Survival Analysis based on Genomic Profiles

Khaled Mubarek A Alqahtani

Department of Statistics

University of Leeds

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

December 2016

Declaration

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others."

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors Dr Arief Gusnanto and Prof Charles Taylor for the continuous support of my Ph.D study and related research, for their patience, motivation, and immense knowledge. They were available with useful guidance which helped me in all the time of research and writing of this thesis. I would also like to thank Professor Walter Gilks for his constructive comments during my annual reviews.

Besides my supervisors, I would like to thank my family: my mother Shara, my wife Jamilah and my two princesses Taleen and Wateen, and to my brothers and sister for supporting me spiritually throughout writing this thesis and my life in general. This thesis is a gift to the soul of my father Mobarak whose guiding hand on my shoulder will remain with me forever.

Last but not the least, I would like to thank all the staff at the the School of Mathematics at University of Leeds for their help and guidances. Also, I would like to thank Prince Sattam bin Abdulaziz university for providing the funding for my PhD.

Abstract

Accurate survival prediction is critical in the management of cancer patients' care and well-being. Previous studies have shown that copy number alterations (CNA) in some key genes are individually associated with disease phenotypes and patients' prognosis. However, in many complex diseases like cancer, it is expected that a large number of genes with such an association span the genome. Furthermore, genome-wide CNA profiles are person-specific. Each patient has their own profile and any differences in the profile between patients may help to explain the differences in the patients' survival. Hence, extracting the relevant information in the genome-wide CNA profile is critical in the prediction of cancer patients' survival. It is currently a modelling challenge to incorporate the genome-wide CNA profiles, in addition to the patients' clinical information, to predict cancer patients survival. Therefore, the focus of this thesis is to establish or develop statistical methods that are able to include CNA (ultra-high dimensional data) in survival Analysis. In order to address this objective, we go throw two main parts.

The first part of the thesis concentrates on CNA estimation. CNA can be estimated using the ratio of a tumour sample to a normal sample. Therefore, we investigate the approximations of the distribution of the ratio of two Poisson random variables.

In the second part of the thesis, we extend the Cox proportional hazard (PH) model for prediction of patients survival probability by incorporating the genome-wide CNA profiles as random predictors. The patients clinical information remains as fixed predictors in the model. In this part three types of distribution of random effect are investigated.

First, the random effects are assumed to be normally distributed with mean zero and diagonal structure covariance matrix which has equal variances and covariances of zero. The diagonal structure of covariance matrix is the simplest possible structure for a variance-covariance matrix. This structure indicates independence between neighbouring genomic windows. However, CNAs have dependencies between neighbouring genomic windows, and spatial characteristics which are ignored with such a covariance structure.

We address the spatial dependence structure of CNAs. In order to achieve this, we start first by discussing other structures of variance-covariance matrices of random effects (Compound symmetry covariance matrix , and Inverse of covariance matrix). Then, we impose smoothness using first and second differences of random effects. Specifically, the random effects are assumed to be correlated random effects that follow a mixture of two distributions, normal and Cauchy, for the first or second differences (SCox). Our approach in these two scenario was a genome-wide approach, in the sense that we took into account all of the CNA information in the genome. In this regard, the model does not include a variable selection mechanism.

Third, as the previous methods employ all predictors regardless of their relevance, which make it difficult to interpret the results, we introduce a novel algorithm based on Sparse-smoothed Cox model (SSCox) within a random effects model-frame work to model the survival time using the patients' clinical characteristics as fixed effects and CNA profiles as random effects. We assumed CNA coefficients to be correlated random effects that follow a mixture of three distributions: normal (to achieve shrinkage around the mean values), Cauchy for the second-order differences (to gain smoothness), and Laplace (to achieve sparsity).

We illustrate each method with a real dataset from a lung cancer cohort as well as simulated data. For the simulation studies, we find that our SSCox method generally preformed better than the sparse partial least-square methods in prediction performance. Our estimator had smaller mean square error, and mean absolute error than its main competitors. For

the real data set, we find that the SSCox model is suitable and has enabled a survival probability prediction based on the patients clinical information and CNA profiles. The results indicate that cancer T- and N-staging are significant factors in affecting the patients survival, and the estimates of random effects allow us to examine the contribution to the survival of some genomic regions across the genome.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Obtaining and interpreting information from many variables measured on patients, with an intent to predict disease-related outcomes, is becoming an increasingly important goal in medical research. Statistical methods which allow for censoring are essential when the outcome of interest is a possibly censored time to event. The majority of classical statistical methods that yield a relationship between covariates and outcome rely on the number of covariates p to be less than the number of observations n. Indeed, for the best results these methods typically need p to be somewhat less than n.

Collecting very large amounts of covariate information, such as microarray, SNP, and Copy number alterations (CNA) data via Next Generation Sequencing technology, has been made a reality through technological advances, whilst still tracking survival information on patients in clinical studies. Using these technologies, it is usually the case that the data structure is *high-dimensional* data, that is p is large relative to n. Going one step further we get to ultra-high dimensional data, often denoted $p >> n$, in which the number of covariates is much larger than the sample size, and in this case most classical statistical methods must be adapted.

The most commonly used model for analyzing survival data is the Cox proportional hazards (PH) model introduced by Cox *et al.* (1972), which will be explained in detail in Chapter 4. However, this approach may be infeasible in the high-dimensional setting

and, as a result, a number of different strategies have been proposed for modifying this method to this setting.

Some methods, discussed in Section 1.3.1, proceed by feature selection, in which only a subset of the covariates are selected for inclusion in the model. There are two different approaches to feature selection: discrete or shrinkage. Discrete feature selection involves developing a system to determine whether individual features should enter the model, whereas feature selection by shrinkage works by penalizing the magnitude of the coefficients in the model leading to some coefficients being set identically to zero (e.g., Tibshirani *et al.* (1997); Fan & Li (2002); Zhang & Lu (2007); Antoniadis *et al.* (2010)). Hybrids of marginal screening and shrinkage, such as sure independence screening (SIS), have been proposed to handle ultra-high dimensional survival data (e.g., Fan *et al.* (2010); Zhao & Li (2012)).

Another group of approaches, discussed in Section 1.3.2, focus on summarizing the feature space with a smaller number of derived variables, allowing the number of features involved in the model to remain unchanged. Constructed covariates are developed in these methods from the information in the original feature space, thus allowing all of this information to be summarized by a few constructed covariates rather than the many original covariates.

In situations where the PH assumption is not satisfied, a wide range of survival models, such as the accelerated failure time model and the semi parametric transformation model, have been proposed as useful alternatives. Similarly, methods for fitting these models have also been augmented to include high-dimensional predictors. However, for the purposes of this thesis we have restricted ourselves to settings where the PH assumption does hold.

This chapter is organized as follows. In Section 1.2, we provide the reader with a brief biological background about the genome and Next Generation Sequencing. After that, in Section 1.3, we review methods that have been developed for relating high-dimensional data to survival outcomes. Data set used in this thesis is presented in Section 1.4. Motivation and contributions of this thesis are given in Section 1.5. Section 1.6 presents the software development for this thesis. Finally, Section 1.7 gives layout of the thesis.

1.2 Biological background

1.2.1 Genome

The genome is made up of deoxyribonucleic acid (DNA), which carries the genetic information in all cellular forms of life (see Alberts, 2008). It consists of long chains of nucleotides which are themselves made up of three components:

- a nitrogenous base which are cytosine (C), guanine (G), adenine (A) and thymine (T);
- a five-carbon sugar molecule (deoxyribose);
- a phosphate molecule;

The DNA in the nucleus is split up into a set of different chromosomes (see National Institutes of Health), and the number of chromosomes can differ from animal to animal. For instance, the human genome ($\approx 3.2 \times 10^9$ nucleotides) is assigned to 24 chromosomes, whereas the mouse genome consists of just 21 chromosomes. Each chromosome normally has two copies of DNA, and each of these copies then has two strands of DNA sequences where the bases are paired (A is paired with T, and C with G) (see Alberts, 2008). For example, if TAACGT is a DNA sequence in one strand then the DNA sequence in the same location in the other strand is ATTGCA. For more details, we refer the reader to Alberts (2008) and National Institutes of Health.

1.2.2 Next Generation Sequencing

With the potential to revolutionize various fields, such as personalized medicine and genetic diseases, and now a fundamental tool in molecular biology and genetics, Next Generation Sequencing (NGS) is an incredibly powerful platform which has allowed the sequencing of thousands to millions of DNA molecules simultaneously. First marketed in 2005, NGS is a young field but now has various technologies such as ILLumina, SOLiD and 454 system. Although these have some different trait, for example run time, quality and even cost, there all produce the same raw data (see Henson *et al.*, 2012).

The NGS process, as shown in Figure 1.1, begins with isolating DNA that is then chopped into short fragments to build a genomic library. Following this, the fragments are sequenced and mapped to human reference genome. These mapped sequencings are then called reads. The result is a quantitative dataset - read count per window across the genome.

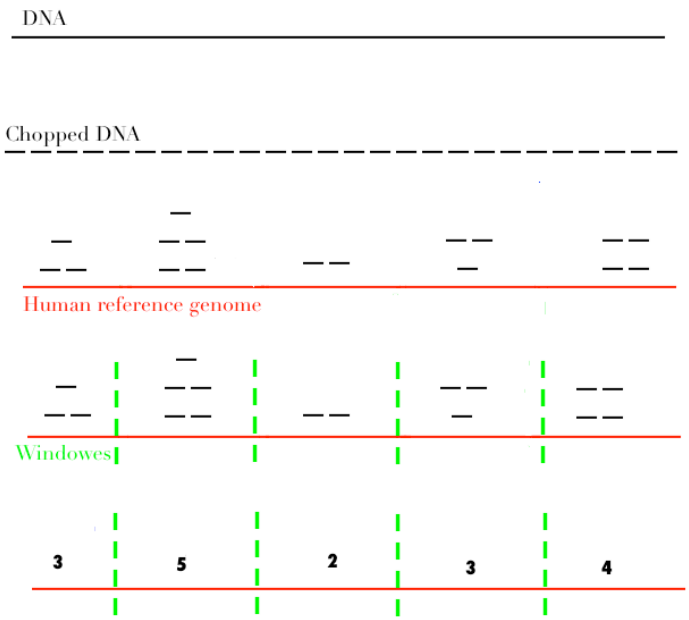


Figure 1.1: Work flow of NGS, which consist of the following main steps, from isolating DNA to the mapped reads. First, isolating the target DNA. Then, the isolated DNA is fragmented. After that, sequenced fragments are mapped to a reference genome and the mapped sequences are called reads. By completing the previous steps, we end up with a quantitative dataset called read count.

1.3 Literature review

1.3.1 Methods based on feature selection

One strategy to deal with high-dimensionality in the covariate space is to use only a subset of the features; one may attempt to select $k < p$ features associated with survival

for inclusion in the model, but of course this is only reasonable if these features are expected to be predictive of survival. Some major approaches for feature selection include discrete feature selection and shrinkage-based feature selection.

Discrete feature selection

A basic approach for selecting features is univariate selection, in which variables are screened individually for association with survival and given some sort of ranking. As an example, we could perform a univariate Cox score test on each feature, and then include the top k features based on the ranking of the corresponding $p-$values. By adjusting the tuning parameter k as appropriate, a certain error rate can be achieved such as the family-wise error rate or the false discovery rate (e.g., Benjamini & Hochberg (1995)). Although this is easy to implement, there are pitfalls; in settings where covariates are correlated, like in CNA data, it may select highly correlated features which do not lead to a multivariate model that improves over the univariate models.

By including genes sequentially in a multivariate model, the correlation between genes can be accounted for, yielding an improvement on the univariate selection described above. This kind of method would be analogous to forward stepwise selection linear regression. In detail, this means beginning with the null model (or the model with clinical covariates alone), then including the feature with the largest score statistic upon computing score statistics for all features. Next, with this feature in the model we use a score test to determine which of the remaining features should be included to best improve the model. Continuing in this way until our model includes k genes, we have an approach which is not only easy to implement but also better accounts for correlation between genes. However, it leads to a locally optimal model rather than the best model with k genes.

In Bøvelstad *et al.* (2007) the performance of these discrete selection methods was compared with the performance of methods based on shrinkage (ridge and Lasso) and summary variables (supervised and unsupervised principal components regression and partial least squares, discussed in Section 1.3.2). They demonstrated that methods based on shrinkage and derived variables tended to outperform discrete variable selection.

Shrinkage methods

As discussed above, discrete feature selection methods may not capture well the joint effects of multiple genes, resulting in prediction models with possibly low prediction accuracy. However, if p is not small relative to n then using a joint model with p features may not be feasible or even stable. To overcome such difficulties, various regularization procedures, aiming to maximize a penalized log partial likelihood with a penalty accounting for the model complexity, have been proposed. An L_2-penalty yields the ridge-regularized estimator (e.g., Verweij & Van Houwelingen (1994)). Importantly, this approach does not do feature selection because all components will in general be non zero.

When a sparse solution is desired, a natural approach is to use an L_1-penalty to regularize log partial likelihood, and this yields the Lasso solution (e.g., Tibshirani *et al.* (1997)).

The inconsistency in variable selection and the bias towards zero of the nonzero coefficients estimated in finite samples are undesirable features of the standard Lasso, and these drawbacks motivated the development of methods in which coefficients receive different amounts of penalization depending on the magnitude of their values. One such approach, introduced by Zou (2006), is the adaptive Lasso which uses weighted L_1-penalties to apply less penalization to larger coefficients and more penalization to variables that are potentially non-informative .

Another potential problem with the Lasso is that when two highly correlated features are associated with the outcome of interest, the Lasso will tend to identify only one of the features, which can be undesirable for interpretability and replicability. To counteract this problem, Zou & Hastie (2005) proposed the elastic net (EN) for linear regression. The EN adds a ridge-type penalty to the Lasso which improves Lasso's ability to identify sets of correlated genes associated with outcome. Engler & Li (2007) applied EN penalty to the Cox model with an algorithm adapted to the high-dimensional setting.

1.3.2 Methods based on derived variables

Feature selection is particularly effective when a subset of the features relate to the outcome of interest, but otherwise an alternative strategy to reduce the complexity of

the feature space is to project the original space to a lower dimensional subspace and derive prediction models within the subspace. These methods are principal components regression (PCR) (e.g, Massy (1965)) and partial least squares (PLS) (e.g, Lee *et al.* (2013)).

One main feature of PC regression is that the dimension reduction is completely unsupervised - the derived variables are constructed using only information on the predictors regardless of relationship between them and the outcome. Hence, it is possible that while the top PCs capture variability in the feature space well, they are not associated with outcome.

The alternative is the PLS method for constructing derived variables, which has been previously proposed for linear regression by Wold *et al.* (1993). For survival analysis, several approaches have been proposed; for example, Nguyen & Rocke (2002) suggest employing PLS using the observed time to event T in place of Y regardless of an individual's censoring status, but if censoring is extensive or related to covariates then it is possible that this approach may produce misleading covariates not associated with survival. Park *et al.* (2002) reformulate the survival problem using Poisson regression in a generalized linear model framework.

In Lee *et al.* (2013), two PLS-based approaches are presented: the Sparse Cox PLS with L_1-penalty (SPLS-L1) and the Sparse Cox PLS with HL penalty (SPLS-HL). They provide a new formulation of the Sparse PLS (SPLS) procedure for survival data to allow for sparse variable selection and dimension reduction at the same time. They showed that that SPLS method performs better than the standard PLS and sparse Cox regression methods in variable selection and prediction, based on numerical studies.

Therefore, we compared our proposed methods with SPLS as it is the more recent method and shows a better performance by comparing with the standard PLS and sparse Cox regression methods.

1.4 Data set

Eighty-nine patients with early-stage lung squamous cell carcinoma (SCC) had surgery at the Department of Thoracic Surgery at Leeds Teaching Hospitals in Leeds, UK between 1994 and 2003. The information available about these patients included age at surgery, sex, stage of disease, and grade of cancer. Details of the clinical sample design

are described in Belvedere *et al.* (2012). We also have the patients DNA information which are described in Chapter 2. The summary of clinical information is as follows

- Survival is the response variable which represents the number of days between surgery and death, the end of the study, or censoring.
- Age is an explanatory numerical variable which identifies the age of the patient at surgery.
- Status is either censored or uncensored. Their status is censored when information on time-to-event is not available because there was no follow-up or the event did not occur before the experiment ended; uncensored, of course, means that this information is available.
- Sex variable simply indicates whether each patient is male or female.
- Grade is a categorical variable based on what the cancer cells look like under a microscope. There are five possible grades: the higher the grade, the faster the cancer is growing.
- Stage of the cancer is a categorical variable which explains how large the cancer is and if it has spread. The system used in this study is the TNM staging system.
  - T is the size of the tumour; there are three possible levels, with 1 being the smallest and 3 being the largest.
  - N indicates whether cancer cells have spread into the lymph nodes close to the original site of the cancer; N can be level 0, 1, or 2, where 0 means that the cancer has not spread.

Table 1.1 shows a summary of these variables.

1.5 Motivation and contribution

As indicated in Sections 1.3.1 and 1.3.2, in the past 10 years survival analysis has been widely used to deal with high dimensional data sets ($p >> n$), but they do have some weaknesses. The feature selection method is easy to implement but selects

Variable	Mean	Min	.25	.50	.75	Max
Survival (days)	1374	34	361	860	2225	4565
Age	66.7	39	61	68	74	84
Status						
	Censored (23)	Uncensored (66)				
Sex						
	F (26)	M (63)				
Grade						
		G1	G2	G3	G4	GX
	Frequency	2	46	36	1	4
Stage T						
	1 (23)	2 (59)	3 (7)			
Stage N	3					
	0 (47)	1 (35)	2 (7)			
Stage TNM						
	1 (44)	2 (35)	3 (10)			

Table 1.1: Summary the lung cancer dataset.

highly correlated features which may result in a poor performance . On the other hand, PLS methods do not automatically lead to the selection of relevant variables. This is because PLS construct latent variables that are linear combinations of all original covariates, so performance is expected to be reduced if a large number of covariates are in fact unrelated (see Lee *et al.*, 2011).

Moreover, CNA data have dependencies between neighboring genomic windows and have a spatial characteristic which would have been ignored if we had used the methods (feature selection and derived variable) described above as explained in Huang *et al.* (2009). These methods can be adapted in survival analysis to model gene expression data; however, they are still unsuitable for CNA data as they ignore its spatial dependence structure. Moreover, the above methods do not include a variable selection which will lead to poorer performance if a large number of predictors are in fact irrelevant. In this thesis, we try to solve all these weaknesses and find an appropriate method that can deal with the dependencies between neighboring genomic windows and the spatial characteristic of CNA data as well as allow for sparse solutions.

Our contribution can be summarized as follows:

- The analysis of copy number alterations (CNA) have been investigated and applied to lung cancer data set.
- We investigated approximations of the distribution of the ratio of two Poisson random variables because there is no known distribution.
- We applied survival analysis methods (non-parametric and semi-parametric (Cox PH)) in the clinical data of lung cancer data set.
- We have extended the Cox proportional hazard (PH) model for prediction of patients survival probability by incorporating the genome-wide CNA profiles as random predictors. The patients clinical information remains as fixed predictors in the model.
- We have devised a more efficient and more accurate method to evaluate AIC using bisection technique or quadratic optimisation technique.

- We have proposed a new extension of Cox PH model to address the spatial dependence structure of CNAs by using different structures of variance-covariance matrices of random effects.
- We have introduced a novel algorithm based on a smooth extended Cox model (SCox) within a random effects model-framework using penalised partial likelihood to model survival time using patients clinical characteristics as fixed effects and their CNA profiles as random effects. We assumed CNA coefficients b to be correlated random effects that followed a mixture of two distributions: normal as in chapter 5 (to achieve shrinkage around the mean values), and Cauchy for the first- or second-order differences of b (to gain smoothness).
- We have introduced a novel algorithm based on Spars-smoothed Cox model (SSCox) within a random effects model-frame work. We assumed CNA coefficients to be correlated random effects that follow a mixture of three distributions: normal (to achieve shrinkage around the mean values), Cauchy for the second-order differences (to gain smoothness), and Laplace (to achieve sparsity).
- We have presented a full gradient algorithm for maximizing the penalized partial likelihood . We generalized the idea of Goeman (2010) which follows the gradient of the likelihood from a given starting value which uses the full gradient at each step.
- I have written 3 R packages from scratch as explained in Section 1.6.

1.6 Software development

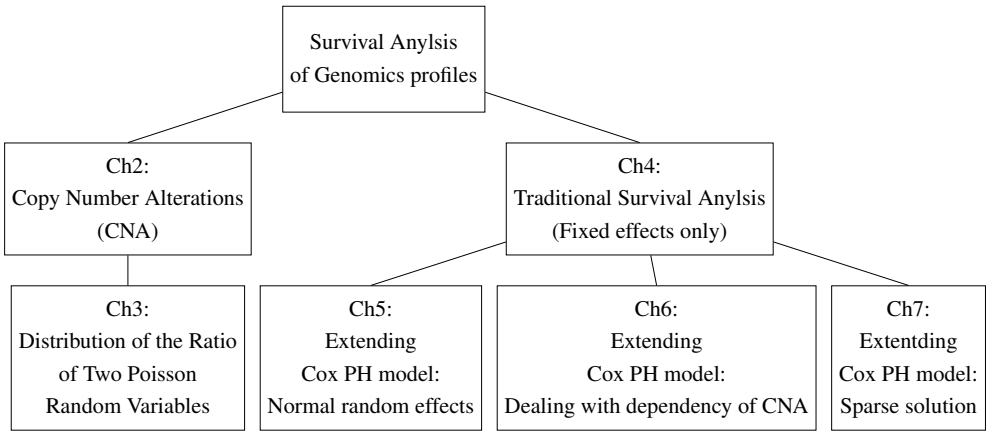
Through our thesis I have written three R package as follows :

1. A CoxCNA package to extend the standard Cox PH model to take into account cancer patients genome-wide copy number alteration (CNA) profiles. This package is used in chapter 5.
2. A SCox package for smooth extended Cox PH model (SCox) which is used in Chapter 6.

3. A SSCox package for Sparse-smoothed extended Cox PH model which is used in Chapter 7 .

1.7 Outline of the thesis

The main body of the thesis is organized into two parts. The first part comprises Chapter 2 and 3 which concentrate on copy number alterations (CNA) and the distribution of the ratio of two Poisson random variables. In the second part, traditional survival analysis, where we only include the clinical data, is presented in Chapter 4, with extensions of Cox PH model to include the CNA addressed in Chapter 5-7. The extension in Chapter 5 is based on normal only, while the extension in chapter 6 deals with dependency between CNA. Finally, the extension of Cox PH model in chapter 7 includes a sparse solution.



In Chapter 2 we review recent CNA detection methods based on next-generation sequencing (NGS). This method, called CNAnorm, was introduced by Gusnanto *et al.* (2012). We applied this method in our lung cancer data set (DNA) and the output is a matrix with dimension 89×13968, where 89 is the number of the patients and 13968 is the number genomic windows. We summarise the flow of this method in details in Chapter 2

In Chapter 2 we discussed the estimation of CNA which can be estimated as the

ratio of a tumour sample to a normal sample. Therefore, in Chapter 3 we examine the distribution of the ratio of two Poisson random variables as the number of reads in a tumour and normal sample assumed to follow Poisson random variables. This chapter can be consider as a separate chapter from the rest of this thesis. In other words, the thesis can stand alone without this chapter. The mean reason for including this chapter is that we advice the main author of CNAnorm to fit a Cauchy−like distribution in the genome wide normalization step instead of using a normal distribution. Also, we did not find in the literatures any discussion about the distribution of the ratio of two Poisson random variables; therefore, we opened the gate for future research.

We start by discussing the approximation of a single Poisson distribution by a normal distribution. Then, the approximated distribution of the ratio of two Poisson random variables by the normal and scaled chi-squared distributions is addressed. similarly, the approximation of the ratio of two Poisson random variables by a Cauchy-like distribution is considered. Finally, we compare the numerical cumulative distribution function (CDF) of distribution of the ratio of two Poisson random variables with the CDF of normal, scaled chi-squared, and the Cauchy-like distributions.

In Chapter 4 we recoup some basic concepts in survival analysis. Then, we describe non-parametric methods for summarising survival data and for comparing two or more groups of survival time (long-rank test) along with the results from and discussion of these methods. The modelling approach is introduced in which the Cox proportional hazards (PH) model is presented. Since model checking is such an important part of the the modelling process, we also include methods for checking the adequacy of a fitted Cox PH model.

In Chapter 5, we propose to extend the Cox proportional hazard (PH) model discussed in Chapter 3, by including the CNA profiles as random predictors. The (standard) Cox PH model has traditionally been used extensively in the prediction of patients survival based on their clinical variables. In this chapter, we extend the model so that the model can incorporate patients genome-wide CNA profiles, in addition to the clinical variables.

We start by discussing the extension of Cox PH model to include the copy number alteration as random effects. CNAs are considered to be random predictors in the model, and the clinical variables as fixed predictors. Specifically, we assumed that the random effects b follow a normal distribution $b \sim N(0, D(\theta))$, and $D(\theta) = \theta I_q$ (I_q

is an identity matrix of size q). The diagonal structure of $D(\theta)$ is the simplest possible structure for a variance-covariance matrix. This structure indicates independence between neighbouring genomic windows.

Then, the estimation of the unknown parameters of the model is discussed. After that, we describes some computational issues. Breslow's estimator of the baseline cumulative hazard rate and the estimates of survivor function are presented. We then discuss residuals for the extended Cox PH model. Simulation studies are described and discussed. Finally, results and evaluation of our lung cancer dataset are presented.

Similar to Chapters 5, in Chapter 6 we propose to extend the Cox proportional hazard (PH) model and at the same time we address the spatial dependence structure of CNAs. In order to achieve this, we start by discussing other structures of variance-covariance matrices of random effects. Then, methods of imposing smoothness using first and second differences of random effects are presented. After that, we discuss the mixture of normal and Cauchy distributions for first or second differences of random effects. We show how to estimate the parameters of the model (fixed effects, random effects, and tuning parameters). Finally, simulation studies and the results of our lung cancer dataset are presented.

