbol{y})$ or the posterior mean $\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\boldsymbol{y})$. When the posterior mode is used as Bayes estimate, there is only the prior distribution making the Bayesian approach different from maximum likelihood estimation since the posterior inference will be based on maximizing *prior* $\times$ *Likelihood*

The focus in this section will be on the standard linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where we assume that $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\sigma}^2)$ and that $\boldsymbol{\sigma}^2$ is known. For fixed values of $\boldsymbol{\beta}$, $\boldsymbol{y}$ thus follows a multivariate normal distribution;

$$p(\boldsymbol{y}|\boldsymbol{\beta}) = (2\pi)^{-n/2}|\sigma^2\boldsymbol{I}|^{-1/2}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})\right), \tag{6.10}$$

that is $\boldsymbol{y}|\boldsymbol{\beta} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$.

Prior Distributions. Different prior distributions may be used depending on what knowledge one has on the behavior of the parameters. Ordinary linear regression may be interpreted in a Bayesian perspective with a non-informative prior for the parameters. This corresponds to having no prior belief on what values of the coefficients $\boldsymbol{\beta}$ are sensible before one looks at the data.

In situations where there are many measured responses and only a few parameters to be estimated, the non-informative prior gives acceptable results, but the choice of prior distributions is however more important when the sample size is small and there is a large number of parameters to be estimated. This is due to the less sharply peaked likelihood in these situations (Gelman et al., 2004, chap. 14).

Genewise Lasso Penalization. To make a Bayesian interpretation of the procedure imposing an individual penalty for each variable, one should assume a prior distribution with the desirable properties. The prior distribution should concentrate more mass in exactly zero for the unimportant variables, while the important variables should have a prior with larger variance to allow for the estimated regression coefficients to take on other values.

The Laplace distribution is appropriate for our purpose, having a peak in zero and fat tails. The variance in the distributions should be different for each variable and a small variance will favor zero as the value for the regression coefficients, while larger variance will not give reason for setting the regression coefficient to zero. The individual penalty parameters should therefore be incorporated in the variance. A large penalty parameter λ_j indicates a small variance which should set the regression coefficients to zero. We assume the variance $2\tau_j^2$ for each regression coefficient and the prior distribution of β_j;

$$p(\beta_j) = \frac{1}{2\tau_j} \exp\left\{-\frac{|\beta_j|}{\tau_j}\right\}.$$

The joint distribution of the β_js may be found to be

$$p(\boldsymbol{\beta}) \propto \exp\left\{-\sum_{j=1}^{p} \frac{|\beta_j|}{\tau_j}\right\}.$$

By assuming a linear model and the distribution given in 6.10, we have by applying Bayes formula in (6.9) that the posterior distribution is found to be

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{y}) &\propto \exp\left\{-\sum_{j=1}^{p} \frac{|\beta_j|}{\tau_j}\right\} \times \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) - \sum_{j=1}^{p} \frac{|\beta_j|}{\tau_j}\right\}, \end{aligned}$$

The regression parameters may be estimated by the posterior mode. Maximizing $p(\boldsymbol{\beta}|\boldsymbol{\theta})$ to find the posterior mode corresponds to maximizing the log-posterior distribution

$$\log p(\boldsymbol{\beta}|\boldsymbol{y}) = -\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) - \sum_{j=1}^{p} \frac{|\beta_j|}{\tau_j}. \tag{6.11}$$

This may be recognized as the part of the log-likelihood relevant for maximization minus a term $\sum_{j=1}^{p} \frac{|\beta_j|}{\tau_j}$ which corresponds to an individual penalty term $\sum_{j=1}^{p} \lambda_j|\beta_j|$ when $\lambda_j = \frac{1}{\tau_j}$. The special case when all variances are equal, that is $\tau_j = \tau \ \forall j$, corresponds to the lasso. Thus both the genewise lasso and the lasso can be interpreted in a Bayesian way by assuming a Laplace prior on the regression coefficients.

Method	Prior distribution
Genewise Lasso Penalization	$p(\beta_j) = \frac{1}{2\tau_j} \exp\left\{-\frac{\lvert\beta_j\rvert}{\tau_j}\right\}$
The Lasso	$p(\beta_j) = \frac{1}{2\tau} \exp\left\{-\frac{\lvert\beta_j\rvert}{\tau}\right\}$
Ridge Regression	$p(\beta_j) = \frac{1}{\sqrt{2\pi\tau}} \exp\left\{-\frac{1}{2\tau^2}\beta^2\right\}$

Table 6.3. The table shows the three penalization methods and the corresponding prior distributions.

Ridge penalization. Ridge regression may also be derived from a Bayesian point of view by adopting a Gaussian prior

$$p(\boldsymbol{\beta}) = (2\pi)^{-p/2}|\tau^2\boldsymbol{I}|^{-1/2} \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right\},$$

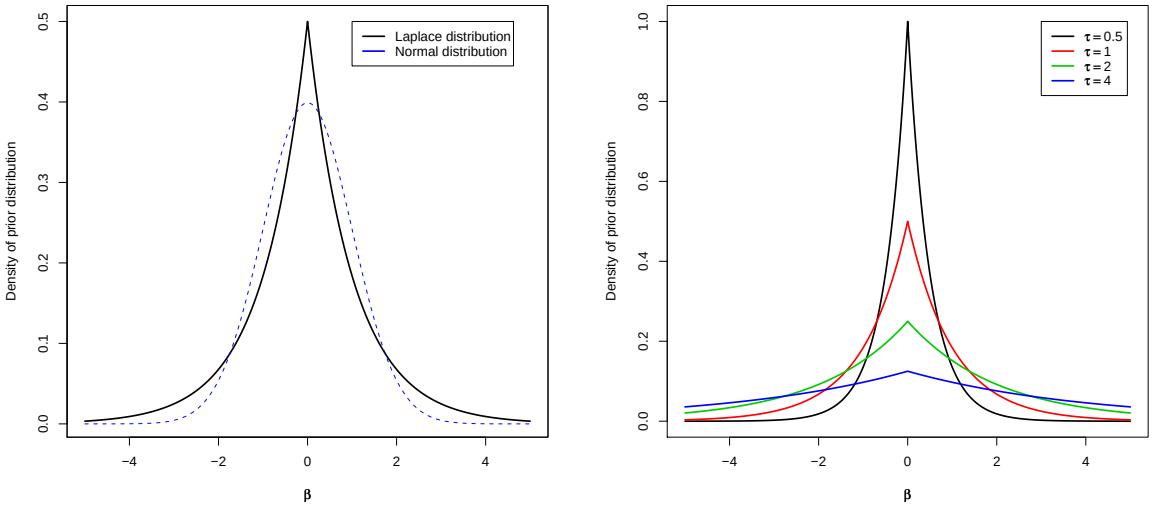
that is $\boldsymbol{\beta} \sim N(\mathbf{0}, \tau^2\boldsymbol{I})$ where $\boldsymbol{I}$ is the $p \times p$ identity matrix. From this it is straightforward to find the posterior distribution by using Bayes's formula:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \quad \propto \quad \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2\tau^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\}. \tag{6.12}$$

The regression parameters may be estimated by the posterior mean or mode. We are only concerned with proportionality when the posterior mode is to be found and it is therefore enough to maximize the expression in (6.12). This is equivalent with maximizing the log-likelihood minus a ridge penalty term with $\lambda = \frac{1}{2\tau^2}$. One may show, by rewriting expression (6.12), that the posterior distribution will be a normal distribution. Because mean and mode are identical in Gaussian distributions, the ridge estimates may be derived both as the mean and the mode of the posterior distribution.

6.5.2. Comments. It is natural to end this section with some comments on the Bayesian interpretation of the penalization methods and how the three different priors in Table 6.3 correspond to the three different penalization methods; the covariatewise lasso, the lasso and ridge regression.

In Figure 6.1(a) a normal and a Laplace distribution are plotted. The normal distribution corresponds to the ridge prior and we see that it differs from the lasso prior, which is the Laplace, by having more mass concentrated around zero and less for higher values. The Laplace distribution seems however to favor zero as a value for the regression coefficients by having a well defined peak at zero. This is in coincidence with what we know about the lasso which is likely to produce regression coefficients exactly equal to zero. We also know that ridge regression



(a) (b)

FIGURE 6.1. Plot of prior distributions for the different regression model.
(a): Illustration of the difference between the prior distribution for the regression coefficients in a ridge regression and in the lasso. The normal distribution corresponds to the prior for ridge regression and the Laplace distribution corresponds to the prior for regression coefficients in the lasso.
(b): Illustration of the difference between prior distributions in the method performing individual/-genewise penalization, the priors on the regression parameters have different variance.

shrinks the regression parameters, but seldom sets the regression coefficients to be exactly zero, and the normal prior therefore has more mass in small values surrounding zero.

In Figure 6.1(b), four illustrations of Laplace distributions are given. This illustrates the different priors for the regression coefficients when individual penalties are used. The regression coefficients for a covariate which is believed to have an influence on the response, will have a less sharply peaked prior than the regression coefficient for a covariate which is not believed to have an influence on the response a priori. We have also seen that there is a connection between the variance of the prior distribution and the amount of shrinkage λ. A small variance will produce many regression coefficients equal to zero corresponding to a large amount of shrinkage. By letting the variance in the Laplace prior distributions vary depending on the external information, some covariates will be more probable of being selected while others are less probable.

It is important to note that this Bayesian interpretation is considered in a linear regression setting and not for survival data and the Cox model. It does, however, just as well illustrate the Bayesian way of thinking where inclusion of prior information is crucial.

CHAPTER 7

Genewise Lasso Penalization Analysis and Results

In this chapter the genewise lasso penalization method is applied to the cervix cancer data. The gene expression data are used as covariates in the Cox-lasso regression model, whereas the aCGH data are used through three different penalization schemes to weight the lasso penalty terms. Each of the three penalization schemes involves estimation of the quantity $\hat{\eta}_j$ and the values of q and λ are decided through two-dimensional cross-validation as described in the previous chapter. The results for the three approaches will be reported and some discussions of the results will be given. The three penalization schemes are based on the following:

PS1 The gene copy number's correlation with gene expression (Spearman correlation coefficients)

PS2 The gene copy number's effect on survival (estimated Ridge regression coefficients)

PS3 The gene copy number's variability (empirical standard deviation)

The three different approaches are considered separately and a motivation for the different schemes will be given before applying them to the data.

7.1. Penalization Scheme 1; Spearman Correlation

In the end of Chapter 5, we considered a method for reducing the data set which utilized the correlation between gene expression and copy number. The method was applied both in the analysis of gene expression and aCGH data. As opposed to the other methods used in Chapter 5, this method utilized a relationship between gene expression and gene copy numbers, and combined the two data sources in the analysis. The motivation was that genes showing high correlation between gene expression and copy number, are considered as possible driving forces for cancer development and progression (Lando et al., 2009). This chapter will concentrate on analyzing the gene expression data, but a similar motivation can be given when applying the genewise lasso penalization model. The correlation between gene expression and copy number $\hat{\rho}_j$ can be used to determine weights on the penalty term for gene j, as explained in the following.

Genes corresponding to high correlations should be more probable of being selected. Negative correlations between gene expression and aCGH will be less interesting in this setting, since gene expression alterations which are negatively correlated with gene copy number are not

considered as consequences of copy number alterations. The genes which show high correlation should therefore be given less penalty than genes which are negatively or not correlated. For the penalties to achieve this effect we consider a weight of the form

$$w_j = \frac{1}{|\eta_j|^q},$$

as described in the previous chapter. We then have to define η_j. As discussed in Chapter 5 the Spearman rank correlation is convenient to use as a measure of correlation, since it measures any monotonic relationship between gene expression and copy number. The Spearman rank correlation is sensible to use in this setting as well, and we may define $\eta_j = \hat{\rho}_j$ for all $\hat{\rho}_j > 0$. For negative or zero correlation, η_j should be some small positive value which will give gene j a large penalty. Genes showing zero correlation should also be adjusted in order to avoid division with zero. One option is to use $\eta_j = \min\{\hat{\boldsymbol{\rho}} : \hat{\boldsymbol{\rho}} > 0\}$, for all genes with $\hat{\rho}_j \leq 0$. Where $\hat{\boldsymbol{\rho}}$ is the vector of estimated Spearman coefficients. That is, all genes which show negative or zero correlation will be given the same weight as the smallest positively correlated gene. The quantity η_j in PS1 can thus be expressed as

$$\eta_j = \begin{cases} \hat{\rho}_j & \text{if } \hat{\rho}_j > 0 \\ \{\min \hat{\boldsymbol{\rho}} : \hat{\boldsymbol{\rho}} > 0\} & \text{if } \hat{\rho}_j \leq 0. \end{cases} \tag{7.1}$$

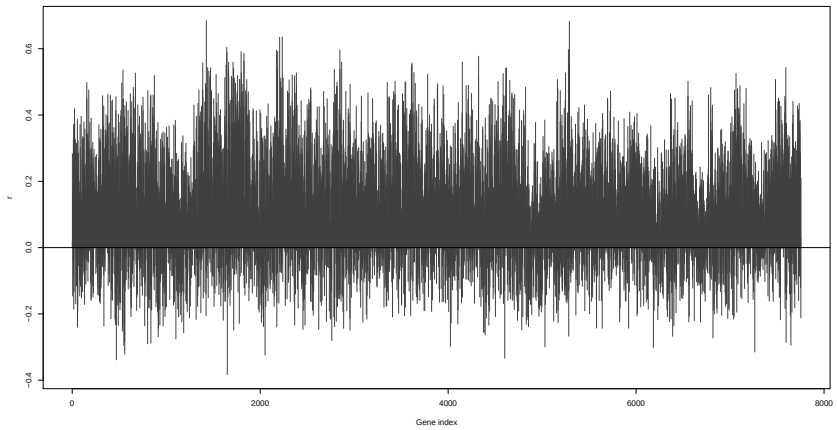
To carry out the analysis of the data, the gene expression vectors were matched together with their corresponding copy number vector and the Spearman correlation coefficients $\hat{\rho}_j$ were calculated for each gene j. The Spearman correlations are plotted in Figure 7.1(a). A plot of η_j for all genes after adjusted for negative and zero correlations is also given in Figure 7.1(b).

