Dissertation

Survival Models with Gene Groups as Covariates

Kai Kammers

Submitted to the Department of Statistics
of the TU Dortmund University

in Fulfillment of the Requirements for the Degree of
Doktor der Naturwissenschaften

Dortmund, May 2012

Referees:

Prof. Dr. Jörg Rahnenführer
Prof. Dr. Katja Ickstadt

Date of Oral Examination: June 26, 2012

Contents

Abstract

An important application of high-dimensional gene expression measurements is the risk prediction and the interpretation of the variables in the resulting survival models. A major problem in this context is the typically large number of genes compared to the number of observations (individuals). Feature selection procedures can generate predictive models with high prediction accuracy and at the same time low model complexity. However, interpretability of the resulting models is still limited due to little knowledge on many of the remaining selected genes. Thus, we summarize genes as gene groups defined by the hierarchically structured Gene Ontology (GO) and include these gene groups as covariates in the hazard regression models. Since expression profiles within GO groups are often heterogeneous, we present a new method to obtain subgroups with coherent patterns. We apply preclustering to genes within GO groups according to the correlation of their gene expression measurements.

We compare Cox models for modeling disease free survival times of breast cancer patients. Besides classical clinical covariates we consider genes, GO groups and preclustered GO groups as additional genomic covariates. Survival models with preclustered gene groups as covariates have improved prediction accuracy in long term survival compared to models built only with single genes or GO groups. We also provide an analysis of frequently chosen covariates and comparisons to models using only clinical information.

The preclustering information enables a more detailed analysis of the biological meaning of covariates selected in the final models. Compared to models built only with single genes there is additional functional information contained in the GO annotation, and compared to models using GO groups as covariates the preclustering yields coherent representative gene expression profiles. For evaluation of fitted survival models, we present prediction error curves revealing that models with preclustered gene groups have improved prediction performance compared to models built with single genes or GO groups.

1 Introduction

Almost 10 % of German women will develop invasive breast cancer over the course of their lifetime. In 2004, approximately 57 000 new cases of breast cancer were diagnosed in women in Germany (Robert Koch Institut, 2010). Statistics for other countries, especially for the United States, are comparable (Ma and Huang, 2007). Despite major progresses in breast cancer treatment, the ability to predict the metastatic behavior of tumor remains limited.

In addition to well-known risk factors (cf. Robert Koch Institut, 2010) like drinking alcohol, getting older, being overweight (increases risk for breast cancer after menopause) and not getting regular exercise, specific genes may have an influence on developing cancer and on patient's survival times. According to Giersiepen *et al.* (2005) about 5 to 10 % of breast cancers can be linked to gene mutations (abnormal changes) inherited from one's mother or father. Mutations of the BRCA1 and BRCA2 genes are the most common. Women with these mutations have up to an 80 % risk of developing breast cancer during their lifetime, and they are more likely to be diagnosed at a younger age (before menopause).

In the last 10 years cancer research has focused on gene expression experiments to detect genes that are responsible for the development of cancer and for its course over time. Thus, the prediction of cancer patient survival based on gene expression profiles is an important application of genome-wide expression data (Rosenwald *et al.*, 2002; van de Vijver *et al.*, 2002; van 't Veer *et al.*, 2002).

In this thesis, our goal is to improve prediction accuracy and interpretability of high-dimensional models for prediction of survival outcomes, by combining gene expression data with prior biological knowledge on groups of genes.

Since 2001 numerous developments have appeared to outcome prediction for several kinds of cancer on the basis of gene expression experiments (Alizadeh *et al.*, 2000; Bair and Tibshirani, 2004; Beer *et al.*, 2002; Khan *et al.*, 2001; Rosenwald *et al.*, 2002; Ramaswamy *et al.*, 2003), with special focus on breast carcinoma (Sørlie *et al.*, 2001; Rosenwald *et al.*, 2002; van de Vijver *et al.*, 2002; van 't Veer *et al.*, 2002; Wang *et al.*, 2005; West *et al.*, 2001). Several of these studies reported considerable predictive success. They allow the discovery of new markers that open the way to more subject-specific treatments with greater efficacy and safety.

Clinical covariates like age, gender, blood pressure, tumor size and grade, as well as smoking and drinking history have been extensively used and shown to have satisfactory predictive power. They are usually easy to measure and of low dimensionality.

By uncovering the relationship between time to event and the tumor gene expression profile, it is hoped to achieve more accurate prognoses and improved treatment strategies. Predicting the prognosis and metastatic potential of cancer at the time of discovery is a major challenge in current clinical research. Numerous recent studies searched for gene expression signatures that outperform traditionally used clinical parameters in outcome prediction (see e.g. Binder and Schumacher, 2008b). A substantial challenge in this context comes from the fact that the number of genomic variables p is usually much larger than the number of individuals n. The goal is to construct models that are complex enough to have high prediction accuracy but that are at the same time simple enough to allow biological interpretation. It is very difficult to select the most powerful genomic variables for prediction, as these may depend on each other in an unknown fashion.

Univariate approaches use single genes as covariates in survival time models, whereas multivariate models need a more elaborate framework. For statistical analysis, the Cox regression model (Cox, 1972) is a well-known method for modeling censored survival data. It can be used for identifying covariates that are significantly correlated with survival times. Due to the high-dimensional nature of microarray data we cannot obtain the parameter estimates directly with the Cox log partial likelihood approach. Techniques have been developed that result in shrunken and/or sparse models, i.e., models where only a small number of covariates is used. The classical ridge-regression (Hoerl and Kennard, 1970) and lasso-regression (Tibshirani, 1996, 1997) are particularly suitable. Efron *et al.* (2004) proposed a highly efficient procedure, called Least Angle Regression (LARS) for variable selection which can be used to perform variable selection with very large matrices. LARS can be modified to provide a solution for the lasso-procedure. Using the connection between LARS and lasso, Gui and Li (2005) proposed LARS-Cox for gene selection in high-dimension and low-sample settings. In this case and in boosting approaches (Bühlmann and Hothorn, 2007), it is avoided to discard covariates before model fitting. Parameter estimation and selection of covariates is performed simultaneously. This is implemented by imposing a penalty on the model parameters for estimation. The structure of this penalty is chosen such that most of the estimated parameters will be equal to zero, i.e., the value of the corresponding covariates does not influence predictions obtained from the fitted model. Schumacher *et al.* (2007) developed techniques for extending a bootstrap approach for estimating prediction error curves (introduced by Gerds and Schumacher, 2007) to high-dimensional gene expression data with survival outcome.

An alternative method was developed by Binder and Schumacher (2008b). They proposed a boosting approach for high-dimensional Cox models (*CoxBoost*). The resulting model is sparse and thus it competes directly with the results from lasso-regression. Binder and Schumacher (2008b) applied the CoxBoost algorithm to gene expression data sets.

In addition, the combination of clinical data and gene expression data is a hot topic of research (cf. Boulesteix *et al.*, 2008; Binder and Schumacher, 2008b). In order to integrate the clinical information and microarray data in survival models properly, it is a common approach to handle the clinical covariates as unpenalized mandatory variables (cf. Binder and Schumacher, 2008b; Bøvelstad *et al.*, 2009). These approaches show that the combination of genomic and clinical information may also improve predictions.

For evaluating a fitted survival model, patients are often divided into subgroups according to their prognoses. Kaplan-Meier curves (Kaplan and Meier, 1958) are calculated for each group and compared with the log-rank test (see, e.g. Rosenwald *et al.*, 2002). It is important that a comparison of groups is performed without the individuals used for model fitting. Otherwise, results would be overoptimistic.

Graf *et al.* (1999) presented an adapted framework for the Brier-Score (Brier, 1950) for survival models with censored observations. At each time point, we compare the estimated probability of being event-free to the observed event-status. Analysis of the prognostic index (Bøvelstad *et al.*, 2007) and the Brier-Score (Graf *et al.*, 1999; Schumacher *et al.*, 2007) can be used to assess the predictive performance of the fitted models.

On the other hand, Haibe-Kains *et al.* (2008) showed in a comparative study of survival models for breast cancer prognostication based on microarray data that the most complex methods are not significantly better than the simplest one, a univariate model relying on a single proliferation gene. This result suggests that proliferation might be the most relevant biological process for breast cancer prognostication and that the loss of interpretability deriving from the use of overcomplex methods may be not sufficiently counterbalanced by an improvement of the quality of prediction of those who really need chemotherapy and benefit from it.

However, due to the large variability in survival times between cancer patients and the amount of genes on the microarrays unrelated to outcome, building accurate prediction models that are easy to interpret remains a challenge. In this thesis, we propose a new approach for improving performance and interpretability of prediction models by integrating gene expression data with prior biological knowledge. To raise the interpretability of prognostic models, we combine genes to gene groups (e.g. according to their biological processes) and use these groups as covariates in the survival models. The hierarchically ordered 'GO groups' (Gene Ontology) are particularly suitable (Ashburner *et al.*, 2000). The Gene Ontology (GO) project provides structured, controlled vocabularies and classifications according to molecular and cellular biology. Gene expression data can be analyzed by summarizing groups of individual gene expression profiles based on GO annotation information. The mean expression profile per group or the first principle component can then be used to identify interesting GO categories in relation to the experimental settings. Another platform is the KEGG data base (Kanehisa *et al.*, 2004) that provides biological pathway information for genes and proteins.

A problem when relating genes to groups is that the genes in each gene group may have different expression profiles: interesting subgroups may not be detected due to heterogeneous or anti-correlated expression profiles within one gene group. We propose to cluster the expression profiles of genes in every gene group to detect homogeneous subclasses within a GO group and preselect relevant clusters (preclustering). The Intra Cluster Correlation (ICC), a measure of cluster tightness, is applied to identify relevant clusters.

In a first step we compared high-dimensional survival models with genes and GO groups as covariates for different variable selection methods (Kammers and Rahnenführer, 2010). Based on this work Lang (2010) constructed in his diploma thesis different aggregation methods for genes within a gene group for high-dimensional data sets. It turned out that summarizing the expression values with the first principle component is the most promising aggregation method. This result is integrated in this thesis with an extensive model building and evaluation procedure. We show comparisons to clinical models and to combinations of genomic and clinical models with different types of evaluation measures as well as the analysis of frequently chosen covariates.

The main results of this thesis are already published in the peer-reviewed journal *BMC Bioinformatics*.

This thesis is organized as follows: Chapter 2 provides the biological background and introduces the data sets. Chapter 3 and 4 describe the statistical methodology for building prognostic models for high-dimensional data sets. Chapter 3 presents methods for summarizing gene expression measurements with a focus on *preclustering* and Chapter 4 introduces the survival framework including the Cox model for high-dimensional data and algorithms for fitting and evaluating it. Chapter 5 shows the main results for two breast cancer data sets. Chapter 6 discusses proper ways for placing the results within the context of recent studies and possibilities for extensions. Concluding remarks are also given in this chapter.

2 Background and Data Sets

In Section 2.1 we present the biological and medical background of breast cancer including risk factors, types of therapy and relation to recent statistical research. In Section 2.2 and Section 2.3 we briefly introduce the microarray technology for gene expression experiments and the Gene Ontology database that provide supplementary information for many genes. Finally, in Section 2.4 we present two well-known breast cancer data sets that are used for all analysis steps in Chapter 5.

2.1 Breast Cancer

Breast cancer is a malignant tumor that starts in the cells of the breast. A malignant tumor is a group of cancer cells that can grow into (invade) surrounding tissues or spread (metastasize) to distant areas of the body. Usually breast cancer either begins in the cells of the lobules, which are the milk-producing glands, or the ducts, the passages that drain milk from the lobules to the nipple (see, e.g. Sariego, 2010). The disease occurs mostly in women, but men can be affected, too.

The size, stage, rate of growth, and other characteristics of the tumor determine the kinds of treatment. It may include surgery, drugs (hormonal therapy and chemotherapy), radiation and/or immunotherapy (Florescu *et al.*, 2011). Surgical removal of the tumor provides the single largest benefit, with surgery alone being capable of producing a cure in many cases. To increase the disease-free survival, several chemotherapy regimens are commonly given in addition to surgery. Radiation is indicated especially after breast conserving surgery and substantially improves local relapse rates and in many circumstances also overall survival (Buchholz, 2009). Some breast cancers are sensitive to hormones such as estrogen and/or progesterone, which makes it possible to treat them by blocking the effects of these hormones.

Research has found several risk factors that may increase the chances of getting breast cancer. According to Robert Koch Institut (2010) an excerpt of risk factors is given by never giving birth, personal history of breast cancer or some non-cancerous breast diseases, family history of breast cancer (mother, sister, daughter), treatment with radiation therapy to the breast/chest, starting menopause at a later age, long-term use of hormone replacement therapy (estrogen and progesterone combined), drinking alcohol and getting older, being overweight (increases risk for breast cancer after menopause) and not getting regular exercise. In addition to these 'clinical' factors, changes in the breast cancer-related genes BRCA1 or BRCA2 may also influence the susceptibility to breast cancer.

For lowering the risk of breast cancer the U.S. Preventive Services Task Force (2009) as well as Boyle and Levin (2008) suggest to be screened for breast cancer regularly and control the risk factors if possible.

Boyle and Levin (2008) point out that worldwide, breast cancer comprises 22.9 % of all cancers (excluding non-melanoma skin cancers) in women. In 2008, breast cancer caused approximately 450 000 deaths worldwide corresponding to 13.7 % of cancer deaths in women. Breast cancer is more than 100 times more common in women than breast cancer in men, although males tend to have poorer outcomes due to delays in diagnosis. Prognosis and survival rate vary greatly depending on cancer type, staging and treatment.

In the article of 2001, Cooper reviews the ways in which microarray technology (see Section 2.2) may be used in breast cancer research (Cooper, 2001). Today, the analysis of gene expression profiles of breast cancer patients and the combination of clinical and gene expression data is a hot topic of research and is very important for risk prediction in survival models (see, e.g., Bøvelstad *et al.*, 2007, 2009; Boulesteix *et al.*, 2008; Binder and Schumacher, 2008b; Binder *et al.*, 2011).

2.2 Microarray Technology

Microarray technology represents a powerful functional genomics technology, which permits the expression profiling of thousands of genes in parallel (Schena *et al.*, 1996) and is based on hybridisation of complementary nucleotide strands (DNA or RNA). Microarray chips consist of thousands of DNA molecules (corresponding to different genes) that are immobilized and girded onto a support such as glass, silicon or nylon membrane. Each spot on the chip is representative for a certain gene or transcript. The expression levels of all genes can be determined by isolating the total amount of mRNA, which is defined to be the transcriptome of the cell at the given time. Fluorescently or radioactively labeled nucleotides (targets) that are complementary to the isolated mRNA are prepared and hybridized to the immobilized molecules. Target molecules that did not bind to the immobilized molecules (probes) during the hybridization process are washed away. The amount of hybridized target molecules is proportional to the amount of isolated mRNA. The relative abundance of hybridized molecules on a defined array spot can be determined by measuring the fluorescent or radioactive signal.

Different types of DNA arrays are designed for mRNA profiling. These types differ by the type of probes that are immobilized on the chip (cDNA or synthetic oligonucleotides) and by the density (probes per square centimetre) of the array. The two basic microarrays variants are *probe cDNA* (0.2 to 5 kb long) that is immobilized to a solid surface using robot spotting and *synthetic DNA fragments* (oligonucleotides, 20 to 80mer long) that are synthesized on-chip (Gene Chip, Affymetrix) or by conventional synthesis followed by on-chip immobilization. This high-density microarray type can carry up to 40 probes per transcript. Half of the probes are designed to perfectly match the nucleotide stretches of the gene, while the other half contains a mismatch (a faulty nucleotide) as a control to test for specificity of the hybridisation signal. Other microarray platforms, such as Illumina, use microscopic beads, instead of the large solid support.

2.3 Gene Ontology Data Base

The Gene Ontology (GO) project is an international bioinformatics initiative to unify the representation of gene and gene product attributes across all species (Ashburner *et al.*, 2000). This project provides a set of structured, controlled vocabularies for community use in annotating genes, gene products and sequences that are available from the GO web site `http://www.geneontology.org`. The ontologies have been extended and refined for several biological areas, and improvements to the structure of the ontologies have been implemented. The current ontologies of the GO project are *biological process*, *molecular function*, and *cellular component.*

GO has a hierarchical structure that forms a directed acyclic graph (DAG). For such a graph we can use the notions of child and parent, where a child can have multiple parents. Every GO term (GO group) is represented by a node in this graph. The nodes are annotated with a set of genes. For an inner node of the GO graph, the corresponding set of genes also comprises all genes annotated to all children of this node. Figure 2.1 represents a part of this graph for the biological process ontology. The arrowhead indicates the direction of the relationship. Child nodes can have different relations to its parents note or its parents notes: a node may have a *part of* relationship to one node or an *is a* relationship to another. The biological process ontology includes terms that represent collections of processes as well as terms that represent a specific, entire process. The former mainly have *is a* relationships to their children, and the latter mainly have *part of* children that represent subprocesses.

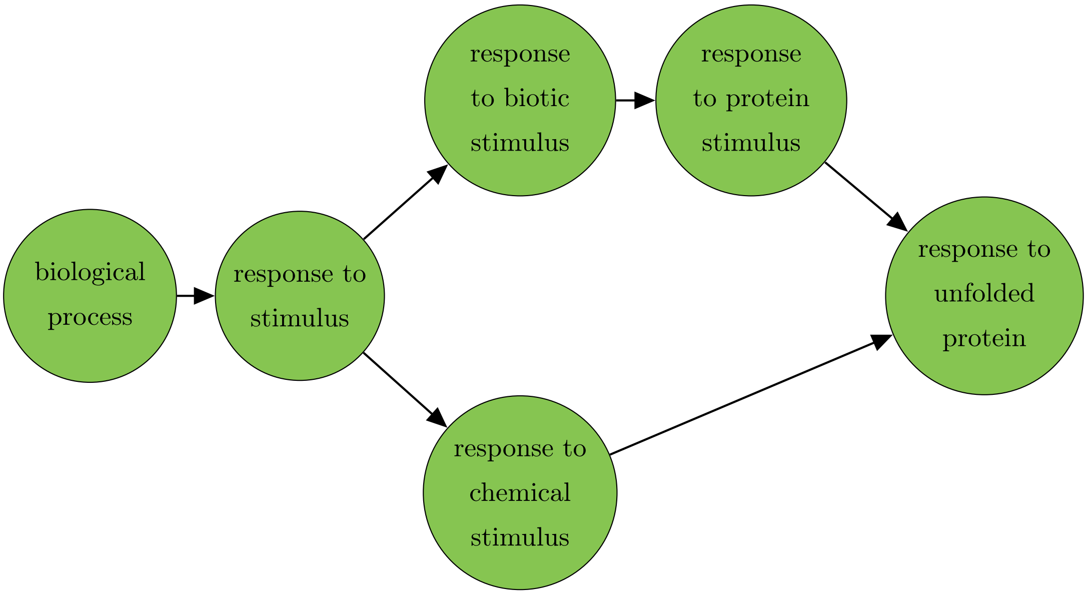
biological process
response to stimulus
response to biotic stimulus
response to protein stimulus
response to chemical stimulus
response to unfolded protein

Figure 2.1: Exemplary part of the directed acyclic graph from Gene Ontology database.