Similar to Chapters 5 and 6, in chapter 7 we propose to extend the Cox proportional hazard (PH) model not only to address dependencies between neighboring genomic windows and a spatial characteristic of CNA but also to be embedded with a variable selection mechanism.

In this chapter we introduce a novel algorithm based on Spars-smoothed Cox model (SSCox) within a random effects model-frame work using penalized partial likelihood to model the survival time using the patients' clinical characteristics as fixed effects and CNA profiles as random effects. We assumed CNA coefficients to be correlated random effects that follow a mixture of three distributions: normal (to achieve shrinkage around the mean values), Cauchy for the second-order differences (to gain smoothness), and Laplace (to achieve sparsity).

This chapter presents a full gradient algorithm for maximizing the penalized partial likelihood . We generalized the idea of Goeman (2010) which follows the gradient of the likelihood from a given starting value which uses the full gradient at each step. Furthermore, the algorithm can automatically switch to a NewtonRaphson algorithm

when it gets close to the optimum to avoid the tendency to slow convergence of gradient ascent algorithms.

Finally, We compared our proposed method Sparse Smoothed Cox PH (SSCox) with sparse Cox PLS with L_1 penalty (SPLS-L1) and sparse Cox PLS with HL penalty (SPLS-HL) presented in Lee *et al.* (2013). We conduct simulations to asses the performance of Sparse Smoothed Cox PH (SSCox). We followed the simulation setting of Bøvelstad *et al.* (2007), Nygård *et al.* (2008) and Lee *et al.* (2013).

Chapter 2

Analysis of Copy Number Alterations in Lung Cancer Data

2.1 Introduction

In this chapter, we focus on copy number alterations (CNAs), which are a type of copy number variation (CNV), or structural variation in the genome (see Redon *et al.*, 2006). Freeman *et al.* (2006) refer the CNVs to the duplication or deletion of DNA segments larger than 1 kbp. According to Gusnanto *et al.* (2012), cancer cells often exhibit severe karyotypic alteration: widespread aneuploidy can result from the loss or gain of an entire chromosome, as well as structured rearrangements such as amplifications, deletions or translocations. Detecting CNAs of cancer cells is an essential way to assess the severity of chromosome rearrangement and to locate chromosomal breakpoints. Moreover, cancer-related genes can be found through the location of commonly duplicated or lost regions by comparing CNAs in tumours from a number of patients. Recently several CNA detection methods based on next-generation sequencing (NGS) have been developed such as CVN-seq (Xie & Tammi (2009)), FREEC (Boeva *et al.* (2011)), ReadDepth (Miller *et al.* (2011)), CNVnator (Abyzov *et al.* (2011)), cn.MOPS (Klambauer *et al.* (2012)), and JointSLM (Magi *et al.* (2011)). Duan *et al.* (2013) compared the previous method and conclude that there are a number of differences between these methods, including the statistical models and parameters used in each method, the input, output, and signature formats they use, and the programming language and operating system each requires.

Our analysis will be heavily based on the method called CNAnorm, introduced in Gusnanto *et al.* (2012). We summarise the flow of this method, which will be explained in detail in the next sections, diagrammatically:

raw data → optimal window size → ratio data → Guanine (G)-Cytosine (C) correction → smooth segmentation → genome-wide normalization → contamination correction.

The organization of this chapter is as follows. Section 2.2 introduces the DNA preparation and next-generation sequencing (NGS) of samples for copy number analysis. Section 2.3 addresses the issue of choosing the optimal window size. Section 2.4 then presents CNAnorm's notation and its steps and is broken down into four subsections. In Section 2.4.1, the Guanine-Cytosine (GC) bias in NGS is discussed and resolved. Next, Section 2.4.2 describes the smooth segmentation of the ratio of the tumour to normal genome. Section 2.4.3 then presents genome-wide normalization . Finally, Section 2.4.4 discusses contamination correction. The results and evaluation are contained in Section 2.5.

2.2 DNA preparation and NGS for copy number analysis

The data for this study were drawn from 89 patients with early-stage lung squamous cell carcinoma (SCC) who had surgery at the Department of Thoracic Surgery at Leeds Teaching Hospitals in Leeds, UK between 1994 and 2003. The available patient information included age at diagnosis, sex, stage of disease, and their grade of cancer. Details of the clinical sample design are described in Belvedere *et al.* (2012) and have been discussed more in Chapter 1.

DNA sequencing is the technique used to determine the order of the nucleotide bases Adenine (A), Guanine (G), Cytosine (C), and Thymine (T) in a DNA molecule. The DNA samples were sequenced to low coverage of the complete genome with the aim of estimating the amount of copy number alterations present. Tumour genomic DNA was prepared from macrodissected of formalin-fixed paraffin-embedded (FFPE) tissue. Briefly, 4 mm thick sections were cut from each FFPE tumour tissue block and

stained with haematoxylin and eosin (HE) using a fine-tipped permanent marker; the most representative tumour areas in each slide were marked. DNA extraction was performed using the QIAamp DNA Mini Kit according to the manufacturer's instructions. More details of the DNA preparation are explained in Belvedere *et al.* (2012).

2.3 Read counts and optimal window size

In order to identify the copy number, the number of reads per fixed-width genomic region (window) were counted. Choosing the optimal window size is a trade-off problem. On the one hand, if the window size is too small, for example an average of only 5-10 reads per window, many of the windows will have zero counts and make the analysis non-informative; in other words, a pattern cannot be observed. On the other hand, using a window that is too wide means that any patterns will be smoothed out. Gusnanto *et al.* (2014) identify a method to estimate the optimal window size for analysis of low-coverage NGS data based on Akaike's information criterion (AIC) and cross-validation (CV) log-likelihood .

2.4 Copy number alteration

As mentioned before, our analysis will be based on the CNAnorm method introduced by Gusnanto *et al.* (2012). Notations and arguments also depend on CNAnorm. To identify the number of reads in a a tumour and normal sample, let x_{jk} represent the number of reads observed in a tumour in chromosome $j = 1, \ldots, h$, and window $k = 1, \ldots, n_j$, where n_j is the number of windows in chromosome j. For instance, x_{12} is the observed number of reads from a tumour in the second window of chromosome 1. In addition, let y_{jk} be the observed number of reads in the normal sample. To identify CNAs in the tumour genome, either as gains or losses, we estimate them as an observed ratio of the tumour to normal genome in each genomic window:

$$\hat{\rho}_{jk} = r_{jk} = \frac{x_{jk}}{y_{jk}}$$

There are two copies of (outosomole) chromosomes in a normal genome, while a tumour genome may have zero, one, two, three or more duplications. As a result, the

ratio r_{jk} ideally takes any value from $G = \{0, 0.5, 1, 1.5, 2, 2.5, \dots\}$ corresponding to the tumour copy number $P = \{0, 1, 2, 3, 4, \dots\}$. In reality, this is not the case, due to errors, different numbers of reads being recorded, different sizes of tumour and normal genomes, and contamination of the tumour sample by a normal cell. The estimates $\hat{\rho}_{jk}$ will thus not necessary belong to G. Also, CNAs corresponding to normal genomic regions might not be centred to a ratio of one.

The steps taken to estimate CNA, which will be explained in detail in the next sections, are :

1. The ratio $r_{jk} = \frac{x_{jk}}{y_{jk}}$ was calculated and corrected for GC content (Section 2.4.1).

2. The ratio r_{jk} after GC correction is smoothed to obtain $\check{r}_{jk}$. The smooth segmentation approach introduced by Huang *et al.* (2007) is used (Section 2.4.2).

3. The distribution of $\check{r}_{jk}$ is normalised so that the most common genomic regions are centred to one. This can be written as

$$\hat{\rho}^a_{jk} = \check{r}_{jk}\hat{\delta},$$

   where $\hat{\delta}$ is a genome-wide alignment. This takes care of the different size of tumour and normal genomes (Section 2.4.3).

4. However, $\hat{\rho}^a_{jk}$ does not take into account the tumour sample contamination explain later in Section 2.4.4. At this stage, the level of contamination $\hat{\psi}$ is estimated, and the distribution of $\hat{\rho}^a_{jk}$ is corrected to obtain the estimate of CNA of $\hat{\rho}_{jk}$.

5. At this stage, by using any segmentation tool, the original data can be segmented and the results are corrected accordingly. In CNAnorm, DNAcopy introduced in Olshen *et al.* (2004) is used .

2.4.1 Guanine (G)-Cytosine (C) correction

The ratio r_{jk} can be influenced by GC content in the window (see Boeva *et al.*, 2011). While there are many advantages of NGS, this GC bias is a big disadvantage. Indeed, it is known that on the Illumina system, GC-poor and GC-rich sequences can lead to uneven or even no coverage of reads across the genome (see Chen *et al.*, 2013). This problem can be solved by modelling the dependency of the ratio on the GC content using the local regression model. Gusnanto *et al.* (2012) use the Loess transformation with multiplicative correction as follows:

$$r_{jk}^{\text{norm}} = \frac{\kappa}{A_{jk}} r_{jk},$$

where κ is the median of r_{jk}, and A_{jk} is the estimated Loess point-wise mean of r_{jk}. For simplicity, we drop the superscript norm in r_{jk}^{norm}. In other words, henceforth r_{jk} denotes the ratio with GC correction.

2.4.2 Smooth segmentation

Smoothing is necessary when there is only a small number of reads in each window, as random variability can bias the normalization and the correction of the ratio distribution. However, this step can be skipped in the case of a large excess of reads, typically > 500 per window.

The smooth segmentation used in CNAnorm follows the smoothing approach explained by Huang *et al.* (2007). In this approach, the genomic spatial structure is taken into consideration. Smooth segmentation employs a linear model under the assumption that the second-order difference of the random-effect parameter follows a Cauchy distribution. The Cauchy distribution is useful for handling jumps in the copy number pattern, while at the same time allowing smooth transitions. The estimates of the random effects are the segmented ratio $\check{r}_{jk}$.

To calculate this ratio, Let $l_1, l_2, ..., l_n$ be fixed genomic locations (positions) where $l_1 \leq ... \leq l_n$, and let $r_1, r_2, \ldots, r_n$ be the observed ratios between test and normal samples. The adapted model is thus

$$r_i = f(l_i) + \epsilon_i, i = 1, ..., n \tag{2.1}$$

where $f(r_i)$ is the unknown random effect parameter and the error $\epsilon \equiv (\epsilon_1, ..., \epsilon_n)$ has independent and identically distributed (iid) t-distribution with a location of zero, unknown variance of σ^2, and k degrees of freedom. Also, it is assumed that the error ϵ and $f(l_i)$ are independent. For more detail, we refer the reader to Huang *et al.* (2007).

2.4.3 Genome-wide normalisation

In order to correct the location of the distribution of the copay number ratio, we have to estimate δ from the segmented ratio data $\check{r}_{jk}$. Because of systematic gains and losses, the ratio r_{jk} shows a multi-modal distribution. However, the multi-modality of the distribution is not clear because of the unwanted random errors. After removing the unwanted random errors in the smoothing step, the segmented ratio $\check{r}_{jk}$ has a clear multi-modal distribution. Each mode of this distribution indicates the position of the CNA in G, which corresponds to a different copy number in P. However, these modes are not yet centered on the expected CNAs in G and thus, in order to estimate δ, we need to characterise the distribution of $\check{r}_{jk}$.

Because of the multi-modality, a mixture normal distribution is fitted to the distribution of the smoothed ratio $\check{r}_{jk}$

$$p(\check{r}_{jk}) = \sum_{m=1}^{M} \pi_m N(\check{r}_{jk} : \mu_m, \sigma^2), \tag{2.2}$$

where π_m are the mixture proportions, $\sum_{m=1}^{M} \pi_m = 1$, $0 \leq \pi_m$, and $m = 1, \ldots, M$. μ_m and σ^2 are the mean and variance of each normal distribution. Each of the means μ_m corresponds to a value in G that represents the ratio of tumour to normal copy numbers, which in turn corresponds to a tumour copy number in P. However, the estimates of the means μ_m are still biased estimates of CNAs in G. We will use the estimates of μ_m in the next steps.

We estimate the mixture component in equation (2.2) using the expectation maximisation (EM) algorithm. The number of components in the model M is chosen according to Akaike's information criterion (AIC) across different plausible values.

After estimating $\partial = \hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_M$, it is important to describe the relationship between ∂ and the corresponding tumour copy number in P. Therefore, a simple linear regression is modelled.

It is important to identify the component $V \in \{1, \ldots, M\}$ in the mixture model (equation 2.2) that corresponds to the normal ploidy ratio (one in G). It is defined as the most common component,

$$V = \text{argmax } \hat{\pi}_m.$$

The Vth component is assigned to have a copy number of $(V - 1)$. For instance, the first component represents a total loss (a copy number of zero).
The genome-wide normalisation coefficient $\hat{\delta}$ is estimated as:

$$\hat{\delta} = \frac{1}{\hat{\mu}_v}.$$

As can be seen from the estimation of δ, the process of genome-wide normalisation involves identifying the component corresponding to the normal ratio and then shifting the whole distribution multiplicatively in order to centre the ratio at one.
After estimating $\hat{\delta}$, $\hat{\rho}^a_{jk} = \check{r}_{jk}\hat{\delta}$ is the estimate of CNAs where contamination is still present. In order to find estimates of CNAs that are comparable between samples, it is necessary to characterise any contamination and to make appropriate corrections.

2.4.4 Contamination correction

If there is no contamination, then the smoothed ratio $\check{r}_{jk}$ is expected to take a value in G. However, this is very rarely the case, especially when dealing with tissue from patients' tumours. When contamination does appear, the smoothed ratio will shrink towards a ratio of one (see Gusnanto *et al.*, 2012).

In order to deal with the contamination, Gusnanto *et al.* (2012) assumed that contamination causes the amount of CNAs to shrink linearly towards a ratio of one. For instance, if $\rho_{jk} = 2$, then the number of CNAs will shrink to a value between 1 and 2, while if $\rho_{jk} = 0.5$, then the CNAs will shrink to a value between 0.5 and 1. Since the normal copy number has been centred at one, we can assume that the estimate of the CNAs has arisen from shrinkage of the non-contaminated ρ_{jk} around a ratio of one.

$$\hat{\rho}^a_{jk} = 1 + (\hat{\rho}_{jk} - 1) \times (1 - \hat{\Psi}), \tag{2.3}$$

where $0 \leq \hat{\Psi} \leq 1$ is the estimate of the proportion of contamination.

We estimate Ψ by investigating how the estimates in ∂ have been shrunk towards $\hat{\mu_v}$, which corresponds to the normal copy number. We first normalise the estimates $\partial = (\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_M)$ to $\partial^c = \{\hat{\mu}_m^c\} = \{\hat{\mu}_m \hat{\delta}\}$, for $m = 1, \ldots, M$. The estimate of $\hat{\Psi}$ is given by

$$\hat{\Psi} = \frac{1}{M-1} \sum_m \left\{ 1 - \frac{|\hat{\mu}_m^c - \hat{\mu}_v^c|}{\hat{\mu}_v^c} \frac{1}{0.5 \times |P_m^* - P_v *|} \right\},$$

where the summation is taken over $m = 1, \ldots, v-1, v+1, \ldots, M$, and p_m^* is the copy number in P^*, excluding p_v^*. The estimate of CNAs can now be written from equation (2.3) as:

$$\hat{\rho}_{jk} = 1 + (\hat{\rho}_{jk}^a - 1) \times \frac{1}{(1 - \hat{\Psi})}$$

. Through Gusnanto *et al.* (2012)'s work, we now have estimates $\hat{\rho}_{jk}$ of CNA that take into account variations in read depths, genome sizes and the presence of contamination. Thus, these estimates can now be compared between pairs of samples.

2.5 Results and Discussion

We applied the CNAnorm method for the data from all 89 of the patients in our study by using the R package CNAnorm introduced by Gusnanto *et al.* (2012). Table A.1 in appendix A shows that there are 82 patients with estimated ploidy equal to 2 (diploid), and only 7 patients with estimated ploidy equal to 4 (tetraploid). However, for simplicity, we will use only one patient (LS199) to illustrate the methods.

2.5.1 DNA preparation

In this study, the Illumina Genome Analyzer IIx system was used to obtain and sequence DNA libraries. DNA sequences were obtained from the tumours of 89 patients in the study. These sequences have been stored at the European Nucleotide Archive under accession number ERP000834. The mean read number was 1,030,660 per sample, ranging from 200,000 to 3,000,000. Sequences were aligned to the human genome (USCS hg19). Only reads with mapping quality scores $\geq$ 37 and unique alignments were used. For each window, the average genomic GC content was calculated. A script (bam2window.pl) that can read sam/bam and calculate GC content is available on the CNAnorm website.

2.5.2 Optimal window size

Our first step was to choose the optimal window size. For patient LS199 and the control sample, the results of estimating the optimal window size can be seen in Figure 2.1, based on AIC. For patient LS199, the minimum AIC was achieved with a window size of 150 kb, equivalent to an average of 50 reads per window, while in the control (normal) sample, the minimum AIC was achieved at the window size 250 kb, equivalent to an average of 70 reads per window. Similarly, we chose the optimal window size for each of the other patients, obtaining a number of different values for optimal window size. However, we had to set one fixed window size for all of the tumour samples and the normal sample. We chose to fix the window size at 200 kb for two reasons. First, most of the tumour samples had an optimal window size around 200 kb. Second, we needed to guarantee that the average number of reads per window was at least 10 so that the pattern can be observed.

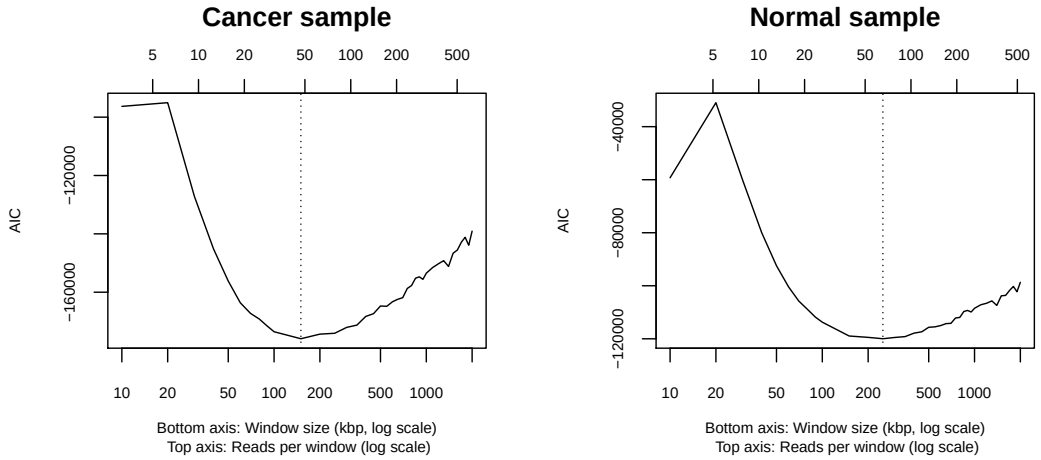


Figure 2.1: AIC as a function of different window sizes (bottom axis) and the corresponding average number of reads per window (top axis) in patient LS199 and in the normal sample. The horizontal axes are in log scale. The vertical dotted lines indicates the optimal window size (optimal number of reads per window).

2.5.3 GC correction

After deciding what the optimal window size was, we calculated the ratio of tumour copy numbers to normal copy numbers in each window. In Figure 2.2, it can be seen from the left panel that the ratio of r_{jk} shows a dependency on the GC content. The red line drawn in the figure is the fitted Loess line. The right panel shows the normalised ratio r_{jk}^{norm} and the fitted Loess line after the correction. The straight line of the fitted Loess line indicates that the the dependency of the ratio on GC content has been removed.

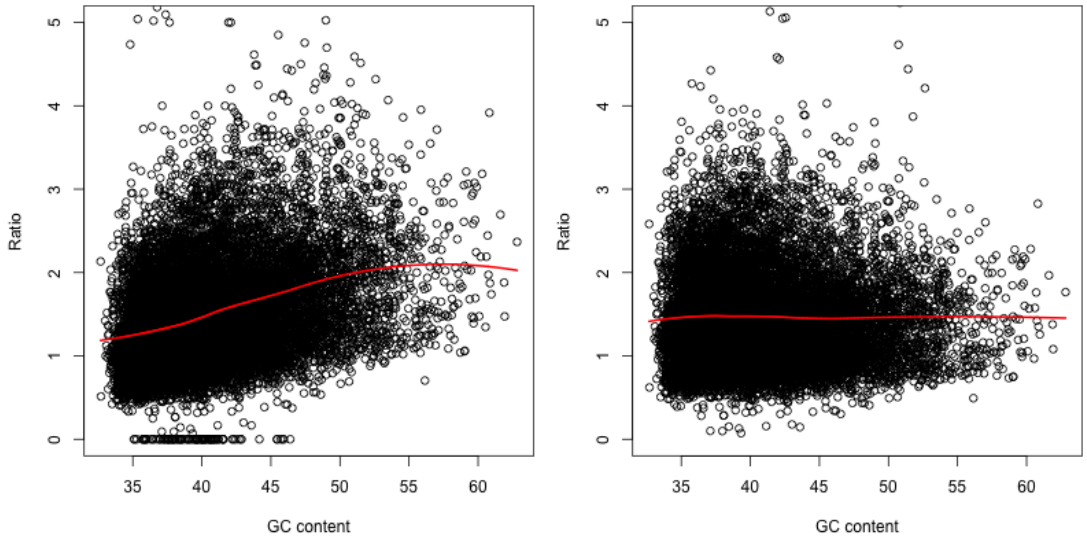


Figure 2.2: The ratio before (left panel) and after (right panel) the GC normalisation on the data from patient LS199. The solid line is the Loess fit line.

2.5.4 Smooth segmentation

We applied smooth segmentation to patient LS199's data. Smooth segmentation was applied to the ratio of tumour to normal sample after GC correction using a 200 kb window size. Looking at the left side of Figure 2.3, it is hard to see the multi-modality in the distribution of the ratio r_{jk}, while on the right we can clearly see the multi-modality in the distribution of the smoothed ratio $\check{r}_{jk}$.

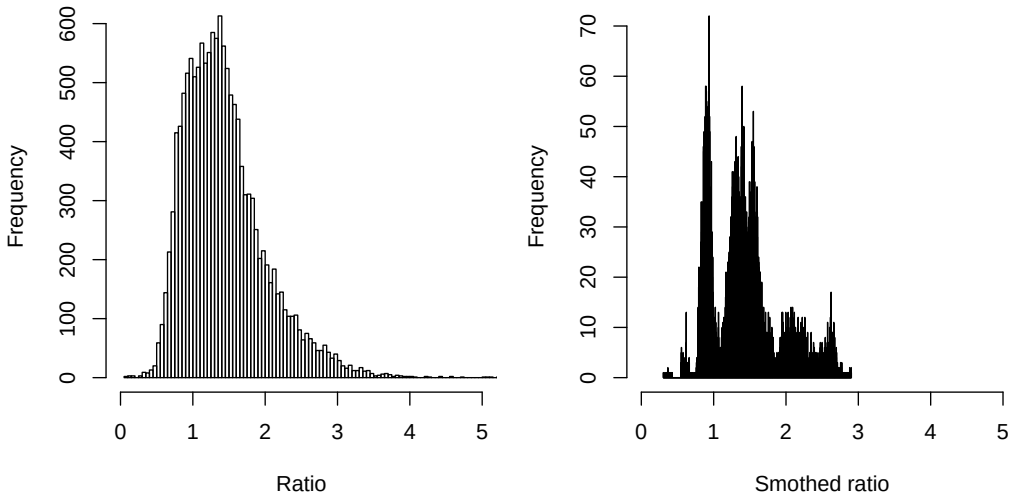


Figure 2.3: Histogram of original ratio (left panel) and smoothed ratio (right panel) for the whole genome of patient LS199.

2.5.5 Genome-wide normalisation

The fit of the mixture model (equation 2.2) is applied to the distribution of the smoothed ratio $\check{r}_{jk}$, as can be seen in Figure 2.4. Based on the AIC, the optimal number of the component is $M = 7$. The estimates of the means are $\hat{\mu_m} = (0.90, 1.12, 1.34, 1.56, 1.96, 2.16, 2.60)$.

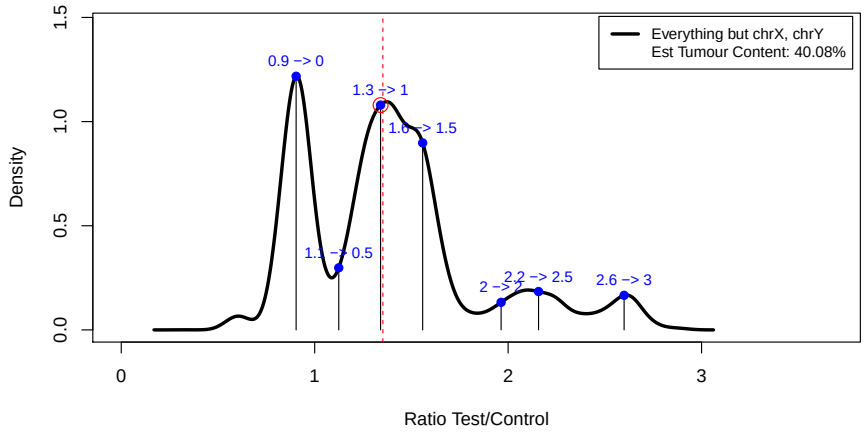


Figure 2.4: The fit of the mixture distribution to the smoothed ratio; the black lines show the estimates of the means for patient LS199.

The estimated proportion of common mixture components is $\hat{\pi}_3 = 0.35$ which indicates that the third mixture component is the most common one $(v = 3)$. This implies that the tumour genome is diploid, since $(3 - 1 = 2)$. The estimates $\hat{\mu_m}$ are plotted against the copy numbers in the left panel of Figure 2.5. We can see from the figure that there is a linear relationship between the estimates of the means $\hat{\mu_m}$ and the copy number. The fitted linear regression is shown as a dotted line. The fitted line has a slope equal to 0.258, which is lower than what was expected (represented by the dashed red line) due to contamination.