To fit the model, the two-dimensional cross-validation procedure was applied to find the best pair of parameters (q, λ). The grid of q-values was set to range from 0 to 5, increasing with 0.25 for each step. For each given value of q, we find the value of λ which maximizes $CV(q, \lambda)$, λ_q^*. More specifically for a given value of q:

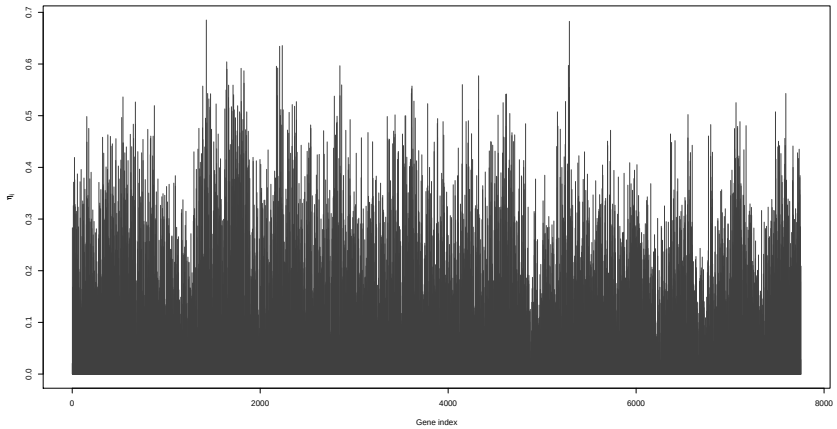
$$\lambda_q^* = \arg\max_{\lambda} CV(q, \lambda).$$

This is done using exactly the same procedure as described in Chapter 4 about K-fold cross-validation. In Figure 7.2(a) λ_q^* is plotted versus q. Defining the cross-validation curve CV as a function of q for given values λ_q^*, will give us the function $CV(q, \lambda_q^*)$ as plotted in Figure 7.2(b). Maximizing this with respect to q will give us the preferable value $\hat{q}$ according to cross-validation and the pair $(\hat{q}, \hat{\lambda})$, where $\hat{\lambda} = \lambda_{\hat{q}}^*$, will be the pair used to fit the final model. The value of q maximizing the cross-validation curve $CV(q, \lambda)$ is marked with a star. In the case of Penalization Scheme 1, $\hat{q} = 1.5$ and global penalty parameter $\hat{\lambda} = 1.6254$.

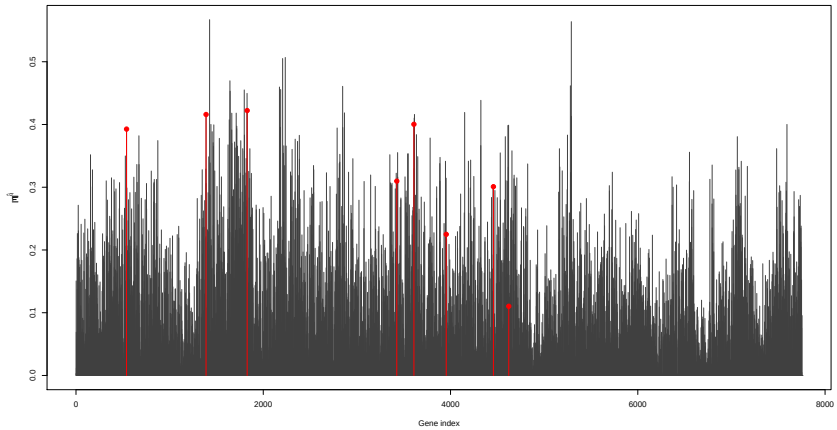
From a statistical point of view, it is interesting to examine the results of the cross-validation. By having a closer look on Figure 7.2(a)-(c) we may get a better understanding of how the



(a)



(b)



(c)

FIGURE 7.1. (a): Plot of the calculated Spearman correlation coefficients $\hat{\rho}_j$ for each gene.
(b): Plot of η_j, that is, when adjusted for non-positive correlations.
(c): Plot of $|\eta_j|^{\hat{q}}$ for the final model. The selected genes are marked.
The genes are ordered by their position on the genome in both panels.

Systematic Description of the Procedure: Penalization Scheme 1

Assume data matrices X and Z, where X is a $n \times p$ matrix consisting of p gene expression vectors for n patients. Z is a $n \times r$ matrix which consists of the r copy number probes for the same n patients. Assume also survival data for the n patients.

(1) Determination of η_j:

a Calculate the Spearman correlation $\hat{\rho}$ between each gene expression vector and corresponding copy number vector.

b Adjust for negative and zero correlations such that

$$\eta_j = \begin{cases} \hat{\rho}_j & \text{if } \hat{\rho}_j > 0 \\ \{\min \hat{\rho} : \hat{\rho} > 0\} & \text{if } \hat{\rho}_j \leq 0. \end{cases}$$

(2) Determination of λ_j:

a Define $\lambda_j = \lambda w_j$, where $w_j = \frac{1}{|\eta_j|^q}$.

b Estimate λ and q through two-dimensional cross-validation.

(3) Fit the model with $X = (X_1, ..., X_p)$ as covariates and the survival data as response. That is, maximize

$$l(\beta) = l_{pl}(\beta) - \lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

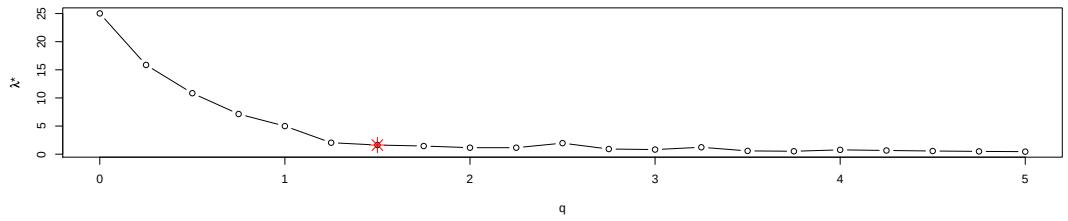
where $l_{pl}(\beta)$ is the Cox log partial likelihood and λ and w_j are estimated as described in Step 2.

Output: p regression coefficients β_j indicating the influence gene j has on survival.

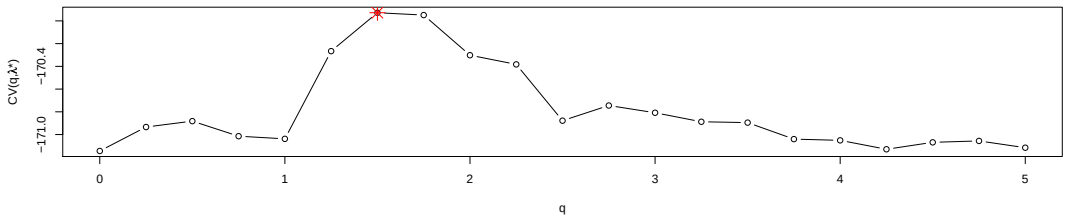
TABLE 7.1. Summary of the procedure using Penalization Scheme 1. The penalty term will be weighted such that the regression coefficients for genes corresponding to a high positive correlation in Step 1, will have a larger probability of being estimated to be nonzero in Step 3.

method actually behaves and performs when applied to data for different combinations of q and λ. There are several interesting aspects regarding the cross-validation results, which can be seen from Figure 7.2.

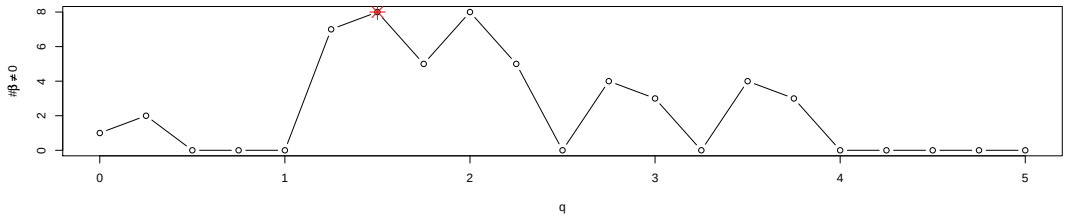
First, by looking at Figure 7.2(a) we may see that by increasing q, the cross-validated value of λ, λ_q^* decreases. When $0 < \eta_j < 1$ (which is the case for all three penalization schemes studied in this thesis), all weights will be larger when q increases. This indicates that in addition to produce weights which reasonably differentiate between the genes based on whether they are believed to be important or not, the penalization schemes will produce overall larger penalties when q increases. Some of the simultaneous shrinking is thus left to the weights.



(a)



(b)



(c)

FIGURE 7.2. Plots for cross-validation of q when using Penalization Scheme1:Spearman correlation.

Figure 7.2(b) shows the cross-validation curve $CV(q, \lambda_q^*)$, that is, where λ_q^* is inserted for λ. We see that the curve has its maximum at $q = 1.5$. The area where $1 < q < 3$ is where the values of $CV(q, \lambda_q^*)$ are largest. Comparing with Figure 7.2(c) we see that the two curves show some similar tendencies, that is when the number of selected genes changes with q, the performance due to the cross-validation changes as well.

Two peaks can be seen in the curve in Figure 7.2(c), for which eight genes are selected in both cases. A considerable drop is seen for the cross-validation curve in Figure 7.2(b) for these q-values. It is interesting to see whether the eight genes are the same in both situations. Investigating the eight genes in both cases, two of the genes are eliminated and replaced by two others for the second peak. So even if the same number of genes are selected by using different q-values, it is not necessarily the case that it is the same eight genes.

For the q-values for which two or less genes are selected, the values for $CV(q, \lambda_q^*)$ are on the same low level and there is a markable difference when a few more genes are included in the model. This indicates that due to cross-validation for PS1, the models corresponding to q-values which select more genes, seem to perform better than for q-values where no genes are selected.

Note also that it is not strange that there are some variations in the values of $CV(q, \lambda_q^*)$, even if all regression coefficients in the final model are estimated to be zero. The estimated regression coefficients when the kth fold is left out, might not be zero and may therefore give different contributions to the cross-validation criterion for different values of q and λ.

The final model was fitted with the estimated parameters, resulting in eight selected genes. In Figure 7.1(c), $|\eta_j|^{\hat{q}}$ is plotted. The genes selected by the lasso are marked red in the plot. Note that the genes corresponding to large values of $|\eta_j|^{\hat{q}}$ in the plot are promoted in the analysis. Genes could still be selected in the analysis if the gene expression shows a strong effect. Large values of $\eta_j^{\hat{q}}$ indicate that gene j is believed to influence survival because of a high positive correlation with gene dosage. From the plot we see that the selected genes correspond to relatively large values of $\hat{\eta}_j$ which indicate that the correlation between gene expression and gene dosage is relatively large for these genes. Note, however, that these are not necessarily the eight genes corresponding to the eight largest values of $|\eta_j|^{\hat{q}}$. MMP10 corresponds to the rightmost marked gene in Figure 7.1(c). Gene MMP10 is chosen although the value of η_j is not as large as for the other chosen variables, indicating that the gene expression for MMP10 is not as correlated with gene dosage as the other chosen variables. The gene probably has a relatively strong influence on survival through its expression alone and is not helped as much by the weighting of the penalty. All selected genes (including MMP10) are, however, subject to much less penalty than the genes with very low correlation. All selected genes are listed in Table 7.2.

Results; Penalization Scheme 1						
Gene Symbol	Gene Identification	Probe Identification	Chromosome	cytoBand	$\hat{\beta}_j$	$\hat{\lambda}_j$
EFNA1	25k_1474684	RP11-307C12	1	1q21-q22	0.385	4.139
PPP1R7	814508	RP11-556H17	2	2q37.3	−0.021	3.908
RFC4	25k_309288	RP11-119E13	3	3q27	−0.541	3.850
FNTA	25k_530359	CTD-2115H11	8	8p22-q11	0.289	5.247
SMARCA2	814636	RP11-48M17	9	9p22.3	−0.045	4.060
ATP5C1	25k_845519	RP4-542G16	10	10p15.1	0.003	7.226
PRDX5	292519	RP11-147G6	11	11q13	0.333	5.402
MMP10	25k_1384851	RP11-750P5	11	11q22.3	−0.003	14.737

Table 7.2. Genes selected when Spearman correlation coefficients are used to determine the weights and two-dimensional cross-validation is used to determine the tuning parameters q and λ.

Four of the selected genes (EFNA1, PRDX5, FNTA, RFC4) were also selected when correlation between gene expression and copy number ($r > 0.4$) was used to reduce the data in Chapter 5. This is reasonable since the two methods are strongly related. The difference lies in that the previous analyses used more or less randomly chosen thresholds to decide the set of explanatory variables and that the genes for which the correlation did not exceed the threshold were completely removed from the analysis. In the weighted analysis, all genes are included in the analysis and genes with gene expression showing a strong relation to survival might get selected although the penalty is relatively large due to a high weight. We also see this from the results. We gain something by applying the weighted analysis compared to the reduction method in Chapter 5 in the sense that the weighted analysis selects four genes which not were selected in Chapter 5.

7.2. Penalization Scheme 2; Ridge Regression Coefficients

In this second penalization scheme we want to find a weight w_j based on whether the gene's copy number explains survival or not. If a gene has changes in copy number that explains survival, the gene should be given less penalty than others. Since a copy number alteration influences survival by first affecting the gene's expression, the genes within aberrated regions are more likely to explain survival through their expressions as well. A quantity that indicates the influence of each genes copy number on survival, could therefore be used as a weight on the penalty terms.