2.4 Description of Data Sets

In this section we introduce two well-known breast cancer data sets that we use for the entire analysis (see Chapter 5): the **Dutch breast cancer (DBC) data set** and the **Mainz cohort (MC) study**. Both consist of genomic and clinical information as well as survival times and are therefore particularly suitable. The DBC data set is analyzed in several publications (see, e.g. Bøvelstad *et al.*, 2007, 2009; Porzelius *et al.*, 2009; van Wieringen *et al.*, 2009) with models for time-to-event endpoints and thus our results are easy to compare to the already published ones. The MC study is a data set that is extensively used in our working group in cooperation with the Leibniz Research Centre for Working Environment and Human Factors (IfADo) at TU Dortmund University. The complete study data is well-known from its surgical origin to final data matrices. In addition, we want to highlight that there is no missing data in both data sets.

The **Dutch breast cancer (DBC) data set** is a subset of the original data set from the fresh-frozen-tissue bank of the Netherlands Cancer Institute with 24 885 gene expression measurements from $n = 295$ women with breast cancer. According to van de Vijver *et al.* (2002) and van Houwelingen *et al.* (2006) the following selection criteria were employed:

- the tumour was a primary invasive breast carcinoma of size less than 5cm in diameter,
- the age at diagnosis was 52 years or younger,
- the diagnosis was between 1984 and 1995, and
- there was no previous history of cancer, except non-melanoma skin cancer.

All patients had been treated by modified radical mastectomy or breast conserving surgery, including dissection of the axillary lymph nodes, followed by radiotherapy if indicated. Among the 295 patients, 151 were lymph node negative and 144 were lymph node positive. All tumors were profiled on cDNA arrays containing 24 885 genes. After data pre-processing as proposed by van Houwelingen *et al.* (2006) the data set was reduced to a set of 4 919 genes. The data, including gene expression measurements, clinical information and survival data for each patient, was obtained from the website `https://www.msbi.nl/dnn/People/Houwelingen.aspx`. Working with this reduced data set makes it easier to compare our results with previous publications (see, e.g. Bøvelstad *et al.*, 2007, 2009). Our analysis is performed with only 1 876 genes, that are annotated to at least one GO group, according to the *biological process* ontology. In total, there are 5 560 GO groups to which at least one of these genes is annotated. The mean number of genes included in these GO groups is approximately 17 genes where 90% of all GO groups contain at most 30 genes. For 79 patients an event was observed. The clinical covariates are age, size, nodes and grade.

The **Mainz cohort (MC) study** consists of $n = 200$ node-negative breast cancer patients who were treated at the Department of Obstetrics and Gynecology of the Johannes Gutenberg University Mainz between the years 1988 and 1998 (Schmidt *et al.*, 2008). All patients underwent surgery and did not receive any systemic therapy in the adjuvant setting. Gene expression profiling of the patients' RNA was performed using the Affymetrix HG-U133A array, containing 22 283 probe sets, and the GeneChip System. These probe sets are identifiers for approximately 14 500 well-characterized human genes that can be used to explore human biology and disease processes. The normalization of the raw data was done using RMA from the Bioconductor package `affy`. The raw .cel files are deposited at the NCBI GEO data repository with accession number GSE11121. For covariates in the survival models, 17 834 probe sets and 8 587 GO groups are available. The mean number of genes included in these GO groups is approximately 102 probe sets where 90 % of all GO groups contain at most 146 probe sets and the number of observed events is 47. The clinical data covers age at diagnosis, tumor size and grade as well as the estrogen receptor status.

Probe sets can code for the same gene. We will not differentiate between *probe sets* and *genes* in the following and we call them genes.

3 Methods for Preclustering

In this chapter we present methods that search for correlated covariates and aggregation procedures that summarize these covariates. Partitioning around Medoids (PAM) is a clustering method that groups positively correlated variables according to their correlation matrix. This approach is introduced in Section 3.1 and followed by a permutation test for correlation in Section 3.2. This test is used for preselecting covariates within PAM-clustering and as an alternative for clustering when only two covariates are present. Further, we introduce simple aggregation methods for summarizing covariate information in Section 3.3. Finally, we present the principal component analysis (PCA) for multivariate dimension reduction in Section 3.4.

3.1 Partitioning Around Medoids Clustering

Partitioning around Medoids (PAM) (cf. Kaufman and Rousseeuw, 1995) is a clustering algorithm related to the K-means algorithm. Both algorithms are partitional (dividing the dataset into groups) and both attempt to minimize the squared error, the sum of squared distances of all data points to their respective cluster centers. PAM is more robust to noise and outliers compared to K-means because it minimizes the sum of pairwise dissimilarities instead of the sum of squared Euclidean distances.

Let X be a data matrix with n observations and p covariates. The PAM procedure is based on the search for K representative objects, the medoids ($\{m_1, \dots, m_K\} \subset \{1, \dots, p\}$), whose sum of average dissimilarity to all objects in the cluster is minimal. Here, the objects are the covariates that should be clustered.

In order to find correlated subgroups, we first calculate the dissimilarity

$$d_{ij} = 1 - \mathrm{Cor}(x_i, x_j)$$

for all pairs (i, j) with values x_i and x_j $(i, j = 1, \ldots, p)$. In matrix notation, the dissimilarity matrix D, given by

$$D = 1 - \mathrm{Cor}\,(X) \ \in \mathbb{R}^{p \times p},$$

is generated by the data matrix X $(X \in \mathbb{R}^{n \times p})$. The distance between two column vectors is calculated via their correlation coefficient: if two column vectors are highly positive correlated, their distance is close to zero, if they are uncorrelated, their distance is one, if they are highly negative correlated, their distance is two. The correlation coefficient is calculated with the empirical Pearson's correlation coefficient which is defined for two objects x and y with n values by

$$\mathrm{Cor}\,(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Let C $(C(i) : \mathbb{N}_p \to \mathbb{N}_K)$ be a configuration function. The grouping of all p objects in K homogeneous clusters is done with the function C that maps the object i $(i = 1, \ldots, p)$ to cluster k $(k \in 1, \ldots, K)$ according to the smallest dissimilarity to the medoids

$$C(i) = \underset{1 \le k \le K}{\mathrm{argmin}}\, d_{im_k}. \tag{3.1}$$

To evaluate a solution with given medoids and the resulting configuration of all objects, we calculate the sum of intra-cluster-dissimilarities (often referred to as *cost-function*):

$$H(m_1, \ldots, m_K) = \sum_{k=1}^{K} \sum_{C(i)=k} d_{im_k}. \tag{3.2}$$

The PAM optimization problem is the minimization of equation (3.2) regarding to the choice of the medoids and considering the mapping rule (3.1).

The two following steps are carried out iteratively until convergence, starting with K randomly selected objects as initial solution $(m_1, \ldots, m_K)$:

1. `Build`: Select sequentially K initial clusters and assign each gene to its closest medoid.
2. `Swap`: For each medoid m_k and all non-medoid objects θ, swap m_k and θ, then calculate the homogeneity measure given in equation (3.2) and finally select the configuration of medoids with the lowest cost.

Note that we have to test $(K \cdot (p - K))$ swaps in each `Swap` step.

The number of clusters K for the PAM algorithm has to be chosen in advance. There are several techniques that provide adequate measures or graphical representation of how well each object lies within its cluster, e.g. the average silhouette width (cf. Rousseeuw, 1987). To find tight clusters of highly correlated objects, De Haan *et al.* (2010) suggest using the Intra Cluster Correlation. For each possible number of clusters K, the normalized sums of all elements of the lower triangle of the correlation matrix are computed for each cluster. Afterwards, the arithmetic mean over all clusters is calculated. Thus, the optimal number of clusters K_{ICC} according to this method is given by

$$K_{\text{ICC}} = \underset{K=1,\ldots,n-1}{\operatorname{argmax}} \left[\frac{1}{K} \sum_{k=1}^{K} \left(\frac{2}{n_k(n_k - 1)} \sum_{\substack{i_k, j_k = 1 \\ j_k > i_k}}^{n_k} \operatorname{Cor}\left(x_{i_k}, x_{j_k}\right) \right) \right],$$

with object indices i_k and j_k containing n_k single objects in cluster k. If the term in the round brackets consists of only one object, the term is zero by definition. The maximum mean K_{ICC} among the $K = 2, \ldots, (N - 1)$ possible cluster configurations corresponds to the optimal number of clusters within the given data set. Alternative calculations for the optimal number of clusters are e.g. provided in Hastie *et al.* (2009) and in Kaufman and Rousseeuw (1995).

3.2 Permutation Test for Correlation

A permutation test is a statistical test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels of the observed data points. Let Ψ be a $(n \times n)$ permutation matrix that has exactly one entry *one* in each row and each column and *zeros* elsewhere. Each such matrix represents a specific permutation of n elements and, when used to multiply with a vector $x \in \mathbb{R}^n$, can produce that permutation within the vector. For each of the permuted vectors, we can calculate the test statistic and the ranking of the observed test statistic among the permuted test statistics provides a p-value.

In our context, we look at a group with p objects (e.g. gene expression measurements) and investigate if object i has a positive correlation with at least one object of the group. We can formulate the following hypotheses:

$$\begin{aligned} &H_0\colon i\text{th object has no correlation with another object} \\ &H_1\colon i\text{th object has a correlation with at least one other object.} \end{aligned}$$

The maximum correlation $C_0 \in \mathbb{R}$ between object i and all other $(p-1)$ objects can be calculated with

$$C_0 = \max_{j \in \{1,\ldots,p\} \setminus i} \left\{ \operatorname{Cor}(x_i, x_j) \right\}, \tag{3.3}$$

in which x_j $(j = 1, \ldots, p)$ represents the n-dimensional vector of object j. In the case $p = 2$, the maximum correlation is reduced to $\operatorname{Cor}(x_1, x_2)$.

The vector $C_{\mathrm{p}} \in \mathbb{R}^{N_p}$ represents the N_p maximum correlations after applying N_p times a permutations-matrix Ψ on x_i:

$$C_{\mathrm{p}} = \left(\max_{j \in \{1,\ldots,p\} \setminus i} \left\{ \operatorname{Cor}(\Psi \cdot x_i, x_j) \right\} \right)_{1,\ldots,N_p}.$$

According to Buening and Trenkler (1994) we can calculate the p-value $p^{(i)}$ for object i with

$$p^{(i)} = 1 - \frac{\Big(1, 0, \ldots, 0\Big) \cdot \operatorname{Rg}\left(\left(C_0, C_{\mathrm{p}}^{\top}\right)^{\top} \right)}{N_p + 1}, \tag{3.4}$$

where $\mathrm{Rg}(\cdot)$ is the rank statistic. If C_0 has a high rank, the ratio in (3.4) is close to one and thus the p-value $p^{(i)}$ is close to zero. For p-values smaller than a given significance level α, we have to reject the null hypothesis.

It is important to note that the number of permutations N_p should be large. Common values are in the range of 10^5 or greater, depending on the size of the data set.

3.3 Simple Aggregation Methods

In this section we present simple aggregation methods for summarizing information of several covariates to one single representative covariate. There are a number of methods for summarizing groups of covariates. Starting with a data matrix $X \in \mathbb{R}^{p \times n}$, we are looking for a dimension reducing function:

$$f : \mathbb{R}^{p \times n} \to \mathbb{R}^n, \quad f(x_1, \ldots, x_p) = y, \qquad x_1, \ldots, x_p, y \in \mathbb{R}^n \tag{3.5}$$

where the generated variable y should be a representative vector for the group of covariates $x_1, \ldots, x_p$ and the loss of information should be minimized.

A simple possibility for summarizing a set of covariates is given by standard measures of location. Component wise computation of the arithmetic mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and of q-quantiles

$$\tilde{x}_q = \begin{cases} x_{\lceil nq \rceil}, & \text{if } nq \notin \mathbb{N} \\ \frac{1}{2}\left(x_{(nq)} + x_{(nq+1)}\right), & \text{if } nq \in \mathbb{N} \end{cases}$$

of the data matrix $X \in \mathbb{R}^{p \times n}$ could be performed. The median $\tilde{x}_{0.5}$ corresponds the 0.5-quantile.

Another possibility to summarize covariate information is directly provided by results of the PAM-clustering approach: medoids of clusters are particularly suitable to represent groups and no further aggregation has to be performed.

3.4 Principle Component Analysis

Correlation structures within high-dimensional data sets could be interpreted as redundant information. Principal components analysis (PCA) is a method that reduces data dimensionality by finding a new set of covariates, smaller than the original set of covariates, that nonetheless retains most of the sample's information, i.e. the variation present in the data set, given by the correlations between the original covariates. The new covariates, called principal components (PCs), are uncorrelated, and are ordered by the fraction of the total information each retains. As such, it is suitable for data sets in multiple dimensions, such as large experiments in gene expression. PCA on genes will find relevant components, or patterns, across gene expression data.

Mathematically, PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

The PCA is in general computed by determining the eigenvectors and eigenvalues of the covariance matrix. To calculate the covariance matrix from a data set given by a $(p \times n)$ matrix X, we first center the data by subtracting the mean of each sample vector. Considering the columns of the centered data matrix X as the sample vectors, we can write the covariance matrix of the samples C_n as:

$$C_n = \frac{1}{n} X X^\top.$$

If we are interested in the covariance matrix for the n p-dimensional covariates (i.e. gene expression measurements), we first center the data for each covariate. The covariance matrix C_p is then given by

$$C_p = \frac{1}{p} X^\top X.$$

Often the scale factors $\frac{1}{n}$ and $\frac{1}{p}$ are included in the matrix and the covariance matrices are simply written as $XX^\top$ and $X^\top X$, respectively. The eigenvectors of the covariance matrix are the axes of maximal variance. Since we are only interested in the PCA for the p-dimensional covariates we present a solution for $X^\top X$ in the following.

With the help of singular value decomposition (SVD) (cf. Wall *et al.*, 2003) the factorization of an arbitrary real-valued $(p \times n)$ matrix X with rank r is obtained by

$$X = U\Sigma V^\top, \qquad U \in \mathbb{R}^{p\times n},\ \Sigma \in \mathbb{R}^{n\times n},\ V \in \mathbb{R}^{n\times n}. \tag{3.6}$$

Here U is a $(p \times n)$ orthogonal matrix ($U^\top U = Id_p$) whose columns u_k are called the *left singular vectors* and V is a $(p \times p)$ orthogonal matrix ($V^\top V = Id_p$) whose columns v_k are called the *right singular vectors.* The elements of the $(n \times n)$ matrix Σ are only nonzero on the diagonal, and are called the *singular values.* The first r singular values of $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ are sorted in decreasing order $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0$.

The advantage of the SVD is that there are a number of algorithms for computing the SVD. One method is to compute V and S by diagonalizing $X^\top X$:

$$X^\top X = V\Sigma^2 V^\top$$

and then to calculate U by ignoring the $(r + 1), \ldots, n$ columns of V with $\sigma_k = 0$

$$U = XV\Sigma^{-1}.$$

The SVD and PCA are closely related. As mentioned before, the covariance matrix for X is $X^\top X$. With the help of the SVD, we can write

$$X^\top X = (U\Sigma V^\top)(U\Sigma V^\top)^\top = U\Sigma V^\top V\Sigma U^\top = U\Sigma^2 U^\top$$

using the fact that V is orthogonal, thus $V^\top = V^{-1}$. Further, U can be calculated as the eigenvectors of $X^\top X$. The diagonal of Σ contains the square roots of the eigenvalues of the covariance matrix and the columns of U form the eigenvectors.

An important result of the SVD of a matrix X is that

$$X^{(l)} = \sum_{k=1}^{l} u_k \sigma_k v_k^\top$$

is the *closest* rank l matrix to X (*closest* means that $X^{(l)}$ minimizes the sum of squares of the differences $\sum_{ij}\left(x_{ij} - x_{ij}^{(l)}\right)^2$ of the elements of X and $X^{(l)}$).

For dimension reduction, we determine in advance the number of covariates $p_0 \leq p$ which should be selected. In most cases the number of selected PCs is due to the proportion of variance accounted, to which the singular values are proportional. Timm (2002) points out that the explained cumulated proportion of the variance in the data α for the first p_0 principal components is

$$\alpha = \frac{\sum_{j=1}^{p_0} \sigma_j^2}{\sum_{j=1}^{p} \sigma_j^2}.$$

If we only consider the first principle component of the data matrix X the aggregated matrix $X^{(l)}$ is reduced to a single vector that represents all p covariates of the observed data matrix X.

4 Methods in Survival Analysis

Survival analysis involves the modeling of time to event data where the response is often referred to as failure time, survival time, or event time. We focus on applying the techniques to biological and medical applications, i.e. the death or progression of cancer patients is the event of interest. A common feature in these data is that censoring is present when we have some information about an individual's event time, but we do not know the exact event time. For the analysis methods we assume that the censoring mechanism is independent of the survival mechanism.

In this thesis, we will focus on right-censoring, where all that is known is that the individual is still event-free at a given time. There are a lot of reasons why right-censoring may occur. A typical censoring circumstance is that an individual does not experience the event before the study ends, an individual is lost to follow-up during the study period or an individual withdraws from the study. Censoring rates of more than 50 % are not unusual in cancer data sets and a main challenge of survival analysis is the integration of the censoring information in the statistical methodology instead of deleting this data.

In this chapter, we introduce the basis functions, the survival function, and the hazard rate, for modeling survival data in Section 4.1. Basic estimates of the survival function and the hazard rate and the corresponding standard errors are discussed in Section 4.2 and Section 4.3. In Section 4.4, we present the log-rank test for detecting differences in survival in a two sample setting. Including covariate information (clinical and genomic variables) of the individuals leads to more detailed survival models. We consider Cox's proportional hazards model (Cox model), introduced by Cox (1972), in Section 4.5 with a detailed description in the high-dimensional setting in Section 4.6, including a model selection and evaluation procedure. Since most methods for dimension reduction or shrinkage require the selection of a tuning parameter that determines the amount of shrinkage, we describe how to choose the optimal tuning parameter for the presented methods.

4.1 Introduction to Survival Analysis

Let T be a positive random variable which represents the time from a well-defined starting point $t = 0$ to an event of interest with density $f(t)$ and corresponding distribution function $F(t)$. We define the survival function (the probability of being event-free up to time t) by

$$S\left(t\right) = P\left(T > t\right).$$

According to the relationship to the distribution function $F\left(t\right) = 1 - S(t)$, it is easy to see that the survival function $S\left(t\right)$ is right-continuous and monotonically decreasing with limits $S(t) = 0$ for $t \to \infty$ and $S(t) = 1$ for $t \to 0$. In Section 4.2, we introduce an estimator for the survival function.

The hazard rate (function) $\lambda(t)$, also called *risk function* or *conditional failure rate*, is the chance that an individual experiences an event in the next instant time, conditional on survival until time t:

$$\lambda(t) = \lim_{h \downarrow 0} \frac{P\left(t \leq T < t + h \mid T > t\right)}{h}, \qquad \lambda(t) \geq 0 \; \forall t \in [0, \infty].$$

If T is a continuous random variable, the relationship between the survival function and the hazard rate is given by

$$\begin{aligned} \lambda(t) &= \lim_{h \downarrow 0} \frac{P\left(t \leq T < t + h\right)}{h} \frac{1}{P\left(T > t\right)} \\ &= \frac{f(t)}{S(t)} = \frac{\frac{\partial}{\partial t} F(t)}{S(t)} = \frac{-\frac{\partial}{\partial t} S(t)}{S(t)} = -\frac{\partial}{\partial t} \ln\left(S(t)\right). \end{aligned}$$

A related quantitiy is the cumulative hazard function $\Lambda(t)$, defined by

$$\Lambda(t) = \int_0^t \lambda(u) \, \mathrm{d}u = \int_0^t \frac{f(u)}{S(u)} \, \mathrm{d}u = \int_0^t \frac{\frac{\delta}{\delta u} F(u)}{S(u)} \, \mathrm{d}u = -\int_0^t \frac{\frac{\delta}{\delta u} S(u)}{S(u)} \, \mathrm{d}u = -\ln S(t).$$

A well-established estimator for the cumulative hazard function is the Nelson-Aalen estimator that is introduced in Section 4.3.