We aligned the whole distribution of $\check{r}_{jk}$ so that the mixture component corresponding to the normal copy number was centred to a ratio of one as can be seen in right panel of Figure 2.5. We obtained the estimates $\hat{\mu_3} = 1.38$ and $\hat{\delta} = 0.72$. After scaling for the diploid component to have a ratio of one, the scaled estimates $\hat{\mu}^c_m$ were $0.65, 0.81, 0.97, 0.12, 1.43, 1.56,$ and 1.88 . We can see that the diploid value is not exactly to one because we used the fitted value.

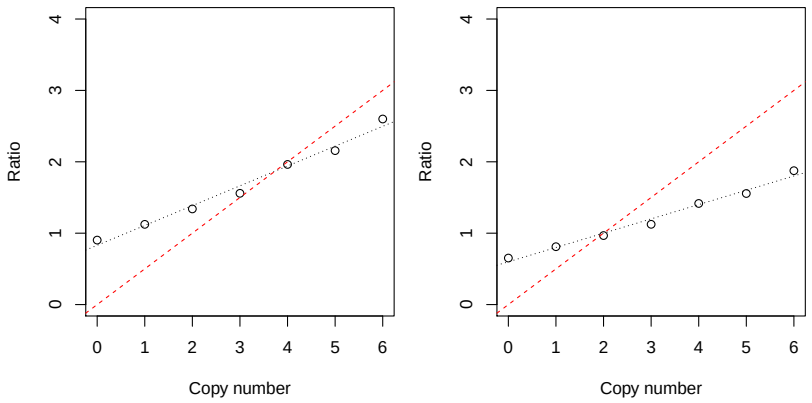


Figure 2.5: Left panel: relationship between the estimates of the means $\hat{\mu_m}$ and copy numbers. Right panel: relationship between the scaled estimates of the means $\hat{\mu}^c_m$ (before the contamination correction) and copy numbers. The dotted line is the fitted linear regression line. The red dashed line, which has a slope of 0.5, is the line we would expect if there were no contamination.

2.5.6 Contamination correction

After scaling for the diploid component to have a ratio of one, the estimate of contamination is $\hat{\Psi} = 0.41$. We corrected the whole distribution of the smoothed ratio, centred to a ratio of one, so that the mean estimates aligned closely to the expected distribution as presented in the left panel of Figure 2.6. In the right panel of Figure 2.6, we can see the histogram of the segmented ratio across the genome after correction for contamination.

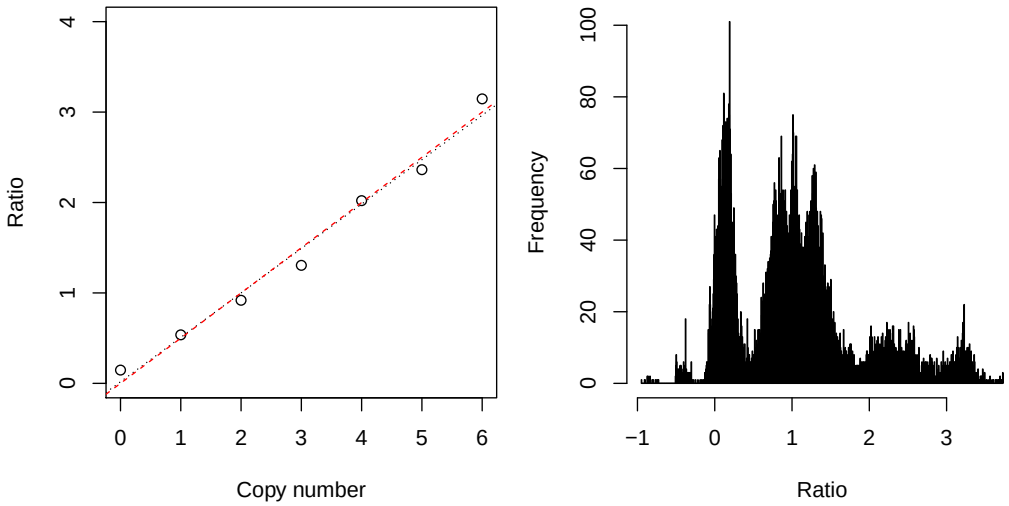


Figure 2.6: Left panel: relationship between estimates of means $\hat{\mu}_m^c$ (after correction for contamination) and copy number in samples LS199. The dotted line is the fitted linear regression line. The red dashed line, which has a slope of 0.5, is the line we would expect if there were no contamination. Right panel : histogram of the segmented ratio across the genome after correction for contamination.

Finally, Figure 2.7 presents the unnormalised ratio r_{jk} for patient LS199, along with the smooth-segmented line, and the bottom panel shows the normalised ratio along with the segmented line (using smooth segmented estimates).

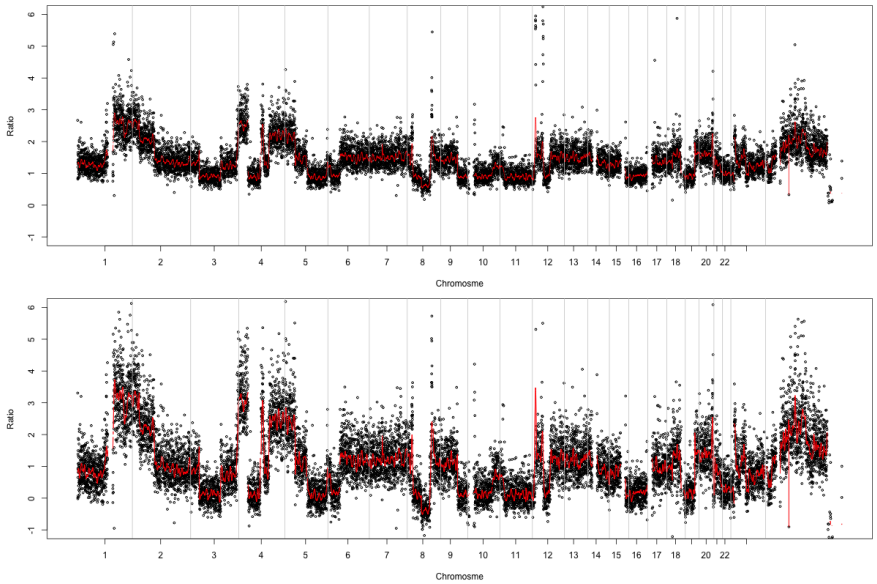


Figure 2.7: Unnormalised (top panel) and normalised (bottom panel) copy number ratios along with the smooth-segmented lines across the genome.

To see the estimated in more details in chromosome, Figure 2.8 shows the estimation of the ratio of the proposed method (CNAnorm) on chromosome 2 for patient LS199.

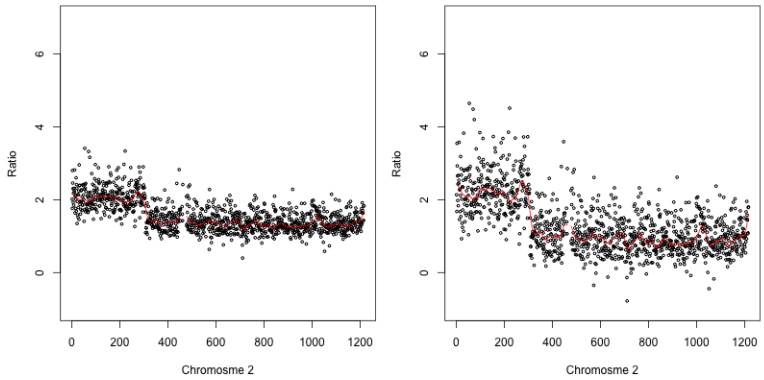


Figure 2.8: Chromosome 2, before (left) and after (right) the normalisation. The solid line is the estimate of CNAs.

To sum up, the genome-wide CNA profile from each patient is calculated by ”depth

of coverage" from their sequences. We estimated that the optimal window size for the patients in our study was 200 kb. The sequence data across patients are not directly comparable because inevitably the tumour samples were contaminated with normal cells to different degrees. To deal with this problem, we performed a normalisation using the CNAnorm package to obtain the CNA estimates. Examples of CNA estimates for the data from patient LS199 can be seen in Figure 2.9, which contains two different forms of estimates: (1) the smooth estimate, where CNAs were estimated as smooth segmented lines as shown in the top panel of Figure 2.9, and (2) the DNACopy estimate, where CNAs were estimated as circular binary segmented lines, as shown in the bottom panel of Figure 2.9. With the 200 kb window size, we have 15,490 genomic windows for each patient. Due to missing data in certain parts of the genome, e.g. centromeres, we consider estimates of CNAs from 13,968 genomic windows from each patient in our analysis.

Finally, Table 2.1 shows the output of the CNAnorm method for patient LS199's data.

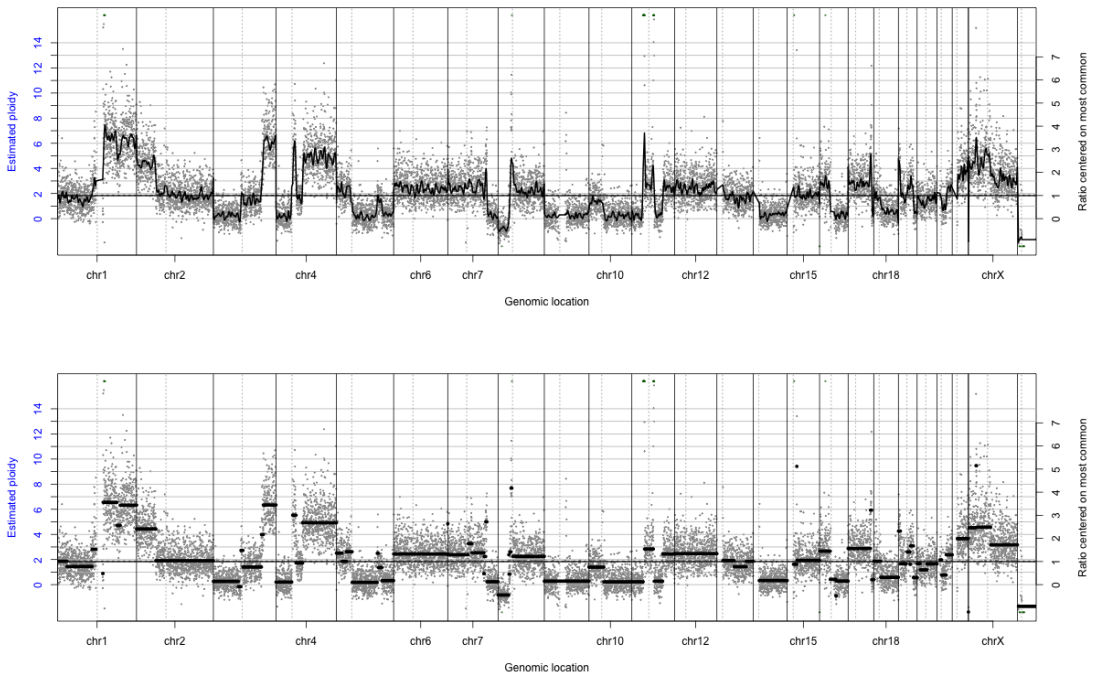


Figure 2.9: Genome-wide CNAs profile of patient LS199 with smooth segmented estimates (top panel) and DNA copy estimates (bottom panel).

	Chr	Pos	Ratio	Ratio.n	Ratio.s.n	SegMean	SegMean.n
1	chr1	1	2.67	6.61	1.56	1.35	1.86
2	chr1	200001	NA	NA	NA	1.35	1.86
3	chr1	400001	0.84	0.04	1.60	1.35	1.86
4	chr1	600001	0.82	-0.04	1.64	1.35	1.86
5	chr1	800001	1.19	1.30	1.67	1.35	1.86
6	chr1	1000001	1.35	1.88	1.71	1.35	1.86
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
15485	chrY	58200001	NA	NA	NA	0.35	-1.73
15486	chrY	58400001	NA	NA	NA	0.35	-1.73
15487	chrY	58600001	NA	NA	NA	0.35	-1.73
15488	chrY	58800001	0.84	0.04	-1.67	0.35	-1.73
15489	chrY	59000001	1.39	2.00	-1.61	0.35	-1.73
15490	chrY	59200001	NA	NA	NA	NA	NA

Table 2.1: The output of the CNAnorm method for data from patient LS199.

Chapter 3

Distribution of the Ratio of Two Poisson Random Variables

3.1 Introduction

In the previous chapter we discussed the estimation of CNA which can be estimated as the ratio of a tumour sample to a normal sample. Therefore, it is worthwhile to examine the distribution of the ratio of two Poisson random variables as the number of reads in a tumour and normal sample are assumed to follow Poisson random variables.

Let $X \sim Pois(\lambda_x)$ and $Y \sim Pois(\lambda_y)$ such that X and Y are independent. In this chapter, we are looking for an approximation of the distribution of the random variable $Z = \frac{X}{Y}$, conditional on $Y \neq 0$. We impose this condition because if $Y = 0$, then $Z = \frac{X}{Y}$ will have some undefined values and so we will not get a proper distribution. It is well-known that

$$X + Y \sim \text{Pois}(\lambda_x + \lambda_y)$$

$$\text{and } X - Y \sim \text{Skellam}(\lambda_x, \lambda_y).$$

However, for the ratio of two Poisson random variables, there is no known distribution. Therefore, in this chapter we derive an approximate distribution for the ratio of two Poisson random variables. In Section 3.2, the approximation of a single Poisson distribution by a normal distribution will be discussed, while in Section 3.3 the distribution of the ratio of two Poisson random variables will be considered as approximated

by the normal and scaled chi-squared distributions. Then in Section 3.4, the approximation of the ratio of two Poisson random variables by a Cauchy-like distribution will be considered. Finally, in Section 3.5, we will compare the numerical cumulative distribution function (CDF) of Z with the CDF of the Cauchy-like, normal, and scaled chi-squared distributions.

3.2 Normal approximation of a single Poisson variable

The Poisson distribution $X \sim Pois(\lambda_x)$ can be approximated with the normal distribution $X_N \sim (\mu = \lambda_x, \sigma^2 = \lambda_x)$ when λ_x is large enough. If λ_x is greater than 10, which is true in our case, the normal distribution is a good approximation if an appropriate continuity correction is performed, i.e., where $P(X \leq x)$ is replaced with $P(X_N \leq x + 0.5)$ (see Makabe & Morimura, 1955). We use the continuity correction as an adjustment because the Poisson variable is discrete, but the normal variable is continuous. The approximation of the CDF of X based on the central limit theorem is

$$F_X(k) \approx \Phi\left(\frac{k + 0.5 - \lambda_x}{\sqrt{\lambda_x}}\right).$$

where Φ is the CDF of a standard normal random variable. However, the Wilson-Hilferty approximation (Lesch & Jeske (2009)) improves on the classical approximation by using a non-linear transformation of the argument k. This approximation uses

$$F_X(k) \approx \Phi(\frac{c - \mu}{\sigma}),$$

where

$$c = \left(\frac{\lambda_x}{1+k}\right)^{\frac{1}{3}}, \mu = 1 - \frac{1}{9k+9}, \text{ and } \sigma = \frac{1}{3\sqrt{1+k}}$$

.

3.2.1 Error in approximating CDFs of the Poisson distribution by the Normal distribution

Here, we look at errors in the normal approximation of the Poisson distribution. Let X be a Poisson random variable with a mean of λ_x, and let X_N be a normal random variable with a mean and variance of λ_x. The CDFs of X and X_N are denoted by F_X

and F_{X_N}, respectively. Now let us look at $F_X - F_{X_N}$ and investigate the improvement that the continuity correction and the Wilson-Hilferty approximations make. Figure3.1 shows values for $F_X(n) - F_{X_N}(n)$ when $n = 0, 1, 2, ..., 20$.

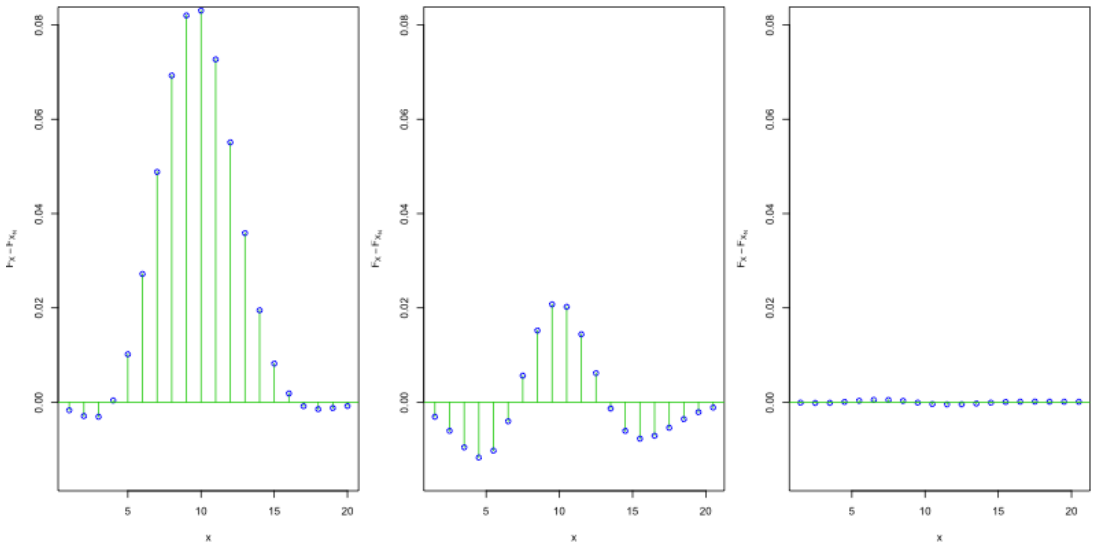


Figure 3.1: From left to right: first panel, $F_X(n) - F_{X_N}(n)$; second panel: $F_X(n) - F_{X_N}(n+1/2)$; third panel: $F_X(n) - F_{X_N}(n)$ using the Wilson-Hilferty approximation. All panels were created using $n = 20$ and $\lambda_x = 10$.

The Berry−Essen theorem identifies the rate at which this convergence takes place by giving a boundary for the maximal error of approximation between the normal distribution and the true distribution of the scaled sample mean; this gives an upper bound of $\frac{C}{\sqrt{\lambda}}$ for all x where C is some constant less than 0.7164 if $N \geq 65$ (see Chen, 2002). Therefore, the Berry−Essen theorem gives an upper bound of $\frac{0.7164}{\sqrt{10}} = 0.2265$ for the approximation error between the normal distribution and the Poisson distribution . By looking to Figure 3.1, The maximum error ($F_X(n) - F_{X_N}(n)$) without the continuity correction is 0.083. With the continuity correction, the maximum error reduces to be 0.021. The maximum error in the Wilson-Hilferty approximation is 0.00049.

3.3 Approximated distribution of the ratio of two Poisson random variables by the normal and scaled chi-squared distributions

The obvious first choice for an approximation of the ratio of two Poisson variables, $Z = \frac{X}{Y}$, is the normal distribution with a mean equal to the expectation of Z and a variance equal to the variance of Z.

A second choice for approximating this ratio is the scaled chi-squared distribution with a scale constant of a and a degree of freedom of v. In order to find a and v, we have to solve two equations:

$$E(Z) = av \tag{3.1}$$

$$Var(Z) = 2a^2v. \tag{3.2}$$

By solving these two equations (3.1, 3.2), we get:

$$a = \left(\frac{Var(Z)}{2 \times E(Z)}\right)$$

$$v = \left(\frac{2 \times E(Z)^2}{Var(Z)}\right)$$

.

3.4 Approximated distribution of the ratio of two Poisson random variables by a Cauchy-like distribution

Feller (2008) showed that when X_N and Y_N are independent and have standard normal distributions, the form of their ratio distribution is a Cauchy distribution. However, when the two distributions for Poisson random variables have non-zero means, which is true in our case, then the form for the distribution of the ratio is much more complicated. Hinkley (1969) found a form for this distribution, when $\text{cor}(X_N, Y_N) = 0$ and the probability density function of the ratio $Z = \frac{X_N}{Y_N}$ of the two normal variables

$X_N = N(\mu_x, \sigma_x^2)$ and $Y_N = N(\mu_y, \sigma_y^2)$ is given by the following expression:

$$f(z) = \frac{b(z).c(z)}{a^3(z)} \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y} \left[2\Phi\left(\frac{b(z)}{a(z)}\right) - 1 \right] + \frac{1}{a^2(z).\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{\mu_x^2}{\sigma_x^2} + \frac{\mu_y^2}{\sigma_y^2}\right)}, \quad (3.3)$$

where

$$\begin{aligned} a(z) &= \sqrt{\tfrac{1}{\sigma_x^2} z^2 + \tfrac{1}{\sigma_y^2}} \\ b(z) &= \tfrac{\mu_x}{\sigma_x^2} z + \tfrac{\mu_y}{\sigma_y^2} \\ c(z) &= \exp\left\{ \tfrac{1}{2} \tfrac{b^2(z)}{a^2(z)} - \tfrac{1}{2} \left(\tfrac{\mu_x^2}{\sigma_x^2} + \tfrac{\mu_y^2}{\sigma_y^2} \right) \right\} \end{aligned}$$

In Section 3.2, we observed that:

$X \sim \text{Pois}(\lambda_x)$ can be approximated by the normal distribution $X_N \sim (\lambda_x, \lambda_x)$.

As we said above, the ratio of two normal distributions is a Cauchy-like distribution. Therefore, the ratio distribution of two random independent Poisson variables may be approximately Cauchy-like distribution. We summarise our argument in the following diagram:

$$\begin{array}{l} X \sim Pois(\lambda_x) \xrightarrow[\text{approximated by}]{\lambda_x \text{ is greater than 10}} X \sim N(\lambda_x, \lambda_x) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \xrightarrow{\text{the ratio of X over Y}} \text{Cauchy-like distribution} \\ Y \sim Pois(\lambda_y) \xrightarrow[\text{approximated by}]{\lambda_y \text{ is greater than 10}} Y \sim N(\lambda_y, \lambda_y) \end{array}$$

3.4.1 Probability density function and cumulative density function of the Cauchy-like distribution

Substituting (3.3) for the mean and variance of X and Y we find that the approximation of the probability density function (PDF) of the Cauchy-like distribution is:

$$f(z) = \frac{b(z).c(z)}{a^3(z)} \frac{1}{\sqrt{2\pi}\sqrt{\lambda_x}\sqrt{\lambda_y}} \left[2\Phi\left(\frac{b(z)}{a(z)}\right) - 1 \right] + \frac{1}{a^2(z).\pi\sqrt{\lambda_x}\sqrt{\lambda_y}} e^{-\frac{1}{2}(\lambda_x + \lambda_y)},$$

where

$$\begin{aligned} a(z) &= \sqrt{\tfrac{1}{\lambda_x} z^2 + \tfrac{1}{\lambda_y}} \\ b(z) &= z + 1 \\ c(z) &= \exp\left\{ \tfrac{1}{2} \tfrac{b^2(z)}{a^2(z)} - \tfrac{1}{2} (\lambda_x + \lambda_y) \right\} \end{aligned}$$

integrating of the PDF is to hard. However, Hinkley (1969) showed that the CDF of a Cauchy-like distribution when setting the correlation coefficient to zero is:

$$F(z) = L\left\{\frac{\lambda_x - \lambda_y z}{\sqrt{\lambda_x}\sqrt{\lambda_y}a(z)}, -\frac{\lambda_y}{\sqrt{\lambda_y}}; \frac{\sqrt{\lambda_y}z}{\sqrt{\lambda_x}\sqrt{\lambda_y}a(z)}\right\} + L\left\{\frac{\lambda_y z - \lambda_x}{\sqrt{\lambda_x}\sqrt{\lambda_y}a(z)}, \frac{\lambda_y}{\sqrt{\lambda_y}}; \frac{\sqrt{\lambda_y}z}{\sqrt{\lambda_x}\sqrt{\lambda_y}a(z)}\right\},$$

where

$$L(h,k;r) = \frac{1}{2\pi\sqrt{1-r^2}} \int_h^\infty \int_k^\infty \exp\left(\frac{x^2 - 2rxy + y^2}{2(1-r^2)}\right) dxdy$$

is the standard bivariate normal integral.

Hinkley (1969) identified a simplified approximation of $F(Z)$:

$$F(Z) \approx \Phi\left\{\frac{\lambda_y z - \lambda_x}{\sqrt{\lambda_x \lambda_y}a(z)}\right\}$$

.

3.5 Comparison of numerical CDF with the CDFs of normal, scaled chi-squared, and Cauchy-like distributions

In Section 3.4, it was shown that the ratio of of two Poisson random variables can be approximated by the Cauchy-like distribution. In this Section, we consider this approximation numerically by comparing the numerical CDF of a ratio of two random variables (given λ_x and λ_y) and the CDF of the Cauchy-like distribution. Then we continue by comparing the numerical CDF of the ratio with the CDF of the normal and scaled chi-squared distributions. Then, the difference between the CDF of the true distribution and the CDF of the approximated distributions is calculated. The steps taken to produce the comparison of the true (numerical) distribution to the approximated distributions are as follows:

- $p(Z = a/b, b \in \mathbb{N}^+, a \in \mathbb{N}) = \sum_{j=1}^{\infty} p(X = a_j)p(Y = b_j)$
- We start by generating $x = (0, 1, 2, 3, \ldots, 500)$ and $y = (1, 2, 3, \ldots, 500)$ and then compute the ratio $Z = \frac{x_i}{y_j}$ for $i = (0, 1, 2, 3, \ldots, 500)$and $j = (1, 2, 3, \ldots, 500)$, so Z will have the dimension 501×501 as can be seen in Table 3.1.