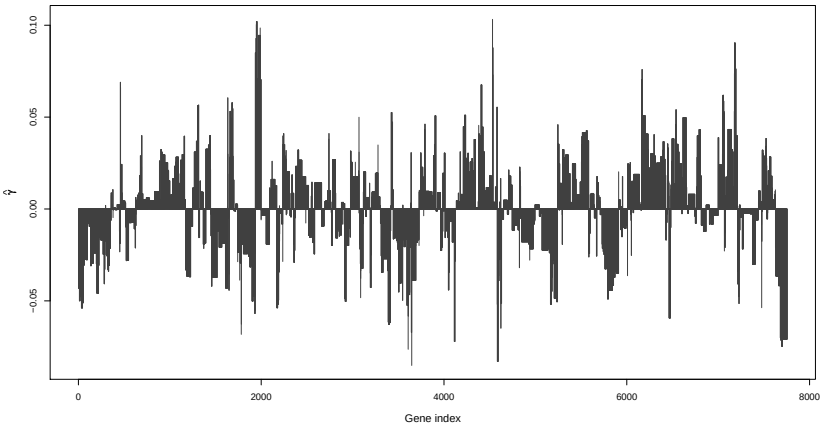
We may find such a quantity by fitting a Cox-ridge regression model to the copy number data and estimate the regression coefficients γ_i in this model. PS2 is performed by doing a Cox-ridge regression on the copy number data where the regression coefficients γ_i are estimated for each copy number probe. The regression coefficients indicate each probe's effect on survival. For gene j the weights may be defined to have the general form

$$w_j = \frac{1}{|\eta_j|^q},$$

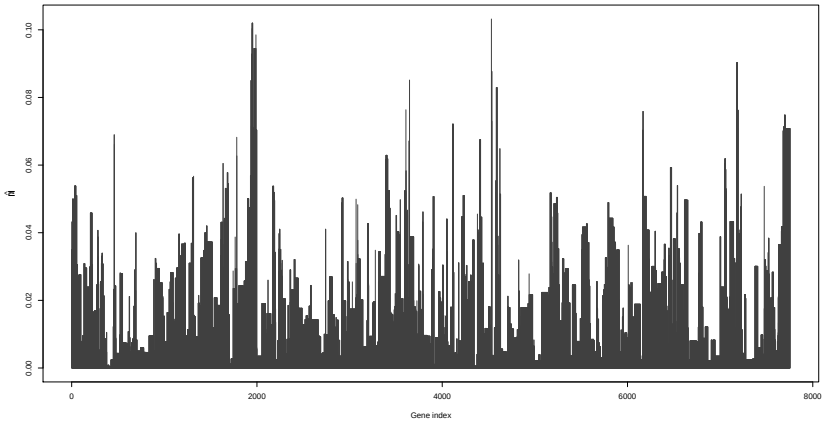
where we insert $\hat{\gamma}_j$ for η_j.

The genes found to influence survival through it's copy number are thus given smaller penalties. Genes corresponding to probes found to have less influence on survival are given larger penalties.

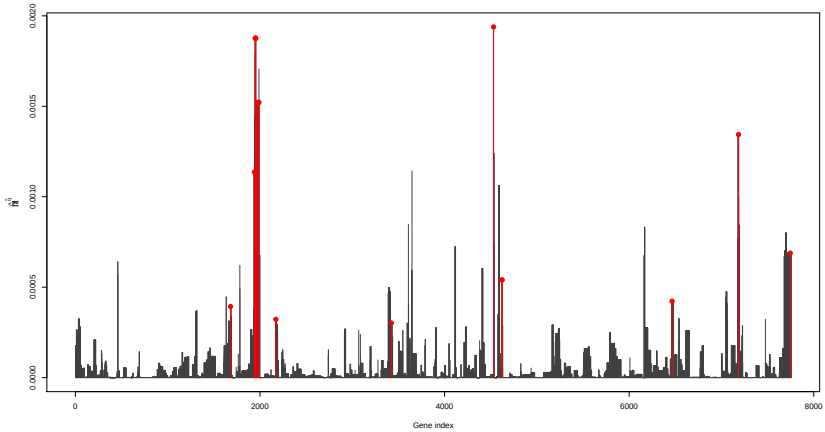
The weights were determined by first fitting a Cox-ridge regression model on the aCGH data with the survival data as response. The estimated regression coefficients $\hat{\gamma}_i$ are plotted in Figure 7.3(a). The absolute value of the regression coefficients are used to calculate the weights. Note that the plot of regression coefficients are given for the 2138 probes in the aCGH data in Figure



(a)



(b)



(c)

FIGURE 7.3. (a): Plot of the estimated ridge coefficients $\hat{\gamma}_j$ for each probe.
(b): Plot of $|\eta_j|$ for all genes j.
(c): Plot of $|\eta_j|^{\hat{q}}$ for all genes j. The genes selected in the final model are marked.
The genes are ordered by their position on the genome.

Systematic Description of the Procedure: Penalization Scheme 2

Assume data matrices X and Z, where X is a $n_1 \times p$ matrix consisting of p gene expression vectors for the n_1 patients. Z is a $n_2 \times r$ matrix which consists of the r copy number probes for n_2 patients. Assume also survival data for the patients.

(1) Determination of η_j:

a Fit a Ridge regression where $Z = (Z_1, ..., Z_r)$ is used as covariates and survival data as response. In the Cox regression setting this corresponds to maximizing;

$$l_{pl}(\gamma) - \lambda_R \sum_{i=1}^{r} \gamma_i^2,$$

and where $l_{pl}(\gamma)$ is the Cox log partial likelihood and λ_R is determined by K-fold cross-validation.

b Let $\eta_j = \hat{\gamma}_j$.

(2) Determination of λ_j:

a Define $\lambda_j = \lambda w_j$ where $w_j = \frac{1}{|\eta_j|^q}$.

b Estimate λ and q through two-dimensional cross-validation.

(3) Fit the model with $X = (X_1, ..., X_p)$ as covariates and the survival data as response. That is, maximize

$$l(\beta) = l_{pl}(\beta) - \lambda \sum_{j=1}^{p} w_j |\beta_j|$$

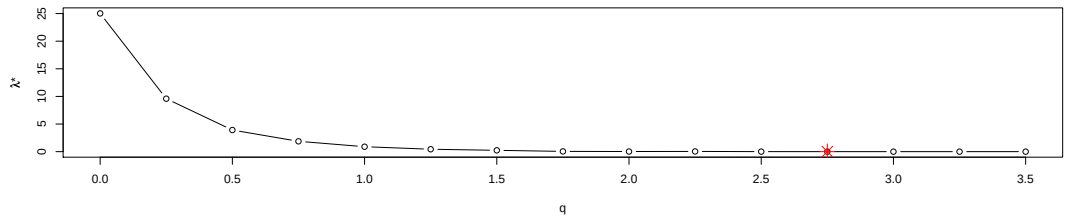
where $l_{pl}(\beta)$ is the Cox log partial likelihood and λ and w_j are estimated as described in Step 2.

Output: p regression coefficients β_j indicating the influence gene j has on survival.

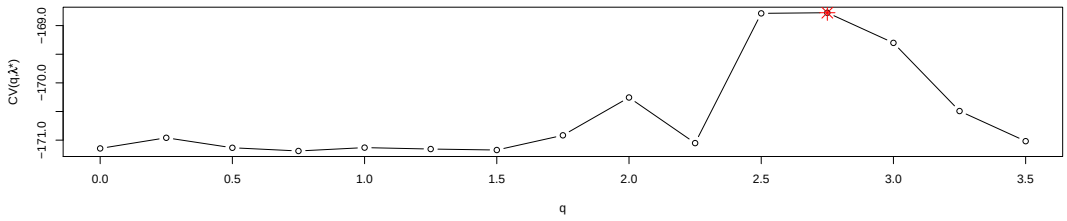
TABLE 7.3. Summary of the procedure using Penalization Scheme 2.

7.3(a), whereas in the plot in Figure 7.3(b), $\eta_j = |\hat{\gamma}_j|$ are plotted for all 7754 genes in the gene expression data after matching the copy number probes with the expression data. We define η_j as the absolute value of the estimated regression coefficient $\hat{\gamma}_j$ since the importance of the covariates depends on the magnitude of the regression coefficient and not its sign.

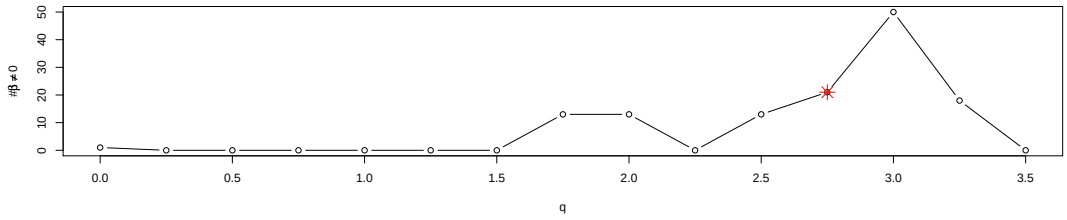
To decide on a proper value of q and λ, the two-dimensional cross-validation procedure is applied. The same grid ranging from 0 to 5 was set for PS2, but for $q > 3.5$ the cross-validation of λ fails. This is probably because λ gets too close to zero, which corresponds to not adding any penalty to the partial likelihood.



(a)



(b)



(c)

FIGURE 7.4. Plots for cross-validation of q when using Penalization Scheme 2: ridge regression coefficients.

Figure 7.4 illustrates the cross-validation results. The three plots are set up in the same way as the corresponding figures for PS1 in Figure 7.2. In the first plot in Figure 7.4(a) λ_q^* found to maximize the cross-validation curve $CV(q, \lambda)$ is plotted for each given q. In Figure 7.4(b) the cross-validation function $CV(q, \lambda_q^*)$ is plotted as a function of q, for which we want to maximize. The maximum is marked with a star to indicate the chosen value of both q in Figure 7.4(b) and λ_q^* in Figure 7.4(a). In the case of PS2, $\hat{q} = 2.75$ and global penalty parameter $\hat{\lambda} = 0.0031$.

We may study the results of the cross-validation a bit closer here as well. As for PS1 we see that as q increase, λ_q^* decreases. The same justification can be applied here as for PS1. The weights will contribute to the simultaneous shrinkage by creating overall larger penalty terms for all genes when q increases.

For $0 < q \leq 1.5$ zero genes are selected and we see that the values of $CV(q, \lambda_q^*)$ are relatively low compared to when a larger number of genes are selected. For the values of q that lead to selection of more than one gene we observe higher values for the cross-validation criterion $CV(q, \lambda^*)$. For different q-values the selected genes seem to be in agreement and all are included in the 50 genes which are selected for $q = 3$.

In Figure 7.4(c) the number of nonzero regression coefficients is plotted for each value of q. Values of q which lead to some nonzero regression coefficients in the final model seem to perform better than the q-values where zero genes are selected, according to the cross-validation curve. The number of genes selected for the preferable value of q, $q = 2.75$ was found to be 21 genes. By comparing Figure 7.4(b) and 7.4(c) one may see that the two curves show some similar behavior for the same values of q. The cross-validation curve $CV(q, \lambda)$ seems to increase when the number of variables included in the model is larger than one. The figures, however, illustrate that including too many genes in the model may not necessarily give a better performance. When 50 genes are included for $q = 3$, the value of the cross-validation curve decreases,thus $q = 3$ seems to overfit the model.

When the weights were chosen and the parameters decided through cross-validation we fitted the final model for which the gene identification numbers and gene symbols are listed in Table 7.4. In Figure 7.3(c) $|\eta_j|^{\hat{q}}$ are plotted for all genes j. In the plot the selected genes are marked in red. The genes corresponding to large values of $|\eta_j|^{\hat{q}}$ in the plot are promoted in the analysis. Genes could still be selected in the analysis if the gene expression shows a strong effect. In fact, we see that not solely the genes with the largest estimated ridge regression coefficients $\hat{\gamma}_j$ were selected in the final analysis.

Two of the selected genes listed in Table 7.4 turned out to be the same gene. Some genes were represented by different gene identifications in the data and therefore represented by two different variables in the analysis. This concerned the two genes with gene symbols SRP72 and CXCL1.

Comparing these results with the results of PS1, we may see that two of the genes selected for PS2 also were selected for PS1. This applies to the two genes FNTA and MMP10. That the two genes were selected in both penalization schemes gives even stronger reason to believe that the two genes are related to survival. Studying the estimated individual penalty parameters for the selected genes, we see that in neither of the two penalization schemes the two genes are among the genes which were given very small penalties. Of the selected genes, these are the two which got less help from the penalization. Studying Figure 7.3(c), it is obvious that there are several genes corresponding to larger values of $|\eta_j|^{\hat{q}}$ that were not selected, and it is reason to believe that the gene expressions for these genes do not explain survival and are therefore not selected.

Results; Penalization scheme 2						
Gene Symbol	Gene Identification	Probe Identification	Chromosome	cytoBand	$\hat{\beta}_j$	$\hat{\lambda}_j$
CSTA	25k_345957	RP11-299J3	3	3q21	0.007	7.853
TXK	148421	RP11-100N21	4	4p12	−0.241	2.717
KIAA1211	272531	RP11-738E22	4	4q12	0.024	1.646
PAICS	273546	RP11-738E22	4	4q12	0.066	1.646
SRP72	811842	RP11-738E22	4	4q11	0.466	1.646
SRP72	25k_814702	RP11-738E22	4	4q11	0.422	1.646
COX18	121420	RP11-447E20	4	4q13.3	0.013	2.030
ANKRD17	179143	RP11-447E20	4	4q13.3	0.051	2.030
RASSF6	282564	RP11-447E20	4	4q13.3	0.230	2.030
CXCL6	2315207	RP11-447E20	4	4q21	0.093	2.030
CXCL1	324437	RP11-447E20	4	4q21	0.244	2.030
CXCL1	25k_324437	RP11-447E20	4	4q21	0.012	2.030
MTHFD2L	701417	RP11-2G10	4	4q13.3	0.305	2.030
SDHA	80915	CTD-2265D9	5	5p15	−0.018	9.574
FNTA	530359	CTD-2115H11	8	8p11	0.022	10.221
MRPL21	809517	RP11-554A11	11	11q13.2	0.235	1.593
MMP10	25k_1384851	RP11-750P5	11	11q22.3	−0.064	5.717
EST	25k_324492	RP11-750P5	11	11q22.3	−0.044	5.717
HGS	25k_264646	RP11-475F12	17	17q25	−0.004	7.305
EST	144797	RP11-15H6	21	21q21.2	−0.098	2.298
MPP1	296880	RP11-296N8	X	Xq28	−0.014	4.485

Table 7.4. Genes selected when ridge regression coefficients are used to determine the weights and two-dimensional cross-validation is used to determine the tuning parameters q and λ.