4.2 Kaplan-Meier Estimator

The standard nonparametric estimator for the survival function $S(t)$ introduced by Kaplan and Meier (1958) is called *Kaplan-Meier estimator* or *Product-Limit estimator*. To allow for possible ties in the data, we consider D distinct event times t_i $(i = 1, \dots, D)$ with $t_i < t_{i+1}$ with d_i events at time t_i. Note that censoring is not an event. Let Y_i be the number of individuals who are at risk at time t_i, e.g. the number of individuals who are alive at time t_i or experienced the event of interest at t_i. The Kaplan-Meier estimator $\hat{S}\left(t\right)$ is defined (for all values of t in the range where there is data) by:

$$\hat{S}\left(t\right) = \begin{cases} 1, & t < t_1 \\ \prod_{t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right), & t \geq t_1. \end{cases}$$

According to Klein and Moeschberger (2003, Chapter 4.2) $\hat{S}(t)$ is a consistent estimator of the survival function $S(t)$. There are analogous relations between the estimates as between the survival function and the distribution function. The Kaplan-Meier estimator is a monotone decreasing step function with jumps at the observed event times and is well defined for all time points less than the largest observation. If the largest time point is an event, the survival curve is zero beyond this point. If the largest point is censored, the value $\hat{S}(t)$ beyond this point is undetermined. There is no information when the last survivor would have died if the survivor had not been censored.

The standard error of the Kaplan-Meier estimator $\hat{S}(t)$ is estimated by Greenwood's formula (see Klein and Moeschberger, 2003, Chapter 4.2):

$$\mathrm{SE}\left(\hat{S}(t)\right) = \hat{S}(t) \cdot \sqrt{\sum_{t_i < t} \frac{d_i}{Y_i(Y_i - d_i)}}.$$

The standard error is especially required for the calculation of pointwise $(1-\alpha)$ confidence intervals, given by

$$\left[\hat{S}(t) \pm u_{1-\alpha/2}\,\mathrm{SE}\left(\hat{S}(t)\right)\right].$$

Here u_α is the α-quantile of a standard normal distribution $\mathcal{N}(0,1)$.

For testing differences in survival of two or more groups, the log-rank test is introduced in Section 4.4.

4.3 Nelson-Aalen Estimator

The Kaplan-Meier estimator can also be used to estimate the cumulative hazard function $\Lambda(t) = -\ln S(t)$; the estimation is obtained by $\hat{\Lambda}(t) = -\ln \hat{S}(t)$. An alternative estimator of the cumulative hazard rate $\Lambda(t)$ was suggested by Nelson (1972) and Aalen (1978). With the notation of Section 4.2 and the assumption that all time points are distinct, the *Nelson-Aalen estimator* of the cumulative hazard rate $\Lambda(t)$ is given by

$$\hat{\Lambda}(t) = \sum_{t_i \leq t} \hat{\lambda}(t_i) = \sum_{t_i \leq t} \frac{d_i}{Y_i}.$$

According to the connection between the survival function $S(t)$ and the cumulative hazard rate $\Lambda(t)$,

$$\hat{S}(t) = \exp\left(-\hat{\Lambda}(t)\right) = \exp\left(-\sum_{t_i \leq t} \frac{d_i}{Y_i}\right)$$

is an alternative estimator for the survival function.

4.4 Log-rank Test

In this section, we focus on hypothesis tests that are based on comparing the Nelson-Aalen estimates for two or more groups. In the survival context the most frequently used test is the log-rank test that compares the differences in hazard rates over time. According to Klein and Moeschberger (2003, Chapter 7.3) we test the following set of hypotheses:

$$H_0\colon \lambda_1(t) = \lambda_2(t) = \ldots = \lambda_K(t), \quad \forall t \leq \tau, \text{vs.}$$

$$H_1\colon \exists i, j \in 1, \ldots, K, \, \exists t \leq \tau\colon \; \lambda_i(t) \neq \lambda_j(t).$$

Here τ is the largest time at which all groups have at least one individual at risk. The test statistic is now constructed with the Nelson-Aalen estimator (see Section 4.3). We consider distinct event times of the pooled population with the notation from Section 4.2. At time t_i we observe d_{ij} events in the jth population out of Y_{ij} individuals at risk. The test statistic is based on a weighted comparison of the estimated hazard rate of the jth population and the estimated pooled hazard rate. Let $W_j(t_i)$ be a positive weight function with the property $W_j(t_i) = 0$ whenever $Y_{ij} = 0$. The general test for the jth group is based on the statistics

$$Z_j(\tau) = \sum_{i=1}^{D} W_j(t_i) \left[\frac{d_{ij}}{Y_{ij}} - \frac{d_i}{Y_i} \right], \quad d_i = \sum_{j=1}^{K} d_{ij}, \; Y_i = \sum_{j=1}^{K} Y_{ij}, \; j \in 1, \ldots, K.$$

By the specific choice of the weight function the influence of early and late observations can be controlled. The log-rank test is defined by $W_j(t_i) = Y_{ij} W(t_i)$ with $W(t_i) \equiv 1$. Thus, the rewritten test statistics

$$Z_j(\tau) = \sum_{i=1}^{D} W(t_i) \left[d_{ij} - Y_{ij} \frac{d_i}{Y_i} \right], \; j \in 1, \ldots, K$$

show that all event times have equal weights. The entries $\hat{\sigma}_{jg}$ of the covariance matrix $\hat{\Sigma}$ for the log-rank test are estimated by

$$\hat{\sigma}^2_{jj} := \mathrm{Var}\left(Z_j(\tau)\right) = \sum_{i=1}^{D} \frac{Y_{ij}}{Y_i}\left(1 - \frac{Y_{ij}}{Y_i}\right)\left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i, \quad j = 1, \ldots, K, \text{ and}$$

$$\hat{\sigma}^2_{jg} := \mathrm{Cov}\left(Z_j(\tau), Z_g(\tau)\right) = \sum_{i=1}^{D} \frac{Y_{ij}}{Y_i}\frac{Y_{ig}}{Y_i}\left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i, \quad j, g = 1, \ldots, K, \ j \neq g.$$

The test statistic is given by

$$\chi^2_{\mathrm{LR}} = \left(Z_i(\tau), \ldots, Z_{K-1}(\tau)\right) \hat{\Sigma}^{-1} \left(Z_i(\tau), \ldots, Z_{K-1}(\tau)\right)^{\top}.$$

If the null hypothesis is true, this statistic follows a χ^2-distribution with $(K-1)$-degrees of freedom. The null hypothesis is rejected at a given α level if χ^2_{LR} is larger than the $(1-\alpha)$-quantile of this distribution.

4.5 Cox Proportional Hazard Model

In addition to the event times there are often covariates observed that might have impact on survival. In order to cope with right-censored survival data we use the Cox model, also referred to as the proportional hazards regression model (see Cox, 1972).

First, we introduce the notation. As before, let T denote the time to event with density $f(t)$, distribution function $F(t)$ and corresponding survival function $S(t)$. We also consider the situation of possibly right-censored time to event data. Let C be the positive random variable that represents the censoring times. A common assumption in survival context (see, e.g., Gerds and Schumacher, 2006) is that T and C are independent. We assume that C is distributed according to $S_C(t) = P(C > t) = 1 - F_C(t)$.

In the usual setup of survival data, we observe for each of the $i = 1, \ldots, n$ individuals the triple $(\tilde{T}_i, \delta_i, Z_i)$, where $\tilde{T}_i = \min(T_i, C_i)$ is the minimum of the event times T_i and the censoring times C_i. In addition, the censoring information is represented by the (non-censoring-)indicator $\delta_i = I_{\{T_i \leq C_i\}}$. δ_i is equal to 1 if $\tilde{T}_i$ is a true event time and equal to 0 if it is right-censored. For each individual i, the covariate information is given by a p-dimensional vector of covariates $Z_i = (Z_{i1}, \dots, Z_{ip})^\top$.

Cox (1972) suggested that the risk of an event at time t for an individual i with given covariate vector $Z_i = (Z_{i1}, \dots, Z_{ip})^\top$ is modeled as

$$\lambda(t \mid Z = Z_i) = \lambda_0(t) \exp\left(\beta^\top Z_i\right), \tag{4.1}$$

where $\lambda_0(\cdot)$ is an arbitrary baseline hazard function and $\beta = (\beta_1, \dots, \beta_p)^\top$ a vector of regression coefficients. The value of $\beta^\top Z_i$ is called *prognostic index* (PI) or *risk score* of individual i. In other words: PI is the sum of the covariate values of a particular individual, weighted by the corresponding (estimated) regression coefficients.

As mentioned before, the Cox model is often called a proportional hazards model and assumes that for two covariate values Z_1 and Z_2 the ratio of their hazard rates is constant over time:

$$\frac{\lambda(t \mid Z_1)}{\lambda(t \mid Z_2)} = \frac{\lambda_0(t) \exp\left(\beta^\top Z_1\right)}{\lambda_0(t) \exp\left(\beta^\top Z_2\right)} = \exp\left(\beta^\top (Z_1 - Z_2)\right) = \text{const.}$$

In the classical setting with $n > p$, the regression coefficients are estimated by Maximum Likelihood Estimation (MLE). We suppose that there are no ties between the ordered event times $t_1 < t_2 < \ldots < t_D$. The likelihood is only calculated for the event times. The risk set $R(t_i)$ at time t_i (that also includes the censoring times) is defined by

$$R(t_i) = \{j \colon t_j \geq t_i\},$$

as the set of all individuals who have not yet failed nor been censored. The partial likelihood can be written as

$$L(\beta) = \left(\prod_{i=1}^{D} L_i\right) = \prod_{i=1}^{D} \frac{\exp\left(\sum_{k=1}^{p} \beta_k Z_{ik}\right)}{\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}\right)}. \tag{4.2}$$

Here, the numerator of the likelihood depends only on information from individuals having an event, whereas the denominator consists of information of all individuals having not yet experienced the event or who may be censored later. Note that Cox's partial likelihood estimation method ignores the actual event times; it takes the ordering of events into account, but not their explicit values. For this reason the likelihood is referred to be *partial*.

Maximizing the partial likelihood is equivalent to maximizing the log partial likelihood

$$\begin{aligned} LL(\beta) &= \ln\left(\prod_{i=1}^{D} L_i\right) = \sum_{i=1}^{D} \ln\left(L_i\right) \\ &= \sum_{i=1}^{D}\sum_{k=1}^{p} \beta_k Z_{ik} - \sum_{i=1}^{D} \ln\left(\sum_{j\in R(t_i)} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}\right)\right), \end{aligned} \tag{4.3}$$

which corresponds to solving the following system of equations: $U_k(\beta) := \frac{\partial}{\partial \beta_k} LL(\beta) = 0$ for all $k = 1, \ldots, p$. The vector $U(\beta)$ with components U_k is called *efficient score vector*.

The estimation of the baseline hazard rate $\lambda_0(t)$ is performed after the estimation of the regression coefficients $\beta_1, \ldots, \beta_p$ with the Breslow estimator (see Klein and Moeschberger, 2003, Chapter 8.8)

$$\hat{\lambda}_0(t) = \sum_{t_i \leq t} \left(\frac{d_i}{\sum_{j\in R(t_i)} \exp\left(\hat{\beta}^\top Z_j\right)} \right).$$

There are several suggestions for constructing the partial likelihood when ties among the event times are present. A widely used adaption of the likelihood was suggested by Efron (1977):

$$L_E(\beta) = \prod_{i=1}^{D} \frac{\exp\left(\beta^\top \left(\sum_{j\in D_i} Z_j\right)\right)}{\prod_{j=1}^{d_i} \left[\sum_{k\in R(t_i)} \exp\left(\beta^\top Z_k\right) - \frac{j-1}{d_i} \sum_{k\in D_i} \exp\left(\beta^\top Z_k\right)\right]}.$$

Here d_i is the number of events at time t_i and D_i the set of individuals who experience an event at time t_i. Even though the partial likelihood (equation 4.2) and its adapted version by Efron show similar results in practice, we utilize Efron's version for our calculations.

Maximizing Cox's partial likelihood does not work for $p \gg n$ and some dimension reduction or regularization methods must be used. The lasso- and ridge-regression are possibilities to jointly optimize over the parameters (see Section 4.6.1).

4.5.1 Hypothesis Testing

There are three main tests for hypotheses for the regression coefficients $\beta = (\beta_1, \dots, \beta_p)^\top$. Let $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^\top$ denote the maximum likelihood estimate of β. All three statistical tests have as null hypothesis that the set of regression parameters β is equal to some particular value β_0 ($H_0\colon \beta = \beta_0$ vs. $H_1\colon \beta \neq \beta_0$). For testing whether the regression coefficients have an effect of individuals' risk to die or not, we choose $\beta_0 = 0$. We start with presenting the (efficient) score test, followed by the Wald test and the likelihood-ratio test. Note that the Wald test and the likelihood-ratio test are asymptotically equivalent.

4.5.2 Score Test

The score test is the most powerful test when the true value of β is close to β_0. The main advantage of the score test is that it does not require an estimate of the information under the alternative hypothesis or unconstrained maximum likelihood. This makes testing feasible when the unconstrained maximum likelihood estimate is a boundary point in the parameter space. Let $U(\beta) = (U_1(\beta), \dots, U_p(\beta))$ be the efficient score vector defined in Section 4.5 and $I(\beta)$ the $(p \times p)$ information matrix evaluated at β:

$$I(\beta) = \left[-\frac{\partial^2}{\partial \beta_j \partial \beta_k} LL(\beta) \right]_{p \times p}, \quad j, k = 1, \dots, p.$$

Here, $LL(\beta)$ is the log partial likelihood (see Section 4.5) evaluated at β. According to Klein and Moeschberger (2003, Chapter 8.3), $U(\beta)$ is asymptotically p-variate normal with mean vector 0 and covariance matrix $I^{-1}(\beta)$ when H_0 is true. The score test for $H_0\colon \beta = \beta_0$ is given by

$$\chi^2_{\text{SC}} = U(\beta_0)^\top I^{-1}(\beta_0) U(\beta_0)$$

having a χ^2-distribution with p degrees of freedom under H_0. We reject the null hypothesis at a significance level α if $\chi^2_{\text{SC}} > \chi^2_{p,1-\alpha}$. Here, $\chi^2_{p,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2-distribution with p degrees of freedom.

4.5.3 Wald Test

The Wald test (first suggested by Wald, 1943) compares the maximum likelihood estimate $\hat{\beta}$ of the parameter vector of interest β with the proposed value β_0, with the assumption that the difference between the two will be approximately normal. Typically the square of the difference weighted with the covariance matrix $I^{-1}(\beta)$ is compared to a chi-squared distribution. With the notation from Section 4.5.2, the Wald test for the global hypothesis $H_0\colon \beta = \beta_0$ is defined by

$$\chi^2_{\text{W}} = (\hat{\beta} - \beta_0)^\top I^{-1}(\hat{\beta})(\hat{\beta} - \beta_0)$$

and has a χ^2-distribution with p degrees of freedom if H_0 is true.

4.5.4 Likelihood-ratio Test

A likelihood ratio test is used to compare the fit of two models: the null model using β_0 and the alternative model using the parameter estimate $\hat{\beta}$. The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than under the other. This likelihood ratio, or equivalently its logarithm, can then be used to compute a p-value to decide whether to reject the null model in favor of the alternative model. Both competing models, the null model and the alternative model, are fitted separately to the data and the log-likelihood is recorded. The test statistic is twice the difference of these log-likelihoods:

$$\chi^2_{\text{LR}} = 2\left(LL(\hat{\beta}) - LL(\beta_0)\right)$$

which has a χ^2-distribution with p degrees of freedom under H_0.

4.5.5 Local Test

In cases of variable selection, we are also interested in testing a hypothesis for single components $\hat{\beta}_i$ of $\hat{\beta}$. These tests are often called *local tests* and are presented in detail in Klein and Moeschberger (2003, Chapter 8.5). Exemplary, we present the Wald local test that we utilize e.g. for forward selection described in Section 4.6.3. In the univariate case, the Wald test statistic for the two-sided test with $\beta_0 \in \mathbb{R}$

$$H_0\colon \hat{\beta}_i = \beta_0 \quad \text{vs.} \quad H_1\colon \hat{\beta}_i \neq \beta_0$$

is given by

$$W = \frac{\hat{\beta}_i - \beta_0}{\mathrm{SE}\left(\hat{\beta}_i\right)},$$

which is compared to a normal distribution. The standard error $\mathrm{SE}\left(\hat{\beta}_i\right)$ is derived by the corresponding value of the information matrix of the maximum likelihood estimate, i.e. square root of the ith diagonal element of $I^{-1}(\beta)$. Typically the square W is compared to a chi-squared distribution.

4.6 Cox Proportional Hazards Model in High-Dimensional Settings

Maximizing Cox's log partial likelihood LL (see equation (4.3)) does not work if the number of covariates is larger than the number of individuals ($p \gg n$), since then a unique maximum for the optimization problem of LL does not exist. Thus, dimension reduction or regularization methods are required. Lasso- and ridge-regression are possibilities to optimize over the parameters.

In the following, we assume that the data consists of two different categories of covariates

(I) p_1 clinical covariates $Z_{cl} = (Z_{cl,1}, ..., Z_{cl,p_1})^\top$: e.g. tumor size, tumor grade, age

(II) p_2 genomic covariates $Z_g = (Z_{g,1}, ..., Z_{g,p_2})^\top$: gene expression values of single genes or combined gene expression values for (preclustered) gene groups.

If a differentiation between the clinical and the genomic covariates is not necessary, we use the standard notation $Z_1, ..., Z_p$.

4.6.1 Penalized Estimation of the Likelihood

By introducing a tuning/complexity parameter $\lambda \in \mathbb{R}_+$, we have the opportunity to shrink large values of the estimated regression coefficients $(\hat{\beta}_1, \ldots, \beta_p)$ towards zero. With the help of a λ-depending function $f_\lambda : \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}_+$, we reformulate the initial maximization problem $\hat{\beta} = \operatorname{argmax}_{\hat{\beta}} \mathrm{LL}(\hat{\beta})$ as

$$\hat{\beta}_\lambda = \underset{\hat{\beta}}{\operatorname{argmax}} \Big(\mathrm{LL}(\hat{\beta}) - f_\lambda(\hat{\beta}) \Big) .$$

We require properties for f_λ: the larger the value of λ, the greater should be the penalty for large values of $\hat{\beta}$. In addition, a value $\lambda = 0$ should be equivalent to the initial optimization problem without a penalty.

By choosing

$$f_\lambda(\beta) = \lambda \sum_{k=1}^{p} \beta_k^2 \;\rightarrow\; \hat{\beta}_\lambda = \underset{\hat{\beta}}{\operatorname{argmax}} \left(\mathrm{LL}(\hat{\beta}) - \lambda \sum_{k=1}^{p} \hat{\beta}_k^2 \right), \tag{4.4}$$

the maximization corresponds to ridge-regression (L_2 regression) introduced by Hoerl and Kennard (1970). Ridge-regression shrinks the regression coefficients by imposing a penalty on their squared values. A result of ridge-regression consists of many small but non-zero regression coefficients.

The lasso- or L_1-regression (Tibshirani, 1996) shrinks the regression coefficients towards zero by penalizing the absolute values instead of their squares. The maximization problem of penalized log partial likelihood thus becomes

$$f_\lambda(\beta) = \lambda \sum_{k=1}^{p} |\beta_k| \;\rightarrow\; \hat{\beta}_\lambda = \underset{\hat{\beta}}{\operatorname{argmax}} \left(\mathrm{LL}(\hat{\beta}) - \lambda \sum_{k=1}^{p} \left|\hat{\beta}_k\right| \right). \tag{4.5}$$

Penalizing with the absolute values has the effect that many regression coefficients are shrunk exactly to zero. Thus, the lasso is a variable selection method.