$Z = \frac{X}{y}$	0	1	2	3	...	500
0	$\frac{0}{0}$	∞	∞	∞	...	∞
1	0	1	2	3	...	500
2	0	$\frac{1}{2}$	1	$\frac{3}{2}$	...	250
3	0	$\frac{1}{3}$	$\frac{3}{2}$	1	...	$\frac{3}{2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots\vdots$	$\vdots$
500	0	$\frac{1}{500}$	$\frac{2}{500}$	$\frac{2}{500}$	...	1

Table 3.1: The ratio $Z = \frac{x_i}{y_j}$

- The CDF of Z was calculated numerically by first finding the PDF of Z $p(Z = z)$ using $p(X = x, \lambda_x = a)$ and $p(Y = y, \lambda_y = b)$ and then summing up to find the CDF.
- First we removed ∞ and non-applicable values (N/A) from Table 3.1. Then we normalised $f(Z)$. After that, the expectation $E(Z)$ and the variance var(Z) were calculated.
- The CDF of the normal approximation was calculated.
- The CDF of the scaled chi-squared approximation with constant a and degree of freedom v for variable $\frac{Z}{a}$ was calculated.
- The CDF of the Cauchy-like distribution was calculated.
- After calculating the CDFs of Z by using the normal, scaled chi-squared, and Cauchy-like distributions, we compared these different CDF with the numerical CDF for different values of λ_x, λ_y, as can be seen in Figure 3.2.

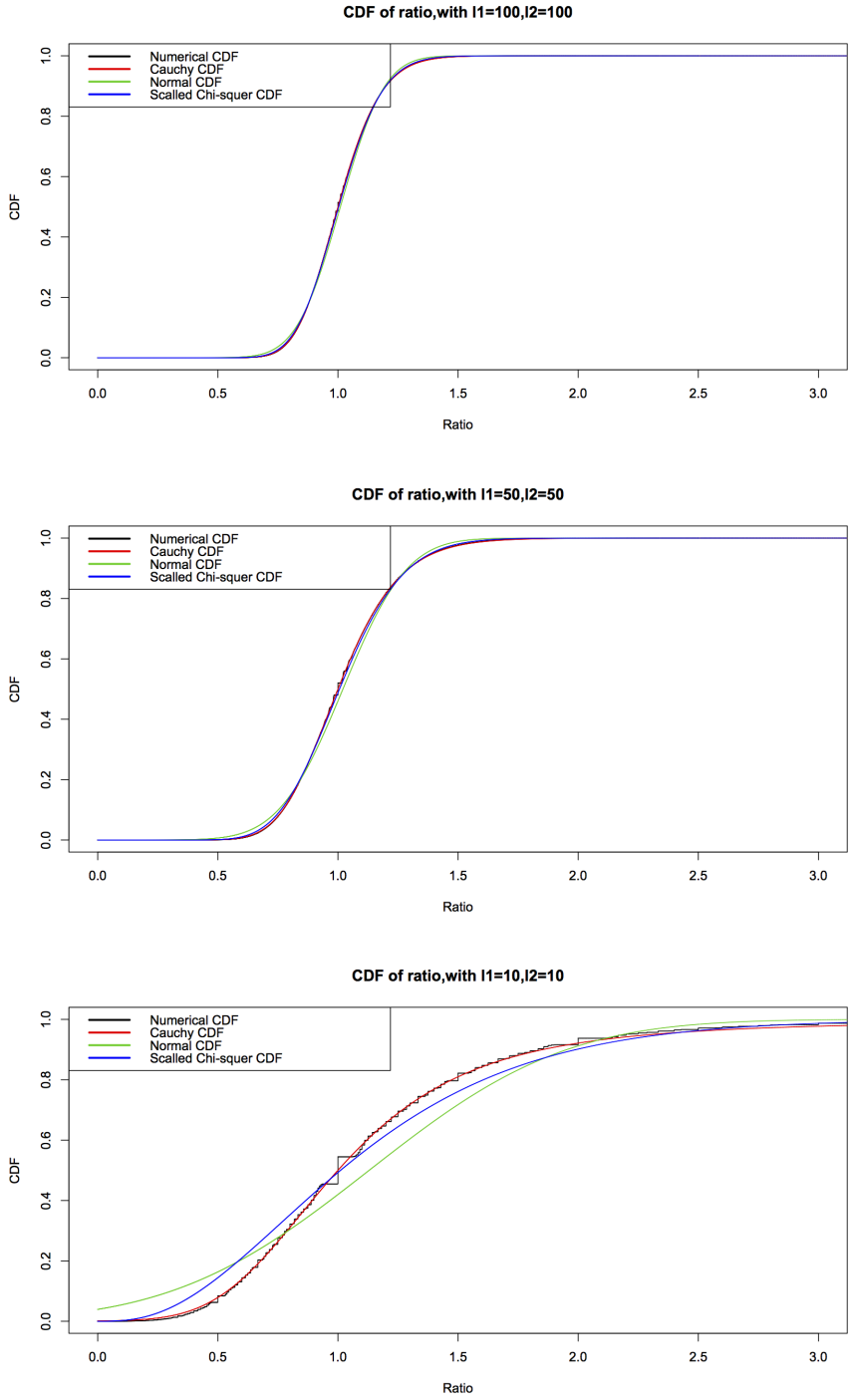


Figure 3.2: The CDFs of the ratio

The differences between the true distribution of the ratio and the approximation distributions are shown in Figure 3.3.

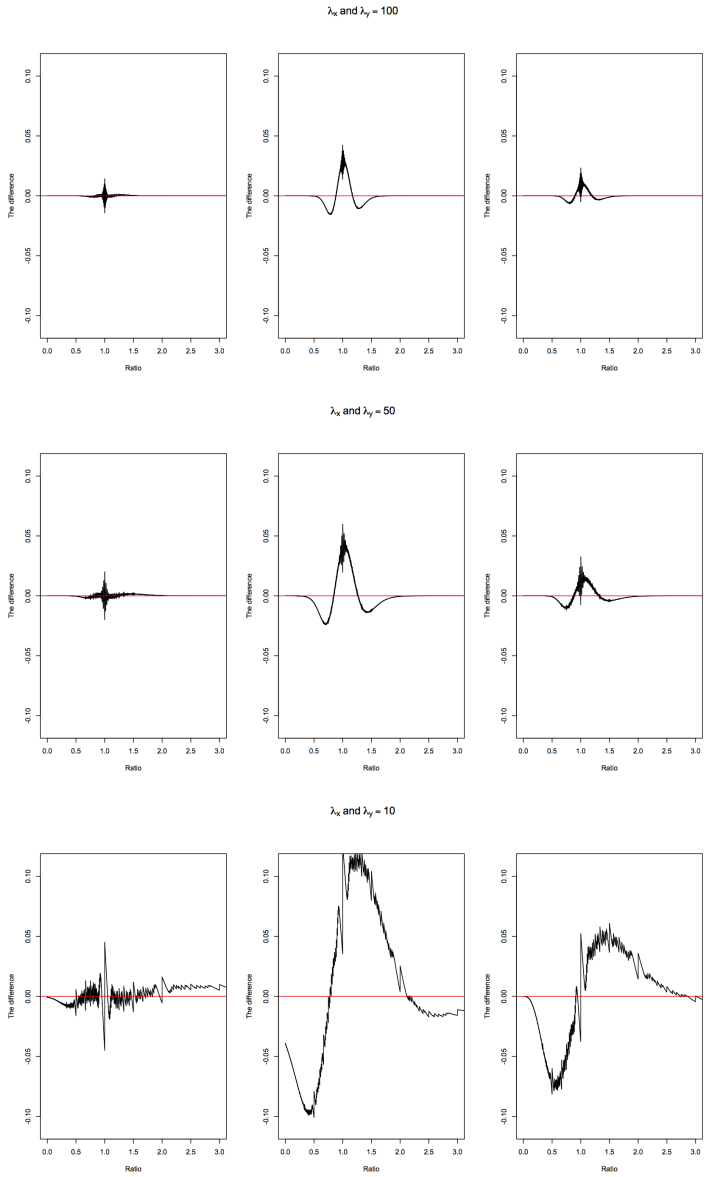


Figure 3.3: The numerical CDF of the the true distribution minus the CDF of the Cauchy-like distribution (left panel), normal distribution (middle panel), and scaled chi-squared distribution (right panel)

To check that we calculated the CDFs correctly, we differentiated the CDF to

obtain the PDF. Then, we multiplied the PDF by Z to get the expectation and the variance, which should match the $E(Z)$ and $Var(Z)$. In Table 3.2, we can see the comparison of the expectation and the variance of the true distribution with those of the approximated distributions.

λ	E(Z) Var(Z)	E(Z_N) Var(Z_N)	E(Z_{Ch}) Var(Z_{Ch})	E(Z_{Cl}) Var(Z_{Cl})
$\lambda = 100$	1.010192 0.02094713	1.010166 0.02094663	1.010166, 0.02094653	No Moment Exist
$\lambda = 50$	1.020823 0.04400519	1.020799 0.04400372	1.020799 0.04400348	No Moment Exist
$\lambda = 10$	1.129967 0.4128282	1.140104 0.3849734	1.12993 0.4127926	No Moment Exist

Table 3.2: Comparison of the expectation and the variance of the numerical distribution (left panel) to the approximated distributions (Z_N: Normal, Z_{Ch}: scaled chi-squared, and Z_{Cl}: Cauchy- like distribution.

3.5.1 Estimation of λ_x and λ_y via the Cauchy-like distribution

Recall that $X \sim$ Pois(λ_x), $Y \sim$ Pois(λ_x), and $Z = \frac{X}{Y}$, and that we have calculated the PDF of the Cauchy-like distribution. Using the optimisation method (Broyden-Fletcher-Goldfarb-Shanno (BFGS) by Head & Zerner (1985)), the parameters (λ_x, λ_y) was estimated via the Cauchy-like distribution. The values (λ_x, λ_y) were estimated 100 times and then the average of these estimations was calculated. The optimisation was carried out using with different values of (λ_x, λ_y) and different sample sizes. The values $\lambda_x = \lambda_y = 100, 50, 10$ and sample sizes of $100, 1000,$ and 10000 were used. The results can be seen in Figure 3.4.

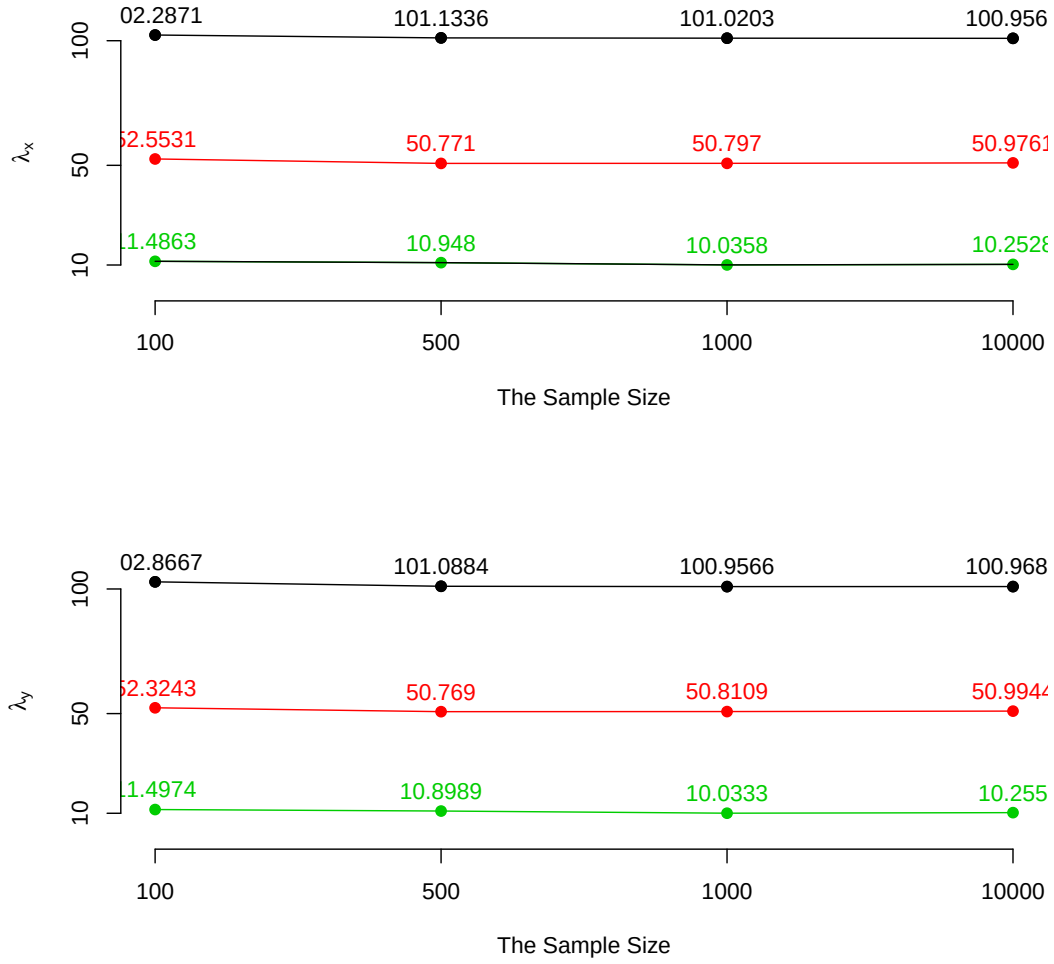


Figure 3.4: Estimation of (λ_x, λ_y) using the optimization method (BFGS)

To see how the estimations of λ_x and λ_y vary, a histogram of the 100 optimisations was plotted with different values for λ_x and λ_y in Figure 3.5.

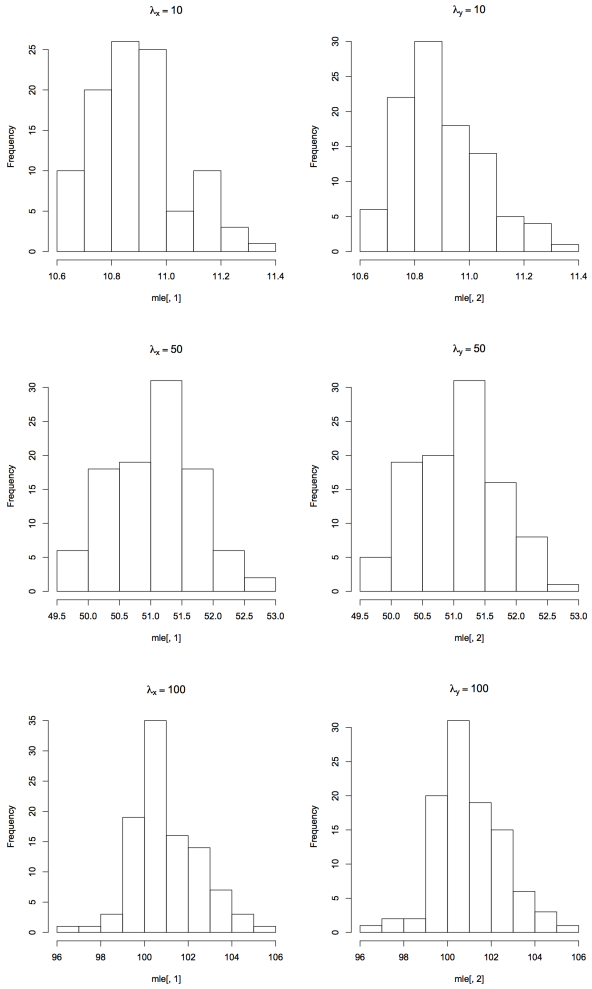


Figure 3.5: Histogram of the estimation of (λ_x, λ_y) using the optimisation method

The maximum likelihood estimation obtained using the optimisation method (BFGS) does not guarantee that the MLE is a global maximum, so there is a possibility that the MLE is only a local maximum and not a global one. Therefore, we did the contour plot to make sure the MLE is actually a global maximum. Figure 3.6 shows the contour plot for $\lambda_x = \lambda_y = 10, 50, 100$, for single sample of size 1000.

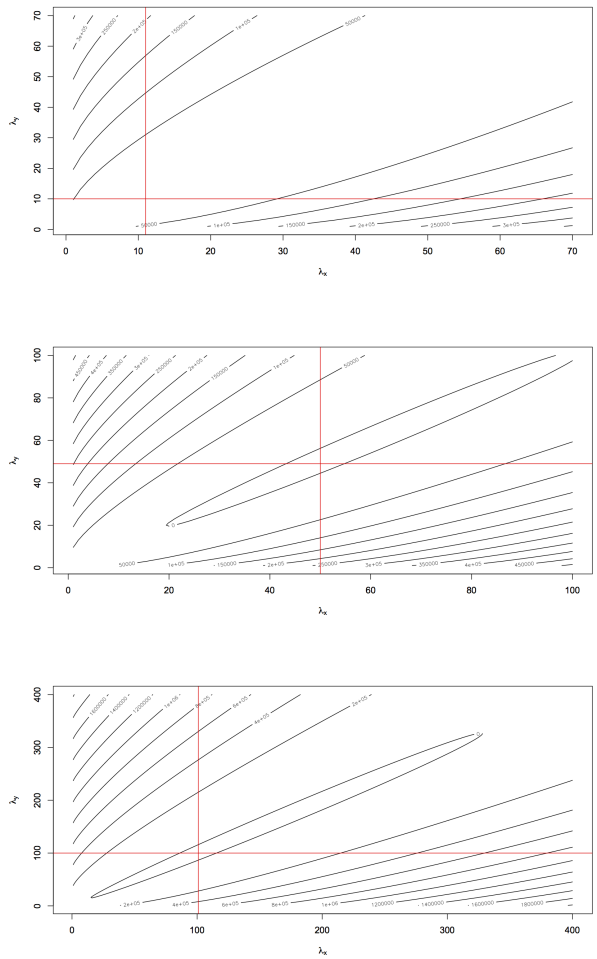


Figure 3.6: Contour plots of the estimations of (λ_x, λ_y) using the optimisation method

Chapter 4

Survival Analysis

4.1 Introduction

A problem frequently encountered by statisticians is the analysis of time-to-event data. This data can be found in diverse fields, such as medicine, public health, engineering, etc. In the course of medical research, such data are mostly referred to as survival data. To date, there have been a number of books on survival data analysis such as Collett (2003), and Moeschberger & Klein (2003). Common features of survival data are censoring and skewness, which require special statistical tools for analysis. Survival data are generally not symmetrically distributed. Their distribution is positively skewed; that is, the histogram will have a longer tail to the right. Therefore, the survival data in general cannot be assumed to have a normal distribution. The survival time of an individual is said to be censored when an event (in this case death) has not been observed or recorded. This censoring may be because the experiment was terminated while individuals were still alive, or because an individual did not attend a follow-up before the end of the study. This is called right censoring. There are other types of censoring (left censoring and interval censoring), but this thesis we will focus only on right censoring: it is the most common type of censoring, and the survival data that we have contains right censoring.

The organization of this chapter is as follows. After an introduction of some basic concepts in survival analysis in Section 4.2, Section 4.3 describes non-parametric methods for summarising survival data and for comparing two or more groups of survival time (long-rank test) along with the results from and discussion of these methods

.The modelling approach is introduced in Section 4.4, where the Cox proportional hazards (PH) model is presented. Since model checking is such an important part of the the modelling process, Section 4.4 also includes methods for checking the adequacy of a fitted Cox model.

4.2 Basic concepts in survival analysis

Survival analysis is the analysis of data representing the time intervals from a well-defined point of origin until the occurrence of an event or a designated end point. An important assumption that will be made in the analysis of censored survival data in general is that the censoring is independent of survival. Therefore, censoring is non-informative, in the sense that it does not give any additional information about patients' outcomes or survival times.

Two important functions in survival analysis are survivor and hazard functions. These functions can be estimated using non-parametric (Kaplan-Meier), semi-parametric (Cox PH model) or parametric methods. In this chapter we will focus on non-parametric and semi-parametric estimation methods.

4.2.1 Functions used in survival analysis

In summarising survival data, the survival function and the hazard function are of central interest. These functions are therefore defined and discussed here.

Survivor Function

The survival time t of an individual can be regarded as a sample value of a random variable $T \geq 0$ (random survival time). Now, suppose that the random variable T has a probability distribution with an underlying probability density function $f(t)$. The distribution function of T is

$$F(t) = P(T < t) = \int_0^t f(u)du.$$

Note that $F(t)$ is non-decreasing and right-continuous, $F(0) = 0$, and $\lim_{t\to\infty} F(t) = 1$.

The survival function $S(t)$ is the probability that the survival time is greater than or equal to t, and so

$$S(t) = P(T \geq t) = 1 - F(t).$$

Note that $S(t)$ is non-increasing, and right-continuous, $S(0) = 1$, and $\lim_{t\to\infty} S(t) = 0$. $S(t)$ can therefore be used to represent the probability that an individual survives from the time origin to some time beyond t.

Hazard Function

The hazard function is commonly used to express the risk (hazard) of death at a given time. The hazard function is also called the *hazard rate*, the *instantaneous death rate*, or the *force of mortality*. In general, the hazard function $h(t)$ is defined to be the probability of death per unit of time (rate) immediately (i.e., in the next instance) after time t, conditional on the patient having survived to time t. The formal definition of the hazard function is

$$h(t) = \lim_{\triangle t\to 0} \frac{P\big[t \leq T < t + \triangle t | T \geq t\big]}{\triangle t}.$$

However, the hazard function is often simply interpreted as the risk of death at time t. If T is a continuous random variable, then

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

Notify that $f(t) = -S'(t)$, it then follows that

$$h(t) = -\frac{d}{dt}(\log S(t)),$$

so

$$S(t) = \exp(-H(t))$$

where

$$H(t) = \int_0^t h(u)\mathrm{d}u.$$

The function $H(t)$ is called the cumulative hazard function and can be obtained from the survival function $S(t)$ as follows:

$$H(t) = -\log S(t).$$

In the analysis of survival data, the survival and hazard functions are estimated from the observed survival times. Methods of estimation that do not require the form of the probability density function of T to be specified, so-called non-parametric methods, are described in Section 4.3, while semi-parametric methods which are a mixed of non-parametric and parametric methods are discussed in Section 4.4. Parametric methods will not be covered in this thesis.

4.3 Non-parametric methods

An initial step in the analysis of survival data is performing numerical or graphical summaries of the survival time for individuals in a specific group. Survival data are summarised through estimates of the survival and hazard functions. Non-parametric methods that can be used to estimate these functions will be discussed first. The term *non-parametric* is used because the method does not require specific assumptions to be made about the distribution of the survival times. The most common non-parametric methods to estimate the survival function are the Life-table, Nelson-Aalen and Kaplan-Meier (K-M) methods. However, we will only discuss the Kaplan-Meier estimate here as it is the most common non-parametric method and it is used to compare two or more groups.

A preliminary way to compare the survival of two or more groups is to draw a Kaplan-Meier plot. For a more precise comparison, numerical hypothesis testing can be used. There are two non-parametric procedures for comparing two or more groups of survival times, the log-rank and Wilcoxon tests. Log-rank test is generally the most appropriate method, while Wilcoxon test is more sensitive when the ratio of hazards is higher at early survival times than at late ones (see Peto & Peto, 1972). As a result, The log-rank test is described in Section 4.3.2, while the Wilcoxn test will not be discussed.

4.3.1 Estimating the survival function with the Kaplan-Meier method

Consider n individuals with observed survival times $t_1, t_2, \ldots, t_n$, where some of these observations may be right-censored. Suppose that there are r death times among the individuals which can be are arranged in ascending order. The jth death time is denoted $t_{(j)}$ and so the ordered death times are $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$. The number of individuals who

were alive just before time $t_{(j)}$, including those who died at this time, is denoted n_j. The term n_j is sometimes referred to as the number of individuals at risk. If we let d_j denote the number of individuals who died at time t_j, the Kaplan-Meier estimator of survival function is

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j} \right)$$

for $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \ldots, r$ with $\hat{S}(t) = 1$ for $t < t_1$, where $t_{(r+1)}$ is taken to be ∞.

4.3.2 Comparison of two or more groups using the log-rank test

We start by considering separately each death time in two groups of survival data.

- First we label the groups Group 1 and Group 2. Suppose there are r distinct death times $t_1 < t_2 < \cdots < t_r$ across these groups.
- Then we suppose that at time t_j there are d_{1j} deaths in Group 1 and d_{2j} deaths in Group 2, for $j = 1, 2, \ldots, r$.
- We also suppose that there are n_{1j} individuals at risk of the death in the first group just before time t_j, while there are n_{2j} at risk in the second group.
- Therefore, at time t_j there are $d_j = d_{1j} + d_{2j}$ deaths in the total out of $n_j = n_{1j} + n_{2j}$ individuals at risk. An explanatory summary can be found in Table 4.1 .

Group	Number of deaths at t_j	Number of surviving beyond t_j	Number at risk just before t_j
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Table 4.1: An explanatory summary of long rank test

Now we consider the null hypothesis, which says that there is no difference in the survival experiences of the individuals in the two groups.

$$H_0 : S_1(t) = S_2(t) \text{ for all } t \in [0, \infty)$$

$$H_a : S_1(t) \neq S_2(t)$$

The log-rank test is a non-parametric test for H_0 based on a comparison of the Kaplan-Meier survival curves $\hat{S}_1(t)$ and $\hat{S}_2(t)$.

Suppose that the marginal totals are fixed. If H_0 is true, we can therefore regard d_{1j} as a random variable, which can take any value from 0 to the minimum of d_j and n_{1j}. In fact, d_{1j} has a hypergeometric distribution, according to which the probability that the random variable associated with the number of deaths in the first group takes the value d_{1j} is

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d1j}}{\binom{n_j}{n_{1j}}}.$$

The mean of the hypergeometric random variable d_{1j} is given by

$$e_{1j} = n_{1j}\frac{d_j}{n_j}.$$

The next step is to combine the information from the table for each death time to give an overall measure of the deviation of the observed values for d_{1j} from their expected values. The clearest way to accomplish this is to sum the differences $d_{1j} - e_{1j}$ over the total number of death times in the two groups. The test statistic is given by

$$U_L = \sum_{j=1}^{r}(d_{1j} - e_{1j}) = \sum_{j=1}^{r}(d_{1j} - n_{1j}\frac{d_j}{n_j}) = O_1 - E_1, \tag{4.1}$$

Note that under H_0 this statistic will have a zero mean ($E(U_L) = 0$), since $E(d_{1j}) = e_{1j}$. Moreover, since d_{1j} has a hypergeometric distribution, the variance of d_{1j} is:

$$Var(d_{1j}) = v_{1j}^2 = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}.$$

And since the death numbers d_{1j} are independent from each other, this gives us

$$Var(U_L) = \sum_{j=1}^{r} v_{1j}^2 = \sum_{j=1}^{r} \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}.$$

The distribution of the statistic U_L is approximately normal:

$$\frac{U_L}{\sqrt{Var(U_L)}} \sim N(0,1).$$

So, the square of the standard normal random variable is a chi-squared distribution with one degree of freedom, denoted χ^2_1, i.e.,

$$\frac{U_L^2}{Var(U_L)} \sim \chi^2_1.$$

This method of combining information over a number of 2×2 tables was constructed by Mantel (1963).