This was important to see since it indicates that the method works well. A gene should not be selected due to a small penalty alone, but due to the gene expression in combination with our prior belief from aCGH data.

7.3. Penalization Scheme 3; Standard Deviation

The last penalization scheme is based on how much variation one observes for the copy number data for a gene. This is as for the two other penalization schemes motivated from the fact that copy number alterations can result in changes of the expression. A gene with a copy number that is constant over the patients is therefore not considered as a possible driving force for cancer progression due to changes in copy number. A gene which shows high variation in its copy number can, however, be more probable of influencing survival through its gene expression. To define the weights

$$w_j = \frac{1}{|\eta_j|^q}, \tag{7.2}$$

Systematic description of the procedure: Penalization Scheme 3

Assume data matrices X and Z, where X is a $n_1 \times p$ matrix consisting of p gene expression vectors for the n_1 patients. Z is a $n_2 \times r$ matrix which consist of the r copy number probes for n_2 patients. Assume also survival data for the patients.

(1) Determination of η_j:
   - **a** Calculate the empirical standard deviation $\hat{\sigma}$ for each copy number probe.
   - **b** Let $\hat{\eta}_j = \hat{\sigma}_j$.

(2) Determination of λ_j:
   - **a** Define $\lambda_j = \lambda w_j$ where $w_j = \frac{1}{|\eta_j|^q}$ and η_j may be replaced by $\hat{\eta}_j$.
   - **b** Estimate λ and q through two-dimensional cross-validation.

(3) Fit the model with $X = (X_1, ..., X_p)$ as covariates and the survival data as response. That is, maximize

$$l(\beta) = l_{pl}(\beta) - \lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

where $l_{pl}(\beta)$ is the Cox log partial likelihood and λ and w_j is estimated as described in Step 2.

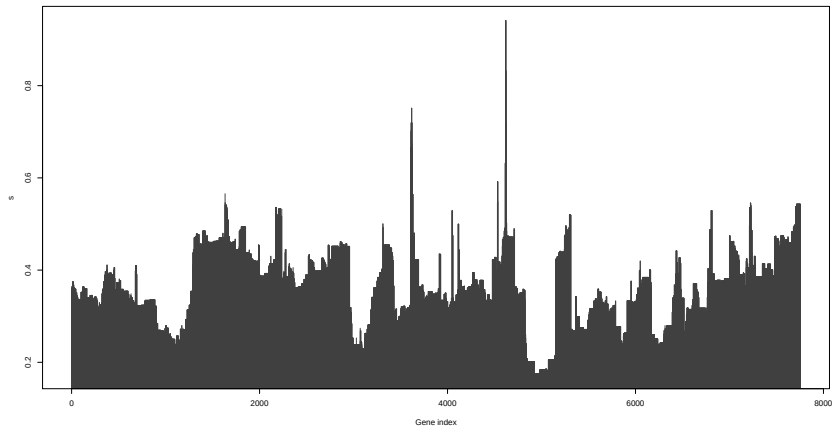
Output: p regression coefficients β_j indicating the influence gene j has on survival.

TABLE 7.5. Summary of the procedure using Penalization Scheme 3. The penalty term will be weighted such that the regression coefficients for genes corresponding to a high standard deviation in Step 1, will have a larger probability of being estimated to be nonzero in Step 3.

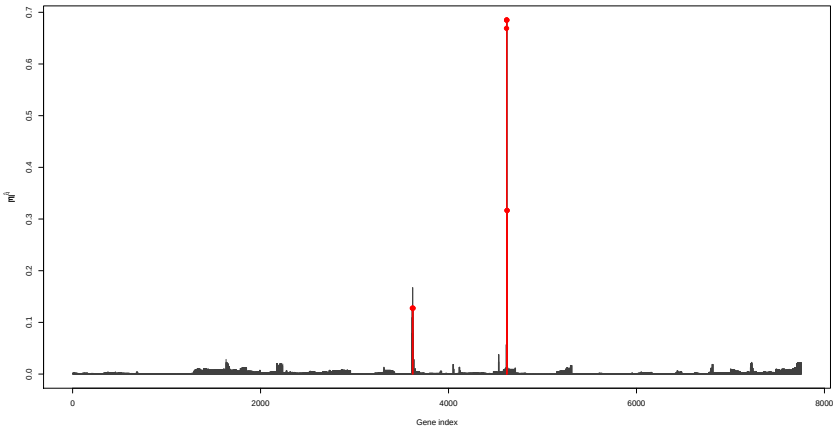
a quantity η_j measuring the variability for the gene's copy number is needed. One option is to use the empirical standard deviation $\hat{\sigma}_j$ for the copy numbers of gene j, hence let $\eta_j = \hat{\sigma}_j$. The weights will be small when there is high variability in the copy number values for gene j, and the gene will be given a smaller penalty which will encourage the gene to be selected in the genewise lasso procedure. If there is low variability in the copy number, the gene will be given a larger penalty.

To apply PS3 to the data the empirical standard deviations, $\hat{\sigma}_j$, were calculated for each probe j and matched together with the corresponding gene expression. These are plotted in Figure 7.5(a). The standard deviations range from 0.17 to 0.94.

The figures illustrating the cross-validation results for PS3 look different than for the two other penalization schemes. There was a need to expand the cross-validation grid for q, since zero genes were selected when the grid was ranging from 0 to 5. By expanding the grid using q-values ranging from 0 to 10, we ensured that the right q-value was selected in the cross-validation. When



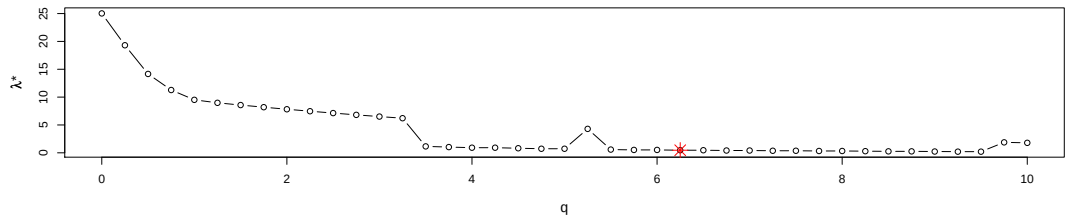
(a)



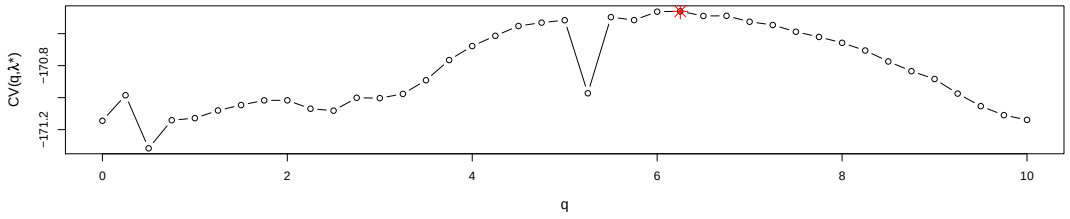
(b)

FIGURE 7.5. (a): Plot of the standard deviations $\hat{\sigma}_j$ for each gene.
(b): Plot of $|\eta_j|^{\hat{q}}$ for each gene and the genes selected by the lasso are marked in red. The genes are ordered by their position on the genome.

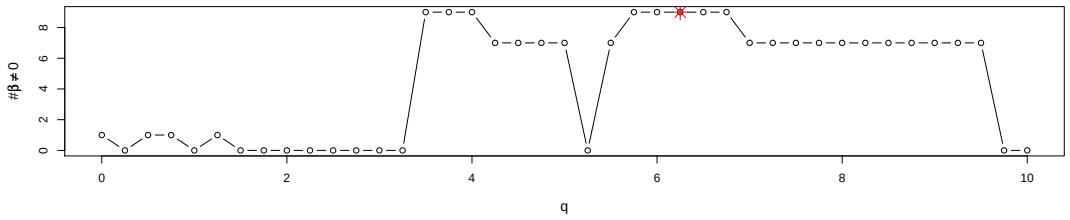
comparing the plots in Figure 7.6 with the two previous penalization schemes, remember that the grid values for q range from 0 to 10. We first consider Figure 7.6(a) where the cross-validated values of λ, λ_q^* are plotted for the different values of q. We see that λ_q^* does not decrease as fast as for the two other penalization schemes. For $q \leq 3.25$ the values of λ_q^* exceeds 5. This is quite large compared to the same values of q in the other penalization schemes. This indicates that the introduction of weights with $q \leq 3.25$ will not differentiate enough between the genes such that a larger common penalty term is still needed. It is reason to believe that this is related to the value of q and η_j, and some brief comments on this is given in the discussion ending this chapter. For $q > 3.25$, λ_q^* is on the same level as for the two other penalization schemes.



(a)



(b)



(c)

FIGURE 7.6. Plots for cross-validation of q when applying Penalization Scheme1: Standard deviation.

For these values of q more genes were selected. The preferred value of q was found to be 6.25. The cross-validation curve seems to behave properly, except for at $q = 5.25$. By first looking at Figure 7.6(a) we see that there is a peak at $q = 5.25$. This indicates that a relatively high value is chosen for λ_q^*. It is difficult to point on any good reason for this to happen for this exact value of q, but it may be due to some computational uncertainties in the cross-validation of λ.

The two plots in Figures 7.6(b)-(c) show much less variable results for different values of q, than for the two previous penalization schemes. In PS1 and PS2 some variables were excluded and some retained for different values of q. For PS3 the same genes seem to be selected for different values of q.

The two-dimensional cross-validation procedure chose the optimal parameters $\hat{q} = 6.25$ and global penalty parameter $\hat{\lambda} = 0.4642$. This corresponds to nine selected genes which are listed

in Table 7.6. The weights are plotted in Figure 7.5(b) where the selected genes are marked in red.

Results; Penalization scheme 3						
Gene Symbol	Gene Identification	Probe Identification	Chromosome	cytoBand	$\hat{\beta}_j$	$\hat{\lambda}_j$
RCL1	125148	RP11-125K10	9	9p24.1-p23	−0.001	3.639
MLANA	266361	RP11-218I7	9	9p24.1	−0.401	3.642
GLDC	248261	RP11-106A1	9	9p24.1	0.011	3.642
YAP1	308163	RP11-21G19	11	11q22	0.037	0.694
BIRC3	201890	RP11-315O6	11	11q22	0.335	0.678
BIRC3	428231	RP11-315O6	11	11q22	0.631	0.678
BIRC2	34852	RP11-315O6	11	11q22	−0.382	0.678
MMP10	25k_1384851	RP11-750P5	11	11q22.3	−0.201	1.466
EST	25k_324492	RP11-750P5	11	11q22.3	−0.132	1.466

Table 7.6. Genes selected when empirical standard deviation is used to determine the weights and two-dimensional cross-validation is used to determine the tuning parameters q and λ.

Another important aspect is related to using the standard deviation for the gene's copy number as a weight. One has to keep in mind that the negative copy numbers represent deletion and the positive values amplifications. Normally the copy number is 2 in each cell, this corresponds to 0 in our data. A deletion of DNA copy number corresponds to loss of genetic material, whereas an amplification corresponds to gain. This is described in Chapter 2. A region could therefore be high-amplified corresponding to a very high copy number. Since 2 is the normal copy number, there is a lower bound for the copy numbers at -2, while there is no upper bound for amplifications. Regions that are subject to high amplification will probably have larger standard deviations as well. This indicates that high-amplified regions will be favored in Penalization Scheme 3, since these regions correspond to high copy numbers and consequently high variance in the data. The penalization scheme is, however, not completely biologically wrong. The selected genes can be interpreted as being related to survival as a consequence of high variance in the copy number, but one should have in mind that we probably loose genes in regions subject to deletion, and that amplified regions will be favored. All genes selected are within regions with high amplification, that is, with five or more copies. This verifies our assumptions.

7.4. Biological Validation

Genewise lasso analyses were in the previous sections carried out using three different penalization schemes. For each of the three penalization schemes a list of the selected genes was presented. It is of interest to study these genes in a biological context and whether the selected

genes are known or previously studied in relation to (cervix) cancer. In the following, a brief review is given to relate some of the genes found in the analyses to other biological findings.

EFNA1, selected in the analysis with PS1, is a known oncogene and high expression of EFNA1 was associated with poor survival of patients with cervical cancer in Holm et al. (2008). Up-regulation of PRDX5 is also previously associated with poor survival: Overexpression of PRDX5 protects apoptosis and loss of cellular function during oxidative stress (Yuan et al., 2004). For another of the selected genes in PS1, upregulation is also associated with poor survival. In Nagase et al. (1999) overexpression of FNTA promoted cell growth, and FNTA-overexpressing cells formed tumors in nude mice. FNTA was also selected when using Penalization Scheme 2.

CSTA is selected by Penalization Scheme 2 and has been shown to be upregulated in metastatic cervical tumors in Lyng et al. (2006). Upregulation of CSTA has also been correlated with poor prognosis of breast cancer (Kuopio et al., 1998). Moreover down-regulation of BIRC2 which was selected in Penalization Scheme 3, has been associated with poor prognosis of renal cell carcinomas (Kempkensteffen et al., 2007).