In applications, where clinical and genomic covariates are integrated in regression models, we perform the penalization only on the high-dimensional genomic covariates, the clinical covariates are handled as unpenalized mandatory covariates. Let Z_{cl} be the vector of clinical, Z_g be the vector of genomic covariates (see Section 4.6) and $\hat{\beta}_{cl} = (\hat{\beta}_{cl,1}, \ldots, \hat{\beta}_{cl,p_1})$ and $\hat{\beta}_g = (\hat{\beta}_{g,1}, \ldots, \hat{\beta}_{g,p_2})$ the corresponding parameter estimates of β_{cl} and β_g. The penalized log partial likelihood thus becomes

$$\mathrm{LL}_{\text{ridge}} = \left(\mathrm{LL}(\hat{\beta}_{cl}, \hat{\beta}_g) - \lambda \sum_{k=1}^{p_2} \hat{\beta}_{g,k}^2 \right) \tag{4.6}$$

for ridge-regression and

$$\mathrm{LL}_{\text{lasso}} = \left(\mathrm{LL}(\hat{\beta}_{cl}, \hat{\beta}_g) - \lambda \sum_{k=1}^{p_2} \left|\hat{\beta}_{g,k}\right| \right) \tag{4.7}$$

for lasso-regression. The regression coefficients of the clinical covariates are not penalized in any way. In both methods the tuning parameter λ controls the amount of shrinkage and is obtained by cross-validation described in the next section.

4.6.2 Choosing the Tuning Parameter

The model complexity of the prediction methods depends on the tuning parameter λ. We use K-fold cross-validation as proposed by Verweij and van Houwelingen (1993) for estimating λ. In K-fold cross-validation, the original data set is randomly partitioned into k subsets that are called *folds*. For easy reference, we utilize $\mathrm{LL}_\lambda \in \{\mathrm{LL}_{\text{lasso}}, \mathrm{LL}_{\text{ridge}}\}$ as a synomym for a penalized log partial likelihood. The K-fold cross-validated log partial likelihood (CVPL) is given by

$$\mathrm{CVPL}\,(\lambda) = \sum_{k=1}^{K} \left[\mathrm{LL}_\lambda \left(\hat{\beta}_{cl}^{(-k)}, \hat{\beta}_{g}^{(-k)}(\lambda) \right) - \mathrm{LL}_\lambda{}^{(-k)} \left(\hat{\beta}_{cl}^{(-k)}, \hat{\beta}_{g}^{(-k)}(\lambda) \right) \right]. \tag{4.8}$$

Here, $\mathrm{LL}_\lambda(\beta_{cl}, \beta_g(\lambda))$ denotes the penalized log partial likelihood given in Section 4.6.1 and $\mathrm{LL}_\lambda^{(-k)}(\beta_{cl}, \beta_g(\lambda))$ the log partial likelihood when the kth fold ($k = 1, ..., K$) is left out. The difference of the two terms compared in the formula is that in the right term the likelihood is evaluated without the kth fold, and the left term is evaluated with all individuals. In both cases the parameters β_{cl} and β_g are estimated without the kth fold. The estimates of β_{cl} and β_g when the kth fold is left out are denoted by $\hat{\beta}_{cl}^{(-k)}$ and $\hat{\beta}_{g}^{(-k)}(\lambda)$. The optimal value of λ is chosen to maximize the sum of the contributions of each fold to the log partial likelihood. Maximizing the cross-validated log partial likelihood with respect to λ yields the optimal penalty/tuning parameter λ_{opt}.

In detail, both terms within the brackets in equation (4.8) are negative with a greater absolute value of the first summand. Thus, the K differences are all negative and consequently the CVPL as well. By adding the kth fold in the first summand $\mathrm{LL}_\lambda \left(\hat{\beta}_{cl}^{(-k)}, \hat{\beta}_{g}^{(-k)}(\lambda) \right)$, the penalized log partial likelihood with complete data should result in a large absolute value. A large value corresponds to a good prediction of the kth fold by the other $(K-1)$ folds. Hence, we can rewrite the maximization problem as

$$\lambda_{\text{opt}} = \underset{\lambda \in \mathbb{R}_+}{\operatorname{argmax}}\ \mathrm{CVPL}(\lambda),$$

which can be interpreted as optimization of the predictive quality.

The maximization problem can be solved with the help of a Newton-Raphson algorithm that alternates between adjusting the parameters $(\beta_{cl}, \beta_g(\lambda))$ and λ and optimizing the

CVPL according to the other parameter. Details are described in the work of Verweij and van Houwelingen (1994).

In the case $K = n$, we perform leave-one-out cross-validation and the value of the CVPL is reproducible. If the number of folds K is smaller than the number of individuals n, the individuals are subdivided into groups (folds) at random and the value of the CVPL is not deterministic.

4.6.3 Variable Selection with CVPL

In addition to the presented L_2-regression in Section 4.6.1 (results in many small but non-zero regression coefficients) and the variable selecting L_1-regression (many regression coefficients are shrunk exactly to zero), we introduce two standard variable selection methods: univariate and forward selection. Both methods also utilize the cross-validated log partial likelihood for model selection. Covariates are added sequentially to the Cox model and the number of covariates is equivalent to the optimal tuning parameter λ_{opt}.

The following descriptions of univariate and forward selection refer to the situation where clinical and genomic information is provided for all individuals and the clinical information is mandatory included in the models. If clinical information should not be modeled, we just start with the empty model.

Univariate Selection Starting with a clinical Cox model, we test the effect each genomic covariate has on survival. For each genomic covariate $Z_{g,j}$, $j = 1, \ldots, p_2$, we fit a Cox model

$$h\left(t \mid Z_{cl}, Z_{g,j}\right) = h_0(t) \exp\left(\beta_{cl}^\top Z_{cl} + \beta_{g,j}^\top Z_{g,j}\right).$$

We then test the null hypothesis $\beta_{g,j} = 0$ versus the alternative $\beta_{g,j} \neq 0$ using the Wald local test (see Section 4.5.5). After testing the genetic covariates one at a time, we arrange them according to increasing p-values and construct the multivariate Cox regression model 4.1 including the λ_{opt} top ranked genomic covariates in addition to the

p_1 clinical covariates. The tuning parameter λ_{opt} thus directly represents the number of genetic covariates in the final model, and it is determined by cross-validation.

Forward Selection We start with the clinical model described above. We iteratively add single genomic covariates to the model by selecting in every step the covariate that yields the most significant model together with the covariates chosen in the steps before. The Likelihood ratio test (see Section 4.5.4) is used for hypothesis testing. The optimal tuning parameter λ_{opt} is determined by cross-validation.

4.6.4 Alternative Approach: CoxBoost

An alternative method to the lasso- and ridge-regression was developed by Binder and Schumacher (2008b). In particluar, they propose a boosting approach (*CoxBoost*) to fit a Cox proportional hazards model by componentwise likelihood based boosting. It is especially suited for models with a large number of predictors and allows for the integration of mandatory clinical covariates. The aim of the CoxBoost approach is to estimate the parameter vector for the covariates in the Cox model. Typical gradient boosting approaches either use all covariates for the fitting of the gradient in each step, e.g. based on regression trees, or, in component-wise boosting, update only one element of the estimate of the parameter vector, corresponding to only one covariate. Binder and Schumacher (2008b) point out that the results for componentwise CoxBoost are similar to those from lasso-like approaches. Let $l = 1, \dots, M$ be the number of boosting steps and $\hat{\beta}_{(l-1)}$ the estimate of the parameter vector β after $(l-1)$ steps of the algorithm. For each of the $k = 1, \dots, p$ covariates a separate update of the corresponding parameter is evaluated. The covariate that improves the overall fit the most will then be selected for the update. The possible updates of $\hat{\beta}_{(l-1)}$ in the lth step after updating covariate k are obtained by the penalized log partial likelihood (c.f. Equations (4.2) and (4.4))

$$\mathrm{LL}_\lambda(\gamma_{k(l)}) = \sum_{i=1}^{D} \beta_{(l-1)}^\top Z_i + \gamma_{k(l)} Z_{ik} - \sum_{i=1}^{D} \ln \left(\sum_{j \in R(t_i)} \exp\left(\beta_{(l-1)}^\top Z_j + \gamma_{k(l)} Z_{ik}\right) \right) - \lambda \gamma_{k(l)}^2,$$

with respect to the parameter $\gamma_{k(l)}$. Note, that Z_{ik} is the covariate for patient i and covariate k. The tuning parameter λ and the parameter estimates $\hat{\gamma}_{k(l)}$ (estimates for the updates of $\hat{\beta}_{k(l-1)}$) are chosen by the cross-validated log partial likelihood (see Section 4.6.3).

The componentwise CoxBoost algorithm for a fixed tuning parameter λ is the following (Binder and Schumacher, 2008b):

(1) Initialize $\hat{\beta}_{(0)} = (0, \ldots, 0)^\top$.

(2) Repeat for $l = 1, \ldots, M$:

  (i) Calculate potential updates $\hat{\gamma}_{k(l)}$ for all covariates $k = 1, \ldots, p$ via penalized log partial likelihood

  (ii) Determine the best update k^* that maximizes the penalized log partial likelihood

  (iii) Calculate the updated parameter vector $\hat{\beta}_{(l)} = (\hat{\beta}_{1(l)}, \ldots, \hat{\beta}_{p(l)})^\top$ via

$$\hat{\beta}_{k(l)} = \begin{cases} \hat{\beta}_{k(l-1)} + \hat{\gamma}_{k(l)}, & \text{if } k = k^* \\ \hat{\beta}_{k(l-1)}, & \text{if } k \neq k^*. \end{cases}$$

Note that the step size for the updates in part (2)(iii) of the algorithm is 1. The modification of the step size is controlled by the tuning parameter λ. Binder and Schumacher (2008b) point out that the resulting model is sparse.

When including mandatory covariates in the componentwise CoxBoost algorithm, Binder and Schumacher (2008b) suggest to update the corresponding parameters before each step of componentwise CoxBoost. In other words, mandatory parameters are updated separately. The CoxBoost algorithm has two tuning parameters, the penalty parameter λ and the number of boosting steps M. To avoid overfitting, Binder and Schumacher (2008b) suggest to choose λ such that the number of boosting steps M is larger than 50. The algorithm is implemented in the `R`-package `CoxBoost` (Binder, 2011).

4.7 Evaluation of Prediction Performance

In this section, we describe how we evaluate the prediction performance of the estimated models. We make use of three different model evaluation criteria, two based on the prognostic index of the individuals and the third one based on a quadratic loss function. The basic idea is to split the data into a training set for model fitting and a test set for model evaluation. It is important to note that we have to consider several splits of the data into training and test sets due to the extreme dependence of the results on such a split (cf. Bøvelstad *et al.*, 2007; Ein-Dor *et al.*, 2006).

4.7.1 Evaluation Procedure

In order to obtain a fair comparison of the prediction methods, we divide the data $S \in \mathbb{N}$ times at random in a training set of $\lceil r \cdot n \rceil$, $r \in (0, 1)$ individuals used for estimation and a test set of $\lfloor (1 - r) \cdot n \rfloor$ individuals used for evaluation, where n is the number of all individuals in the study. After computing the optimal tuning parameter $\hat{\lambda}_{\text{train}}$ by K-fold cross-validation using the training data, we estimate the regression coefficients $\hat{\beta}_{\text{train}}$ on the whole training data set. For each of the S splits into training data and test data, we calculate on the test set the three evaluation criteria explained in the following Sections 4.7.2, 4.7.3 and 4.7.4.

4.7.2 Log-rank Test for two Prognostic Groups

For each individual i with covariate information $Z_i = (Z_{i1}, ..., Z_{ip})^\top$ in the test set, we calculate its individual *prognostic index* (also called *risk score*)

$$RS_i = \hat{\beta}_{\text{train}}^\top Z_i.$$

The individuals in the test set are assigned to two subgroups based on their prognosis, into one with *good* and one with *bad* prognosis. If the prognostic index $\hat{\beta}_{\text{train}}^\top Z_i$ of individual i is higher, the risk of the event of interest is expected to be increased and thus the survival time is expected to be shorter. In order to obtain equally sized subgroups, individual i in the test set is assigned to the *high-risk* group if its prognostic index is above the median of all prognostic indices calculated on the test set. We apply a log-rank test (see Section 4.4) on the two prognostic groups and use the p-value as an evaluation criterion for the usefulness of the grouping.

Bøvelstad *et al.* (2007) point out that a disadvantage of this criterion is that it does not consider the ranking of the patients within the groups and it may not be biologically meaningful.

4.7.3 Prognostic Index

The prognostic index $\hat{\beta}_{\text{train}}^{\top} Z_i$ for individual i in the test set is used as a single continuous covariate in a Cox model. For $\alpha \in \mathbb{R}$ we fit the Cox model $\lambda_i(t \mid Z = Z_i) = \lambda_0(t) \exp(\alpha RS_i)$. Using the likelihood ratio test (see Section 4.5.4), we test the null hypothesis $\alpha = 0$ versus $\alpha \neq 0$ and assess the prediction performance with the obtained p-value. A small p-value indicates the calculated prognostic indices have an effect on survival.

4.7.4 Brier-Score

The Brier-Score is a proper score function that measures the accuracy of a set of probability assessments. It was proposed by Brier (1950) and measures the average squared deviation between predicted probabilities for a set of events and their outcomes. Originally, it was mostly used for weather forecasts in the setting of binary events. Graf *et al.* (1999) proposed a framework where the goodness of a predicted survival function can also be assessed based on the (integrated) Brier-Score. The Brier-Score $BS(t)$ is defined as a function of time $t > 0$ by

$$BS(t) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{(0 - \hat{S}(t|Z_i))^2 \mathbf{1}(\tilde{T}_i \leq t)\delta_i}{\hat{S}_C(\tilde{T}_i)} + \frac{(1 - \hat{S}(t|Z_i))^2 \mathbf{1}(\tilde{T}_i > t)}{\hat{S}_C(t)} \right], \tag{4.9}$$

where $\hat{S}_C(\cdot)$ denotes the Kaplan-Meier estimate of the censoring distribution which is only based on the censored observations (cf. Section 4.2). The values of the Brier-Score range between 0 and 1. Small Brier-Scores at time t reflect good predictions. For a fixed time point t^*, the contributions to the Brier-Score can be splitted into the three following categories:

(1) $\tilde{T}_i \leq t^*$ and $\delta_i = 1$,

(2) $\tilde{T}_i > t^*$,

(3) $\tilde{T}_i \leq t^*$ and $\delta_i = 0$.

The observations in the first category are uncensored observations that experience their events before time t^*. Thus the event status at time t^* is equal to $\mathbf{1}(T_i > t^*) = 0$ and the contribution to the Brier-Score is $(0 - \hat{S}(t|Z_i))^2$. In the second category, we observe individuals that are event-free before and at time t^*, their event status is equal to 1 at time t^* and thus the contribution to the Brier-Score is $(1 - \hat{S}(t|Z_i))^2$. All observations in the third category are censored and occur before t^*. Thus, the event status at t^* is unknown and the contribution to the Brier-Score is not defined and set to 0.

In order to compensate the loss of information due to censoring, the individual contributions have to be reweighted in a similar way as in the calculation of the Kaplan-Meier estimator (cf. Kaplan and Meier, 1958). The division by $\hat{S}_C(\cdot)$ in both terms of (4.9) displays that weighting scheme. Note that the Brier-Score is only defined for t with $S_C(t) > 0$. A detailed derivation of the Brier-Score is described in Graf *et al.* (1999).

Note that the Brier-Score is equal to 0.25 when the trivial prediction $\hat{S}(t) = 0.5$ is made for all individuals.

The Brier-Score as defined in (4.9) is calculated for all $t > 0$. In order to obtain an averaged value, we make use of the integrated Brier-Score (IBS), given by

$$IBS(t) = \frac{1}{\max(t_i)} \int_0^{\max(t_i)} BS(t) \, \mathrm{d}t, \tag{4.10}$$

as a score to assess the goodness of the predicted survival functions of all observations at every time t between 0 and $\max(t_i)$, $i = 1, ..., n$.

Note that the IBS is also appropriate for prediction methods that do not involve Cox regression models: it is more general than the R^2 and the p-value criteria (cf. Graf *et al.*, 1999) and has thus become a standard evaluation measure for survival prediction methods (Binder and Schumacher, 2008a; Gerds and Schumacher, 2006, 2007; van Wieringen *et al.*, 2009).

For assessing the prediction performance in terms of the Brier-Score, it is a common agreement in the literature to present the values of the IBS as well as the run of the Brier-Score curve over time. These curves are also known as *prediction error curves* (see, e.g., Binder and Schumacher, 2008a; Gerds and Schumacher, 2007).

4.7.5 Example for Calculating the Brier-Score

The following artificial example illustrates the calculation of the Brier-Score in the classical survival setting with censored observations. For seven individuals, we observe the following minimum values of the event and censoring times $\tilde{T}_i = \min(T_i, C_i)$ for $i = 1, \dots, 7$:

$$2 \qquad 3+ \qquad 5+ \qquad 6 \qquad 7+ \qquad 9 \qquad 10+,$$

where + indicates a censored observation.