An alternative version of the test using chi-squared is

$$W_L := \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \cdots + \frac{(O_K - E_K)^2}{E_K},$$

where

$$O_K := \sum_{j=1}^{r} d_{Kj} (\text{total observed in Group } K)$$

and

$$E_K := \sum_{j=1}^{r} e_{Kj} (\text{total expected in Group } K).$$

In general, the degree of freedom equals the number of groups minus the number of constraints under H_0. The distribution of W_L is approximated by a chi-squared distribution with $df = K - 1$.

H_0 is rejected if $W_l > k_{(1-\alpha)} = (1-\alpha)$ quantile of χ^2_1.

4.3.3 Results and discussion of the Kaplan-Meier estimator and log-rank test

The K-M estimator was first calculated for lung cancer data without covariates. The K-M plot of these calculations is shown in Figure 4.1, while Table 4.2 represents the numerical summary, where the median of failure time (death) is defined as the time at which the survival function is equal to 0.5. Note that these CIs are constructed by using Greenwoods estimate Greenwood *et al.* (1926), to construct asymptotic confidence

intervals for $S(t)$. From Figure 4.1 and Table 4.2, the median survival time is estimated to be 860 days, with its 95% confidence interval$[669, 1547]$.

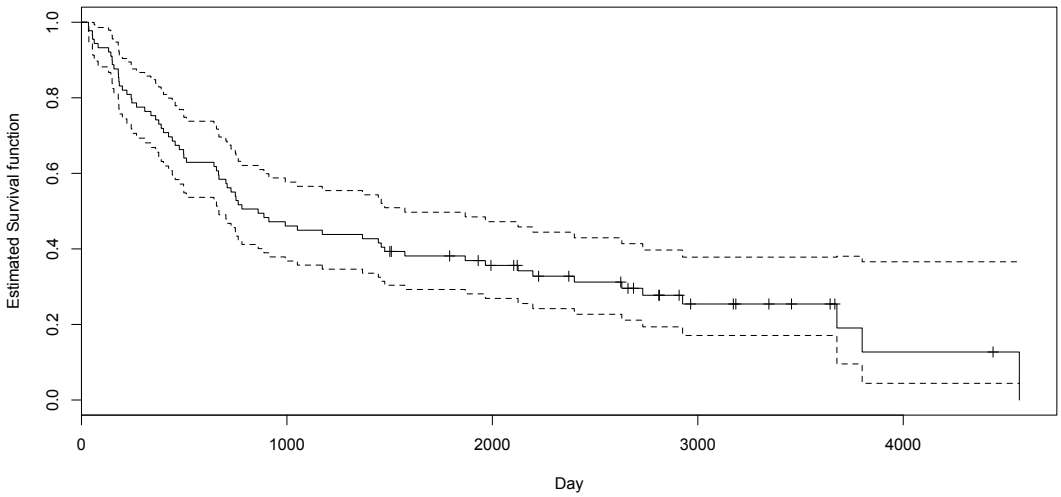


Figure 4.1: The Kaplan-Meier estimate (solid line) and its 95% confidence intervals (dotted lines) for lung cancer's data without covariates (null model)

Records	Event	Median	0.95LCL	0.95UCL
89	66	860	669	1574

Table 4.2: Summary of K-M estimator for lung cancer's data without covariates (null model)

The K-M estimators was then calculated for each covariate. There are five factors in the data, and each will be discussed in turn. Also, in order to decide the importance of a factor, the log-rank test was used, which tests whether there is difference between survival curves for different levels. Recall that H_0 is that there is no difference between groups. *Age* is the only continuous variable; however, we will transform it into a categorical covariate by defining a cut point equal to 65 years, which is the retirement age, so that we can calculate K-M estimators and log-rank test.

1) *Sex*

Figure 4.2 shows the K-M estimators for the covariate of *Sex*, while Table 4.3 shows

the numerical summary. The log-rank test shows no significant difference between male and female (p-value = 0.419).

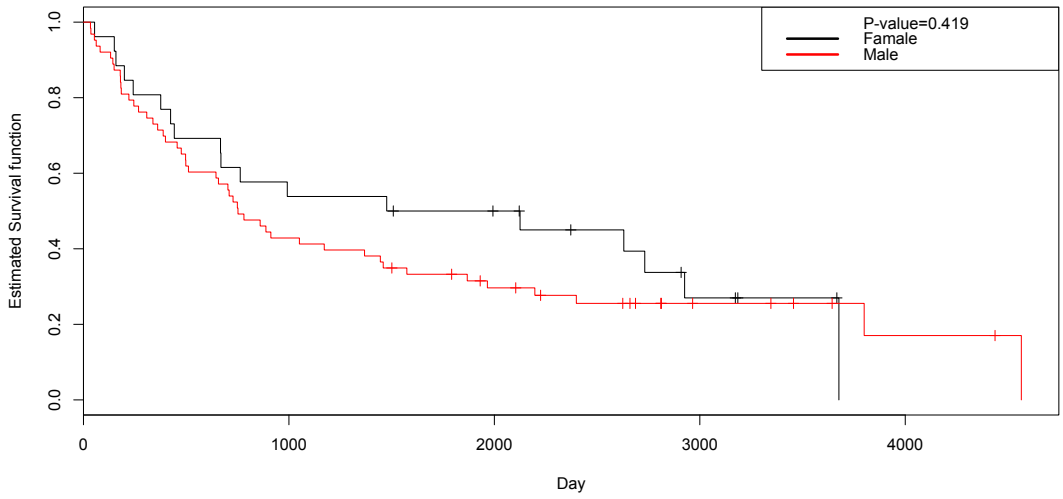


Figure 4.2: The Kaplan-Meier estimators for covariate of *Sex* (Male (red line), Female (black line)) along with the p-value of the log-rank test to compare these two group.

	Records	Event	Median	0.95LCL	0.95UCL
Sex=F	26	18	1800	667	
Sex=M	63	48	752	511	1495

Table 4.3: Summary of K-M estimators of *Sex*

2) *Grade*

This factor has four levels, $1-5$, with higher grades indicating faster cancer growth. In our data, five patients' grades were marked as GX, indicating a missing grade, so they were removed from the sample. Also, only one patient had a grade of 4, and so grades 3 and 4 were combined. Figure 4.3 shows the K-M estimators data for the *Grade* covariates, while Table 4.4 shows the numerical summary. The log-rank test shows no significant difference between the different levels of the grade (p-value= 0.405).

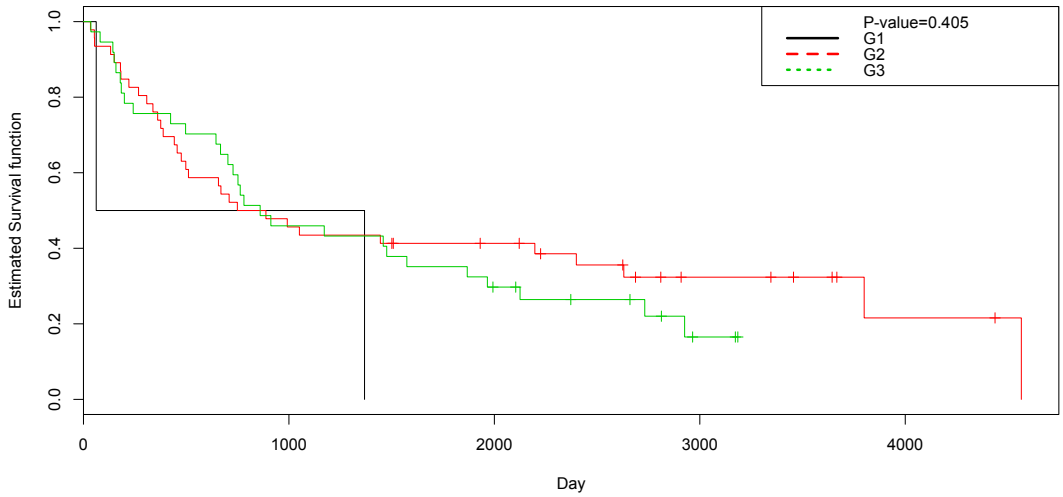


Figure 4.3: K-M estimators for the *Grade* covariate along with the p-value of the log-rank test

	Records	Event	Median	0.95LCL	0.95UCL
Grade=1	2	2	715	62	
Grade=2	46	32	818	498	
Grade=3	137	860	703	1966	

Table 4.4: Summary of K-M estimators of *Grade*

3) *Stage T*

This factor has three levels, $1-3$, in increasing order of tumour size. Figure 4.4 shows the K-M estimators for the covariate *Stage T*, while Table 4.5 shows the numerical summary. The log-rank test shows no significant difference between the stages of T (p-value= 0.283).

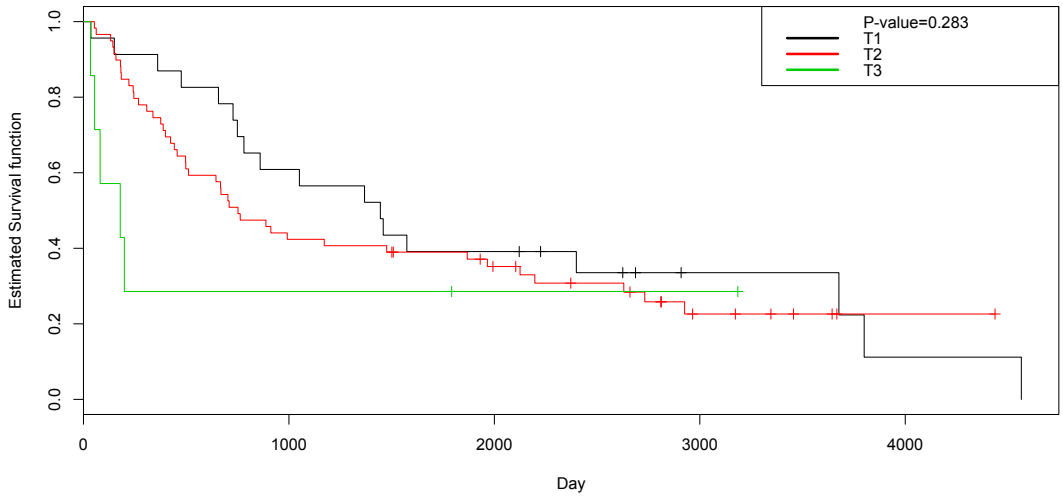


Figure 4.4: K-M estimators for the covariate *Stage T* along with the p-value of the log-rank test

	Records	Event	Median	0.95LCL	0.95UCL
Stage T=1	23	18	1445	781	
Stage T=2	59	43	752	498	1966
Stage T=3	7	5	179	54	

Table 4.5: Summary of K-M estimators of *Stage T*

4) *Stage N*

This factor has three levels, $0-2$, indicating the degree to which the cancer cells have spread into the lymph nodes close to original cancer site. Figure 4.5 shows the K-M estimators for the covariate *Stage N*, while Table 4.6 shows the numerical summary. The log-rank test shows no significant difference between the stages of N (p-value = 0.115).

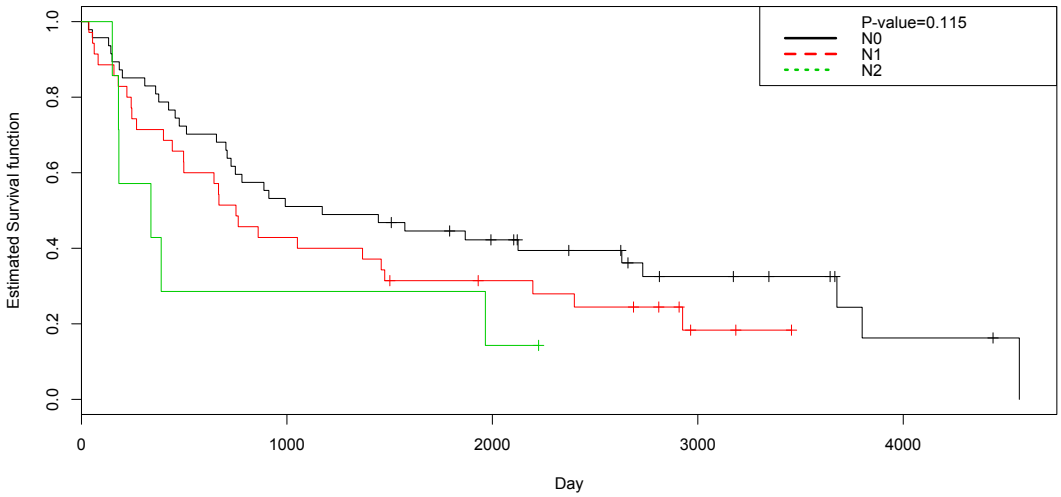


Figure 4.5: K-M estimators for the covariate *Stage N* along with the p-value of the log-rank test

	Records	Event	Median	0.95LCL	0.95UCL
Stage N=0	47	33	1172	728	3800
Stage N=1	35	27	752	497	2197
Stage N=2	7	6	338	180	

Table 4.6: Summary of K-M estimators of *Stage N*

5) *Stage TNM*

This factor also has three levels, 1-3, in an increasing order. Figure 4.6 shows the K-M estimators for this covariate, while Table 4.7 shows the numerical summary. The log-rank test shows no significant difference between the stages of TNM (p-value = 0.093).

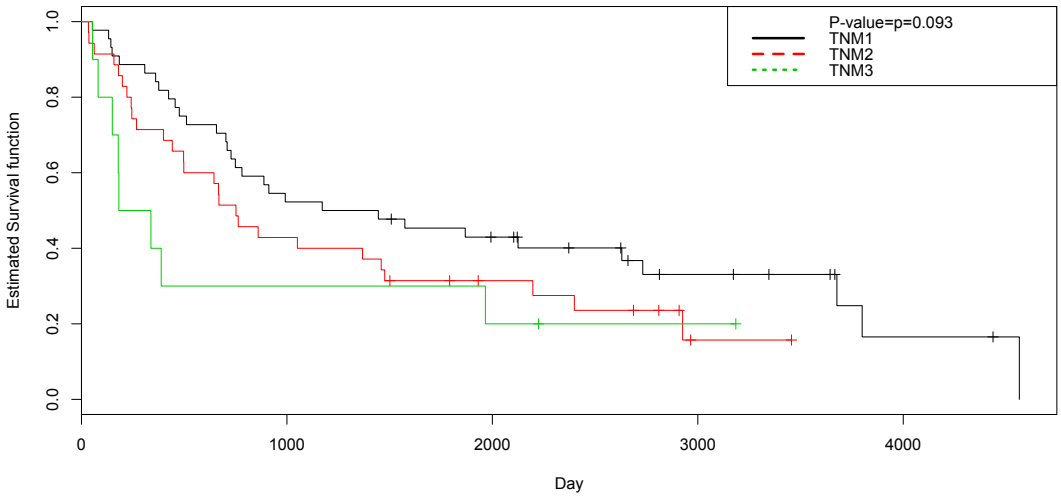


Figure 4.6: K-M estimators for the covariate *Stage TNM* along with the p-value of the log-rank test

	Records	Event	Median	0.95LCL	0.95UCL
Stage TNM=1	44	31	1308	749	3800
Stage TNM=2	35	27	752	497	2197
Stage TNM=3	10	8	260	150	

Table 4.7: Summary of K-M estimators of *Stage TNM*

6) *Age*

Age is a continuous covariate. However, in order to use K-M estimators, Age was treated as a categorical variable by using a cut point of 65 years. Many possible cut points were tried and the log-rank test was applied to each of them. Only for the cut point of retirement age (65) did the log-rank test show a significant difference between the patients over and under age 65 (p-value = 0.0438). Figure 4.7 shows the K-M estimators for the covariate *Age*, while Table 4.8 shows the numerical summary.

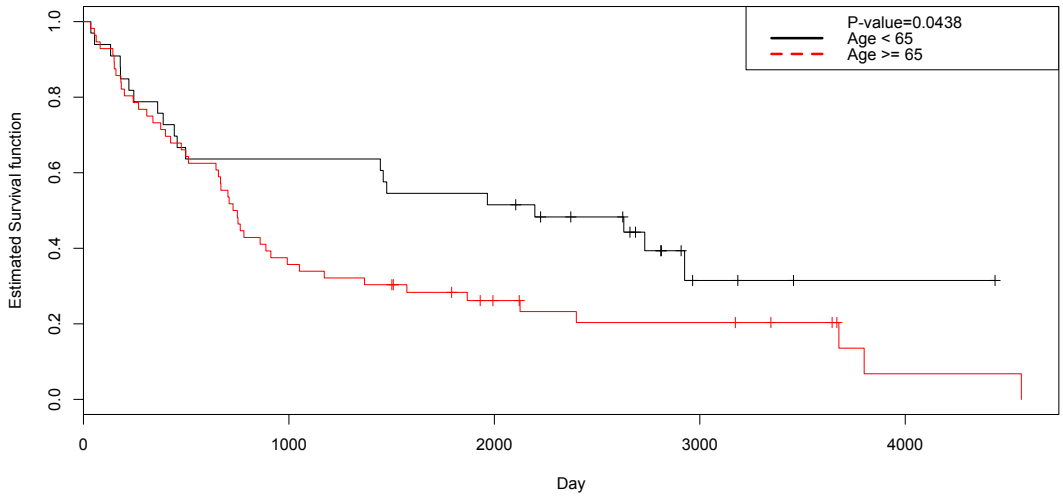


Figure 4.7: The Kaplan-Meier estimators for the covariate *Age* (Age<65 (black line), Age≥65 (red line)) along with the p-value of the log-rank test to compare these two groups.

	Records	Event	Median	0.95LCL	0.95UCL
Age<65	33	20	2197	497	
Age≥65	56	46	738	645	1051

Table 4.8: Summary of K-M estimators of *Age*

After the above analyses, we found that only *Age* made a statistically significant contribution to the survival of patients.

4.4 The semi-parametric method and the Cox proportional hazards model

Whilst non-parametric methods provide an easy way to compare the survival experience of two or more groups, they are limited to testing one covariate at a time. Furthermore, non-parametric methods cannot test continuous variables. Many studies include

data from a number variables, both categorical and continuous, and thus more sophisticated methods are needed. One such method is regression analysis, which falls into two types: fully parametric and semi-parametric. This section will cover one particular semi-parametric method, the Cox proportional hazard (Cox PH) model .

4.4.1 Model and assumptions

The proportional hazards model was first proposed by Cox *et al.* (1972), who made further improvements in Cox (1975). First, we denote δ_i as the event indicator for the the ith patient, $i = 1, 2, ..., n$, where $\delta_i = 1$ if the survival time of the ith patient, t_i, is uncensored and $\delta_i = 0$ if the survival time t_i is censored. A survival time is censored if the actual survival time was longer than the observed survival time t_i or if the ith patient died due to some cause other than cancer. We define X to be a matrix of size $n \times p$, where the columns of X correspond to the different clinical information as fixed predictors, and the rows of X correspond to the different individuals or patients. We denote the rows of X as X_i, which is a p vector of fixed predictors for the ith patient. We also denote $h_0(t)$ to be the baseline hazard function, which denotes the baseline hazard rate for all of the patients in the group across time and does not depend on any predictors. Cox *et al.* (1972) proposed that the hazard rate at time t can be modelled as

$$h_i(t|X) = h_0(t)c(X_i\beta),$$

where β is a p-vector of the models parameters (fixed effects), and $c(X_i\beta)$ is a known function. This is called a semi-parametric model because a parametric form is assumed only for the covariates' effects, while the baseline hazard rate is treated non-parametrically. In this case, $h_i(t|X)$ must be positive. As a result, a common model for $c(X_i\beta)$ is $\exp(X_i\beta)$. This yields

$$h_i(t|X) = h_0(t) \exp(X_i\beta). \tag{4.2}$$

With this formulation, the ratio of hazard rates between two individuals does not depend on t, but only on the difference in the predictors, hence the term proportional hazard.

There are two components of this model that need to be estimated. The first is the unknown coefficient of the risk factor (or explanatory variable) in the linear component

of the model $\beta_1, \beta_2, \ldots, \beta_p$. The second is the baseline hazard $h_0(t)$, whose estimation is especially important if we need to make a prediction for new data. If only inferences about the effects of p explanatory variables $X_1, X_2, \ldots, X_p$ on relative hazard $(h_j(t)/h_0(t))$ need to be made, then $h_0(t)$ does not need to be estimated. The unknown β-coefficients in the proportional hazard model can be estimated using the method of maximum likelihood (or partial likelihood), which will be discussed more in the next section.

4.4.2 Estimation of model parameters

The full likelihood for the proportional hazard model takes the form

$$L(\beta, h_0(t)) = \prod_{i=1}^{n} h(T_i|X_i)^{\delta_i} S(T_i|X_i), \tag{4.3}$$

where $h(T_i|X_i) = h_0(T_i)^{\delta_i}[\exp(X_i\beta)]^{\delta_i}$ and $S(T_i|X_i) = \exp\big(-H_0(T_i)\exp(X_i\beta')\big)$

Maximising $L(\beta, h_0(t))$ jointly with respect to β and $h_0(t)$ is difficult, and for the purpose of estimating β it is preferable to treat $h_0(t)$ as a nuisance parameter to be eliminated from the likelihood. Cox *et al.* (1972) showed how this could be achieved by using a *partial likelihood* for the proportional hazard model. He showed that the usual maximum likelihood theory would still apply to the estimate of β obtained by maximising the partial likelihood function.

4.4.3 Partial likelihood for the case of no tied failure times

Suppose that there are n individuals in the study, and let $t_1 < t_2 < \cdots < t_n$ be the ordered, observed follow-up times (either the actual survival time or the time of right-censoring). Considering the case where only one individual fails at each failure time, so that there are no tied failure times, let δ_i be the event indicator (1 if individual i is observed to fail and 0 otherwise), and let x_i be the covariate vector for individual i, for $i = 1, \ldots, n$. Define the risk set $R(t_i)$ just prior to time t_i as the set of all individuals at risk of failure (still alive and under follow-up) at that point in time. In order to derive the partial likelihood, suppose that t_i is a failure time. Then the conditional

probability that an individual with covariate vector X_i dies at time t_i, given that one of the individuals in $R(t_i)$ dies at this time, is given by:

$$\begin{aligned} &P[\text{individual with covariate } x_i \text{ dies at } t_i | \text{ exactly one death at } t_i] \\ &= \frac{P[\text{individual with covariate } x_i \text{ dies at } t_i]}{P[\text{exactly one death at } t_i]} \\ &= \frac{h[t_i|X_i]}{\sum_{j\in R(t_i)} h[t_i|X_j]} \\ &= \frac{h_0(t_i)\exp[X_i\beta]}{\sum_{j\in R(t_i)} h_0(t_i)\exp[X_j\beta]} \\ &= \frac{\exp[X_i\beta]}{\sum_{j\in R(t_i)} \exp[X_j\beta]}. \end{aligned}$$

The partial likelihood is formed by the product of these conditional probabilities over all deaths (uncensored individuals),

$$L(\beta) = \prod_{i=1}^{r} \frac{\exp[X_i\beta]}{\sum_{j\in R(t_i)} \exp[X_j\beta]},$$

which can be written as:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp[X_i\beta]}{\sum_{j\in R(t_i)} \exp[X_j\beta]} \right)^{\delta_i}.$$

The log partial likelihood then is

$$l(\beta) = \sum_{i=1}^{n} \delta_i (X_i\beta) - \sum_{i=1}^{n} \delta_i \log \left(\sum_{j\in R(t_i)} \exp(X_j\beta) \right). \tag{4.4}$$

The (partial) maximum likelihood estimates are found by maximising $l(\beta)$. First the efficient score equations $U(\beta)$ are found by taking the partial derivative of $l(\beta)$ with respect to each β. The information matrix $I(\beta)$ is the negative of the matrix of the second derivative of $l(\beta)$ with respect to β_i.

The (partial) maximum likelihood estimator is then found by solving the set of p nonlinear equations $U(\beta) = 0$, $h = 1, \ldots, p$. This can be done numerically using the Newton-Raphson technique, which involves iteratively updating β as follows:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + I(\hat{\beta}^{(k)})^{-1} U(\hat{\beta}^{(k)}).$$

Other iterative methods can also be used. The Newton-Raphson technique is discussed more in Section 4.4.4 below.

4.4.4 The Newton-Raphson algorithm

In survival analysis, the Newton-Raphson algorithm is usually used to maximise the partial likelihood function. However, dealing with the partial log-likelihood function $l(\beta)$ (E.q 4.4) is usually easier than dealing with the likelihood function itself. When β is a parameter vector of dimension p, the Newton-Raphson algorithm to find $\hat{\beta}$ that maximises the $l(\beta)$ is as follows:

1. Set $K = 0$.
2. Choose the initial values $\hat{\beta}_0$.
3. solve $\hat{\beta}_{k+1} = \hat{\beta}_k + I^{-1}(\hat{\beta}_k)U(\hat{\beta}_k)$
4. Increase k by one.
5. Go back to step 3 and repeat until $\hat{\beta}_k$ converges.

The terms $U(\hat{\beta}_k)$ and $I^{-1}(\hat{\beta}_k)$ are defined as:

- $U(\beta)$ is the $p \times 1$ vector of the first derivatives of the log-likelihood function with respect to β,

$$U(\beta) = \frac{\partial l(b)}{\partial \beta} = \delta' X - \sum_{i=1}^{n} \delta_i \frac{\sum_{j \in R(t_i)} X_j \exp(X_j \beta)}{\sum_{j \in R(t_i)} \exp(X_j \beta)}$$

  , the term $U(\beta)$ is called the score function.