These known results are in agreement with the results found using the genewise lasso procedure when comparing the signs of the regression coefficient. An upregulation of a gene associated with poor survival should have a positive regression coefficient and a downregulated gene associated with poor survival should have a negative regression coefficient in the analyses. The description of all selected genes are given in Appendix A. For most of the genes the function is in agreement with the signs of our estimated coefficients, but some of the genes have not been extensively studied.

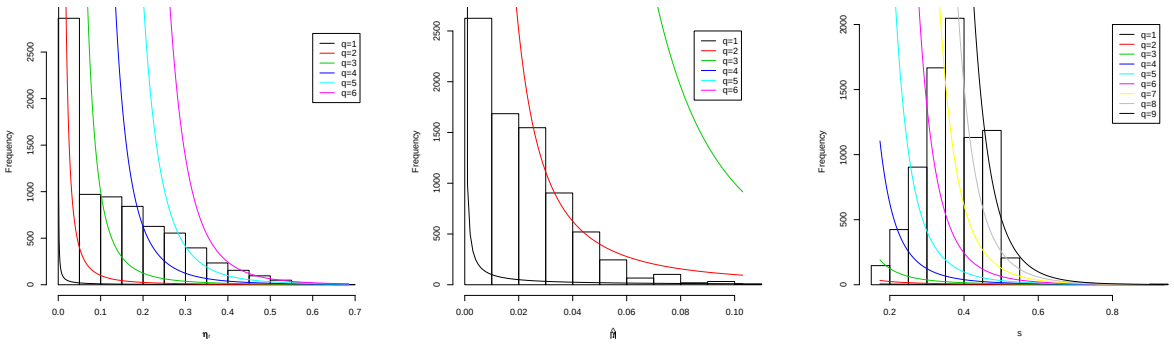
7.5. Further Comments on the Weights

As briefly commented, effective values of q will depend on the size of the η_j's. We may study the weight functions

$$w_j = \frac{1}{|\eta_j|^q}, \tag{7.3}$$

to get a better understanding.

For Penalization Scheme 1 we observe $0 < \eta_j < 0.7$. For Penalization Scheme 2 we observe $0 < \eta_j < 0.11$, whereas for Weighting scheme 3 we observe $0.2 < \eta_j < 0.9$. If we plot the weight functions in (7.3) in the respective intervals and for different values of q we will experience that different values of q are needed in the three different schemes, depending on the relevant intervals for which we observe η_j.



(a) Penalization Scheme 1 (b) Penalization Scheme 2 (c) Penalization Scheme 3

FIGURE 7.7. Plot of the weight functions in the relevant intervals for each penalization scheme. We may see that different q-values is effective in different intervals of $|\eta_j|$ and thus for the different penalization schemes.

Plots of these functions for the three penalization schemes are given in Figure 7.7 (a)-(c). For Penalization Scheme 1, $q = 2, ..., 6$ seem to be able to distinguish between high and low values of η_j. With $q = 1$ the genes corresponding to very low values of η_j (zero/negative correlations), will be given large penalties in the genewise lasso penalization. Increasing q to $q = 5$ seems to give almost all genes a large penalty. As seen in the cross-validation results in Figure 7.22(c), $q = 5$ will result in zero genes selected.

Figure 7.7(b) illustrates the choices of q for Penalization Scheme 2. Since the values of η_j are distributed in a lower interval, we see that for $q > 3$, all genes will be given large penalties. This is in coincidence with what we experienced in the cross-validation for Penalization Scheme 2 which failed for $q > 3.5$. This could be due to the fact that when $q > 3.5$ all genes will be given a very large penalty as a consequence of the small values of $|\eta_j|$.

For the third penalization scheme we see from Figure 7.7(c), that the values of η_j are not as close to 0 as in the other penalization schemes and larger values of q are needed to find effective weights. This was found in the cross-validation as well, where the grid was expanded to range from 0 to 10 instead of 0 to 5. Setting $q = 1, 2, 3$ for PS3 corresponds to giving many genes small penalties and it seems that weights using larger q-values will be more effective.

Note that the histogram in the background of Figures 7.7 (a)-(c) illustrates how the η_j-values are distributed in the relevant interval, but should not be confused to be related to the weight functions other than by indicating how many genes will be given a small penalty and how many will be given a large penalty. It is also important that these figures do not consider λ which also will influence the penalization when it is estimated. For the two first penalization schemes, this

is not of major concern since the estimated λ_q^* for $q \geq 1$ seem to be on fairly the same level. The same can be seen for the estimated λ_q^* values for $q > 3.5$ in the last penalization scheme.

CHAPTER 8

Validation on New and Independent Data

In this chapter we evaluate the performance of the selected genes as biomarkers on a new and independent data set. As is previously commented, especially in Chapter 4, a complex model may overfit the data. This means that although variables are found to be important for the data used to train the model, they may not perform well as predictors for new and independent data.

We concentrate on the genes selected in the three penalization schemes of Chapter 7 and the genes selected when correlation between gene expression and aCGH data was used to reduce the data in Chapter 5. Gene expression data have been collected for 41 new and completely independent cervical cancer patients and the validity of the genelists was evaluated by clustering the patients into two groups based on the listed genes and testing whether survival in the two groups is significantly different through a log-rank test.

8.1. Data

A new data set for 41 patients in an independent cohort was provided by the biologists at the Radium Hospital. The data contain survival data and gene expression data for the 41 patients. The expression data were measured by a totally different method than cDNA microarrays, called Illumnia. See Lando et al. (2009) for a description of how the data were extracted by Illumnia. In the data sets there are many isoforms for the same gene. We chose to include all of these in the following analyses.

Tumor samples were collected at the time of diagnosis, and all patients received the same treatment as the patients in the cohort studied in the previous chapters. The survival times are the time between diagnosis, and relapse or cancer related death. Patients that have not experienced a relapse of the disease, or are dead of other reasons not related to cancer, are censored. In Figure 8.1 the observation times for all patients are plotted when ordered from high to low. A summary of the survival data is also given in Table 8.1. We may see that for about 30% of the patients there is observed a relapse of the disease. The survival times range from 2.033 to 46.131 months. The median survival time is also reported, and we see that the survival times are relatively much longer for the censored compared to the observed.

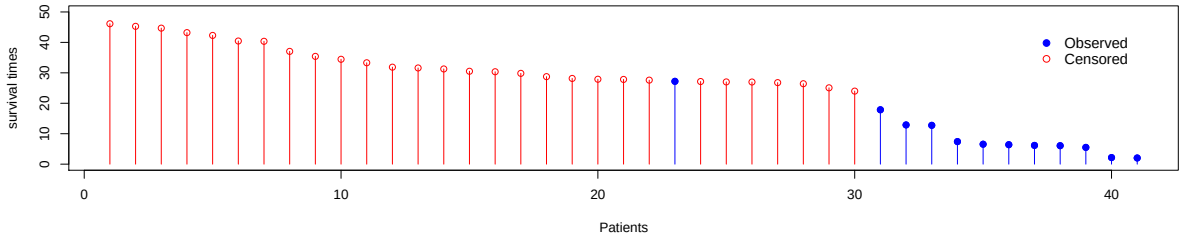


FIGURE 8.1. Plot of the survival times for the patients in the new cohort.

SUMMARY OF SURVIVAL DATA				
Data	Number of Patients	Median Survival Time	Minimum Survival Time	Maximum Survival Time
All Patients	41	27.836	2.033	46.131
Observed	12	30.525	24.000	46.131
Censored	29	6.475	2.033	27.213

TABLE 8.1. Overview of the survival data for the new cohort.

8.2. K-Means Clustering

We wanted to investigate whether the selected genes were able to separate the patients into groups for which the survival is significantly different. One way of doing this is to cluster the patients into groups based on the gene expression data for the relevant genes. Various methods for clustering the data exist. For simplicity we chose to cluster the patients into two groups (imagine possibly one positive and one negative for tumor progression). This could for instance be done by K-means clustering, which cluster the data into K groups. To obtain two groups we set $K = 2$. Based on an initial set of K cluster centers, the K-means algorithm alternate through the following steps until convergence is reached;

- for each data point, identify the closest cluster center (in Euclidean distance), and assign the data point to that cluster,
- recalculate the cluster means based on the data points currently assigned to the cluster and define these as the new cluster centers.

When the assignments do not change, the K groups are defined. The procedure is described in more detail in for instance Hastie et al. (2001, chap. 14). The clustering was done with the standard routine for K-means clustering in R; `kmeans()`.

K-means clustering was performed for the new cohort of 41 patients based on the genes selected in the three penalization schemes in genewise lasso penalization and the genes selected in the method using correlation ($r > 0.4$) to reduce the data set. For all four methods we were able to separate the patients into two groups based on the selected genes and there are several patients in both groups. Table 8.2 shows the number of patients in the two groups for each of the four approaches.

	Number of Patients	
Method	Group 1	Group 2
PS1	23 (4)	18 (8)
PS2	25 (3)	16 (9)
PS3	9 (2)	32 (10)
R04	23 (4)	18 (8)

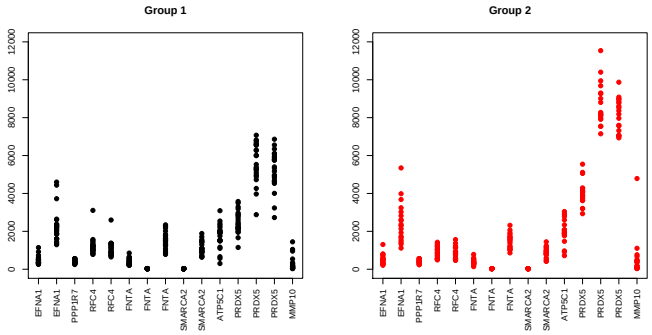
TABLE 8.2. The number of patients in the two groups for each of the four approaches. The number of observed cases of relapse in each group is given in parenthesis.

For each of the four approaches the gene expression values for the two groups are plotted in Figure 8.2. Studying these plots one may see that the gene expression values for some genes seem to differentiate better between the two groups than others. For instance, for PS1 we see that the gene expression values of PRDX5 seem to be very different in the two groups, taking on much larger values in Group 2. For the other genes it is harder to distinguish the possible differences in the gene expression values by eye. For PS2 we see a similar difference between Group 1 and 2 for the gene PAICS. For PS3 it is difficult to distinguish from the plot, but BIRC2 may show some higher values for Group 1. For the reduction method where $r > 0.4$ we see the same as for PS1. This is reasonable since PRDX5 is selected in both approaches, and is the gene showing the largest difference in expression between the two groups.

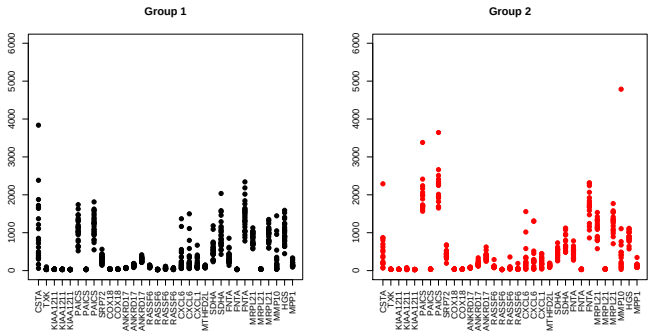
8.3. Kaplan-Meier Analysis and Log-Rank Test

When the patients were clustered into two separate groups for each of the four approaches, we wanted to look for differences in survival in the two groups. This could be done by estimating the survival curves for the two groups by the Kaplan-Meier estimator and visualize these in Kaplan-Meier plots. The Kaplan-Meier estimator was briefly discussed in Chapter 3. Kaplan-Meier plots are shown for the groupings in each of the four approaches in Figure 8.3. In addition a log-rank test was performed to test whether the groups show significant differences in survival.

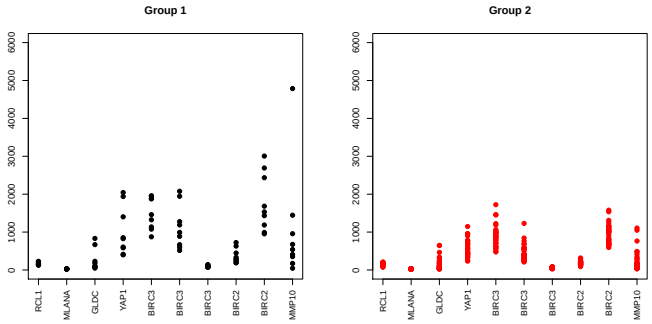
The Kaplan-Meier plot for PS1 indicates a clear difference between the two estimated survival curves. The p-value of the log-rank test is 0.0346 and indicates that the hazards of the two groups