For easy reference, we estimate the survival function for the event times and the censoring times with the Kaplan-Meier estimate by ignoring the covariate information (cf. Section 4.2 and Graf *et al.* (1999)). In Figure 4.1 we calculate the basic survival measures and provide a graphical presentation of the Kaplan-Meier estimator $\hat{S}(t)$ on the right panel. To assess the quality of the Kaplan-Meier estimator at time $t > 0$, we calculate the Brier-Scores at each time point by measuring the average discrepancy between the true event status and the estimated predicted value. In order to get an impression of how the calculation of the Brier-Score for each time t is done, in Table 4.1 we present exemplarily the calculation of the Brier-Score at time $t^* = 6$. Columns 6 and 9 are set in green and red color to discriminate between events that already occurred before t^* and that may occur later. The corresponding values are calculated

t_i	Y_i	d_i	$\hat{S}(t)$
2	7	1	$\frac{6}{7}$
3	6	0	$\frac{6}{7}$
5	5	0	$\frac{6}{7}$
6	4	1	$\frac{6}{7} \cdot \frac{3}{4} = \frac{9}{14}$
7	3	0	$\frac{9}{14}$
9	2	1	$\frac{9}{14} \cdot \frac{1}{2} = \frac{9}{28}$
10	1	0	$\frac{9}{28}$

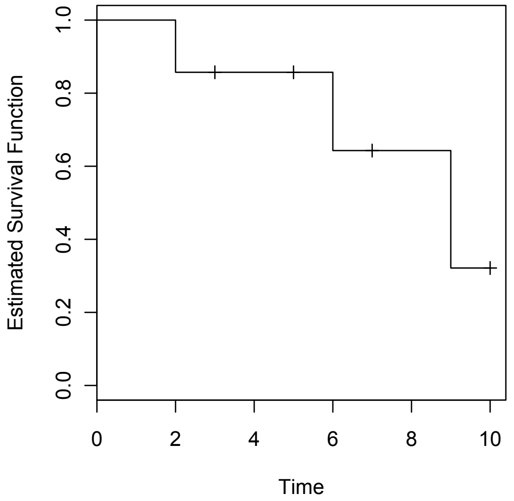

Figure 4.1: Table of basic survival measures: event times t_i, number of individuals at risk Y_i and number of events d_i at time t_i (left panel), and the Kaplan-Meier estimator with its run of the curve (right panel). A + indicates a censored observation.

with the following formula (see equation 4.9):

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(0-\hat{S}(t|Z_i))^2\mathbf{1}(\tilde{T}_i \leq t)\delta_i}{\hat{S}_C(\tilde{T}_i)} + \frac{(1-\hat{S}(t|Z_i))^2\mathbf{1}(\tilde{T}_i > t)}{\hat{S}_C(t)}\right].$$

According to the splitting in the three different categories (see Section 4.7.4), we only need to calculate one summand of the Brier-Score because the other one is zero. In the case where the minimum of event and censoring time $\tilde{T}_i$ is less or equal than the time point t^*, the first, green marked summand is calculated for individual i. If $\tilde{T}_i$ is greater than t^*, the second, red marked summand is calculated for individual i. For all individuals with known event status at t^*, we are able to calculate the squared differences of the estimated survival probability and the event status. The corresponding values are shown in column six of Table 4.1. The individuals number 2 and 3 are censored before t^*, thus their event status is unknown at t^* and their actual weight according to the estimated censoring distribution $\hat{S}_C$ is uniformly distributed across all successive observations. The contributions of all seven individuals are summed up and divided by the number of individuals. The resulting value of the Brier-Score at t^* is $BS(t^*) = BS(6) \approx 0.2296$ and thus less than 0.25, corresponding

Table 4.1: Calculation of the Brier-Score $BS(6)$ at time point $t^* = 6$.

i	$\tilde{T}_i$	δ_i	$\hat{S}(\tilde{T}_i)$	$\hat{S}_C(\tilde{T}_i)$	$\left(\mathbf{1}(\tilde{T}_i > t) - \hat{S}(t)\right)^2$	weights		ith summand
1	2	1	$\frac{6}{7}$	1	$(0 - \frac{9}{14})^2$			$(0 - \frac{9}{14})^2 \cdot \frac{1}{1}$
2	3+	0	$\frac{6}{7}$	$\frac{5}{6}$	-	↴		0
3	5+	0	$\frac{6}{7}$	$\frac{2}{3}$	-	↓	↴	0
4	6	1	$\frac{9}{14}$	$\frac{2}{3}$	$(0 - \frac{9}{14})^2$	$+\frac{1}{4}$	$+\frac{1}{4}$	$(0 - \frac{9}{14})^2 \cdot \frac{1}{2/3}$
5	7+	0	$\frac{9}{14}$	$\frac{4}{9}$	$(1 - \frac{9}{14})^2$	$+\frac{1}{4}$	$+\frac{1}{4}$	$(1 - \frac{9}{14})^2 \cdot \frac{1}{2/3}$
6	9	1	$\frac{9}{28}$	$\frac{4}{9}$	$(1 - \frac{9}{14})^2$	$+\frac{1}{4}$	$+\frac{1}{4}$	$(1 - \frac{9}{14})^2 \cdot \frac{1}{2/3}$
7	10+	0	$\frac{9}{28}$	0	$(1 - \frac{9}{14})^2$	$+\frac{1}{4}$	$+\frac{1}{4}$	$(1 - \frac{9}{14})^2 \cdot \frac{1}{2/3}$
								$\mathbf{BS(6) = \frac{1}{7}\sum_{i=1}^{7} \dots \approx 0.2296}$

to the trivial prediction. The Brier-Scores for all other time points are calculated analogously and are given by

$$0.1224 \qquad 0.1224 \qquad 0.1224 \qquad 0.2296 \qquad 0.2296 \qquad 0.2181 \qquad 0.0701.$$

For the first three time points and the last one, the Kaplan-Meier estimator shows a better prediction performance than for time points in the middle. The average prediction performance is calculated with the help of the integrated Brier-Score IBS, given in equation (4.10):

$$IBS = \frac{1}{t_{\max}} \int_0^{t_{\max}} BS(t)\,\mathrm{d}t = \frac{1}{10} \int_0^{10} BS(t)\,\mathrm{d}t \approx 0.1578.$$

5 Results

We performed the analysis of two high-dimensional breast cancer data sets with the help of the free software environment for statistical computing R (R Development Core Team, 2011) in version 2.14.0. At this point we want to highlight that the proposed methods are computationally intensive. Due to the preclustering approach, the 100 splits into training and test data and the cross-validation procedure for obtaining the optimal tuning parameter λ, all computations were performed on the LiDOng high performance computing cluster of TU Dortmund University with 432 nodes and up to 64 GB RAM per node. The calculation takes several weeks to accumulate all results for one high-dimensional data set. When we make use of additional R-packages, we refer to them at the corresponding points in the following sections.

At the beginning of this chapter, in Section 5.1 we present a descriptive analysis of the two data sets including survival and censoring times as well as mappings from genes to Gene Ontology (GO) groups. Section 5.2 gives an overview of the course of the extensive evaluation procedure with a focus on cluster analysis and aggregation of (preclustered) gene groups to representative covariates. A comparison of the developed models is presented for both data sets in Section 5.4. The main focus is on the analysis and assessment of the most promising combinations of selection and aggregation methods. Finally, in Section 5.5 we show an exemplary result according to the annotation of genes within GO groups and their biological functions for the most frequently chosen covariates.

5.1 Descriptive Analysis of Data Sets

With the help of the R-package `survival` (Therneau and Lumley, 2009) we show in Figure 5.1 Kaplan-Meier estimators with pointwise 95 % confidence intervals for the two high-dimensional breast cancer data sets. The beginning of both studies is defined by the date of surgery and censoring is due to missing information from follow up examinations. Thus, the survival times are right-censored. The censoring rate is 73.2 % for the DBC data set and 76.5 % for the MC study. Due to the large amount of censored observations a median survival time cannot be calculated.

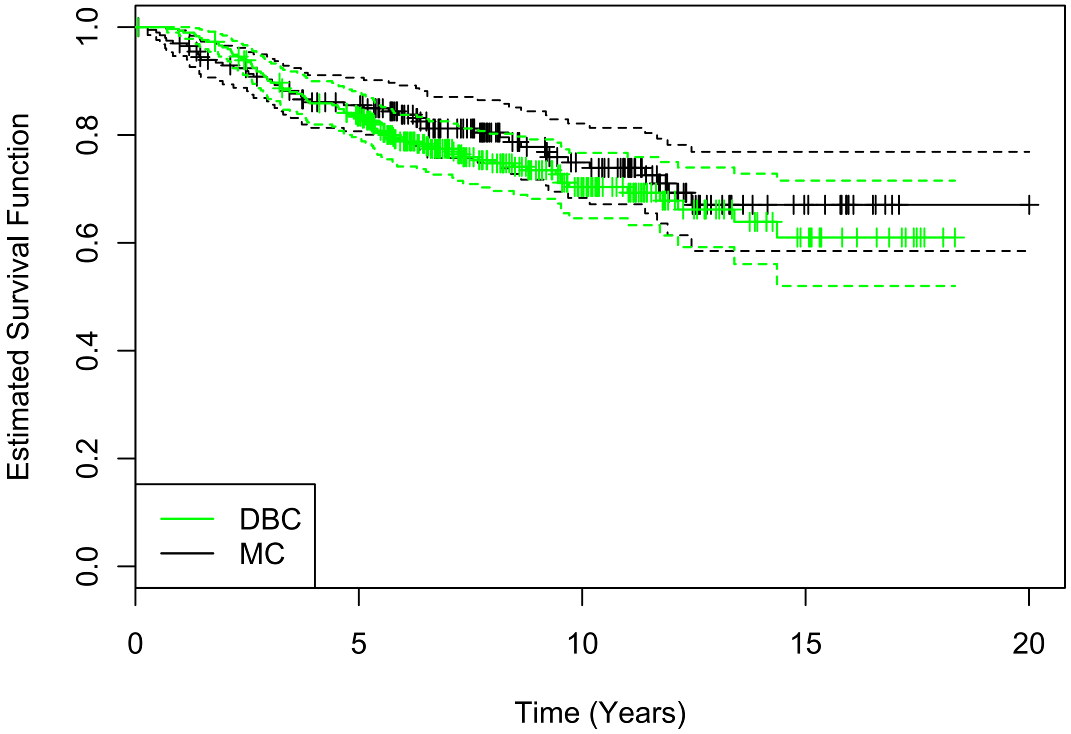

Figure 5.1: Kaplan-Meier estimators $\hat{S}(t)$ for survival functions with corresponding 95 % confidence intervals for DBC data set and MC study.

We utilize the R-package `topGO` (Alexa and Rahnenführer, 2009) in combination with the package `hgu133a.db` (Carlson *et al.*, 2009), both available from the Bioconductor-Repository, for mapping genes to GO groups. Genes that are not yet annotated to at least one GO group are eliminated from the data sets. The resulting 1876 genes of the DBC data set are assigned to 5560 GO groups, and 17 643 genes of the MC study are annotated to 8587 GO groups.

Table 5.1: Basic data information of the two data sets.

data set	patients	events	all genes	annotated genes	GO groups
DBC	295	79	4919	1876	5560
MC	200	47	22283	17643	8587

Figure 5.2 shows with boxplots the number of genes contained in each GO group for the two data sets. The median group size of the DBC data (2 genes) is considerably smaller than the median group size of the MC study (8 genes). The differences become even greater when considering the mean group size: 16.5 for the DBC data set and 98.9 for the MC study. The smaller number of genes in the GO groups in the DBC data set are a direct consequence of the data preprocessing (see Section 2.4): we use less than one fifth of the size of the original data set. Thus small GO groups are over-represented. Table 5.1 summarizes this basic data information.

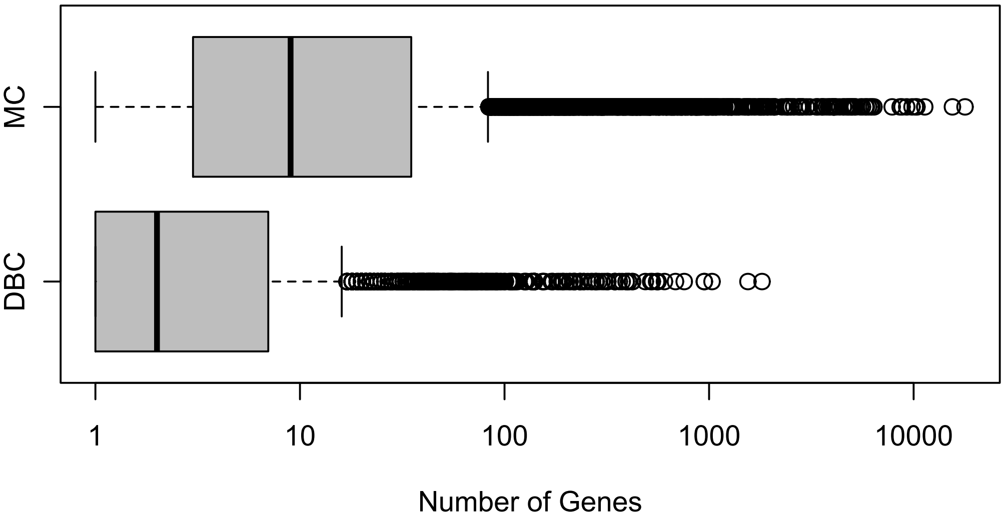

Figure 5.2: Boxplots of the number of genes included in GO groups in the DBC data set and the MC study

At this point, we have to mention that the complete analysis is performed on the data sets separately. A joint analysis of both data sets is not feasible due to different data preprocessing procedures (see Section 2.4) and the consequential different representation with GO groups.

5.2 Analysis Steps

At first, we present the chronology of the analysis which is identical for both data sets. The starting point is a $(p \times n)$ gene expression matrix Z. The entry in the jth row and ith column corresponds to the expression value of gene j from patient i. In addition, the two data sets contain clinical information for each patient. The clinical covariates for the DBC data set are age at diagnosis, tumor size and tumor grade as well as the number of nodes. For the MC study the clinical covariates are age at diagnosis, tumor size and tumor grade as well as the estrogen receptor status. For all patients, we have information concerning their survival times and event status. In the following, when referring to the 'clinical model', we make use of models that contain the four covariates for each data set, respectively. In the 'genetic models' single genes, gene groups as well as preclustered gene groups are used as covariates. The last type of model is the 'clinical-genomic model' where we combine the genomic models with the clinical model. In order to calculate a representative gene expression matrix for GO groups and preclustered GO groups, we present clustering and aggregation steps in the following paragraphs.

Cluster Analysis At first, gene expression data is annotated to GO groups as described in Section 5.1. With the help of Partioning Around Medoids clustering (see Section 3.1), we search for correlated subgroups within the GO groups. The `R`-package `cluster` (Maechler *et al.*, 2005) provides the corresponding algorithms. We calculate the dissimilarity matrix D (see Section 3.1)

$$D = 1 - \mathrm{Cor}\left(Z^\top\right) \ \in \mathbb{R}^{p \times p},$$

with the help of the gene expression matrix Z ($Z \in \mathbb{R}^{p \times n}$). The distance between two column vectors is calculated with the empirical Pearson correlation coefficient (see Section 3.1): if two gene vectors (columns of $Z^\top$) are highly positive correlated, their distance is close to zero, if they are uncorrelated, their distance is one, and if they are highly negative correlated, their distance is two. The optimal number of clusters K is chosen by the maximum Intra Cluster Correlation. This method is also introduced in

Section 3.1. Due to high computational costs the maximum number of clusters within one GO group is limited to 20 clusters. Without a limit, if we consider a GO group with more than 10 000 genes, we have to calculate more than $1.66 \cdot 10^{11}$ swaps in one iteration of the PAM-clustering algorithm.

Permutation Test The permutation test (see Section 3.2) with $N_p = 10^4$ permutations is performed in advance of the clustering. All genes that do not show a positive correlation with at least one gene of the same GO group at the 5 % significance level are defined as single clusters and the PAM-clustering algorithm is performed on the remaining genes. If no or one gene remains in a GO group, this group cannot be summarized. If a GO group contains two genes, a clustering is also not feasible. To solve this problem, we suggest an additional permutation test with $N_p = 10^4$ permutations. This procedure is heuristic and should only be used as an alternative approach for PAM-clustering if the number of genes within a GO group is small (necessary for one or two genes and recommended for up to 5 genes).

Aggregation A (preclustered) gene group must be appropriately summarized in order to obtain one representative value for each patient and each group. These aggregated covariates are particularly suitable for survival models in the further analysis. Simple methods for aggregation, e.g. the arithmetic mean, the median or the medoid gene, are described in Section 3.3. They yield similar results due to the standardized data sets.

Another method to reduce data dimensionality is Principal Component Analysis (see Section 3.4) that finds a new (smaller) set of covariates that represent the variation in the data set, given by the correlations between the original covariates. The Principal Component Analysis (PCA) is based on singular value decomposition and is available via the function `prcomp()` within the `R` standard package `stats`.

An extensive comparison of these aggregation methods is performed in Lang's diploma thesis (Lang, 2010). He showed that the most promising and suitable method for summarizing gene expression values within one gene group is PCA with the first principal component. Thus we summarize the gene expression measurements from all genes belonging to one GO group or to one cluster via the first principle component of all genes that belong to this gene group or cluster. Note that the first principal component and thus the parameter estimates must be interpreted with caution. It is uniquely determined according to its direction but not to its sign. In our analysis we select positively correlated covariates. We choose the sign of the first principal component such that the direction is in accordance with the correlated covariates.

Reference Models In order to assess the merit of the preclustering approach, we present results for models using only genes or only GO groups as explanatory variables and also combine the genomic information with the clinical data. In order to obtain a fair comparison of models with different types of genomic covariates, we only use those genes that are annotated to GO groups. Figure 5.3 illustrates the paths from a given data matrix via the clustering and aggregation methods to the final different types of covariates that are used in Cox models for further analysis.

Duplicated covariates Prior to fitting survival models it is necessary to remove duplicated covariates in order to avoid failures when inverting matrices with linear dependent rows. Duplicates result from identical GO groups or from identical subgroups after preclustering. If two or more identical covariates occur, we remove the duplicates from the covariate matrix and save the removed covariates separately for further investigation (if necessary).

Variable selection and evaluation After clustering, aggregating and removing duplicated covariates, we obtain an appropriate data matrix for each type of genomic covariates (genes, GO groups, preclustered GO groups).

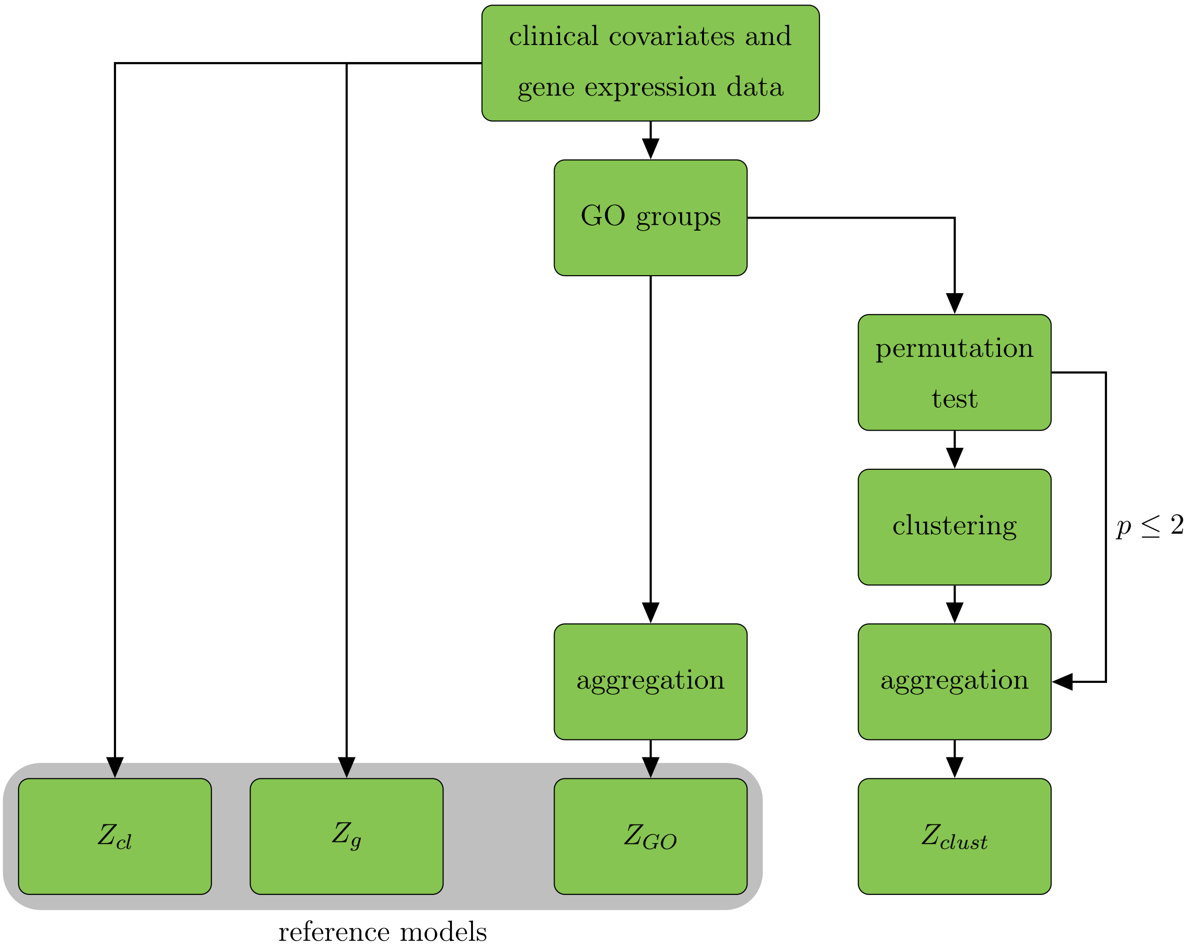

Figure 5.3: Analysis steps: Z_{cl} $\widehat{=}$ clinical covariates, Z_g $\widehat{=}$ genes as covariates, Z_{GO} $\widehat{=}$ aggregated GO groups as covariates, Z_{clust} $\widehat{=}$ preclustered and aggregated GO groups as covariates; p $\widehat{=}$ number of covariates.