- $I(\beta)$ is the $p \times p$ matrix of the negative second derivatives of the log-likelihood function, $I(\beta)$ is called the information matrix.

In this study, even though, there is a built-in function in R (>coxph), we wrote our own code to carry out the Newton-Raphson algorithm because it will be easier to extend it later on. This code consists of four steps. First, the ordered design matrix is generated. Then the score function is evaluated, followed by the calculation of the information matrix. Finally, the Newton-Raphson iteration is carried out.

Cox (1975) showed that the $\hat{\beta}$ has nice statistical properties. These include:

- Consistency: That is, $\hat{\beta}$ will converge to the true value of β which generated the data as the sample size gets larger.

- Asymptotic Normality: $\hat{\beta}$ will be approximately normally distributed with mean β_t and a variance which can be estimated from the data. This approximation will be better as the sample size gets larger. This result is useful in making inference for the true β.
- Efficiency: Among all other competing estimators for β, the $\hat{\beta}$ has the smallest variance, at least, when the sample size gets larger.

4.4.5 Breslow's estimator of the baseline cumulative hazard rate

Breslow (1974) starts with the full likelihood equation (Eq. 4.3) with β being replaced by $\hat{\beta}$ in order to obtain the baseline hazards function

$$L(\beta, h_0(t)) = \prod_{i=1}^{n} h_0(T_i)^{\delta_i} [exp(X_i\beta)]^{\delta_i} \exp\left(- \int_0^{t_i} h_0(u) \exp(X_i\beta) du \right). \quad (4.5)$$

In the likelihood function (4.5), there are two main components of $h_0(t)$. In order to maximise the second component, $\exp\big(- \int_0^{t_i} h_0(u) \exp(X_i\beta) du\big)$, $h_0(t)$ should be made as small as possible. However, the first component $h_0(T_i)^{\delta_i}$ indicates that, where $\delta_i = 1$, a larger value of $h_0(t_i)$ will give a larger value of the likelihood function (4.5). Therefore, this lead to $\hat{h}_0(t) = 0$ for all $t \notin (t_1, \ldots, t_r)$. However, values $\hat{h}_0(t_1), \ldots, \hat{h}_0(t_r)$ that maximise (4.5) will also maximise the log of (4.5):

$$\ell(h_0(t_i)) = \sum_{i=1}^{r} [\log h_0(t_i) + X_i\hat{\beta}] - \sum_{i=1}^{r} h_0(t_i) \sum_{j \in R(t_i)} \exp[X_j\hat{\beta}]. \quad (4.6)$$

By differentiating (4.6) with respect to $h_0(t_i)$, setting this to zero and then solving the resulting equation, it can be seen that the maximum likelihood estimate of $h_0(t_i)$ is

$$\hat{h}_0(t_i) = \frac{1}{\sum_{j \in R(t_i)} \exp(X_j\hat{\beta})}.$$

To estimate $H_0(t)$, these estimates of $\hat{h}_0(t_i)$ need to be combined as follows:

$$\hat{H}_0(t) = \sum_{t_i \leq T_j} \frac{1}{\sum_{j \in R(t_i)} \exp(X_j\hat{\beta})}.$$

This is Breslow's estimator of the baseline cumulative hazard rate (with no tied failure time).

Note that $\hat{H}_0(t)$ is a step function, with jumps at the observed death times. This estimator is reduced to the Nelson-Aalen estimator when there are no covariates present

$$\hat{H}(t) = \left\{ \begin{array}{ll} 0 & \text{if } t \leq t_1 \\ \sum_{t_i \leq t} \frac{di}{Y_i} & \text{if } t_1 \leq t \end{array} \right\}$$

, where Y_i is the number of individuals who are at risk at time t_j.

4.4.6 Result of the Cox PH model

Four patients' records have missing data for the *Grade* covariate. Therefore, these four records were discarded and we only used the remaining 85 observations in my analyses. Also, the factor *Stage TNM* was discarded because it can be predicted from *Stage T* and *Stage N*. To examine this relationship, an ordinal logistic model of *Stage TNM* based on the main effect of *Stage T* and *Stage N*, as well as their interaction effect was fitted. Moreover, a classification tree of *Stage TNM* based on *Stage T* and *Stage N* was constructed.

Based on the fitted model of *Stage TNM* on *Stage T* and *Stage N*, and the classification tree, the model fitted 88 out of 89 patients in the true class of *Stage TNM*, as we can see in Table 4.9. The only patient which was misclassified into *Stage TNM2* instead of *Stage TNM3* is the patient coded with 16 .

Stage TNM	1	2	3
True Class	44	35	10
Predicted Class	44	34	11

Table 4.9: Fitted ordinal logistic model of *Stage TNM* based on *Stage T* and *Stage N*

The Cox PH model was first fitted with the covariates *Age*, *Sex*, *Grade*, *Stage T*, and *Stage N*. The results can be seen in Table 4.10.

	coef	exp(coef)	se(coef)	z	p
Age	0.05	1.05	0.02	3.33	0.0008
Sex M	0.24	1.27	0.32	0.75	0.46
Grade G2	-1.00	0.37	0.78	-1.28	0.20
Grade G3	-0.80	0.45	0.80	-1.00	0.32
Stage T2	0.23	1.26	0.32	0.70	0.48
Stage T3	1.93	6.89	0.65	2.97	0.0029
Stage N1	0.28	1.32	0.30	0.94	0.35
Stage N2	1.33	3.78	0.50	2.67	0.007
Rsquare = 0.221 (max possible = 0.996)					
Likelihood ratio test = 21.28 on 8 df, p=0.00644					
Wald test = 21.87 on 8 df, p=0.005167					
Score (log-rank) test = 22.56 on 8 df, p=0.003975					

Table 4.10: Hazard ratios from the Cox PH model for the lung cancer dataset

For model selection, an automatic variable selection procedure, stepwise selection, was used. The final model of survival for lung cancer patients after applying the stepwise selection procedure can be seen in Table 4.11.

	coef	exp(coef)	se(coef)	z	p
Age	0.0531	1.0545	0.0156	3.4087	0.0007
Stage T2	0.1505	1.1624	0.3016	0.4990	0.6178
Stage T3	1.8004	6.0522	0.5765	3.1230	0.0018
Stage N1	0.3445	1.4113	0.2842	1.2121	0.2255
Stage N2	1.3360	3.8040	0.4776	2.7973	0.0052
Rsquare = 0.201 (max possible= 0.996)					
Likelihood ratio test = 19.1 on 5 df, p=0.00183					
Wald test = 20.06 on 5 df, p=0.0012					
Score (log-rank) test = 20.65 on 5 df, p=0.00094					

Table 4.11: Hazard ratios from the Cox PH model for the lung cancer dataset with the significant covariates

Based on the stepwise selection, the most significant explanatory variables in this

study of the survival rates of lung cancer patients were found to be *Age*, *Stage T*, and *Stage N*.

The final model for the ith individual can be expressed in the form:

$$h_i(t) = h_0(t)\exp\{0.053 \text{ Age}_i + 0.15 \text{ Stage T2}_i + 1.8 \text{ Stage T3}_i + 0.345 \text{ Stage N1}_i + 1.34 \text{ Stage N2}_i \},$$

where $i = 1, 2, \ldots, 85$.

An explanation of this model is given. The estimated log-hazard ratio for an individual at stage T2 relative to an individual at stage T1, when both are the same age and at the same level of stage N, is $\hat{\beta}_2 = 0.15$. Consequently, the estimated hazard ratio is $e^{0.15} = 1.16$. Similarly, the estimated log-hazard ratio for an individual at stage T3, relative to an individual at stage T1of the same age and at the same level of stage N, is $\hat{\beta}_3 = 1.8$; in this case, the estimated hazard ratio is $e^{1.8} = 6.08$.

The hazard ratio for two patients who are the same age and at the same level of stage T, one at stage N1 and the other at stage N0, is $e^{0.34} = 1.41$, while for a patient at stage N2, relative to one at stage N0, again of the same age and at the same level of stage T, is $e^{1.34} = 3.8$.

Finally, the hazard ratio for an individual at a given level of stages T and N, relative to another patient at the same levels of stages T and N whose age is one unit less, is $e^{0.053} = 1.05$. Since this is greater than unity, we conclude that, other things being equal, any given patient has a 5% greater hazard of death than a patient one year younger.

4.4.7 Residuals for the Cox model (model diagnostic)

To check whether the Cox PH model is suitable for the data, we examine different type of residuals as explained in next subsections.

Cox-Snell residuals

The residuals given by Cox & Snell (1968) is

$$r_j = \hat{H}_0(t_j)\exp\big(X_j\hat{\beta} + Z_j\hat{b}\big); j = 1, 2, \ldots, n.$$

Here, $\hat{H}_0(t)$ is an estimate of the baseline cumulative hazard function at time t_j. In practice, the Nelson-Allen estimate is used which is given by

$$\hat{H}_0(t) = -\log \hat{S}_0(t) = \sum_{i=1}^{k} \frac{d_j}{\sum_{j \in R(t_{(i)})} \exp\left[\sum_{k=1}^{p} \beta_k X_{jk}\right]}.$$

The Cox-Snell residual can also be expressed as

$$r_j = \hat{H}_j(t_j) = -\log \hat{S}_j(t_j),$$

where $\hat{H}_j(t_j)$ is the estimated cumulative hazard and $\hat{S}_j(t_j)$ is the survival function of the jth individual at t_j (based on the model).

If T is a random variable associated with survival time for an individual and $S(t)$ is the survival function, then the random variable $r = -\log S(T)$ has an exponential distribution with a mean equal to one, irrespective of the form of $S(t)$. Therefore, if the model is correct, the Cox-Snell residuals should approximately follow an exponential distribution with with a mean equal to one, and the cumulative hazard function of the residuals will be $H_r(t) = t$. As a result, the plot of the cumulative hazard estimate of residuals by the Nelson-Aalen versus the Cox-Snell residuals should be a straight line through the origin (intercept=0) with a slope of 1.

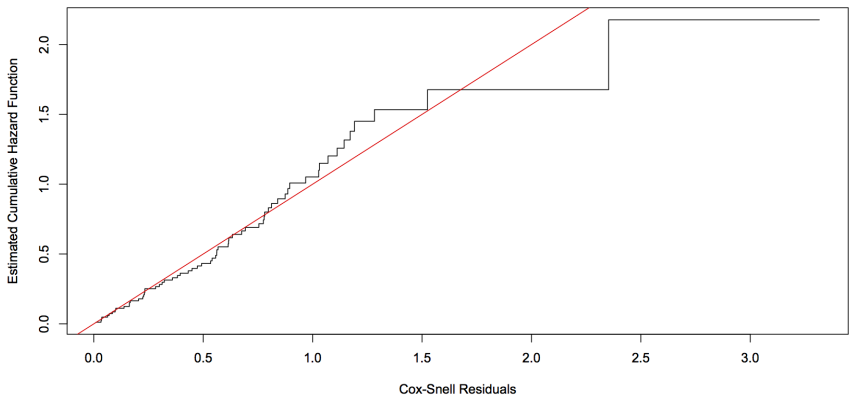


Figure 4.8: Cumulative hazard of Cox-Snell residuals (solid black line) from the Cox PH model fit, compared to the identity line (grey dashed line).

Martingale residuals

The Martingale residuals are defined as Cox & Snell (1968)

$$r_{Mj} = \delta_j - r_j = \delta_j - \hat{H}_j(t_j), \tag{4.7}$$

where r_j is the Cox-Snell residual. The expression (4.7) can be interpreted as the number of observed deaths minus the number of expected deaths. The Martingale residuals are different from the Cox-Snell residuals. Not only do they check the model assumption but they also suggest the form of the covariates in the model. In other words, the Martingale residuals determine the functional form of a covariate, which can be seen by plotting the Martingale residuals against the new covariate. The points are then fitted using some smoothing technique, such as the Lowess method. The appropriate functional form of the new covariate can be determined by the smoothed curve.

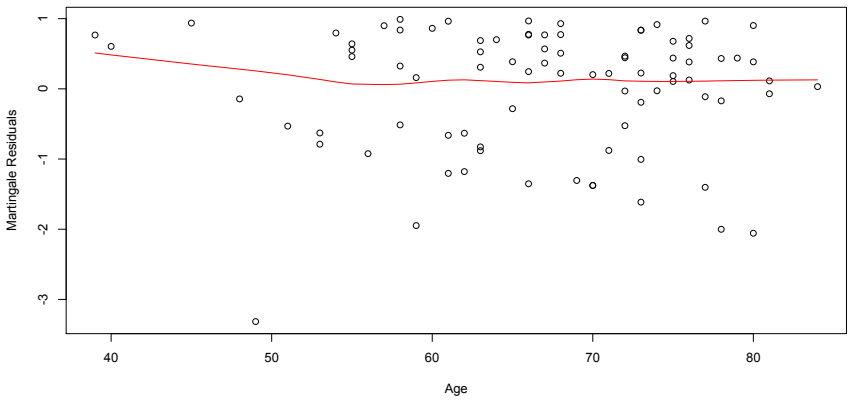


Figure 4.9: Plot of the Martingale residuals against age with a smoothed curve (solid red line).

Deviance residuals

The Martingale residuals are not symmetric, having a maximum value of 1 and a minimum value of $-\infty$, which makes plots based on the residuals hard to interpret. The deviance residual, which was introduced by Therneau *et al.* (1990) is used to obtain a

residual that is more symmetrically distributed around zero. The deviance residual is defined as

$$D_j = \text{sign}[r_{Mj}]\big[-2\big(r_{Mj} + \delta_j \log(\delta_j r_{Mj})\big)\big]^{\frac{1}{2}},$$

where r_{Mj} is the Martingale residual for the jth individual.

The deviance residuals can be plotted against the risk scores $R_i = X_i\hat{\beta}$, which provide information about whether an individual might be expected to survive for a short or long period of time. Individuals who have large negative risk scores will have a lower than average risk of death, and visa versa. The deviance residuals whose absolute values are too large relative the other deviance residuals indicate potential outliers.

From looking to Figure 4.10, the patients who have the largest deviance residuals are patient number 45 and number 21, with a deviance of 2.68 and -2.57 respectively. The first group of patients had a covariate vector of (*Age*=58, *Stage T*=2, and *Stage N*=1) and died at 36 days, whereas the second group of patients had a covariate vector of (*Age*=49, *Stage T*=3, *Stage N*=1) and died at 3185 days. Based on the risk profile of patient 45 (risk score=-0.59), the patient should have had a relatively long survival time, but he was in fact the shortest-lived patient. On the other hand, based on her risk profile (risk score=0.73), patient 21 should have had a relatively short survival time, but she was in fact one of the longest lived patients, and is censored.

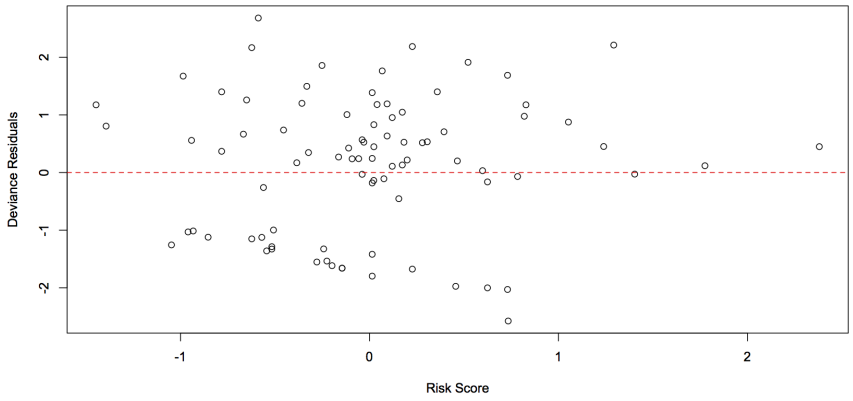


Figure 4.10: Plot of the deviance residuals versus the risk scores.

Evaluation of the proportional hazard assumption

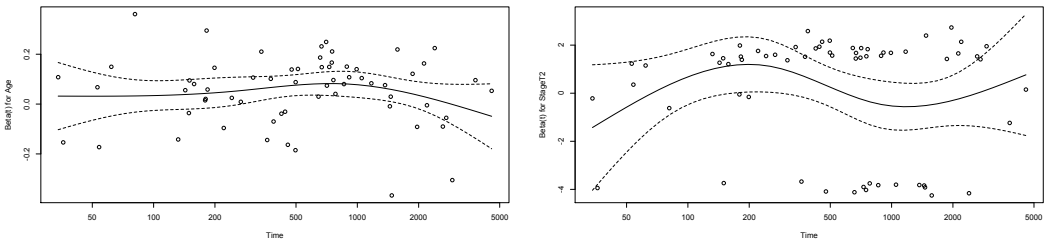
The proportional hazard assumption can be evaluated using tests and graphical diagnostics based on the scaled Schoenfeld residuals. The tests of the proportional hazard assumption are calculated for each covariate by correlating the corresponding set of scaled Schoenfeld residuals with a suitable transformation of time.

	rho	chisq	p
Age	-0.0162	0.0201	0.8872
Stage T2	-0.0712	0.3155	0.5743
Stage T3	-0.3609	8.9826	0.0027
Stage N1	0.0869	0.4864	0.4856
Stage N2	0.0060	0.0023	0.9621
GLOBAL		10.4380	0.0637

Table 4.12: Proportional hazard assumption for each covariate along with a global test for the model as a whole, with the null hypothesis that the Cox proportional hazard assumption is valid

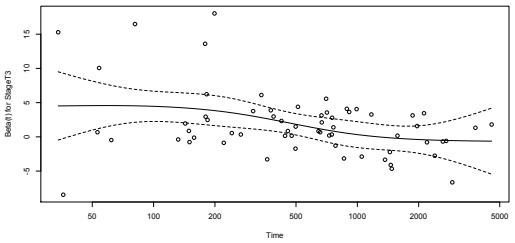
The three columns of Table 4.12 are rho, chisq, and p. The rho column shows the Pearson product-moment correlation between the scaled Schoenfeld residuals and the time for each covariate The chisq column shows the test statistics, and p column gives the p-value. The last row, Global, gives the global test's overall covariates.

There is no evidence for nonproportionality, as shown by the p-value of 0.06 for the global test. All of the covariates, except *Stage T3*, show strong evidence for proportionality as shown by the p-values in Table 4.12. *Stage T3* shows some evidence of nonproportionality. However, this is not of concern, because *Stage T3* represents the most severe cases, when the size of the tumour is larger than 7 cm.

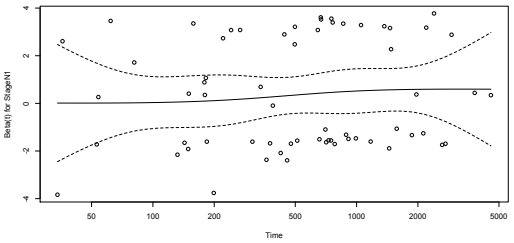


(a) Age

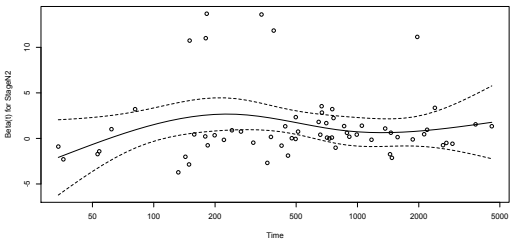
(b) Stage T2



(c) Stage T3



(d) Stage N1



(e) Stage N2

Figure 4.11: Plots of scaled Schoenfeld residuals against transformed time for each covariate in a model of the lung cancer data. The solid line is a smoothing spline fitted to the plot, with the dotted lines representing a ± 2-standard-error band around the fit.

Chapter 5

Extending Cox PH model : Normal random effects

5.1 Introduction

Accurate survival prediction is critical in the management of cancer patients' care and well-being. Previous studies have shown that copy number alterations (CNA) in some key genes are individually associated with disease phenotypes and the patient's prognosis. For example, Haan *et al.* (2014) investigated CNA in 194 regions in the genome that are associated with colorectal cancer. The study also showed that there are some significant differences in patients' survival probability between those who have gains and those without gains in some key genes. Chiu *et al.* (2014) showed there is significant difference in melanoma patients' survival between high and low risk groups based on the CNA in five marker genes. Gatza *et al.* (2014) also indicated that CNA in some key genes essential for cell proliferation significantly affect the survival of breast cancer patients. Lu *et al.* (2015) suggested that CNA in MYC and BCL2 is significantly associated with the survival probability of patients with diffuse B-cell lymphoma. Mampaey *et al.* (2015) showed that colon cancer patients who have a loss in chromosome four have lower survival probability than those without a loss.

All of the above studies suggest that CNA in some key regions in the genome hold information that is relevant in the estimation of patients survival probability. However, cancer is a complex disease where the pattern of CNA across the genome exhibits complex gains and losses. Given the complicated network of genes in the development

and proliferation of cancer, the whole genome-wide pattern of CNA contains a considerable portion of the genomic information relevant to patients survival. However, many of these genomic variations are passenger events caused by the disease, rather than driver events which are influencing the disease. Furthermore, genome-wide CNA profile is patient-specific and any differences in the profile between patients may help to explain the differences in the patients survival. Although some of the information in the genome-wide CNA profile is critical in the prediction of cancer patients survival, extracting that information from the background noise remains challenging.

In terms of modelling, the main challenge is how to incorporate genome-wide CNA profiles, in addition to clinical information, to predict cancer patients survival. In this chapter, we propose to extend the Cox proportional hazard (PH) model discussed in Chapter 4, by including the CNA profiles as random effects predictors. The (standard) Cox PH model has been used extensively in the prediction of patients survival based on their clinical variables. In our thesis, we extend the model to incorporate patients' genome-wide CNA profiles, in addition to the clinical variables.

The organization of this chapter is as follows. Section 5.2 discusses the extension of Cox PH model to include the copy number alteration as random effects. In Section 5.3, the estimation of the unknown parameters of the model is discussed. Section 5.4 describes some computational issues. In Section 5.5, Breslow's estimator of the baseline cumulative hazard rate and the estimates of survivor function are presented. Section 5.6 discusses residuals for the extended Cox PH model. Simulation studies are described and discussed in Section 5.7. Finally, The results and evaluation of the lung cancer dataset are found in Section 5.8.

5.2 Cox proportional hazards model

In this section, we discuss incorporating genom-wide CNA profiles into the basic Cox PH model, discussed in Chapter 4, as proposed by Cox *et al.* (1972). However, the main challenge lies in the dimension of the matrix of CNA profiles, denoted Z. The matrix Z is of size $n \times q$ with $n \ll q$, where q is the number of genomic regions in the CNA profiles (in our lung cancer cohort, $n = 80$ and $q = 13,968$). It is clear that incorporating the CNA profiles as fixed predictors as in the original Cox PH model will make the model parameters β unestimable.

To deal with this problem, we extend the Cox PH model by including the CNA profiles as random effects in the original model such that the hazard rate at time t can be modelled as:

$$h_i(t|X, Z) = h_0(t) \exp\{X_i\beta + Z_i b\} \tag{5.1}$$

where $h_0(t)$ is the baseline hazard function, denoting the baseline hazard rate for all the patients in the group across time and does not depend on any predictors, X_i, which is $i-$th row (vector) of X, is a p vector of fixed predictors for the $i-$th patient. Z_i is the $i-$th row (vector) of Z (of length q), b is a $q-$vector of random effects that we assume to follow a normal distribution $b \sim N(0, D(\theta))$, and $D(\theta) = \theta I_q$ (I_q is an identity matrix of size q). We consider the approach by Ripatti & Palmgren (2000) which linked together the likelihood approximation of Breslow & Clayton (1993) and the penalised likelihood concept, to derive a generalization of the model estimation. The marginal (integrated) likelihood $L(h_0(t), \beta, \theta)$ for model (5.1) is

$$\begin{aligned} L(h_0(t), \beta, \theta) &= \int \prod_{i=1}^{n} h_i(t|b)^{\delta_i} S_i(t|b) p(b; D(\theta)) db \\ &= \int \prod_{i=1}^{n} [h_0(t) \exp(X_i\beta + Z_i b)]^{\delta_i} \exp[-H_0(t) \exp(X_i\beta + Z_i b)] p(b; D(\theta)) db, \end{aligned} \tag{5.2}$$

where the random effects b are integrated out. As b is restricted to follow a multivariate normal distribution with mean 0 and covariance variance matrix $D(\theta)$, the probability density function (PDF) of this multivariate normal distribution is

$$p(b; D(\theta)) = (2\pi)^{-\frac{q}{2}} |D(\theta)|^{-\frac{1}{2}} \exp\{-\frac{1}{2} b' D(\theta)^{-1} b\}.$$

Unfortunately, the above integral Eq. (5.2) is difficult to solve . Therefore, we follow Breslow & Clayton (1993) in their approach for the GLMM and use a Laplace approximation for the integral in (5.2).