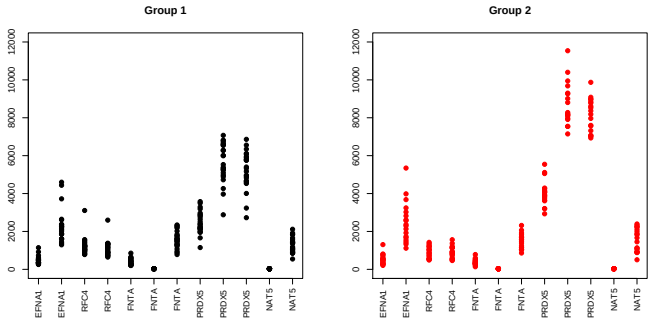
(a) Penalization Scheme 1



(b) Penalization Scheme 2

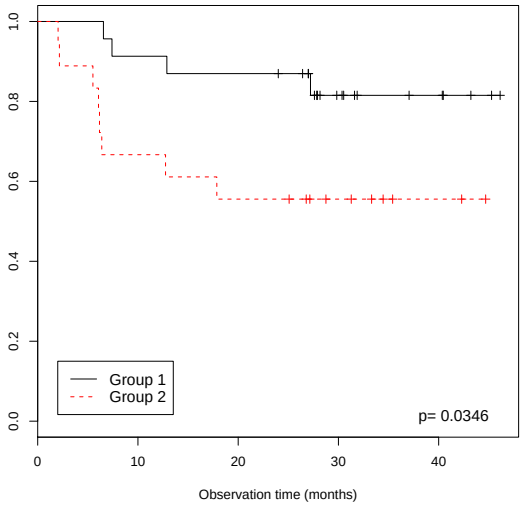


(c) Penalization Scheme 3

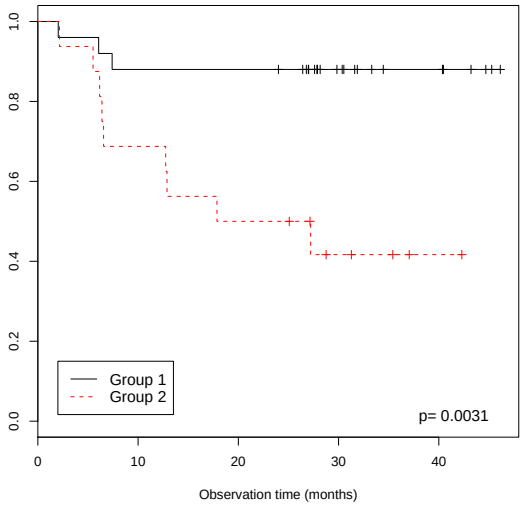


(d) Reduction, $r > 0.4$

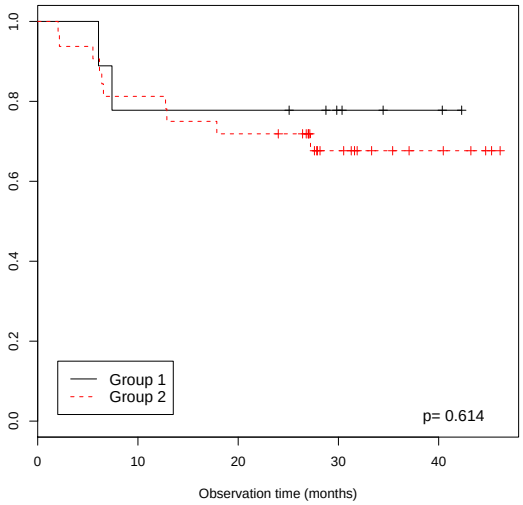
FIGURE 8.2. Plots of gene expression data for when the patients are clustered in two groups by K-means clustering.



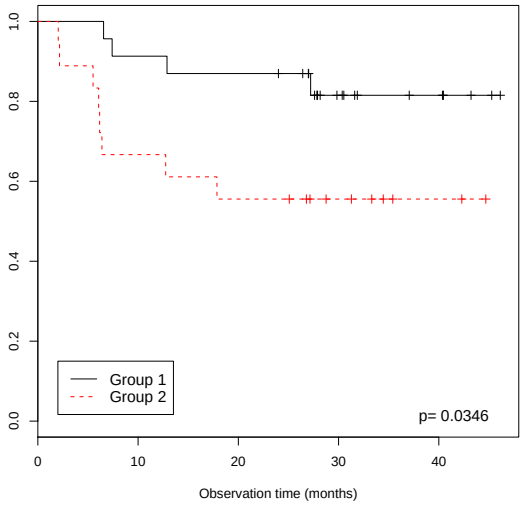
(a) Penalization Scheme 1



(b) Penalization Scheme 2



(c) Penalization Scheme 3



(d) Reduction, $r > 0.4$

FIGURE 8.3. Kaplan Meier plots for the groups in each of the four methods under study. The p-value of the log-rank test is indicated.

are significantly different. For PS2 there is an even more obvious separation between the two curves, indicating an even clearer difference in survival. The log-rank test implies significantly different hazards for the two groups with $p = 0.0031$.

The last penalization scheme seems to show less difference in the survival curves, indicating that the genes selected in PS3 are not able to separate the patients into two groups with significant differences in survival (p=0.614). This is, however, in accordance with our suspicions in Chapter 7, where it was indicated that the penalization scheme favors genes with high amplification (high

copy number), and that PS3 may not be as good an approach as the two previous penalization schemes. In Figure 8.2(c) the gene expression data are plotted for the selected genes in PS3. Compared to the other plots in Figure 8.2 it is more difficult to see differences in the expression values between the two groups. Based on this and the results of PS3 in the previous chapter, the poor performance may not come as a surprise.

The Kaplan-Meier plot for the method reducing the gene expression data based on correlation ($r > 0.4$) is given in Figure 8.3(d). From the Kaplan-Meier plot and the indicated p-value there is a significant difference in the survival of the two groups. The p-value is 0.0346. The Kaplan-Meier plots for SP1 and the reduction method ($r > 0.4$) looks quite similar. Studying the plots of the gene expression data in Figure 8.2, we see that the gene PRDX5 was selected in both of these two approaches. By simply studying the plots, it seems that PRDX5 is important when the genes are clustered into two groups since the expression values of PRDX5 are very different in the two groups.

8.4. Discussion

The data set for the new cohort was generated and ready for analysis just at the end of this study, which made it possible to do an independent testing. The results of Chapter 7 are considerably strengthened by the validation results in this chapter. Based on the genes selected in PS1 and PS2, we were able to divide the patients into two groups showing significantly differences in survival. The validation indicated robust results even for the new data set which contained gene expressions measured with a totally different teqhnique compared to the original gene expression data.

Time permitting, there are, however, several things which could be analyzed with more care when evaluating the performance in this chapter. More attention could be drawn to the method used to cluster the patients. We used K-means clustering and chose to cluster the patients into two groups by convenience. Other approaches, for instance hierarchical clustering, could be applied and a more careful determination of different alternatives for distance and linkage measures could be done, as well as methods for estimating the number of clusters in the data.

In the gene expression data in this chapter, different isoforms for the same gene was included in the clustering. It could be investigated further which of these should be used in the clustering. Whether or not other preprocessing procedures should be applied to the data should also be evaluated to the full extent.

Although the validation approach discussed in this chapter could be improved as indicated above, the approach is convenient since only gene expression data are available for the new cohort. If

aCGH data were available, one way of validating the results on the new cohort could be to apply genewise lasso penalization on the independent data set and check whether the same genes are selected or not.

CHAPTER 9

Concluding Remarks

Data integration is a highly actual issue in connection with high-dimensional data sets. This is particularly relevant in genomics where different types of measurements are related to the same biological features. Such relations should be taken advantage of, by combining different data sources properly in a statistical model. The main objective of this thesis was to study the possibilities of data integration for genomic high-dimensional data in a Cox regression setting and to propose a novel method for data integration, aiming on selection of genes that are important for survival. The proposed method is a new version of a shrinkage method in a regression setting with $p > n$, which imposes individual amounts of shrinkage to the regression coefficients based on external information from another data source.

A high-dimensional data set, *The Radium Hospital Cervix Cancer Cohort Data*, was described in Chapter 2. This data set contains both gene expression and copy number data for cervix cancer patients, along with survival data. Before presenting the method for data integration, the statistical background theory was presented, and analyses using the standard lasso in a Cox regression model without data integration were carried out. For the original data sets, the lasso selected one gene for the gene expression data and one region for the aCGH data. As often, when analyzing data where the number of explanatory variables is very large, a reduction of the data set was needed prior to the analyses, even when the lasso or other shrinkage methods were applied. Some simple approaches for reducing the number of covariates were studied, where different criteria were applied for both data sets. A first step toward integration of the two data sources was to select variables depending on the size of the correlation coefficients between gene expression and aCGH data.

Both some of the simple reduction methods and the integrated reduction method did lead to a higher number of selected genes by the lasso. This is, however, not necessarily an optimal approach for analyzing the data. In many settings it is difficult to decide on an effective criterion to exclude some variables from the analysis. One also has to make a choice of a threshold value. These thresholds are often difficult to decide, and variables are either excluded or retained in the model exclusively depending on the choice. To circumvent these issues, the method for genewise lasso penalization was introduced in Chapter 6. The idea was to suggest a regression model which integrates the information in the two data sources in a suitable way. By modulating the penalization terms in a lasso regression model, the explanatory variables in the model could be

given penalty terms specific for each regression coefficient. By introducing a Bayesian interpretation, it was illustrated that the method could be viewed in an (empirical) Bayesian framework where inclusion of external knowledge is crucial.

In Chapter 7 the proposed method for data integration was applied to the data. Three different genewise penalization schemes were suggested. For the three different schemes, the method selected three sets of genes which are considered as possible biomarkers for the disease. An independent gene expression data set for 41 new patients was introduced in the last chapter to test whether the selected genes were able to predict survival for new patients. The genes selected in two of the penalization schemes (PS1 and PS2) and the method reducing the gene expression data based on correlation ($r > 0.4$) with the copy number data, were able to divide the patients into two groups showing significantly different survival.

There are several subjects for discussion connected to the data and the analyses performed, and comments and discussions are given in the appropriate chapters. One issue which not has been discussed previously is related to the data at hand and relatively low number of observed cases of relapse (or cancer related death). There is observed a relapse for only 1/3 of the patients, while 2/3 are censored. Most of the patients that are censored will most probably not experience a relapse. This is specific for the disease under study, for which the patients most probably experience a relapse of the disease within 30 months after the diagnosis/treatment. When fitting the Cox-lasso model with a large number of covariates, it would of course had been preferable to have a larger number of observed cases to improve the fit. The heavy censoring also makes it difficult to divide the data into a training and a test set for validation, since the fit would probably suffer from having too few observations in the training data. It was therefore necessary to introduce a completely new and independent data set in order to evaluate the prediction ability of the selected genes.

As mentioned in Chapter 5, there has been much less attention directed to aCGH data in connection with survival, than there has been on gene expression data. Although the main focus in this thesis has been to the analysis of the gene expression data, Chapter 5 also describes some analyses done of the aCGH data. The analyses of the aCGH data did however not give any clear conclusions. There are a number of reasons for this. First and foremost, the aCGH data are very correlated, hence the covariates in the regression model will be very correlated. This is not infeasible for the lasso method, but there are reasons for this to be ineffective. One reason is that the lasso in blocks of correlated covariates, tends to select only one or some of these variables, disregarding the rest of the block. In further analyses one may then be focusing on "wrong" regions. One way of resolving this particular problem is as we did in Lando et al. (2009). When the lasso analysis is done, one may find the correlation between the selected regions and all other regions. Regions which are highly correlated with the selected regions,

should then be studied with the same importance as the regions selected by the lasso. This worked well in Lando et al. (2009), where both selected and correlated regions were studied and found to be in correspondence with biological findings. Another option also discussed in Chapter 5 is to cluster the data and fit the regression model with representatives of the clusters as covariates. This did not work out in our analysis, but has been showed to be successful in for instance Park et al. (2007).

Another aspect concerning the separate analysis of the aCGH data which could be of interest, is the fact that the data in addition to report the relative gene dosage, reflects three different states: "normal", "loss" and "gain" in copy number. It could be interesting to do a regression analysis defining categorical variables for the copy number data. This is also a common way of analyzing variables in the Cox regression model if they do not show a log-linear effect on the hazard. This may lead to a more effective analysis and may improve upon the results.

We have seen that the new method applied to the data studied in this thesis, was able to select groups of genes which can be considered as possible new biomarkers of cervix cancer. The results of Chapter 8 strengthen the importance of most of the selected genes even more, when validated on an independent data set. Hence the advantages in regression analyses of genomic data are obvious, and one could imagine various other high-dimensional settings where data integration in (lasso) regression models is sensible, for instance in climate on finance research. Although we use the term *Genewise lasso penalization* and discuss the method in a biological context motivated from the data at hand, the method could as mentioned be applied to any other type of data where external information on the covariates are available. The external information could be additional data collected on the covariates or any expert knowledge indicating whether some variables may be of larger importance than others.

In other applications, analyzing other than survival data, one may be interested in fitting other types of regression models. The new method could just as well be transformed to apply in other regression designs, for instance in the family of glm-models. In addition to applying the method to other types of data and in other applications and regression models, the new idea of giving each variable an individual penalty could also be used in combination with different penalty terms, such as for example ridge regression.

For the data analyzed in this thesis, the results suggest that there might be something to gain by applying *Genewise lasso penalization*. With the possible extensions to other scenarios and regression models, as well as other penalization terms, there are several interesting topics for further research. As was the motive for this thesis, it is reason to believe that an increased interest of data integration will evolve in the future. The proposed method may then be an appropriate suggestion when data integration in penalized regression models is desired.

Bibliography

Aalen, O., Borgan, O. & Gjessing, H. K. (2008). *Survival and Event History Analysis.* Springer.

Bejjani, B. A., Theisen, A., Ballif, B. & Shaffer, L. (2005). Array-based comparitive genomic hybridization in clinical diagnosis. *Expert Review of Molecular Diagnostics* **5**, 421–429.

Berrar, D., Dubitzky, W. & Granzow, M. (2003). *A practical approach to microarray data analysis.* Springer.

Birkes, D. & Dodge, Y. (1993). *Alternative Methods of Regression.* John Wiley & Sons, Inc.

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, O., Frigessi, A. & Lingjærde, O. C. (2007). Predicting survival from microarray data - a comparative study. *Bioinformatics* **23**, 2080–2087.

Donoho, D. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. Aide-Memoire. Department of Statistics, Stanford University.

Fan, J. & Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* **30**, 74–99.

Fan, J. & Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *Proceedings of the Madrid International Congress of Mathematicians.*

Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G. & Kong, A. (2008). Unsupervised empirical Bayesian multiple testing with external covariates. *Annals of Applied Statistics* **2**, 714–735.

Gelman, A., Carlin, J. B., Stern, H. & Rubin, D. (2004). *Bayesian Data Analysis.* Chapman & Hall/CRC.

Gruber, M. (1998). *Improving Efficiency by Shrinkage.* Marcel Dekker, Inc.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning.* Springer-Verlag.