The results for each of the two data sets after applying the evaluation procedure described in Section 4.7.1 are presented in the following sections. To reduce bias, we have to consider several splits of the data into training and test set due to the dependence of the results on such a split. In Section 5.3, we present detailed results for one specific random split and in Section 5.4 a comprehensive analysis summarizing 100 random splits. We split n breast cancer patients into training set and test set, where 2/3 of the patients are assigned to the training set and 1/3 to the test set. We use the training data for estimating the tuning parameter $\hat{\lambda}_{\text{train}}$ and the regression coefficients $\hat{\beta}_{\text{train}}$ and $\hat{\gamma}_{\text{train}}$ and the test data for evaluation. For each split into training data and test data, we calculate the three evaluation criteria on the test set (see Section 4.7). The results are compared with boxplots and prediction error curves.

In detail, we apply two classical variable selection methods, univariate and forward selection (see Section 4.6.3), and three shrinkage and dimension reduction procedures, lasso- and ridge-regression (see Section 4.6.1) as well as the CoxBoost algorithm (see Section 4.6.4), to the different types of genomic covariates of the training data sets. The optimal tuning parameter $\hat{\lambda}_{\text{train}}$ is determined with the K-fold cross-validated log partial likelihood which is described in Section 4.6.2. According to Bøvelstad *et al.* (2007) we apply $K = 10$ fold cross-validation. Due to the small number of clinical covariates, the shrinkage and dimension reduction procedures are only applied to the genomic covariates when considering the combined *clinical-genomic* models and the pure *genomic* models. For evaluating the prediction performance, we calculate the prognostic indices/risk scores (see Section 4.5) for all patients in the test set. With the help of the prognostic indices, we apply the log-rank test (see Section 4.4 and Section 4.7.2) on the test set and use the p-value as an evaluation criterion for the usefulness of the grouping. The prognostic index is also used as a single continuous covariate to assess the prediction performance of the fitted Cox model (see Section 4.7.3). In addition to these two test-based evaluation types, we calculate the integrated Brier-Score for each split into training and test data (see Section 4.7.4). We present the results for 100 splits with boxplots and with prediction error curves that show the course of the Brier-Scores over time.

For lasso- and ridge-regression we make use of several functions from the `R`-package `penalized` (Goeman, 2010b) that are described in detail in Goeman (2010a). The CoxBoost algorithm implemented in the `R`-package `CoxBoost` (Binder, 2011) and the `R`-package `survival` (Therneau and Lumley, 2009) provides the basis functions for univariate and forward selection in the survival context. The Brier-Score is implemented in the `R`-package `ipred` (Haibe-Kains *et al.*, 2010).

In the following sections, we present at first an exemplary analysis of one split into training and test data and afterwards a comprehensive analysis of 100 splits. Analysis of frequently chosen covariates across the 100 splits is essential to assess the stability of the fitted survival models.

5.3 Exemplary Analysis: One Split into Training and Test Data

We apply model selection methods and three evaluation criteria to one specific random split of the Mainz cohort study into training and test data to illustrate how the model building and evaluation are performed (as explained in Section 4.7.1). We split the 200 breast cancer patients into training set and test set, where 2/3 of the patients (in this case 133) are assigned to the training set and 1/3 (here 67) to the test set. We use the training data for estimating the tuning parameter $\hat{\lambda}_{\text{train}}$ and the regression coefficients $\hat{\beta}_{\text{train}}$ and $\hat{\gamma}_{\text{train}}$ and the test data for evaluation. Table 5.2 shows the results for evaluation criteria when using genes, GO groups, or preclustered gene groups as covariates. The values p_{LR} and p_{PI} correspond to the p-values derived from the log-rank test and the prognostic index, respectively. The IBS is the value of the integrated Brier-Score.

This example indicates that the predictive performance of models built with GO groups alone and of models with preclustered GO groups is comparable with classical models using only genes as covariates. The p-values for model assessment are similar, but in addition, we have more information in the final model; annotations of preclustered GO groups can help clinicians to investigate the selected genes according to their biological function.

For all three types of genomic covariates the two prognostic groups are clearly separated on the test data, with significant differences in overall survival ($p < 0.02$) between the high-risk group and the low-risk group for lasso-regression (see p_{LR} and p_{PI}). The separation between the two groups is best when using a model containing preclustered GO groups ($p = 0.0092$).

Due to the high censoring, especially at the end of the studies, the integrated Brier-Scores are calculated up to 10 years follow-up. They result in a value of approximately 0.10 for all methods with favoring the preclustered models.

For ridge-regression, all covariates are kept in the model since parameter estimates are unlikely to get shrunken exactly to 0. The number of covariates for lasso-regression

Table 5.2: One random split into training and test data for the Mainz cohort study: Results for the prediction methods using (i) genes, (ii) GO groups, and (iii) preclustered GO groups. For ridge-regression, all covariates are kept in the model since parameter estimates are unlikely to get shrunken exactly to 0. LR $\widehat{=}$ log-rank test, PI $\widehat{=}$ prognostic index, IBS $\widehat{=}$ integrated Brier-Score, λ $\widehat{=}$ tuning parameter, sel.cov $\widehat{=}$ number of selected covariates; L_1 $\widehat{=}$ lasso-regression, L_2 $\widehat{=}$ ridge-regression.

Method	**Covariates**	$\boldsymbol{p}_{\mathrm{LR}}$	$\boldsymbol{p}_{\mathrm{PI}}$	$\boldsymbol{IBS}$	λ	**sel.cov**
L_1	genes	0.0190	0.0017	0.1042	11.72	19
L_1	GO	0.0176	0.0018	0.1103	10.75	16
L_1	clustered	0.0092	0.0002	0.0830	28.53	5
L_2	genes	0.0098	0.0003	0.0877	5112.08	17834
L_2	GO	0.0541	0.0097	0.1022	11749.16	6530
L_2	clustered	0.0690	0.0006	0.0896	96499.04	31229

range from 5 for preclustered GO groups to 19 for genes. These models fulfill the requirement of sparseness.

From only one split of the data into training and test sets, we will not know to which extent the resulting criteria values depend on the actual training/test randomization. In the next sections, we present a comprehensive analysis of 100 random splits.

For illustration of the results presented in Table 5.2 we show Kaplan-Meier curves for two prognostic groups of patients derived by dividing all patients according to the median prognostic index of the patients in the test set. Here we use lasso-regression for model selection and the log-rank test for evaluation. We compare models with genes, GO groups, and preclustered GO groups as covariates (see Figure 5.4).

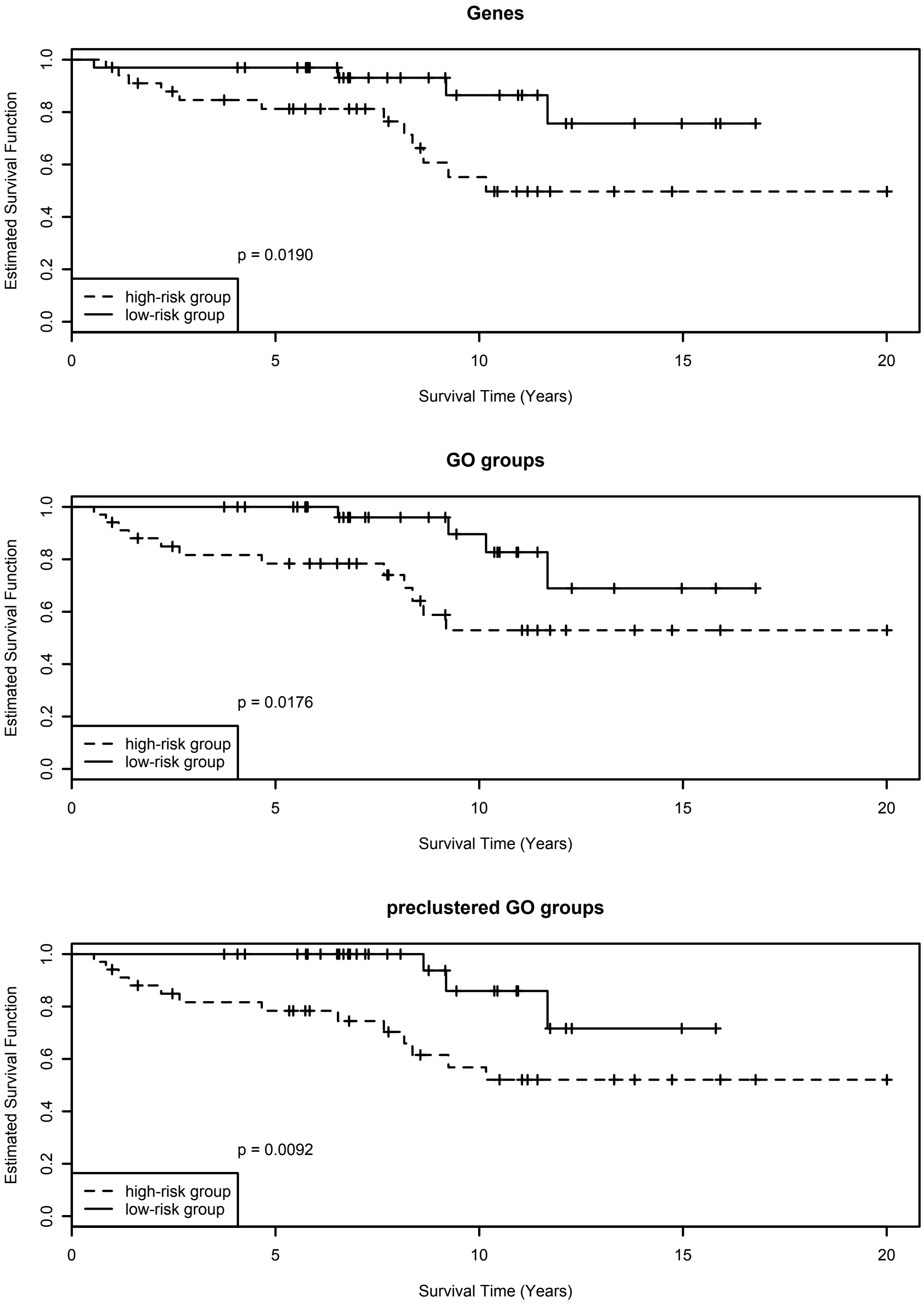

Figure 5.4: **Mainz cohort study:** Kaplan-Meier curves for the high-risk and low-risk groups defined by the estimated prognostic indices of the 67 patients in the test data set, the cutoff is defined as the median prognostic index on the test data. Genes, GO groups, and preclustered GO groups are used as covariates, respectively, and lasso-regression is applied as model selection method.

5.4 Comparison of Selection Methods

We observe high variability of the chosen tuning parameters and the parameter estimates depending on the split into training and test data. In order to quantify which covariates are consistently selected in different splits and how stable the evaluation measures are, we calculate results for 100 random splits and compare the selected genes and GO groups.

In Figures 5.5 (DBC) and 5.6 (MC), we present boxplots for the results for the two breast cancer data sets, after applying the evaluation procedure to the five model building procedures (lasso- and ridge-regression, CoxBoost, univariate and forward selection) for each of the three types of genomic covariates (genes, GO groups, preclustered GO groups). Results for the clinical model are presented as a reference. We consider the median of the 100 values obtained from our prediction performance criteria as the outcome of main interest. For easy reference we present all median values in Tables 5.3 and 5.4 for the DBC data set and the MC study, respectively. Best performance values are highlighted in boldface.

Rows of the figures correspond to two model evaluation criteria, the prognostic index and the integrated Brier-Score, and the columns correspond to two types of models: the genomic model and the genomic model combined with clinical covariates that are mandatory. Results for the log-rank test are nearly the same as for the prognostic index and therefore not shown here. In both figures we show the results for the five model selection methods. The p-values for the prognostic index (cf. Section 4.7.3) are shown on the $-\log_{10}$ scale (a value of 2, e.g., corresponds to a p-value of 0.01). Thus, large values correspond to good prediction performance. For the integrated Brier-Score small values correspond to good prediction performance. The reference is a Brier-Score of 0.25, for a random estimation (cf. Section 4.7.4). Due to the high censoring, especially at the end of the studies, the integrated Brier-Scores are calculated up to 10 years follow-up. For both evaluation criteria in all plots the horizontal line at the median indicates the reference model containing only clinical information.

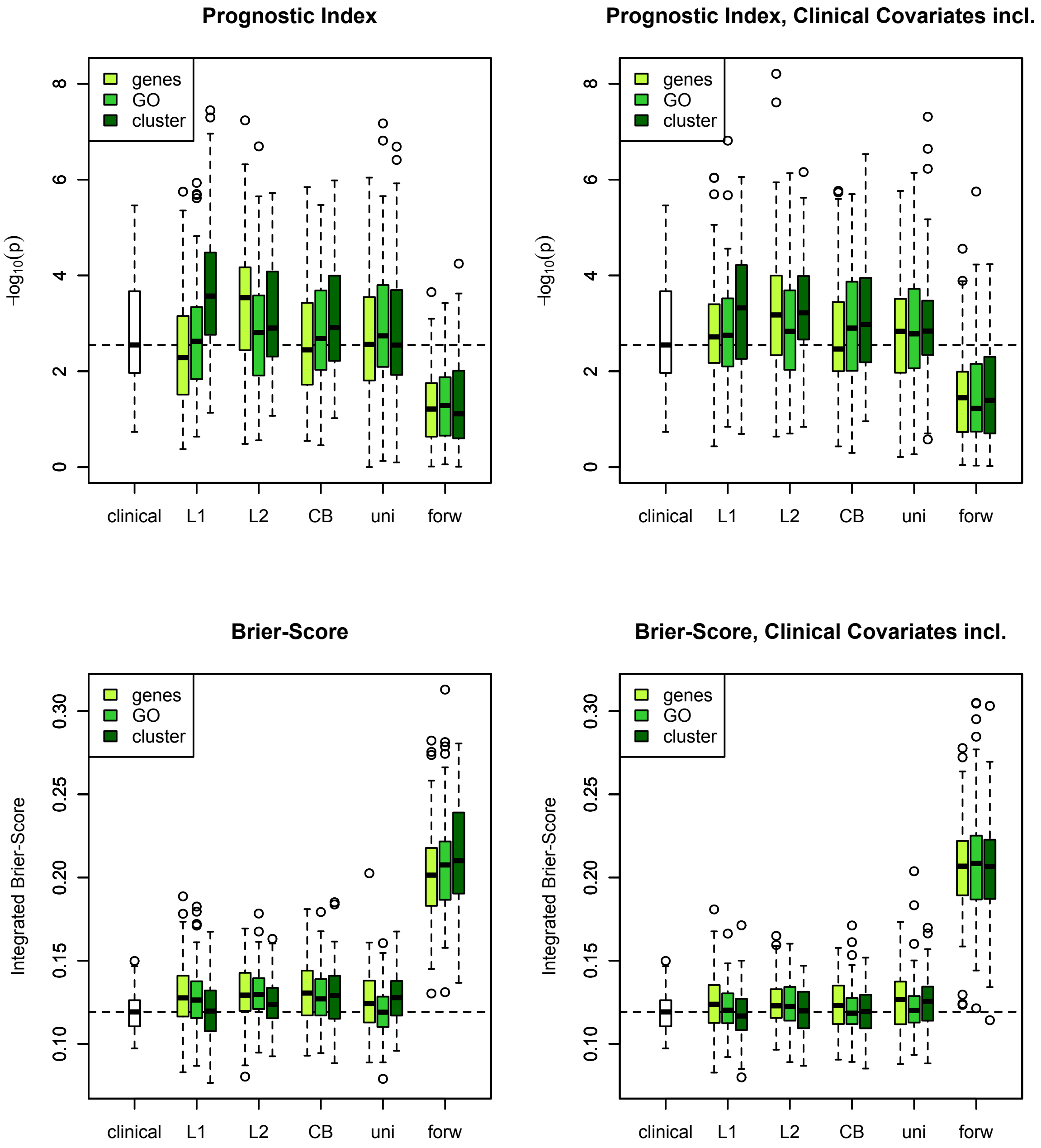

Figure 5.5: **Dutch breast cancer data set:** The boxplots show results for all model building procedures applied to 100 training/test splits for genes, GO groups (GO), and preclustered GO groups (cluster) for the Dutch breast cancer data set. P-values of the prognostic index are presented on $-\log_{10}$ scale such that large values correspond to good prediction performance. The Brier-Scores are calculated for 10 years follow-up. Small values of the integrated Brier-Score correspond to good prediction performance. $L_1 \,\widehat{=}\,$ lasso-regression, $L_2 \,\widehat{=}\,$ ridge-regression, $CB \,\widehat{=}\,$ CoxBoost, $uni \,\widehat{=}\,$ univariate selection, $forw \,\widehat{=}\,$ forward selection.

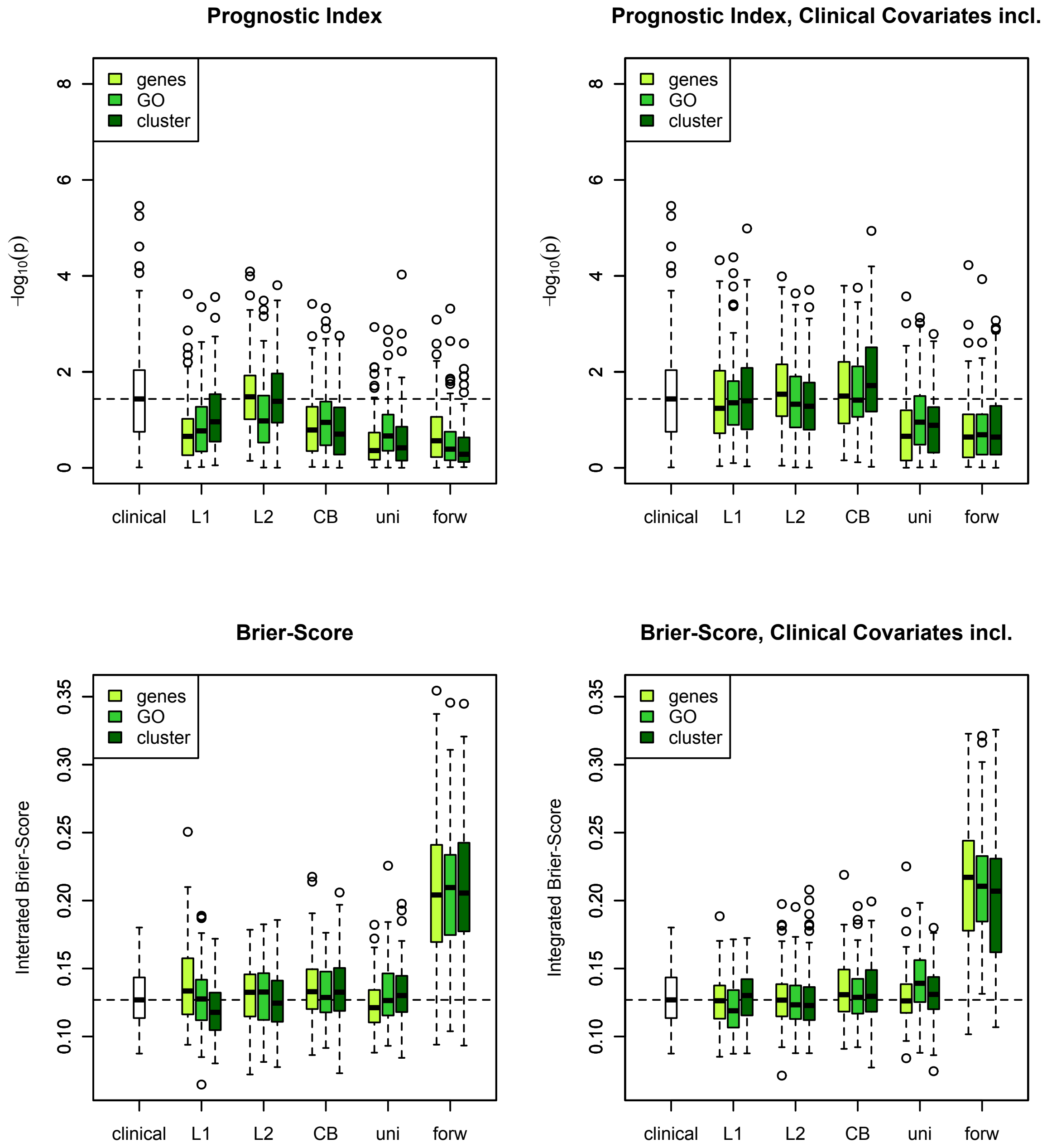

Figure 5.6: **Mainz cohort study:** The boxplots show results for all model building procedures applied to 100 training/test splits for genes, GO groups (GO), and preclustered GO groups (cluster) for the Mainz cohort study. P-values of the prognostic index are presented on $-\log_{10}$ scale such that large values correspond to good prediction performance. The integrated Brier-Scores are calculated for 10 years follow-up. Small values of the integrated Brier-Score correspond to good prediction performance. $L_1 \widehat{=}$ lasso-regression, $L_2 \widehat{=}$ ridge-regression, $CB \widehat{=}$ CoxBoost, $uni \widehat{=}$ univariate selection, $forw$ $\widehat{=}$ forward selection.