First, instead of using the marginal likelihood $L(h_0(t), \beta, \theta)$ (Eq. (5.2)), we use the log of the marginal likelihood $\log(L(h_0(t), \beta, \theta)) = \ell(h_0(t), \beta, \theta)$ which is

$$\begin{aligned} \ell(h_0(t), \beta, \theta) = \int \sum_{i=1}^{n} \Big[\delta_i \big[\log(h_0(t)) + X_i\beta + Z_i b \big] - H_0(t) \exp(X_i\beta + Z_i b) \Big] \\ - \frac{q}{2} \log(2\pi) + \log |D(\theta)|^{-\frac{1}{2}} - \frac{1}{2} b' D(\theta)^{-1} b db. \end{aligned} \tag{5.3}$$

Then, we write Eq. (5.3) in the form :

$$e^{\ell(h_0(t),\beta,\theta)} \propto C|D(\theta)|^{-\frac{1}{2}} \int e^{-k(b)},$$

where C are the constant terms that are not related to the parameters. k' and k'' denote the q vector and $q \times q$ dimensional matrix of first-and second-order partial derivatives of k with respect to b. Ignoring the multiplicative constant C, the approximation returns

$$\ell(h_0(t), \beta, \theta) \approx -\frac{1}{2}\log|D(\theta)| - \frac{1}{2}\log|k''(\tilde{b})| - k(\tilde{b})$$

where

$$k(\tilde{b}) = -\left\{\left[\sum_{i=1}^{n}[\delta_i\big[\log(h_0(t)) + X_i\beta + Z_i\tilde{b}\big] - H_0(t)\exp(X_i\beta + Z_i\tilde{b})\right] - \frac{1}{2}\tilde{b}'D^{-1}(\theta)\tilde{b}\right\}$$

and $\tilde{b} = \tilde{b}(\beta, \theta)$ denotes the solution to the partial derivatives of $k(b)$ with respect to b. i.e., $\tilde{b}$ satisfies

$$k'(\tilde{b}) = -\sum_{i=1}^{n} Z_i[\delta_i - H_0(t)\exp(X_i\beta + Z_i\tilde{b})] - D(\theta)^{-1}\tilde{b} = 0.$$

The set of second partial derivative of $k(b)$ with respect to b denoted $k''(b)$ has the form

$$k''(b) = \sum_{i=1}^{n} H_0(t)\exp(X_i\beta + Z_i\tilde{b})Z_iZ_i' + D(\theta)^{-1}.$$

Therefore, as shown above, the approximate marginal log likelihood by using the Laplace approximation leads to :

$$\begin{aligned}\ell(h_0(t), \beta, \theta) \approx & -\frac{1}{2}\log|D(\theta)| \\ & - \frac{1}{2}\log\left|\sum_{i=1}^{n} H_0(t)\exp(X_i\beta + Z_i\tilde{b})Z_iZ_i' - D(\theta)^{-1}\right| \\ & + \sum_{i=1}^{n}\Big[\delta_i[\log(h_0(t)) + X_i\beta + Z_i\tilde{b}] \\ & - H_0(t)\exp(X_i\beta + Z_i\tilde{b})\Big] - \frac{1}{2}\tilde{b}'D(\theta)^{-1}\tilde{b}.\end{aligned} \tag{5.4}$$

If θ were known and b were considered a fixed effects parameter, the first two terms are ignored and β can be chosen to maximize the second two terms which is a

penalized log likelihood as shown in Green (1987). Thus $(\hat{\beta}, \hat{b}) = (\hat{\beta}(\theta), \hat{b}(\theta))$, where $\hat{b}(\theta) = \tilde{b}(\hat{\beta}(\theta))$, jointly maximize

$$\sum_{i=1}^{n} \left[\delta_i [\log(h_0(t)) + X_i\beta + Z_i\tilde{b}] - H_0(t)\exp(X_i\beta + Z_i\tilde{b}) \right] - \frac{1}{2}\tilde{b}' D(\theta)^{-1}\tilde{b}. \quad (5.5)$$

Equation (5.5) is the full log likelihood for Cox model with b as another set of parameters and penalty term. It turns out that it can be maximized using penalized fixed effect partial likelihood as Cox showed in Cox *et al.* (1972):

$$\ell_P(\beta, \theta, b) = \sum_{i=1}^{n} \left[\delta_i (X_i\beta + Z_i b) - \delta_i \log\left(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \right) \right] - \frac{1}{2} b' D(\theta)^{-1} b, \quad (5.6)$$

where $\frac{1}{2}b' D(\theta)^{-1}b$ is the penalty term penalizing for extreme value of b.

5.3 Parameter Estimation

5.3.1 Estimation of β and b

We can derive an estimation of β and b at fixed θ by partial differentiation of the log partial likelihood $\ell_P(\beta, b)$ in Eq. (5.6) with respect to each of β and b. The resulting estimating equations for β and b, respectively, are

$$u(\beta) = \sum_{i=1}^{n} \delta_i \left[X_i - \frac{\sum_{j \in R(t_i)} X_j \exp(X_j\beta + Z_j b)}{\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b)} \right] \quad (5.7)$$

and

$$u(b) = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j \in R(t_i)} Z_j \exp(X_j\beta + Z_j b)}{\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b)} \right] - D(\theta)^{-1} b. \quad (5.8)$$

The second derivatives with respect to β , required for the information matrix, take the form

$$\frac{\partial^2 \ell_p}{\partial\beta_l \partial\beta_m} = \frac{\sum_{i=1}^{n} \left(\sum_{j \in R(t_i)} x_{jl} \exp(X_j\beta + Z_j b) \right) \left(\sum_{j \in R(t_i)} x_{jm} \exp(X_j\beta + Z_j b) \right)}{\left(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \right)^2} - \frac{\sum_{i=1}^{n} \left(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \right) \left(\sum_{j \in R(t_i)} x_{jl} x_{jm} \exp(X_j\beta + Z_j b) \right)}{\left(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \right)^2},$$

and the second derivatives with respect to b take the form

$$\begin{aligned}\frac{\partial^2 \ell_p}{\partial b_l \partial b_m} =& \frac{\sum_{i=1}^{n} \big(\sum_{j \in R(t_i)} z_{jl} \exp(X_j\beta + Z_j b) \big) \big(\sum_{j \in R(t_i)} z_{jm} \exp(X_j\beta + Z_j b) \big)}{\big(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \big)^2} \\ & - \frac{\sum_{i=1}^{n} \big(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \big) \big(\sum_{j \in R(t_i)} z_{jl} z_{jm} \exp(X_j\beta + Z_j b) \big)}{\big(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \big)^2} \\ & - D(\theta)^{-1}.\end{aligned}$$

The information matrix of β and b, respectively, are

$$I(\beta) = -\ell''_P(\beta) = \left[- \frac{\partial^2 \ell_p}{\partial \beta_l \partial \beta_m} \right],$$

and

$$I(b) = -\ell''_P(b) = \left[- \frac{\partial^2 \ell_p}{\partial b_l \partial b_m} \right].$$

For later use we can write $I(b)$ as

$$I(b) = A + D(\theta)^{-1}, \tag{5.9}$$

where

$$\begin{aligned}A =& \frac{\sum_{i=1}^{n} \big(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \big) \big(\sum_{j \in R(t_i)} z_{jl} z_{jm} \exp(X_j\beta + Z_j b) \big)}{\big(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \big)^2} \\ & - \frac{\sum_{i=1}^{n} \big(\sum_{j \in R(t_i)} z_{jl} \exp(X_j\beta + Z_j b) \big) \big(\sum_{j \in R(t_i)} z_{jm} \exp(X_j\beta + Z_j b) \big)}{\big(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \big)^2}\end{aligned}$$

The estimates $(\hat{\beta}(\theta), \hat{b}(\theta))$ can be found by alternating between solving (5.7) and (5.8) at fixed θ using a Newton-Raphson algorithm which is summarized in these steps:

1. Given a value θ
2. Let $k = 0$
3. Choose the initial values ($\hat{\beta}^{(0)}$ and $\hat{b}^{(0)}$).
4. $\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + I^{-1}(\hat{\beta}^{(k)}) u(\hat{\beta}^{(k)})$
   and $\hat{b}^{(k+1)} = \hat{b}^{(k)} + I^{-1}(\hat{b}^{(k)}) u(\hat{b}^{(k)})$
5. increase k by one.
6. Go back to step 4 and repeat until convergence.

5.3.2 Computational issues: The calculation of the inverse of high dimensional matrix $(I(b))$

When we try to apply our method to estimate the parameters (β, b), we need to take into consideration the time and the memory. Our estimation based on Newton-Raphson method which require the inverse of information matrix $(I(b))$ in (E.q (5.9)) which is a huge matrix, in our case we need to make an inverse of 13986×13986. To deal with this problem, we consider two different ways:

First

The first solution is based in Pawitan (2001). If we can write the information matrix $I(b)$ in (E.q (5.9)) in this form,

$$V = \Sigma + ZDZ',$$

it can be inverted by :

$$V^{-1} = \Sigma^{-1} - \Sigma^{-1}Z(Z'\Sigma^{-1}Z + D^{-1})^{-1}Z'\Sigma^{-1}.$$

Let $V = I(b)$, and $\Sigma = D(\theta)^{-1}$. Therefore, if we can write A in (E.q (5.9)) in this form $Z'DZ$, we will end up with inverting 80×80 instead of 139680×13986.

Therefore, we are left with converting A to $Z'DZ$; as shown in the next two points

- We can write $A = Z'DZ$ by using singular value decomposition and take the first 80 singular values and the corresponding left and right-singular vectors. One draw back of this way is the time consumption to calculate singular value decomposition especially with a big matrix.
- In order to save time and memory we can use the method called (IRLBA) introduced by Baglama & Reichel which is a fast and memory-efficient method for computing a few approximate singular values and singular vectors of large matrices.

Second

The second solution is the sparse computation based on Therneau *et al.* (2003).

When we look carefully to the information matrix $I(b)$, we found that the the information matrix $I(b)$ is a diagonally dominant matrix, where adding the penalty further increases the dominance of the diagonal. Therefore, using a sparse option, where only the diagonal of $I(b)$ is retained, should not have a large impact on the estimation procedure.

Ignoring the off diagonal of $I(b)$ has a number of implications:

1. The speed is increased dramatically and saving in space (we take 13986 elements of diagonal of $I(b)$ instead of 13986*13986)
2. The solution points is identical; ignoring trivial difference due to distinct iteration paths; because the score vector $U(b)$ and likelihood are not changed.
3. The Newton-Raphson iteration may undergo a slight loss of efficiency so that 1-3 more iteration required

For comparison consideration, we compare the four different methods which differ in the way that we make the inverse of $I(b)$.

1. The original method with the full information matrix (FULL).
2. The method which used Pawitan hint using default SVD (SVD).
3. The method which used Pawitan hint using IRLBA SVD (IRLAB).
4. The method based in Sparse solution (Sparse).

Next table shows the absolute mean difference between the four methods for $\theta = 1 \times 10^{-5}$ and number of windows $= 5000$.

Method	mean of absolute difference
FULL	-
SVD	7.59e-20
IRLAB	8.66e-20
Sparse	1.638678e-07

Figures 5.1, 5.2, and 5.3 shows the comparison between $\hat{b}$ across the four methods.

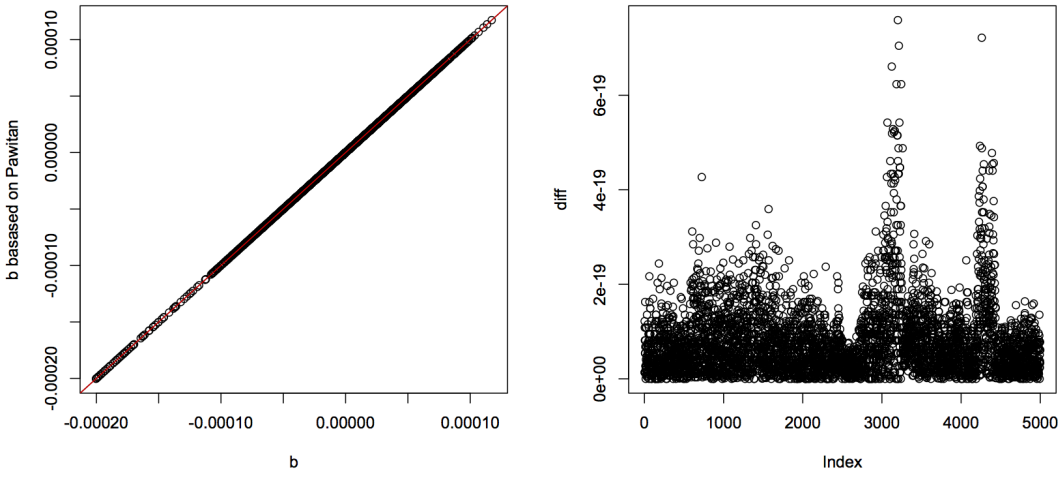


Figure 5.1: Left panel is $\hat{b}$ based on full information matrix VS $\hat{b}$ based on Pawitan and SVD; and the right panel is the absolute difference

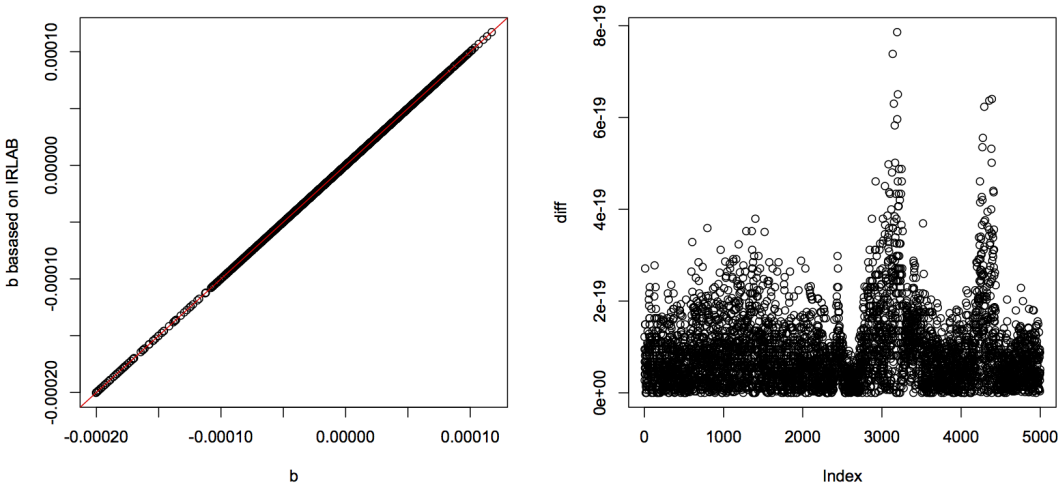


Figure 5.2: Left panel is $\hat{b}$ based on full information matrix VS $\hat{b}$ based on Pawitan and IRLAB; and the right panel is the absolute difference

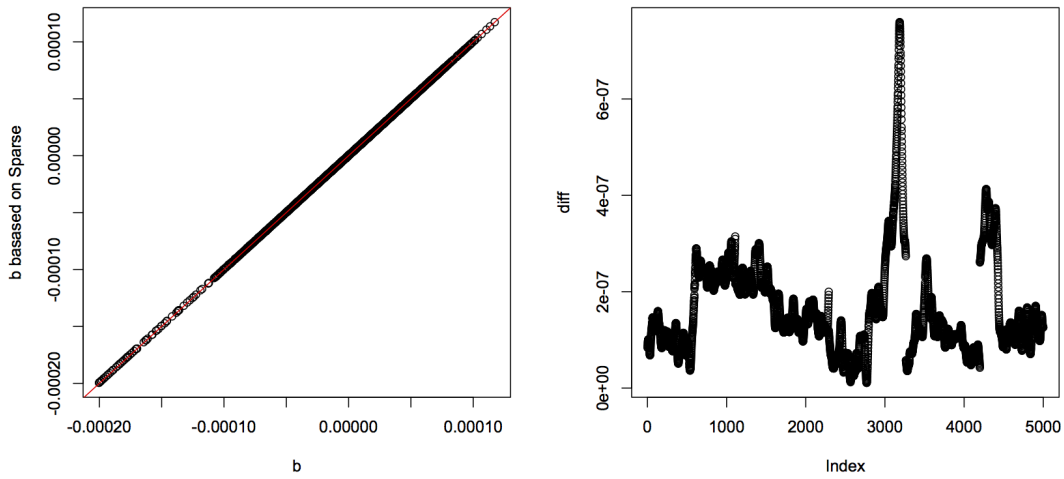


Figure 5.3: Left panel is $\hat{b}$ based on full information matrix VS $\hat{b}$ based on sparse information matrix; and the right panel is the absolute difference

5.3.3 Estimation of θ

To estimate θ, we will require the marginal log likelihood $\ell(h_0(t), \beta, \theta)$ in Eq. (5.3), where no constant has been dropped. The estimate of θ is then obtained as the one that minimizes the Akaike's Information Criterion (AIC)

$$\text{AIC}(\theta) = -2 \times \ell(h_0(t), \beta, \theta) + 2 \times \text{df} \tag{5.10}$$

across different values of θ evaluated, where $\ell(h_0(t), \beta, \theta)$ is the log of the marginal likelihood (E.q (5.4)) and df is the degrees of freedom of fit, which is calculated as (Gray (1992))

$$\text{df} = q - \text{trace}[D^{-1}H^{-1}] \tag{5.11}$$

where H is the Hessian matrix from the estimation of β and b in Section 5.3.1.

The evaluation of AIC is not the only way to estimate θ; there are other alternatives. Firstly, Ripatti & Palmgren (2000) obtained an expression for an approximate estimate of θ based on the log marginal likelihood $\ell(h_0(t), \beta, \theta)$. In other words, this approach maximizes a profile quasi-likelihood function. The expression, although feasible in their context with a handful random predictors, is not computationally practical for

large datasets. Secondly, θ can be estimated via cross-validation (van Houwelingen *et al.* (2006)) and this will be discussed in details in Chapter 6. We find in our data that these three methods are equivalent in estimating θ. Also, the simulation results in Section 5.7 shows that the estimation of the variance component θ is similar among these approaches. Our choice to use AIC is for practical reasons, since it is somewhat faster. Moreover, we can speed up this method as described in the next section.

5.4 Computational considerations

The AIC approach above to estimate θ requires us to evaluate AIC (and estimation of β and b) across a range of θ (grid search). Although we find that this method is still faster than the estimation proposed in Ripatti & Palmgren (2000), we have found it more efficient and accurate to minimize AIC using bisection or a quadratic optimization technique. The latter is found to be the fastest technique to converge to the optimal θ.

The steps in the bisection technique can be briefly describe as follows.

1. Determine the largest and smallest plausible values of θ, say, θ_a and θ_b
2. Calculate AIC at $\theta = \theta_a$ and $\theta = \theta_b$.
3. Retain θ which give a lower AIC, and replace the other θ by $\frac{\theta_a+\theta_b}{2}$
4. Repeat steps $2-3$ until convergence to obtain $\hat{\theta}$.

Similarly, the steps taken in quadratic optimization technique are

1. Determine three values for θ, say $\theta_S, \theta_M, \theta_L$.
2. Calculate AIC for these θs, say A_S, A_M, A_L.
3. Solve a system of three equations with three unknowns $(c_1, c_2, \text{and}, c_3)$

$$A_S = c_1 + c_2\theta_S + c_3\theta_S^2$$

$$A_M = c_1 + c_2\theta_M + c_3\theta_M^2$$

$$A_L = c_1 + c_2\theta_L + c_3\theta_L^2$$

4. Calculate $\theta_N = \frac{-c_2}{2c_3}$.

5. Replace $\text{argmax}_\theta\ A$ by θ_N

6. Repeat steps $2-5$ until convergence.

5.5 Estimation of $h_0(t)$ and $S(t)$

Our main interest in this modelling is to predict the survivor function $\hat{S}(t)$ for a patient with certain clinical characteristics and CNA profile. To do this, we first estimate the baseline hazard function $h_0(t)$ after we obtain $\hat{\beta}$ and $\hat{b}$ by using an extension of the Breslow (1974)'s estimator by

$$\hat{h}_0(t_i) = \frac{1}{\sum_{j \in R(t_i)} \exp(X_j\hat{\beta} + Z_j\hat{b})}. \tag{5.12}$$

The cumulative hazard function $H_0(t)$ can be similarly estimated as:

$$\hat{H}_0(t) = \sum_{t_i \leq T_j} \frac{1}{\sum_{j \in R(t_i)} \exp(X_j\hat{\beta} + Z_j\hat{b})}, \tag{5.13}$$

and the baseline survivor function $S_0(t)$ is

$$\hat{S}_0(t) = \exp\{-\hat{H}_0(t)\}. \tag{5.14}$$

The predicted survivor function for a new patient with a known clinical characteristic x' (a $p-$vector) and CNA alteration profiles z' (a $q-$vector) is

$$\hat{S}_i(t; x, z) = \hat{S}_0(t)^{\exp\{x'\hat{\beta} + z'\hat{b}\}},$$

where $\hat{S}_i(t)$ are from Eq. (5.14) above with $\hat{\beta}$ and $\hat{b}$ fitted from the (current) data.

In a general context, the quantity $R_i = X_i\hat{\beta} + Z_i\hat{b}$ is considered as the $i-$th individual risk score; higher R_i is associated with higher risk or hazard and lower R_i with lower hazard.

To show that the results of the estimation of the survivor function are consistent, we follow the approach of Pawitan *et al.* (2004). We use all data set to estimate the parameters $\hat{\beta}$ and $\hat{b}$. After that :

- we compute the risk score for every patient in the data

$$R_i = X_i\hat{\beta} + Z_i\hat{b}$$

- we estimate the survival curve for each patient

$$\hat{S}_i = \{\hat{S}_0(t)\}^{\exp\{R_i\}}$$

.

Then we divided the data into a high and low risk groups based on the risk score R_i's. There are more than one way to choose the cut off point to split the group and the result would be comparable. To be more specific, we chose the median of risk score R_i's to be the cut off point so the that the number of events is equal in the two groups. Finally, we compare the Kaplan-Meier survival curves for each group with model-based average survival function where the average is taken pointwise over time.

5.6 Model diagnostics

To check whether the Cox PH model is suitable for the data, we re-calculated the residuals for the Cox PH model, discussed in Chapter 4, to include the random effects estimates. We calculated the Cox-Snell residuals to include the random effects estimates as

$$r_j = \hat{H}_0(t_j)\text{exp}\big(X_j\hat{\beta} + Z_j\hat{b}\big); j = 1, 2, \ldots, n.$$

The Cox Snell residual can also be expressed as

$$r_j = \hat{H}_j(t_j) = -\text{log } \hat{S}_j(t_j),$$

where $\hat{H}_j(t_j)$ is the estimated cumulative hazard and $\hat{S}_j(t_j)$ is the estimated survivor function of the $j-$th individual at t_j.

Also, Martingale residuals are defined as

$$r_{Mj} = \delta_j - r_j = \delta_j - \hat{H}_j(t_j), \tag{5.15}$$

where r_j is the Cox-Snell residual defined above.

5.7 Simulation results

We generated a CNA matrix of dimension 85×100 .The first 20 columns have a relation with patients' survival times while the rest 80 are similar for all patients.

- For the first 10 columns, we generated the CNA for the patients who exceed the median survival time from $N(6, 1)$, while the others have CNA from $N(2, 1)$.
- For the columns from $11-20$, we generated the CNA for the patients who exceed the median survival time from $N(2, 1)$, while the others have CNA from $N(6, 1)$.
- For the columns from $21 - 100$, we generated the CNA for all patients from $N(4, 1)$

We applied our method using the AIC (grid search) to find the optimal θ, and $\hat{b}$. Figure 5.4 shows that the optimal θ is 0.0015 ($\log(\theta) = 6.5$) with AIC equal to 502.3207. However, by using Bisection and Quadratic equations techniques, the optimal θ is 0.0022 with AIC equals to 501.9824. They are more accurate and faster than the grid search. Quadratic equations techniques is the fastest one to converge.

Moreover, when we used the expression for approximate estimate of θ in Ripatti & Palmgren (2000), θ converged to a similar value of previous $\theta = 0.0020$. If we start by $\theta_0 = 0.01$ or $\theta_0 = 0.00001$, both have converged to $\theta = 0.00201$

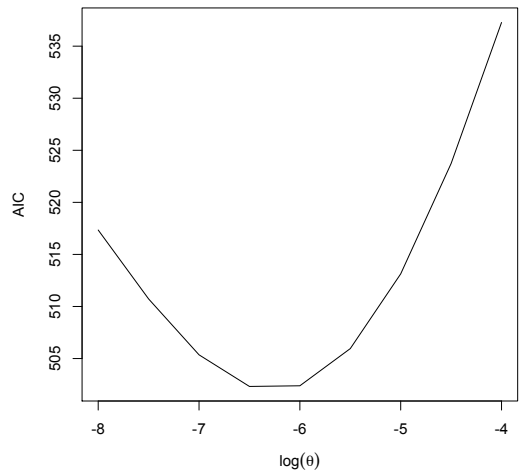


Figure 5.4: AIC for the simulated data.

The estimation of random effects b can be seen in Figure 5.5. As we anticipated, the first twenty values reveal a signal. The first 10 values have negative signals, while the values from $11 - 20$ have positive signals.

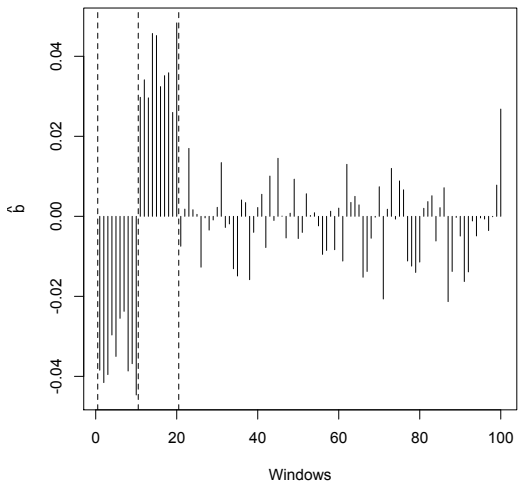


Figure 5.5: Estimation of random effects b based on the optimal θ.

Finally, we have repeated the simulation 1000 times and Figure 5.6 shows 6 different simulations, while Figure 5.7 shows the box-plot in each of 1000 simulations.