Hofvind, S.-H., Nygård, J., Olsen, A., Sauer, T., G.B., S., Skjeldestad, F. & S.Ø., T. (2001). Masseundersøkelsen mot livmorhalskreft i norge, evaluering av programmet 1992-98. Tech. Rep. 1-2000, Kreftregisteret.

Holm, R., de Putte, G. V., Suo, Z., Lie, A. K. & Kristensen, G. B. (2008). Expressions of EphA2 and EphrinA-1 in early squamous cell cervical carcinomas and their relation to prognosis. *International Journal of Medical Sciences* **5**, 121–126.

Huang, J., Shuangge, M. & Zhang, C.-H. (2006). Adaptive lasso for sparse high-dimensional regression models. Tech. rep., The department of Statistics and Actuarial Science, The university of Iowa.

Kempkensteffen, C., Hinz, S., Christoph, F., Köllermann, J., Krause, H., Schrader, M., Schostak, M., Miller, K. & Weikert, S. (2007). Expression parameters of the inhibitors of apoptosis cIAP1 and cIAP2 in renal cell carcinomas and their prognostic relevance. *International Journal of Cancer* **120**, 1081–1086.

Kreftregisteret (2009). Masseundersøkelsen mot livmorhalskreft. Webpage, available at: `http://www.kreftregisteret.no/no/Forebyggende/Masseundersokelsen-mot-livmorhalskreft`. Visited in January 2009.

Kuopio, T., Kankaanranta, A., Jalava, P., Kronqvist, P., Kotkansalo, T., Weber, E. & Collan, Y. (1998). Cysteine proteinase inhibitor cystatin A in breast cancer. *Cancer Research* **58**, 432–436.

Lando, M., Holden, M., Bergersen, L. C., Svendsrud, D. H., Stokke, T., Sundfør, K., Glad, I. K., Kristensen, G. B. & Lyng, H. (2009). Gene dosage, expression, and ontology analysis identifies driver genes in the carcinogenesis and chemoradioresistance of cervical cancer. *PLOS Genetics* (submitted).

Lyng, H., Brøvig, R. S., Svensrud, D. H., Holm, R., Kaalhus, O., Knutstad, K., Oksefjell, H., Sundfør, K., Kristensen, G. B. & Stokke, T. (2006). Gene expressions and copy numbers associated with metastatic phenotypes of uterine cervical cancer. *BMC Genomics* **7**, 268. `http://biomedcentral.com/1471-2164/7/268`.

Lyng, H., Lando, M., Brøvig, R. S., Svensrud, D. H., Johansen, M., Galteland, E., Brustugun, O., L.A., M.-Z., Myklebost, O., Kristensen, G., Hovig, E. & Stokke, T. (2008). Genecount: genome-wide calculation from array comparative genomic hybridization data. *Genome Biology* **9**, R86.1–16.

Marubini, E. & Valsecchi, M. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. John Wiley & Sons Ltd.

Nagase, T., Kawata, S., Nakajima, H., Tamura, S., Yamasaki, E., Fukui, K., Yamamoto, K., Miyagawa, J., Matsumura, I., Matsuda, Y. & Matsuzawa, Y. (1999). Effect of farnesyltransferase overexpression on cell growth and transformation. *International Journal of Cancer* **80**, 126–133.

Nygård, S., Borgan, O., Lingjærde, O. C. & Størvold, H. L. (2008). Partial least squares Cox regression for genome-wide data. *Lifetime Data Anal* **14**, 179–195.

Park, M. & Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B* **69**, 659–677.

Park, M., Hastie, T. & Tibshirani, R. (2007). Averaged gene expressions for regression. *Biostatistics* **8**, 212–227.

Pinkel, D. & Albertson, D. (2005). Comparative genomic hybridization. *Annual Review of Genomics and Human Genetics* **6**, 331–354.

Quackenbush, J. (2006). Microarray analysis and tumor classification. *The New England Journal of Medicine* **354**, 2463–2472.

Ruddon, R. (2007). *Cancer biology*. Oxford University Press, 4th ed.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **58**, 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine* **16**, 385–395.

van de Wiel, M. & Wieringen, W. N. V. (2007). CGHregions: Dimension reduction for array CGH data with minimal information loss. *Cancer Informatics* **2**, 55–63.

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van't Veer, L. J. & Wessels, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine* **25**, 3201–3216.

Verweij, P. J. M. & Houwelingen, H. C. V. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine* **13**, 2427–2436.

Verweij, P. J. M. & van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in Medicine* **12**, 2305–2314.

Wieringen, W. N. V., van de Wiel, M. & Ylstra, B. (2007). Weighted clustering of called array CGH data. *Biostatistics* **9**, 484–500.

Xiong, J. (2006). *Essential Bioinformatics*. Cambridge University Press.

Yuan, J., Murrell, G. A. C., Trickett, A., Landtmeters, M., Knoops, B. & Wang, M.-X. (2004). Overexpression of antioxidant enzyme peroxiredoxin 5 protects human tendon cells against apoptosis and loss of cellular function during oxidative stress. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1693**, 37–45.

Zhang, H. H. & Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika* **94**, 691–703.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the America Statistical Association* **101**, 1418–1429.

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

APPENDIX A

Lists of Selected Genes and Gene Ontology

A.1. Lists of Selected Genes

Univariate Cox Gene Expression Data, $p \leq 0.15$				
Probe Identification	Gene Identification	chromosome	cytoBand	$\hat{\beta}$
RP11-34M16	307660	8	8q21	0.0094
Univariate Cox Gene expression data, $p \leq 0.1$				
Probe Identification	Gene Identification	chromosome	cytoBand	$\hat{\beta}$
RP11-506C8	45542	2	2q33-q36	−0.0061
RP11-34M16	307660	8	8q21	0.0717
Univariate Cox Gene expression data, $p \leq 0.05$				
Probe Identification	Gene Identification	chromosome	cytoBand	$\hat{\beta}$
RP11-223L18	200814	3	3q25.1-q25.2	0.0001
RP11-506C8	45542	2	2q33-q36	−0.0056
RP11-34M16	307660	8	8q21	0.0716
Univariate Cox Gene expression data, $p \leq 0.025$				
Probe Identification	Gene Identification	chromosome	cytoBand	$\hat{\beta}$
RP11-99J16	740027	1	1q42.1	0.3716
RP11-343C9	139835	4	4p15.1	−0.0019
RP4-542G16	25k_845519	10	10p15.1	0.1428
RP11-223L18	200814	3	3q25.1-q25.2	0.2022
CTC-307M15	1473471	5	5q33.1	0.2321
RP11-105I12	212198	1	1q42.1	0.1704
RP11-738E22	811842	4	4q11	0.0343
RP11-506C8	45542	2	2q33-q36	−0.0732
RP11-34M16	307660	8	8q21	0.1191
RP11-36J15	120881	18	18p11.3	0.1377
RP11-141E12	25k_45376	18	18q21.1	−0.1863
RP4-669K10	509588	1	1p35.3	−0.0424
RP11-71A24	855624	9	9q21.13	0.2196
Correlation Gene expression data, $r \geq 0$				
Probe Identification	Gene Identification	chromosome	cytoBand	$\hat{\beta}$
RP11-34M16	307660	8	8q21	0.0091
RP11-545E17	741497	9	9q34	0.0005
Correlation Gene expression data, $r \geq 0.4$				
Probe Identification	Gene Identification	chromosome	cytoBand	$\hat{\beta}$
RP11-119E13	25k_309288	3	3q27	−0.2407
RP11-307C12	25k_1474684	1	1q21-q22	0.2511
CTD-2115H11	25k_530359	8	8p22-q11	0.2264
RP11-147G6	292519	11	11q13	0.4289
RP5-1027G4	731073	20	20p11.23	0.0460

TABLE A.1. Genes selected by the lasso after reducing the data by Cox univariate regression and by correlation with aCGH data.

A.2. Lists of Selected Regions

Variance aCGH data, 200 top ranked probes					
Probe Identification	chromosome	$\hat{\beta}$	Probe Identification	chromosome	$\hat{\beta}$
RP11-81P15	3	-0.454	RP1-128M19	21	-0.306
RP11-48M17	9	-0.12	RP11-102E10	21	-0.158
RP11-750P5	11	-0.105	RP1-128M19	21	-0.306
Univariate Cox aCGH data, $p \geq 0.025$					
Probe Identification	chromosome	$\hat{\beta}$	Probe Identification	chromosome	$\hat{\beta}$
RP1-128M19	21	-0.417	RP11-118O11	3	-0.0222
RP11-172O13	7	0.167	RP11-154H23	3	-0.0176
RP11-344B7	18	0.0312	RP11-522N9	3	-0.0176
RP11-39D15	18	0.0307	RP11-252O10	3	-0.0176
RP11-51B9	18	0.0307	RP11-11L10	3	-0.0176
RP11-36J15	18	0.0307	RP11-220O14	3	-0.0176
RP11-99M10	18	0.0307	RP11-16M12	3	-0.0176
RP11-411B10	18	0.0307	RP1-62O9	17	0.109
RP4-635E18	1	-0.0754	RP11-81K2	17	0.109
RP4-539L13	1	-0.0754	RP5-875H18	17	0.109
RP11-196P5	1	-0.0754	RP11-94C24	17	0.109
RP4-636F13	1	-0.0754	RP11-506D12	17	0.101
RP11-738E22	4	0.975	RP11-481C4	17	0.101
RP11-81P15	3	-0.285	RP11-515O17	17	0.101
RP11-45D17	X	-0.465	RP11-436F9	7	0.0375
RP4-783C10	1	-0.2	RP11-556N21	13	-0.0788
RP5-884C9	1	-0.2	RP11-111G7	13	-0.0787
RP11-113D13	1	-0.2	RP11-570F6	13	-0.0787
RP1-118J21	1	-0.2	RP11-44J9	13	-0.0787
RP1-179D3	X	-0.163	RP11-125I23	13	-0.0787
RP3-463A9	X	-0.163	RP11-153M24	13	-0.0787
RP11-24O17	3	-0.0179	RP11-40A8	13	-0.156
RP11-79C12	3	-0.0176	RP11-327P2	13	-0.156
RP11-152N21	3	-0.0176			
Univariate Cox aCGH data, $p \geq 0.05$					
Probe Identification	chromosome	$\hat{\beta}$	Probe Identification	chromosome	$\hat{\beta}$
RP1-128M19	21	-0.272	RP3-463A9	X	-0.0583
RP11-172O13	7	0.00139	RP11-556N21	13	-0.0422
RP11-344B7	18	0.0063	RP11-111G7	13	-0.0419
RP11-39D15	18	0.00679	RP11-570F6	13	-0.0419
RP11-51B9	18	0.00627	RP11-44J9	13	-0.0419
RP11-36J15	18	0.00627	RP11-125I23	13	-0.0419
RP11-99M10	18	0.00627	RP11-153M24	13	-0.0419
RP11-411B10	18	0.00627	RP11-40A8	13	-0.193
RP4-635E18	1	-0.0623	RP11-327P2	13	-0.193
RP4-539L13	1	-0.0623	RP11-108L12	11	0.348
RP11-196P5	1	-0.0623	GS1-77L23	9	-0.35
RP4-636F13	1	-0.0623	RP4-796I11	20	0.0122
RP11-738E22	4	0.704	RP1-232N11	20	0.0122
RP11-102E10	21	-0.0213	RP4-781B1	20	0.0122
RP11-81P15	3	-0.342	RP5-1005L2	20	0.0122
RP11-45D17	X	-0.423	RP5-1049G16	20	0.0122
RP4-783C10	1	-0.0855	RP11-347D21	20	0.0122
RP5-884C9	1	-0.0839	RP1-155G6	20	0.0122
RP11-113D13	1	-0.0839	RP3-470L14	20	0.0122
RP1-118J21	1	-0.0839	RP4-791K14	20	0.0122
RP1-179D3	X	-0.058	RP5-894K16	20	0.0122

TABLE A.2. Probes selected by the lasso when variance and univariate Cox regression is used as selection criterion.

A.3. Gene Ontology; Genes Selected by the Lasso

Univariate Cox Gene Expression data, $p \leq 0.15$		
Gene Identification	Gene Symbol	Gene Ontology
307660	FABP4	Fatty acid binding protein 4, adipocyte
Univariate Cox Gene Expression data, $p \leq 0.1$		
Gene Identification	Gene Symbol	Gene Ontology
45542	IGFBP5	Insulin-like growth factor binding protein 5
307660	FABP4	Fatty acid binding protein 4, adipocyte
Univariate Cox Gene Expression data, $p \leq 0.05$		
Gene Identification	Gene Symbol	Gene Ontology
200814	MME	Membrane metallo-endopeptidase
45542	IGFBP5	Insulin-like growth factor binding protein 5
307660	FABP4	Fatty acid binding protein 4, adipocyte
Univariate Cox Gene Expression data, $p \leq 0.025$		
Gene Identification	Gene Symbol	Gene Ontology
740027	TSNAX	Translin-associated factor X
139835	UGDH	UDP-glucose dehydrogenase
845519	ATP5C1	ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1
200814	MME	Membrane metallo-endopeptidase
1473471	KIAA0194	KIAA0194 protein
212198	TP53BP2	Tumor protein p53 binding protein, 2
811842	SRP72	Signal recognition particle 72kDa
45542	IGFBP5	Insulin-like growth factor binding protein 5
307660	FABP4	Fatty acid binding protein 4, adipocyte
120881	RAB31	RAB31, member RAS oncogene family
45376	MYO5B	Acetyl-Coenzyme A acyltransferase 2
509588	TAF12	TAF12 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 20kDa
855624	ALDH1A1	Aldehyde dehydrogenase 1 family, member A1
Correlation Gene Expression data, $r > 0$		
Gene Identification	Gene Symbol	Gene Ontology
307660	FABP4	Fatty acid binding protein 4, adipocyte
741497	LCN2	Lipocalin 2
Correlation Gene Expression data, $r > 0.4$		
Gene Identification	Gene Symbol	Gene Ontology
309288	RFC4	Replication factor C (activator 1) 4, 37kDa
1474684	EFNA1	Ephrin-A1
530359	FNTA	Farnesyltransferase, CAAX box, alpha
292519	PRDX5	Peroxiredoxin 5
731073	NAT5	N-acetyltransferase 5

TABLE A.3. Gene ontology of the genes selected by the lasso after reducing the data by Cox univariate regression and by correlation with aCGH data.