Investigating Figure 5.5 (DBC) and Figure 5.6 (MC) we first note, that there is a fairly large spread of values over the 100 splits. This is due to the variation caused by splitting the data at random into training and test sets as well as to the variation in the performance of the prediction methods for given splits.

For all different model building and evaluation settings, forward selection has the poorest prediction performance. It has been shown in several publications that forward selection has problematic performance (see, e.g. Bøvelstad *et al.*, 2007, 2009). Lasso- and ridge-regression often outperform this standard selection method. Univariate selection is in most of the cases slightly inferior compared to penalty models.

In the boxplots of Figure 5.5, in terms of the likelihood-ratio test for the prognostic index, all models (except lasso-regression and CoxBoost with genes as covariates) have an increased prediction performance compared to the clinical model and are in median significant at the 0.01 significance level. The upper left and upper right panel of Figure 5.5 show that lasso-regression with preclustered GO groups (median p-value across 100 splits is 0.0003) and in combination with clinical covariates (median p-value of 0.0006) has the best prediction performance for the DBC data set. Only ridge-regression with genes as covariates has similar results.

This result does not hold for the integrated Brier-Score for this data set. Here, all methods provide comparable prediction performances (the median of the IBS is approximately 0.12), even though the results indicate an advantage of the clinical-genomic model for lasso-regression with preclustered GO groups ($IBS = 0.1168$). We also note that variance for clinical models is lower than for any integration of genomic covariates. As the integrated Brier-Score is the average across the Brier-Scores calculated at all time points a further analysis of the Brier-Score at different time points is necessary and presented at the end of this section.

In the Mainz cohort study, we see the same result for the genomic models using the integrated Brier-Score for evaluation (see the lower panels of Figure 5.5). In fact, the Brier-Score favors the lasso-regression with genetic covariates ($IBS = 0.1170$). It is noticeable that for the MC study and prognostic index as performance measure, the model using only genomic information is worse than the clinical model (Figure 5.5, upper left), but the clinical-genomic model is comparable to the clinical model. The

Table 5.3: **Dutch breast cancer data set:** Summary of Figure 5.5. The table shows median values for results for all model building procedures applied to 100 training/test splits for all types of covariates. Notation is according to Figure 5.5. Best performance values are highlighted in boldface.

Method	**Covariates**	**L_1**	**L_2**	**CB**	**uni**	**forw**
PI	clinical	0.0028	0.0028	0.0028	0.0028	0.0028
PI	genes	0.0052	**0.0003**	0.0036	0.0027	0.0615
PI	GO	0.0024	0.0016	0.0021	0.0018	0.0515
PI	cluster	**0.0003**	0.0013	0.0012	0.0028	0.0772
PI	clinical+genes	0.0019	0.0007	0.0034	0.0015	0.0356
PI	clinical+GO	0.0018	0.0015	0.0013	0.0017	0.0595
PI	clinical+cluster	0.0005	0.0011	0.0006	0.0015	0.0403
IBS	clinical	0.1192	0.1192	0.1192	0.1192	0.1192
IBS	genes	0.1277	0.1293	0.1306	0.1243	0.2014
IBS	GO	0.1264	0.1297	0.1271	0.1190	0.2076
IBS	cluster	0.1197	0.1236	0.1291	0.1278	0.2101
IBS	clinical+genes	0.1238	0.1229	0.1232	0.1267	0.2068
IBS	clinical+GO	0.1203	0.1225	0.1185	0.1202	0.2085
IBS	clinical+cluster	**0.1168**	0.1199	0.1194	0.1256	0.2067

Table 5.4: **Mainz cohort study:** Summary of Figure 5.6. The table shows median values for results for all model building procedures applied to 100 training/test splits for all types of covariates. Notation is according to Figure 5.6. Best performance values are highlighted in boldface.

Method	**Covariates**	**L_1**	**L_2**	**CB**	**uni**	**forw**
PI	clinical	0.0367	0.0367	0.0367	0.0367	0.0367
PI	genes	0.2208	0.0331	0.1627	0.4369	0.2730
PI	GO	0.1700	0.1053	0.1128	0.2166	0.4077
PI	cluster	0.1098	0.0411	0.1984	0.3831	0.5212
PI	clinical+genes	0.0575	0.0292	0.0320	0.2199	0.2281
PI	clinical+GO	0.0441	0.0472	0.0388	0.1117	0.2043
PI	clinical+cluster	0.0402	0.0520	**0.0193**	0.1298	0.2288
IBS	clinical	0.1270	0.1270	0.1270	0.1270	0.1270
IBS	genes	0.1335	0.1326	0.1330	0.1212	0.2042
IBS	GO	0.1276	0.1327	0.1287	0.1265	0.2096
IBS	cluster	**0.1178**	0.1246	0.1326	0.1302	0.2056
IBS	clinical+genes	0.1263	0.1268	0.1308	0.1262	0.2171
IBS	clinical+GO	0.1190	0.1233	0.1288	0.1392	0.2106
IBS	clinical+cluster	0.1302	0.1229	0.1296	0.1310	0.2070

best median prediction performance for the prognostic index is also provided by a model using preclustering information: CoxBoost based on the clinical-genomic model (median p-value of 0.0193).

For both data sets, we observe that methods built from preclustered GO groups as covariates perform better than models using only genes. The combination of clinical and genomic information does not always show better results than using the genetic covariates alone. By comparing the results for lasso- and ridge-regression, we observe a slightly better prediction performance for Cox models using the lasso-regression, especially for models with preclustered gene groups as covariates. Results for the CoxBoost algorithm and for univariate selection are slightly inferior to the lasso- and ridge-regression models.

The solutions for all methods except ridge-regression are always sparse, but the optimal tuning parameter varies considerably between the splits and thus the number of chosen covariates. In Figure 5.7 we present for both data sets boxplots for the number of chosen covariates for the different variable selection methods with different types of genetic covariates. The results for the clinical-genomic models are similar and not shown here. All methods provide sparse solutions and the numbers of chosen covariates are less than 50. The interquartile range for the number of chosen covariates for lasso-regression and for all three different types of covariates ranges approximately from 3 and 20 for the DBC data set and from 5 to 12 for the Mainz cohort study. In terms of the median, results from the CoxBoost approach consist of more covariates than lasso-regression for all types of covariates. There is a higher variance on the number of chosen covariates for the DBC data set for lasso models and for the MC study for the CoxBoost models. For both data sets univariate selection make use of less that 10 covariates in most of the splits across all three types of covariates. Forward selection choses in median 25 covariates for the DBC data set and 13 for the MC study. Across all different types of variable selection methods there is a tendency that the numbers of selected genes and preclustered GO groups are similar, whereas the numbers of selected GO groups are smaller. In average, lasso regression and the CoxBoost approach select more covariates compared to univariate selection, with a higher variance on the number of chosen covariates.

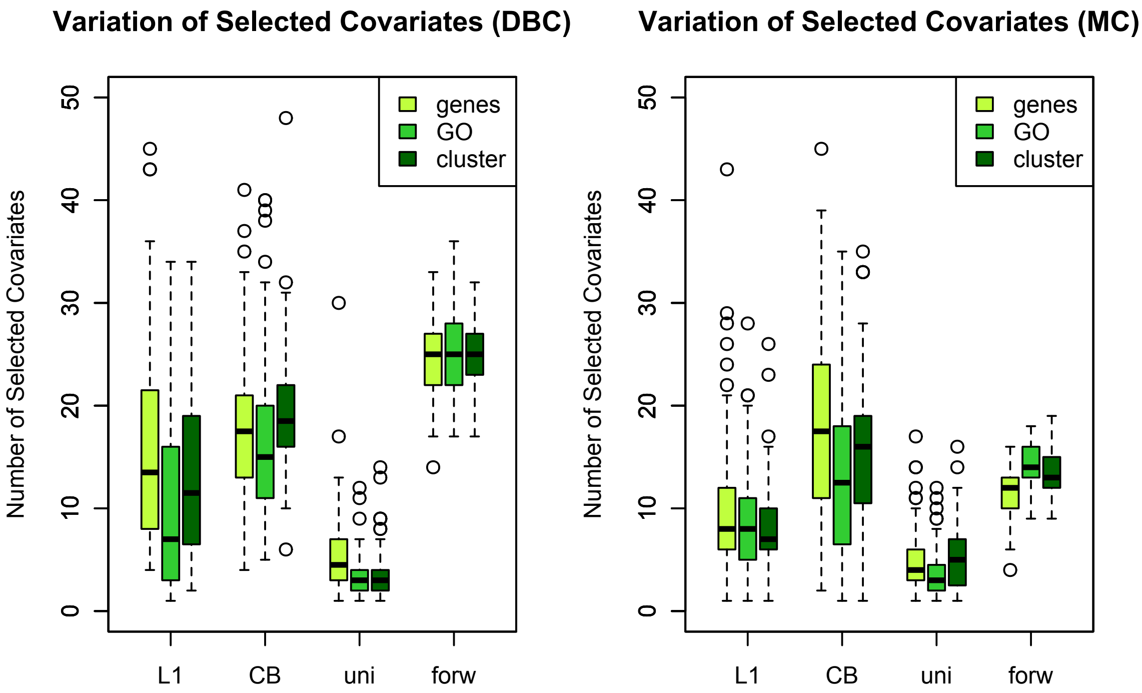

Figure 5.7: Boxplots showing the number of selected covariates for 100 training/test splits, models with genes, GO groups and preclustered GO groups, applied to the Mainz cohort study (MC) and the Dutch breast cancer data set (DBC). $L_1 \hat{=}$ lasso-regression, $CB \hat{=}$ CoxBoost, $uni \hat{=}$ univariate selection, $forw \hat{=}$ forward selection.

For a more detailed analysis of the results in terms of the Brier-Score, we have a closer look at the run of the curves of the Brier-Score over time for lasso models with preclustered GO groups in comparison to the other models. Prediction error curves (see, e.g. Gerds and Schumacher, 2006; Graf *et al.*, 1999) (averaged values for the Brier-Score calculated at each time point for 100 splits) for models with the three different types of genomic covariates are shown in Figure 5.8 and 5.9 for the DBC data set and the MC study, respectively.

The performance of the clinical model serves as reference. For both data sets, the model with preclustered GO groups has a better prediction performance over time in comparison to clinical models. The preclustered models outperform the clinical models, starting at four years for the DBC data set and at three years for the MC study. The other two genomic models are also inferior to the preclustered models.

This result shows that there is highly relevant information inside the preclustered gene groups for risk prediction of breast cancer patients, starting at 3 to 5 years follow-up. An analysis of frequently chosen covariates is provided in the next section.

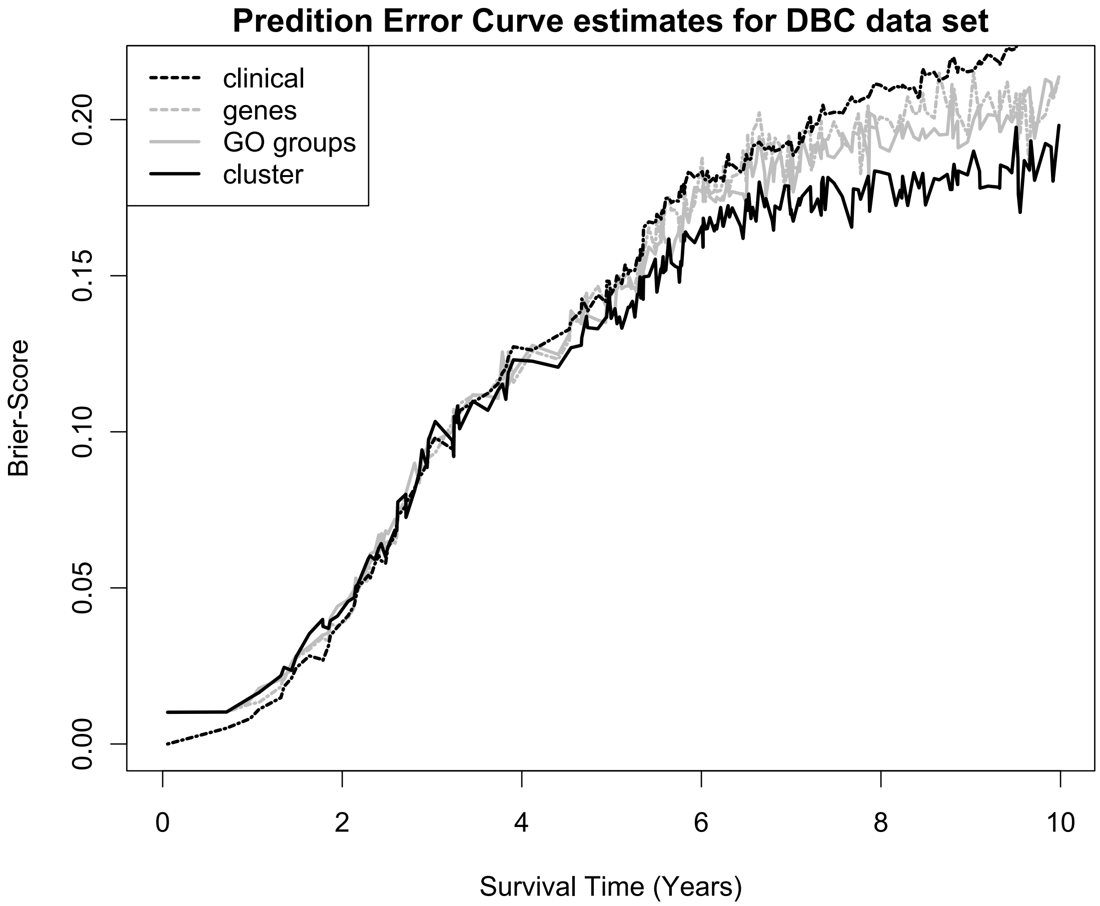

Figure 5.8: Prediction error curves for the DBC data set for the lasso evaluation procedure. We show averaged values for the Brier-Score calculated at each time point for 100 splits for models with the three different types of genomic covariates and the clinical model. A better prediction performance leads to lower curves.

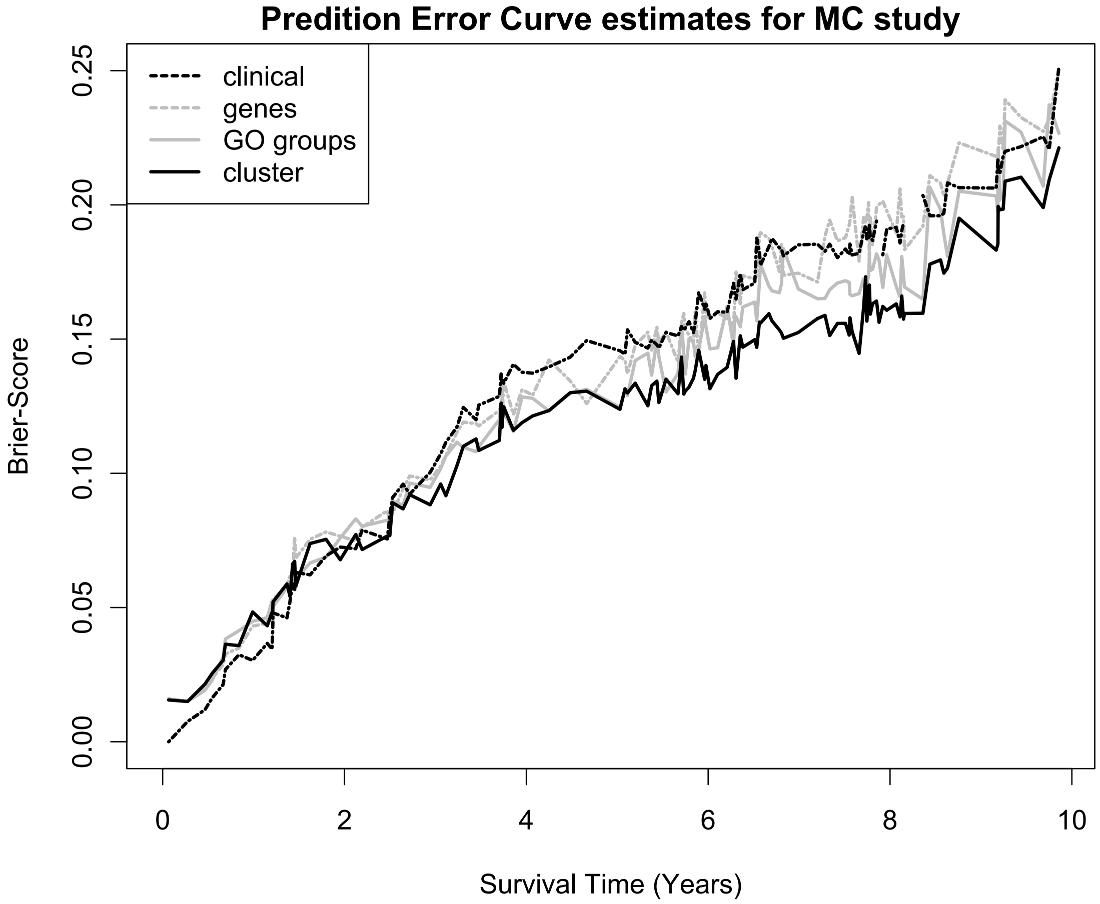

Figure 5.9: Prediction error curves for the MC for the lasso evaluation procedure. We show averaged values for the Brier-Score calculated at each time point for 100 splits for models with the three different types of genomic covariates and the clinical model. A better prediction performance leads to lower curves.

5.5 Analysis of Important GO Groups

In this paragraph, we present for one of the best models an exemplary analysis of most frequently selected covariates across all 100 splits. We consider the results for lasso-regression with preclustered gene groups as covariates for the Mainz Cohort study. The Due to the sparse solution for lasso models compared to ridge-regression we provide results for this method. Table 5.5 contains the numbers of the most frequently selected covariates, the corresponding GO groups with GO IDs (Ashburner *et al.*, 2000) and further information concerning the medoid gene, the cluster size and the

Table 5.5: Top 10 selected covariates for preclustered GO-groups according to 100 splits into training and test data: Probe set names for the medoid genes and GO IDs for GO groups. The first column corresponds to the selected number for the covariate across 100 splits into training and test data for L_1 regression on the Mainz cohort study. The value of the *effect* indicates whether the covariate has an increasing (+1) or decreasing (−1) effect on patients' risk to die.

count	GO	effect	medoid	clustersize	annotation
85	GO:0043170	+1	209258_s_at	410	macromolecule metabolic process
81	GO:0007049	+1	210052_s_at	222	cell cycle
74	GO:0050896	−1	211908_x_at	102	response to stimulus
52	GO:0032501	−1	212195_at	310	multicellular organismal process
40	GO:0032501	−1	210935_s_at	362	multicellular organismal process
21	GO:0050794	−1	210417_s_at	312	regulation of cellular process
18	GO:0043170	+1	211693_at	434	macromolecule metabolic process
18	GO:0050896	−1	204118_at	230	response to stimulus
16	GO:0006952	−1	203535_at	27	defense response
15	GO:0042221	+1	219140_s_at	39	response to chemical stimulus

annotation for the GO groups that are helpful for the biologist.

We observe that most of the chosen clusters are subgroups of large GO groups and consist of more than 100 genes. The value of the *effect* indicates whether a high value of the corresponding covariate has an increasing (+1) or decreasing (−1) influence on patients' risk to die. For a detailed analysis of the effects the boxplots in Figure 5.10 show the variation of the estimated regression coefficients in the Cox regression model for the most frequently chosen clusters, represented via medoid genes. First of all, the direction of the effect among all splits into training and test data is stable. From this it follows that a detected cluster has a consistent effect on patients' survival - either positive or negative. The first two clusters (from GO:0043170 and GO:0007049) shown in Table 5.5 are chosen in more than 80 % of the splits into training and test data. Their parameter estimates are positive, i.e. high expression values of the included genes lead to increased risk to die and thereby to shorter survival. In addition, the clusters at fourth and fifth position are contained in the same GO group. Thus the top 5 frequently chosen covariates for preclustered gene groups as covariates underline a very stable model selection procedure. At first view the direction of the effects are in accordance with the biological interpretation, e.g. high expression values of the genes within the cluster from GO group GO:0007049 (`cell cycle`) lead to shorter survival and high expression values of the genes within the cluster from GO:0006952 (`defense response`) to longer survival.

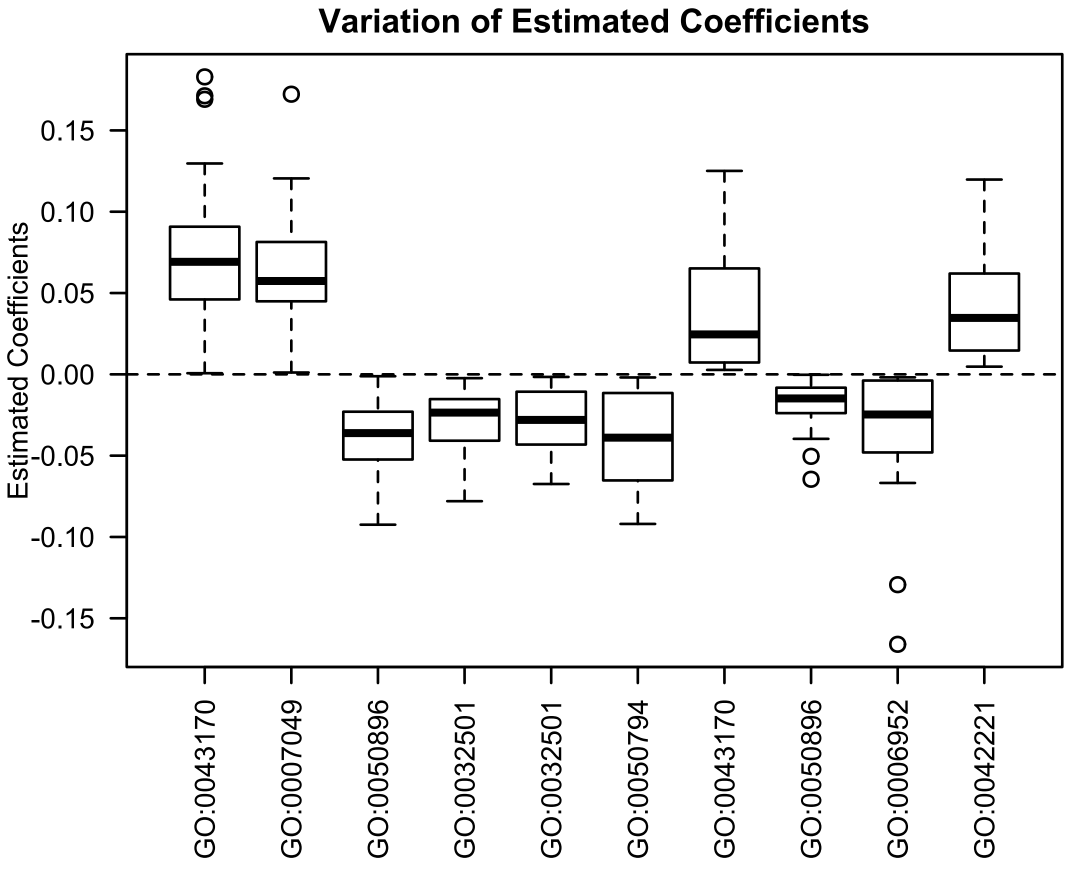

Figure 5.10: Boxplots show variation of estimated regression coefficients in the Cox regression model for the most frequently chosen clusters from Table 5.5, represented via medoid genes (probe sets).

6 Discussion and Conclusion

The typical challenge when relating survival times to gene expression measurements is a relatively small number of individuals compared to a large number of predictors. In this case the use of classical approaches is not possible.

For investigating the relationship between microarray gene expression data and censored survival data, we analyzed two published breast cancer data sets. We introduced methods for summarizing gene expression measurements with a focus on *preclustering* (Chapter 3) and the survival framework including the Cox model for high-dimensional data and algorithms for fitting and evaluating it (Chapter 4). We presented results for the evaluation procedure applied to these two data sets. Standard approaches focus on single genes as covariates (cf. Bøvelstad *et al.*, 2007; Gui and Li, 2005; Haibe-Kains *et al.*, 2008). We integrate additional biological knowledge by building models with preclustered GO groups as covariates.

In accordance with Bøvelstad *et al.* (2007), the lasso-regression method seems most suitable and promising: its prediction performance is slightly better compared to ridge-regression and the solution is sparse. Bøvelstad *et al.* (2007, 2009) show that ridge-regression performs better than all the other methods. In our analysis, ridge-regression leads in general to comparable but not better results compared to the lasso-regression. However, an important disadvantage of this method is that it does not select variables. The CoxBoost approach (Binder and Schumacher, 2008b) provides comparable results to lasso- and ridge-regression. Its solution is sparse and this procedure should always be considered as an alternative to lasso-regression. We observe relevant differences between high-risk and low-risk patients, but there are too many genes or GO groups to be further investigated.

The preclustering approach is beneficial concerning prediction performance in the lasso setting and leads to improved or comparable results in the other models. However, a main benefit of preclustering is that we detect genes with similar expression patterns and that these subgroups are correlated with survival. In addition, we can have a detailed view on the GO groups containing the preclustered subgroups. Table 5.5 shows that the cluster sizes as well as the corresponding GO groups are quite large. However, in this case the selection of the top 5 clusters is quite stable. For gaining further biological insight a more detailed analysis of the composition of these clusters is required and promising.

In terms of the Brier-Score, we showed that the prediction performance of models using clinical, genomic or both information is comparable. It seems that these different kind of covariates contain an overlap of information for predicting survival.

This work shows that different model selection procedures can be used to identify genes and (preclustered) GO groups related to survival outcomes and to build models for predicting survival times of future patients.

The integration of GO groups is useful, since they contain aggregated information of biological function and thus are often more informative than single genes. It is encouraging that in terms of prediction performance, our results obtained with preclustered GO groups as predictors are comparable to those using only genes as predictors. We demonstrated that this result holds true also for models using GO groups and not only genes. Especially, the analysis of prediction error curves reveals an improved prediction performance for the new preclustering approach. Thus the potentially improved interpretability makes these models with preclustered GO groups competitive. The agenda in the present work was:

- Constructing models with a relatively small subset of relevant covariates that are enriched with additional gene group information in terms of the Gene Ontology.
- Presenting a new approach of preclustering genes from one functional group due to different expression profiles within one GO group.
- Comparing prediction rules and prediction error curves for the three types of covariates (genes, gene groups, preclustered gene groups).

- Adding clinical information and comparing the results to single use of genomic data.

For future work, there are several opportunities for extending the presented approach. The options can be divided into two categories: integrating alternative model selection procedures (internal extensions) and transfer to other settings or data sets (external extensions).

For the internal options, we can investigate other possibilities for integrating group information in survival models. Binder and Schumacher (2008b) applied the CoxBoost algorithm to gene expression data sets without using gene group information. They point out that the main benefit from combining clinical and microarray information was increased prediction performance. This boosting approach may also allow flexible regularization for groups of covariates. Biological prior information from Gene Ontology may be integrated as group information.

The group lasso (Yuan and Lin, 2006) is an extension of the lasso to do variable selection on (predefined) groups of variables in linear regression models. Meier *et al.* (2008) extended the group lasso to logistic regression models, especially for high-dimensional problems. The method can also be applied to generalized linear models and survival models (Simon *et al.*, 2012).

Another possible step for improving our models could be the integration of more detailed information concerning the hierarchically structured gene ontology. For coping with high correlations between GO groups one can follow the approach of Alexa *et al.* (2006) where correlations between neighboring GO groups in the GO graph are iteratively removed.

Finally, we can enrich our preclustering approach with other biological data bases, e.g. the KEGG data base (Kanehisa *et al.*, 2004), or apply the approach without any biological prior knowledge. This has the advantage that there will be no duplicates in the processed data. But we are confronted with the problem that performing clustering on more than 10 000 covariates is computationally intensive. The data set has to be reduced in advance to get results in finite time.

For external transfer of the approach we think of different genomic data like SNP data or array-CGH data in combination with survival outcome. These types of data are of higher dimensions than classical microarray data sets and an analysis with the proposed methods would be challenging.

Finally, when thinking of a classical high-dimensional classification problem with binary outcome, e.g. disease and non-disease, there are often informative genes that are selected according to a two-sample statistical test combined with multiple testing procedures. Instead of keeping the top ranked genes in the models, one could think of a clustering approach in advance and keep medoid genes of clusters and construct models with these covariates.

Bibliography

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, **6**(4), 701–726.

Alexa, A. and Rahnenführer, J. (2009). *R package topGO: Enrichment analysis for Gene Ontology, version 1.14.0*.

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, **22**(13), 1600–1607.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, Jr, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**(6769), 503–511.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**(1), 25–29.

Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, **2**(4), E108.

Beer, D. G., Kardia, S. L. R., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M. G., Iannettoni, M. D., Orringer, M. B., and Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, **8**(8), 816–824.

Binder, H. (2011). *CoxBoost: Cox models by likelihood based boosting for a single survivalendpoint or competing risks*. R package version 1.3.

Binder, H. and Schumacher, M. (2008a). Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. *Statistical Applications in Genetics and Molecular Biology*, **7**(1), Article 12.

Binder, H. and Schumacher, M. (2008b). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, **9**, 14.

Binder, H., Porzelius, C., and Schumacher, M. (2011). An overview of techniques for linking high-dimensional molecular data to time-to-event endpoints by risk prediction models. *Biometrical Journal*, **53**(2), 170–189.

Boulesteix, A.-L., Porzelius, C., and Daumer, M. (2008). Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, **24**(15), 1698–1706.

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjaerde, O. C. (2007). Predicting survival from microarray data - a comparative study. *Bioinformatics*, **23**(16), 2080–2087.

Bøvelstad, H. M., Nygård, S., and Borgan, Ø. (2009). Survival prediction from clinico-genomic models - a comparative study. *BMC Bioinformatics*, **10**, 413.

Boyle, P. and Levin, B. (2008). World cancer report.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, **78**, 1–3.

Buchholz, T. A. (2009). Radiation therapy for early-stage breast cancer after breast-conserving surgery. *The New England Journal of Medicine*, **360**(1), 63–70.

Buening, H. and Trenkler, G. (1994). *Nichtparametrische statistische Methoden*. de Gruyter Lehrbuch. de Gruyter.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22**(4), 477–505.

Carlson, M., Falcon, S., Pages, H., and Li, N. (2009). *R package version hgu133a.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133a), version 2.3.5*.

Cooper, C. S. (2001). Applications of microarray technology in breast cancer research. *Breast Cancer Research*, **3**(3), 158–175.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society. Series B*, **34**(2), 187–220.

De Haan, J. R., Piek, E., van Schaik, R. C., de Vlieg, J., Bauerschmidt, S., Buydens, L. M. C., and Wehrens, R. (2010). Integrating gene expression and GO classification for PCA by preclustering. *BMC Bioinformatics*, **11**, 158.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72**(359), 557–565.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32**(2), 407–499.

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *PNAS*, **103**(15), 5923–5928.

Florescu, A., Amir, E., Bouganim, N., and Clemons, M. (2011). Immune therapy for breast cancer in 2010 - hype or hope? *Current Oncology*, **18**(1), e9–e18.

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, **48**(6), 1029–1040.

Gerds, T. A. and Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics*, **63**(4), 1283–1287.

Giersiepen, K., Heitmann, C., Janhsen, K., and Lange, C. (2005). *Gesundheitsberichterstattung des Bundes, Brustkrebs.*, volume Heft 25. Robert Koch-Institut.

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

Goeman, J. J. (2010b). *R package penalized, version 0.9-31*.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, **18**(17-18), 2529–2545.

Gui, J. and Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21**(13), 3001–3008.

Haibe-Kains, B., Desmedt, C., Sotiriou, C., and Bontempi, G. (2008). A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics*, **24**(19), 2200–2208.

Haibe-Kains, B., Sotiriou, C., and Bontempi, G. (2010). *R package survcomp: Performance Assessment and Comparison for Survival Analysis, version 1.1.6*.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer series in statistics. Springer.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation of nonorthogonal problems. *Technometrics*, **12**(1), 55–67.

Kammers, K. and Rahnenführer, J. (2010). Improving interpretability of survival models with gene groups as covariates. Technical Report 2, TU Dortmund, Fakulä Statistik.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The kegg resource for deciphering the genome. *Nucleic Acids Research*, **32**(Database issue), D277–D280.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(282), 457–481.

Kaufman, L. and Rousseeuw, P. J. (1995). *Finding Groups in Data - An introduction to cluster analysis.* Wiley, New York.

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, **7**(6), 673–679.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis Techniques for Censored and Truncated Data.* Second edition.

Lang, M. (2010). *Korrelierte Gengruppen als Kovariablen in Überlebenszeitmodellen.* Diploma thesis, TU Dortmund University.

Ma, S. and Huang, J. (2007). Additive risk survival model with microarray data. *BMC Bioinformatics*, **8**, 192.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2005). *R package cluster: Cluster Analysis Basics and Extensions*.

Meier, L., Geer, S. v. d., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society. Series B*, **70**(1), 53–71.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**(4), 945–966.

Porzelius, C., Binder, H., and Schumacher, M. (2009). Parallelized prediction error estimation for evaluation of high-dimensional models. *Bioinformatics*, **25**(6), 827–829.

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

Ramaswamy, S., Ross, K. N., Lander, E. S., and Golub, T. R. (2003). A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, **33**(1), 49–54.

Robert Koch Institut (2010). Ausgewählte Einzellokalisationen aus dem Beitrag zur Gesundheitsberichterstattung des Bundes: Verbreitung von Krebserkrankungen in Deutschland. Brustdrüse der Frau.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., López-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T., Staudt, L. M., and Project, L. M. P. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine*, **346**(25), 1937–1947.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53 – 65.

Sariego, J. (2010). Breast cancer in the young patient. *The American Journal of Surgery*, **76**(12), 1397–1400.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *PNAS*, **93**(20), 10614–10619.

Schmidt, M., Hasenclever, D., Schaeffer, M., Boehm, D., Cotarelo, C., Steiner, E., Lebrecht, A., Siggelkow, W., Weikel, W., Schiffer-Petry, I., Gebhard, S., Pilch, H., Gehrmann, M., Lehr, H.-A., Koelbl, H., Hengstler, J. G., and Schuler, M. (2008). Prognostic effect of epithelial cell adhesion molecule overexpression in untreated node-negative breast cancer. *Clinical Cancer Research*, **14**(18), 5849–5855.

Schumacher, M., Binder, H., and Gerds, T. (2007). Assessment of survival prediction models based on microarray data. *Bioinformatics*, **23**(14), 1768–1774.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2012). The sparse group lasso. *Journal of Computational and Graphical Statistics*, **accepted**.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lønning, P., and Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *PNAS*, **98**(19), 10869–10874.

Therneau, T. and Lumley, T. (2009). *R package survival: Survival analysis, including penalised likelihood, version 2.35-8*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, **58**(1), 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**(4), 385–395.

Timm, N. (2002). *Applied multivariate analysis.* Springer texts in statistics. Springer, New York.

U.S. Preventive Services Task Force (2009). Screening for breast cancer.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, **347**(25), 1999–2009.

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., Van't Veer, L. J., and Wessels,

L. F. A. (2006). Cross-validated cox regression on microarray gene expression data. *Statistics in Medicine*, **25**(18), 3201–3216.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**(6871), 530–536.

van Wieringen, W. N., Kun, D., Hampel, R., and Boulesteix, A.-L. (2009). Survival prediction using gene expression data: a review and comparison. *Computational Statistics and Data Analysis*, **53**, 1590–1603.

Verweij, P. J. and van Houwelingen, H. C. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, **12**(24), 2305–2314.

Verweij, P. J. and van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine*, **13**(23-24), 2427–2436.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54**(3), 426–482.

Wall, M. E., Rechtsteiner, A., and Rocha, L. M. (2003). *Singular value decomposition and principal component analysis.*, chapter 5, pages 91–109. Kluwer: Norwell, MA.

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., van Gelder, M. E. M., Yu, J., Jatkoe, T., Berns, E. M. J. J., Atkins, D., and Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**(9460), 671–679.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, Jr, J., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS*, **98**(20), 11462–11467.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B*, **68**, 49–67.

Acknowledgements

First and foremost I want to thank my advisor Prof. Dr. Jörg Rahnenführer. It has been an honor to be his first Ph.D. student at TU Dortmund University. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. The joy and enthusiasm he has for his research was motivational for me. I am also thankful for the excellent example he has provided as a successful professor in statistics.

My special thanks are also due to Prof. Dr. Katja Ickstadt for co-advising this work.

I am very grateful to Michel Lang for his improvement of my code to make it usable for the high performance computing cluster.

My time at TU Dortmund University was made enjoyable in large part due to the many friends and groups that became a part of my life. I am grateful for time spent with my colleagues.

Part of the work has been supported by the German Research Foundation (DFG) within the Research Training Group "Statistical Modelling", project C2 and the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project A3.

Last but not least, I would like to express my special appreciation to my family for giving me love, constant support, and always demonstrating confidence in me.

Declaration

I declare that this thesis is written by myself and that I exclusively used the indicated literature and resources. The thoughts taken directly or indirectly from external sources are proper marked as such.