A.4. Gene Ontology; Genes Selected by Genewise Lasso Penalization

Genewise Penalization Analysis, Penalization Scheme 1		
Gene Identification	Gene Symbol	Gene Ontology
1474684	EFNA1	Ephrin-A1
814508	PPP1R7	Protein phosphatase 1, regulatory (inhibitor) subunit 7
309288	RFC4	Replication factor C (activator 1) 4, 37kDa
530359	FNTA	Farnesyltransferase, CAAX box, alpha
814636	SMARCA2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2
845519	ATP5C1	ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1
292519	PRDX5	Peroxiredoxin 5
1384851	MMP10	Matrix metallopeptidase 10 (stromelysin 2)
Genewise Penalization Analysis, Penalization Scheme 2		
Gene Identification	Gene Symbol	Gene Ontology
345957	CSTA	Cystatin A (stefin A)
148421	TXK	TXK tyrosine kinase
272531	KIAA1211	KIAA1211 protein
273546	PAICS	Phosphoribosylaminoimidazole carboxylase
811842	SRP72	Signal recognition particle 72kDa
814702	SRP72	Signal recognition particle 72kDa
121420	COX18	COX18 cytochrome c oxidase assembly homolog (S. cerevisiae)
179143	ANKRD17	Ankyrin repeat domain 17
282564	RASSF6	Ras association (RalGDS/AF-6) domain family member 6
2315207	CXCL6	Chemokine (C-X-C motif) ligand 6 (granulocyte chemotactic protein 2)
324437	CXCL1	Chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)
324437	CXCL1	Chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)
701417	MTHFD2L	Methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2-like
80915	SDHA	Succinate dehydrogenase complex, subunit A, flavoprotein (Fp)
530359	FNTA	Farnesyltransferase, CAAX box, alpha
809517	MRPL21	Mitochondrial ribosomal protein L21
1384851	MMP10	Matrix metallopeptidase 10 (stromelysin 2)
324492		Transcribed locus
264646	HGS	Hepatocyte growth factor-regulated tyrosine kinase substrate
144797		Transcribed locus
296880	MPP1	Membrane protein, palmitoylated 1, 55kDa
Genewise Penalization Analysis, Penalization Scheme 3		
Gene Identification	Gene Symbol	Gene Ontology
125148	RCL1	RNA terminal phosphate cyclase-like 1
266361	MLANA	Melan-A
248261	GLDC	Glycine dehydrogenase (decarboxylating)
308163	YAP1	Yes-associated protein 1, 65kDa
201890	BIRC3	Baculoviral IAP repeat-containing 3
428231	BIRC3	Baculoviral IAP repeat-containing 3
34852	BIRC2	Baculoviral IAP repeat-containing 2
1384851	MMP10	Matrix metallopeptidase 10 (stromelysin 2)
324492		Transcribed locus

TABLE A.4. Gene ontology of the genes selected by the Genwise lasso penalization procedure.

APPENDIX B

Regularity Conditions

The theory of counting processes is used to describe the regularity conditions needed in Section 6.2. We define $N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$ ad $Y_i(t) = I\{T_i \geq t, C_i \geq t\}$ and allow for the covariates to be time dependent, that is $\boldsymbol{x}(t)$ depending on t. We concentrate on the finite time interval $[0, \tau]$ and assume without loss of generality that $\tau = 1$.

Conditions.

A $\int_0^1 h_0(t)dt < \infty$

B *The processes* $\boldsymbol{x}(t)$ *and Y(t) are left-continuous with right hand limits, and*

$$P\{Y(t) \ \forall \ t \in [0,1]\} > 0.$$

C *There exists a neighborhood* $\mathcal{B}$ *of* β_0 *such that*

$$\sup_{t\in[0,1],\boldsymbol{\beta}\in\mathcal{B}} \{Y(t)\boldsymbol{x}(t)^T\boldsymbol{x}(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x}(t))\} < \infty.$$

D *Define*

$$s^{(0)}(\boldsymbol{\beta}, t) = EY(t)exp(\boldsymbol{\beta}^T\boldsymbol{x}(t))$$
$$s^{(1)}(\boldsymbol{\beta}, t) = EY(t)\boldsymbol{x}(t)exp(\boldsymbol{\beta}^T\boldsymbol{x}(t))$$
$$s^{(2)}(\boldsymbol{\beta}, t) = EY(t)\boldsymbol{x}(t)\boldsymbol{x}(t)^Texp(\boldsymbol{\beta}^T\boldsymbol{x}(t)),$$

where $s^{(0)}(\cdot, t)$, $s^{(1)}(\cdot, t)$ *and* $s^{(2)}(\cdot, t)$ *are continuous in* $\boldsymbol{\beta} \in \mathcal{B}$, *uniformly in* $t \in [0, 1]$. $s^{(0)}, s^{(1)}$ *and* $s^{(2)}$ *are bounded on* $\mathcal{B} \times [0, 1]$; $s^{(0)}$ *is bounded away from zero on* $\mathcal{B} \times [0, 1]$. *The matrix*

$$I(\boldsymbol{\beta}_0) = \int_0^1 v(\boldsymbol{\beta}_0, t)s^{(0)}(\boldsymbol{\beta}_0, t)h_0(t)dt$$

is finite positive definite where

$$v(\boldsymbol{\beta}, t) = \frac{s^{(2)}}{s^{(0)}} - \frac{s^{(1)}}{s^{(0)}}\frac{s^{(1)\,T}}{s^{(0)}} \ .$$

Conditions A-D entail that the local asymptotic quadratic property for the partial likelihood is guaranteed, which implies asymptotic normality of the maximum partial likelihood estimates.

APPENDIX C

R-Scripts

All R-scripts are available at: `http://folk.uio.no/linncb/R-scripts/`.
The R-script performing genewise lasso penalization is a modified version of `computeTuningParameterLasso()` of Bøvelstad et al. (2007). The source code of Bøvelstad et al. (2007) can be found on `http://www.med.uio.no/imb/stat/bmms/software/microsurv/`. The script `genewiseLassoPenalization()` performs the two-dimensional cross-validation for two tuningparameters q and λ.

```
#=============================================================================#
genewiseLassoPenalization <- function(datafull, weighttype, response){
library(glmpath)
#initial parameters set by the default values in the program of Bøvelstad et al. (2007)
K = 10
M = 1000
maxsteps = 1000
n = nrow(datafull)
p = ncol(datafull)
blockSize = n/K
lambda.max = 10^6
log.lambda.max = log(lambda.max)
log.lambda.min = -100
log.lambda.vec = seq(log.lambda.min, log.lambda.max, length.out=M)
lambda.vec = exp(log.lambda.vec)
timefull = as.numeric(response[,1])
statusfull = as.numeric(response[,2])
loglik = rep(0,M)

qgrid <- seq(0,10, 0.25)
CVlambda <- rep(0,length(qgrid))
CVbeta <- rep(0, length(qgrid))
CVcurve <- matrix(0,nrow = length(qgrid),ncol = length(loglik))

# Two-dimensional Cross-Validation of q and lambda
for(j in 1:length(qgrid)){
  q <- qgrid[j]
  loglik = rep(0,M)
  print(paste("q=", q))
  source("defineWeights.R")
  w <- defineWeights(weighttype)
  wj <- abs(w)^q
  weighted.genedata <- matrix(0, ncol=p,nrow=n)
  for(i in 1:length(wj)){
    weighted.genedata[,i] <- datafull[,i]/(abs(wj[i]))
  }
  print(dim(weighted.genedata))
  Xweighted <- matrix(as.numeric(as.matrix(weighted.genedata)),
                      ncol = ncol(weighted.genedata), byrow = F)
  print(dim(Xweighted))
  Z.all = list(x=Xweighted,time=timefull,status=statusfull)
```

```
  rm(Xweighted)

   for(k in 1:K){
        print(paste("k=",k))
        test = ((k-1)*blockSize+1):(k*blockSize);
        train = (1:n)[-test]
        data.train = weighted.genedata[train,]
        time = as.numeric(response[train,1])
        status = as.numeric(response[train,2])
        Z = list(x=data.train,time=time,status=status)
        rm(data.train,time,status)
        fit = coxpath(Z,standardize=FALSE,max.steps=maxsteps)
        l = which(lambda.vec<=max(fit$lambda) & lambda.vec>=min(fit$lambda))
        lmin = min(l);
        lmax = max(l);
        if (lmin>1) loglik[1:(lmin-1)] = NA
        loglik.train = predict.coxpath(fit,Z,s=lambda.vec[l],type="loglik",mode="lambda")
        loglik.all = predict.coxpath(fit,Z.all,s=lambda.vec[l],type="loglik",mode="lambda")
        loglik[l] = loglik[l]+loglik.all-loglik.train
        loglik[(lmax+1):M] = loglik[(lmax+1):M]+loglik.all[lmax-lmin+1]-loglik.train[lmax-lmin+1]
   }
rm(fit,Z)

#fit final model
fit.all = coxpath(Z.all,standardize=FALSE,max.steps=maxsteps)
lambda = min(lambda.vec[which.max(loglik)],max(fit.all$lambda))

#results
opt.beta=predict.coxpath(fit.all,Z.all,s=lambda,type = "coef",mode="lambda")
beta = t(as.vector(opt.beta))
CVlambda[j] <- lambda
CVcurve[j,] <- loglik
CVbeta[j] <- length(which(beta!=0))

#write results to file
file1 = paste(weighttype, "/ResultaterCrossVal_", sep = "")
filename = paste(file1, j, sep="")
save(lambda,beta,CVlambda, CVcurve,CVbeta, file = paste(filename,".RData", sep = ""))
rm(loglik, fit.all,Z.all)
}
}



#=============================================================================#
defineWeights<-function(method){
if(method == "Spearman"){
        print(paste("Weighttype: ", method))
        load("SpearmanWeights.RData")
        wj <- 1/gamma
}else if(method == "StandardDeviation"){
        print(paste("Weighttype: ", method))
        load("SDWeights.RData")
        wj <- 1/gamma
}else if(method == "Ridge"){
        print(paste("Weighttype: ", method))
        load("RidgeWeights.RData")
        wj <- 1/gamma
}
return(wj)
}
#=============================================================================#
```

In order to run the genewise lasso penalization procedure on the data, the two data sources have to be matched, such that the variables are given the correct weights in the penalization. The script `MATCHING.R` below, illustrates how this was done on the data studied in this thesis.

```
#=============================================================================#
#MATCHING.R

#read in data:

#response
response <- read.table("Respons_genekspr.txt", dec = ",", header = T)
responseCopy <- read.table("Respons_kopitall.txt", dec = ",")


#files containing the matching of the gene identifications and probe identifications
match <- read.csv2("cDNA_m_genomiskID.csv", stringsAsFactors=F)
matchGENE <- match[,2]
matchCOPYNUMBER <- match[,1]

#Informationvector for copy number data
#(probe identifications indexed as the ordering in the datafile)
probeInfo <- read.csv("probeInfo.cvs", stringsAsFactors = F)
probeID <- probeInfo[,1]

#Information vector for genedata
#(gene identifications indexed as the ordering in the datafile)
geneInfo <- read.csv("probeInfo_genekspr.cvs", stringsAsFactors = F)
geneID <- geneInfo[,1]

#the data are contained in two separate files;
# ``genedata.csv'' and ``Kopitalldata.csv''

#genedata
geneexpr <- read.csv("genekspr_imp.csv")
genedata <- t(geneexpr)

#copy number data
copin <- read.csv("Kopitalldata.csv")
copydata <- t(copin)

#in some settings we need to remove the patients
#which are not included in both datasets
#(for example when correlations between the data are calculated)
nGene <- 1:nrow(genedata)
rmGene <- c(nGene[rownames(genedata)== "MM051"],nGene[rownames(genedata)== "MM058"],
nGene[rownames(genedata)== "MM067"], nGene[rownames(genedata)== "MM085"],
nGene[rownames(genedata)== "MM155"])

nCopy <- 1:nrow(copydata)
rmCopy <- c(nCopy[rownames(copydata) == "MM016"],nCopy[rownames(copydata) == "MM045"])

#Matching of the two data sources;

#find the genes and copy numbers for which we have
#measurements from both data sources available.
index <- rep(0,length(matchCOPYNUMBER))
for(i in 1:length(probeID)){
  index[matchCOPYNUMBER == probeID[i]]= i
}

index2 <- rep(0, length(matchGENE))
for(i in 1:length(geneID)){
```

```
  index2[matchGENE == geneID[i]] = i
}

#The macthing (geneID, probeID, chrom, cytoband) is kept in the matrix MATCH
check <-rep(0,length(index))
for( i in 1:length(index)){
  check[i] <- index[i] !=0 && index2[i] != 0
}
MATCH <- match[check==1, 1:4]
MATCH

#select the correct measurements and make gene data sets, ready for analysis
Indexgenes <- rep(0, nrow(MATCH))
for(i in 1:length(geneID)){
  Indexgenes[MATCH[,2]==geneID[i]] = i
}
gendata2 <- genedata[-rmGene,Indexgenes]  #for correlation
rmrespons <- response[-rmGene]
gendata3 <- genedata[,Indexgenes] #for regression analysis


#select the correct measurements and make copy number data sets, ready for analysis
Indexcopy <- rep(0, nrow(MATCH))
for(i in 1:length(probeID)){
  Indexcopy[MATCH[,1]==probeID[i]] = i
}
copydata2 <- copydata[-rmCopy,Indexcopy] #for correlation
copydata3 <- copydata[,Indexcopy] #for determinations of weights
#=============================================================================#
```