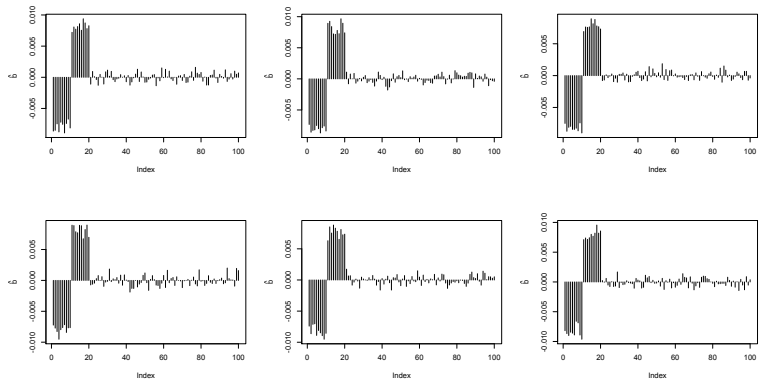


Figure 5.6: Estimation of random effects b for 6 different simulations

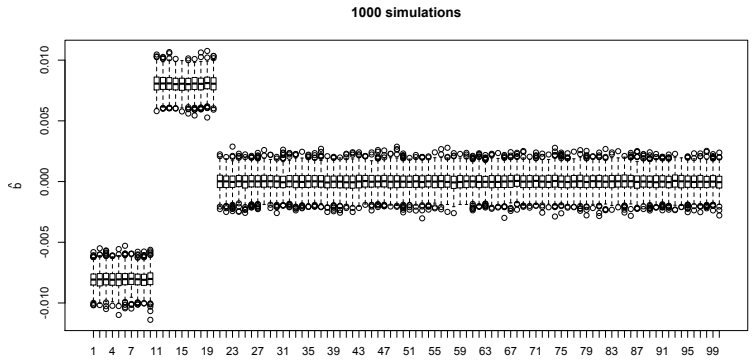


Figure 5.7: box-plot of the estimation of random effects b for 1000 simulations

5.8 Lung cancer dataset analysis

5.8.1 Model fit: estimation of θ

An important parameter to be estimated from (extended) Cox PH model is θ, which is the variance of the distribution of random effects. This parameter is important in the interpretation: as θ goes to zero (in limit terms), the estimates of the random effects

will be zero and no information in CNA are taken into account in the model. Figure 5.8 shows the process of estimating θ, under two different segmentation of CNA profiles in the data. The figure indicates that AIC (solid line) is not U shaped and decreases as $\log(\theta)$ also decreases. The figure also indicates that $2\ell(h_0(t), \beta, \theta)$ decreases as $\log(\theta)$ decreases.

To estimate θ, we use the principle described in Pawitan *et al.* (2004), by identifying the (one-sided) confidence interval for $\log(\theta)$. Looking at the value of $\{2\ell(h_0(t), \beta, \theta)\} + 3.84$ ($95-$ th percentile of the χ^2 distribution), the figure gives a one-sided confidence interval of $\log(\theta) \leq 11.5$, corresponding to $\theta \leq 10^5$. Naturally, we consider maximum θ in this confidence interval as an optimal θ because a higher value of θ corresponds to more information in CNA taken into account in the model in terms of degrees of freedom of fit. We therefore estimate θ as $\hat{\theta} = 10^5$, which corresponds to approximately 4 degrees of freedom of fit for the CNA profiles.

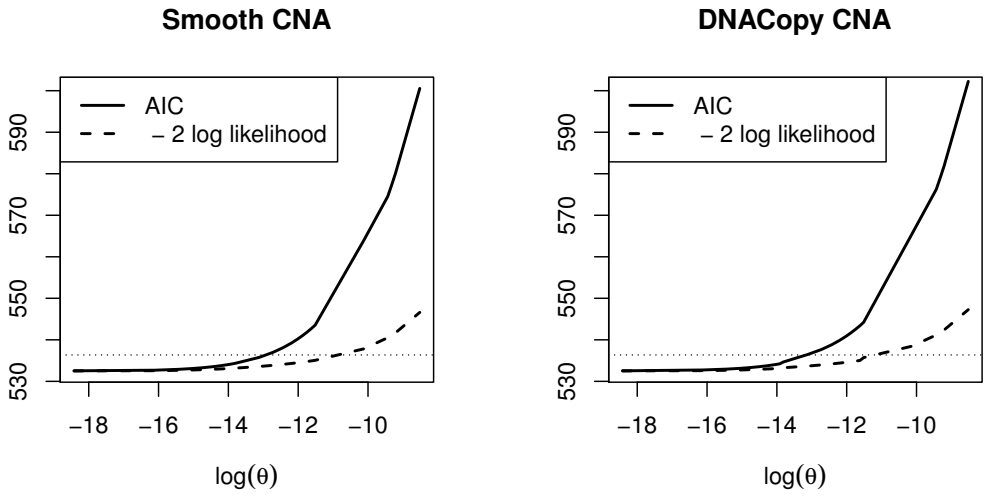


Figure 5.8: Akaike's information criterion (AIC, solid line) and $2\ell(h_0(t), \beta, \theta)$ (dashed line). The horizontal dotted line indicates $\min\{2\ell(h_0(t), \beta, \theta)\} + 3.84$ to create 95% confidence interval for θ

5.8.2 Model fit: fixed predictors

Using the optimal $\hat{\theta}$ the estimation of the fixed effects and their inference can be seen in Table 7.3. For comparison, we estimate the fixed effects under both conditions of with and without the CNA profiles in the model. The table indicates that Age, Stage-T, and Stage-N are statistically significant (p-value< 0.05). The estimates indicate that the hazard ratio increases by about six percent ($e^{0.055} \approx 1.06$) as age-at-operation increases by one year (everything else being equal). The positive estimates of Stage-T3 indicate that larger tumour size is associated with significant increase in the hazard (relative to Stage-T1 as baseline). Similarly, the estimates of Stage-N2 indicates that a wider spread of cancer cells to the lymph nodes increases significantly the hazard (relative to the Stage-N0 as baseline).

Predictor	Estimate	Exp	Std.Error	z values	p-value
(Without CNA profiles)					
Age	0.0551	1.06	0.0164	3.37	0.0008
StageT2	0.1818	1.20	0.3215	0.57	0.5700
StageT3	1.7623	5.83	0.6392	2.76	0.0058
StageN1	0.3616	1.44	0.3019	1.20	0.2300
StageN2	1.3653	3.92	0.4824	2.83	0.0047
(With smooth CNA profiles)					
Age	0.0565	1.06	0.0164	3.44	0.0006
StageT2	0.1837	1.20	0.3215	0.57	0.5679
StageT3	1.8382	6.28	0.6392	2.88	0.0040
StageN1	0.3521	1.42	0.3019	1.17	0.2435
StageN2	1.3282	3.77	0.4824	2.75	0.0059
(With DNACopy CNA profiles)					
Age	0.0568	1.06	0.0164	3.46	0.0005
StageT2	0.1961	1.22	0.3215	0.61	0.5420
StageT3	1.8459	6.33	0.6392	2.89	0.0038
StageN1	0.3487	1.42	0.3019	1.15	0.2480
StageN2	1.3556	3.88	0.4824	2.81	0.0049

Table 5.1: Summary of the fixed predictors (from left to right columns): estimates $\hat{\beta}, \exp(\hat{\beta})$, standard error of $\hat{\beta}$, test statistic $z($ under $H_0 : \beta = 0)$, and $p-$values. Stage-T1 and Stage-N0 are part of the baseline

5.8.3 Model fit: random effects

The random effect estimates b of the full Cox PH model, using CNA profile from smooth and CBS (DNACopy) segmentation are presented in the top and bottom panel of Figure 5.9, respectively. The magnitude of the estimates is relatively small (compared to the fixed effects estimates for example). This is due to the shrinkage effect on the estimation of random effects: 80 observations to estimate almost 14 thousand variables. Positive estimates of random effect indicate that the relevant windows are associated with the increase of the hazard, while negative estimates of random effects indicate the opposite. In this regard, almost all of chromosome 7, for example, are associated with the increase of hazard, while some regions in chromosome 12 are associated with a reduction of hazard.

With regard to the inference of the random effects, this is not easy because none of the individual random effects are statistically significant from zero. This is not to say that none of the random effects are associated with the hazard, but the limited number of observations in the data are not able to provide inference for thousands of parameters. We may still have information from the data, but the information is spread across all of the windows in the genome, as will be described in the next section.

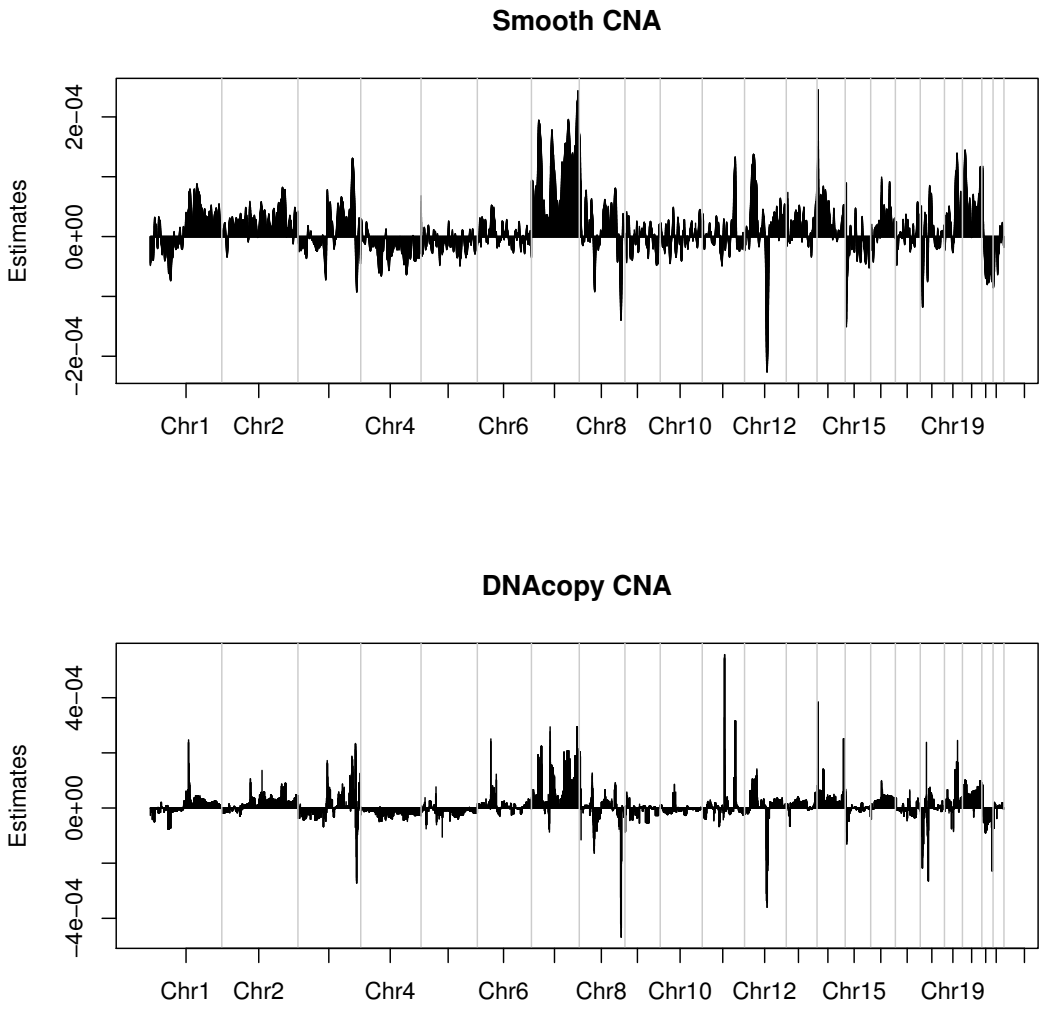


Figure 5.9: Random effects estimate b in the full model, using CNA profiles from smooth and CBS (DNACopy) segmentation (top and bottom panel, respectively). Genomic windows with missing values (for example in the centromere regions) were excluded from analysis, hence are not plotted. A more detailed view of the random effects estimates in each chromosome is presented Figure 5.10 and 5.11.

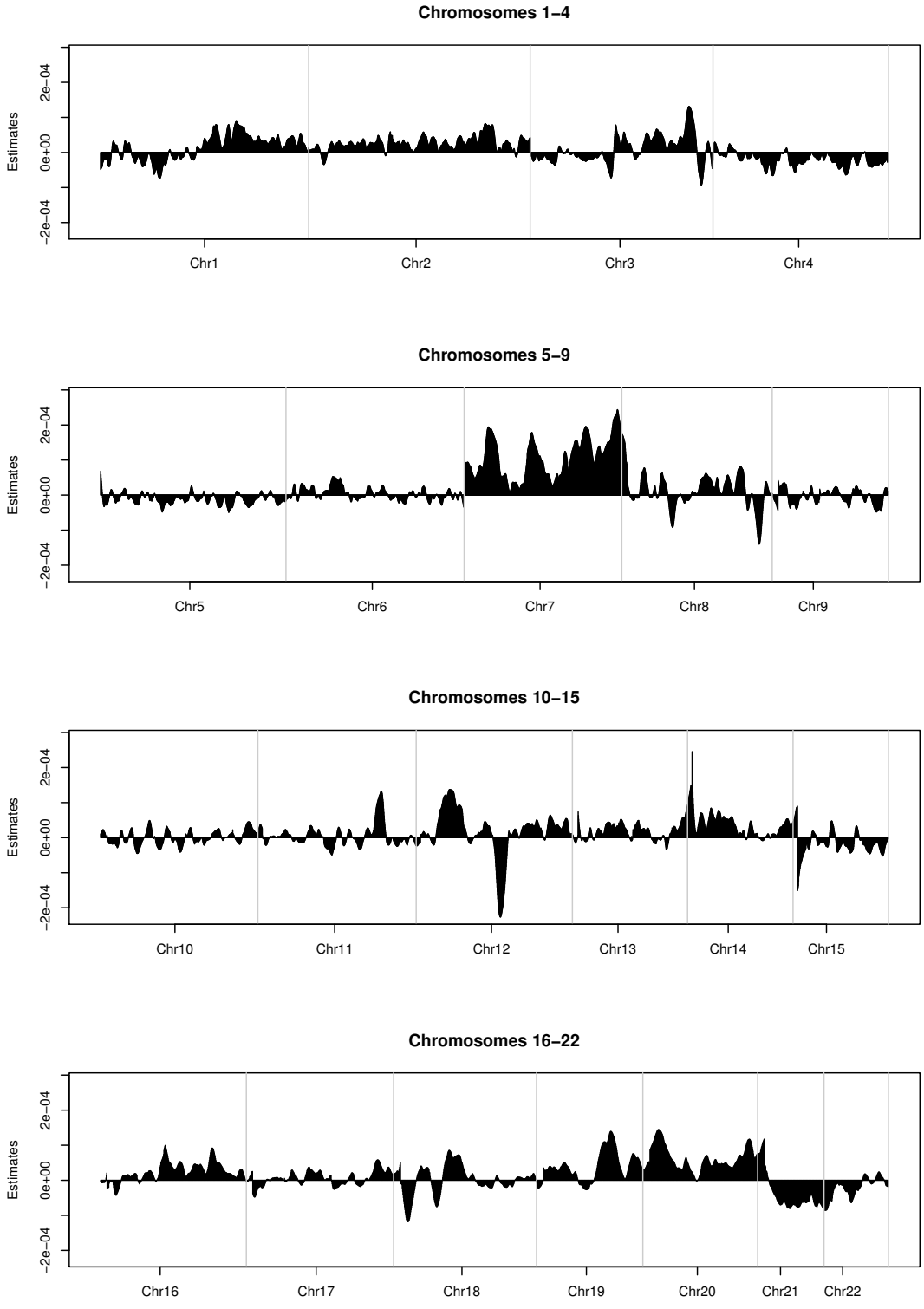


Figure 5.10: A more detailed view of the random effects estimates b in each chromosome, using CNA profiles from smooth segmentation.

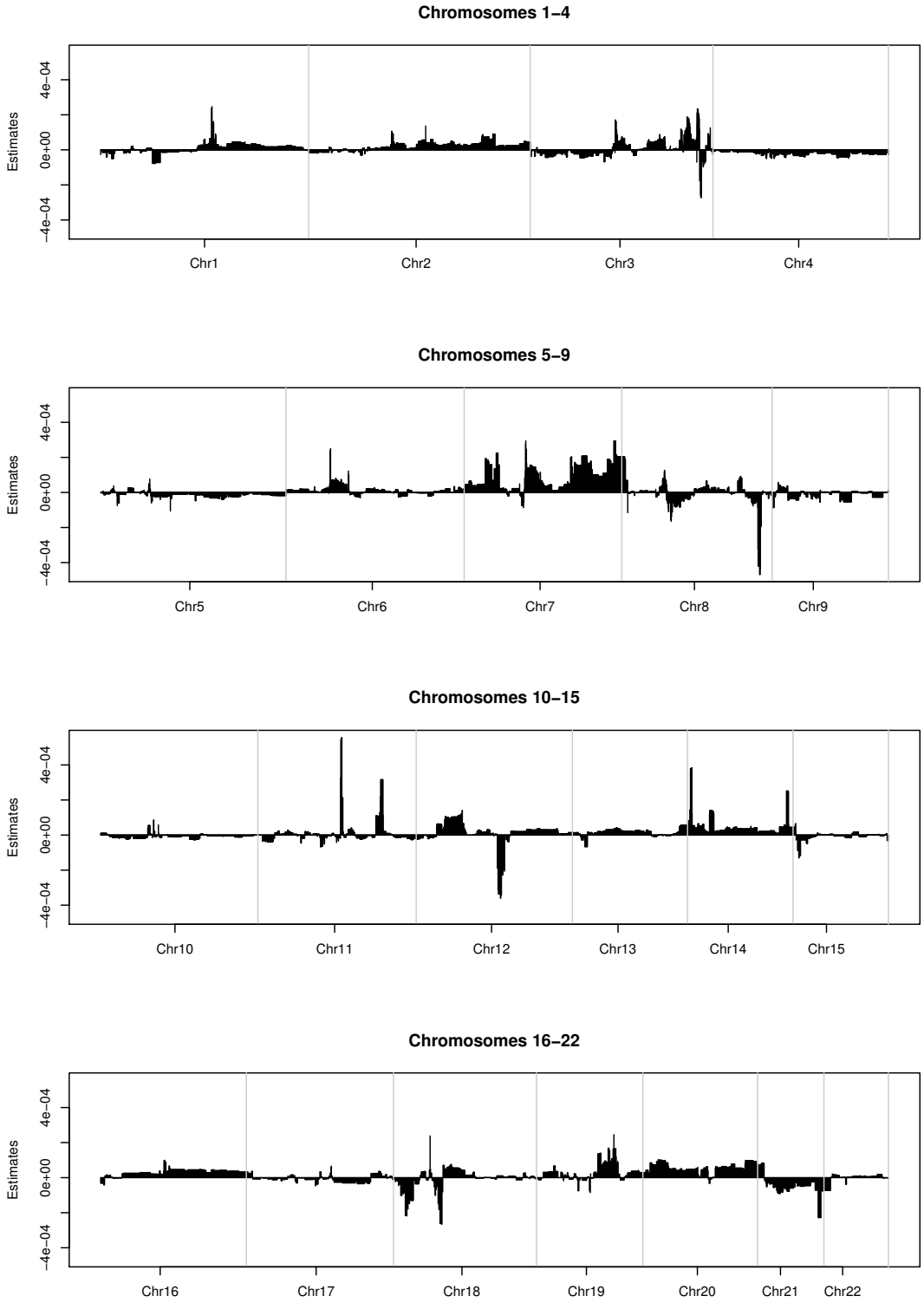


Figure 5.11: A more detailed view of the random effects estimates b in each chromosome, using CNA profiles from CBS (DNACopy) segmentation.

5.8.4 Cumulative hazard rate and the estimates of survival function

To show that the Cox PH modelling with CNA profiles is able to distinguish individuals at different levels of risk, we estimate the survivor functions for three individuals in the lung cancer dataset. They correspond to low, medium, and high risk individuals based on their risk scores R_i, corresponding to the 10th, 50th, and 90th percentile of the distribution of R_i in the dataset.

Figure 5.12 shows the estimated survivor functions for the three individuals using smooth-segmented and DNAcopy CNA profiles as random predictors ((a) and (b), respectively). The figures indicate that the median survival times for the low, medium, and high risk individuals are approximately 7.5 years, 2.5 years, and 8 months.

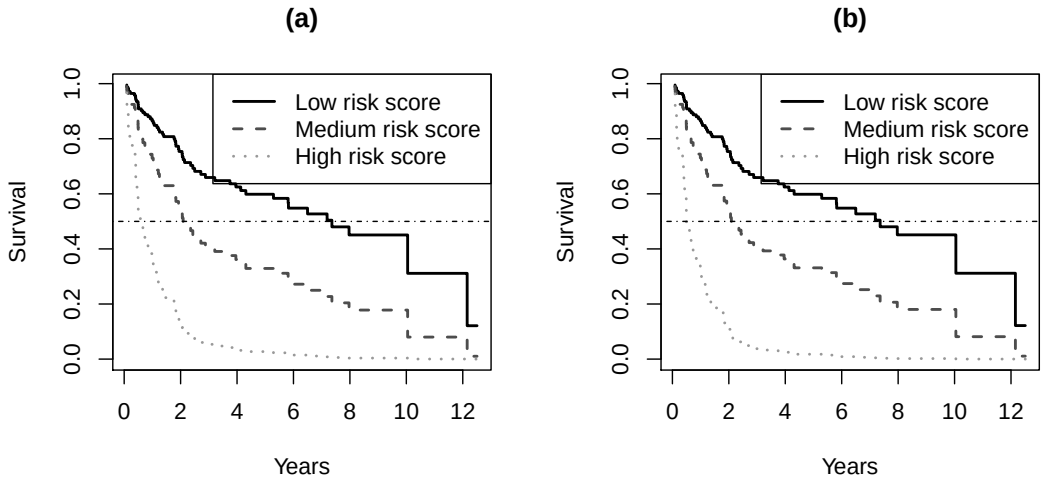


Figure 5.12: Estimated survival function from the extended Cox PH model for three individuals who are in the 10th, 50th, and 90th percentile of risk set R_i, representing low, medium, and high risk individuals respectively base on smooth-segmented and DNAcopy CNA profiles as random predictors ((a) and (b), respectively).

Figure 5.13 (a) and (b) present a comparison of estimated survivor functions based on the extended Cox PH model with the Kaplan- Meier estimates, in two groups of individuals using smooth-segmented and DNAcopy CNA profiles respectively. The

two groups are the top 10% and bottom 10% of individuals based on their risk scores R_i, corresponding to high and low risk groups respectively, and the Cox PH model-based curves are point-wise averages from those individuals. The figure indicates that the estimated survivor functions based on the model are relatively close to the estimates based on the Kaplan-Meier estimates although there are some differences.

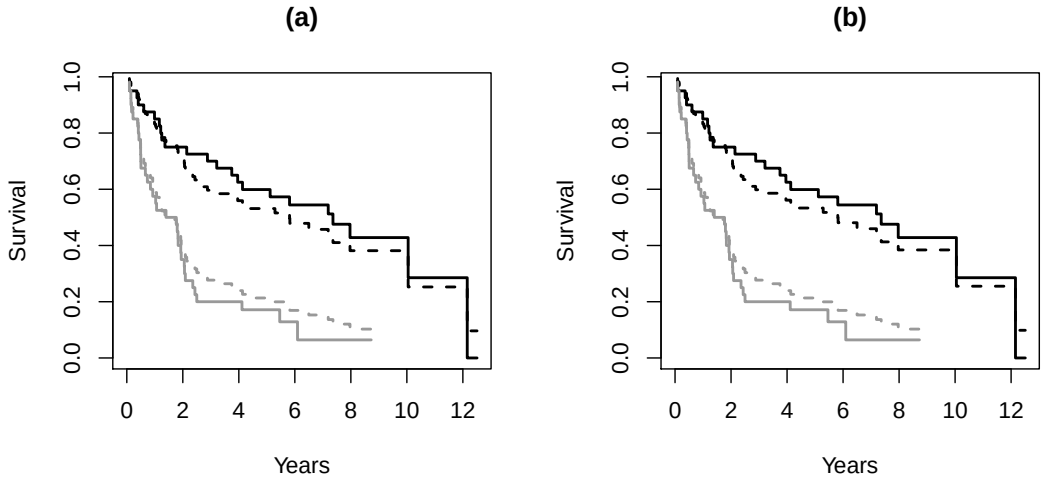


Figure 5.13: Comparison of survival function based on Kaplan- Meier estimate (solid lines) and based on the extended Cox PH model (dashed lines), in the high-risk (black lines) and low-risk (grey lines) groups. The groups are based on the top and bottom 10% of individuals based on their risk scores R_i. The figure based on the Cox PH model is pointwise average between individuals within the risk group. The horizontal dotted line marks the 50% survival probability. Panel (a) is based on smooth CNA estimates; whereas panel (b) is based on DNAcopy CNA estimates.

5.8.5 Model diagnostics

As part of the model diagnostics, we plot the cumulative hazard of the Cox-Snell residuals from the model fitting based on smooth segmented CNA (a) and DNAcopy CNA (b) as shown in Figure 5.14 (solid black line). The figure indicates that the cumulative hazard line is very close to the identity line, which suggests that the extended Cox PH model is suitable and has a reasonably good fit for the CNA profile data. The cumula-

tive hazard line near the top right corner of the figure is slightly jagged, as expected, due to rare event (death) near the upper end of the survival time distribution.

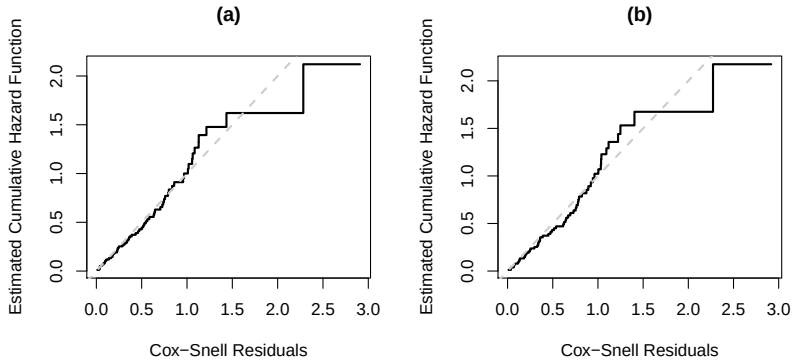


Figure 5.14: Cumulative hazard of Cox-Snell residuals (solid black line) from the Cox PH model fit, compared to the identity line (grey dashed line), based on smooth-segmented CNA (a) and DNAcopy CNA (b) profiles.

The Martingale residuals determine the functional form of a covariate, which can be seen by plotting the Martingale residuals against the new covariate. Then the points are fitted by using some smoothing technique such as the Lowess method. The appropriate functional form of the new covariate can be determined by the smooth curve.

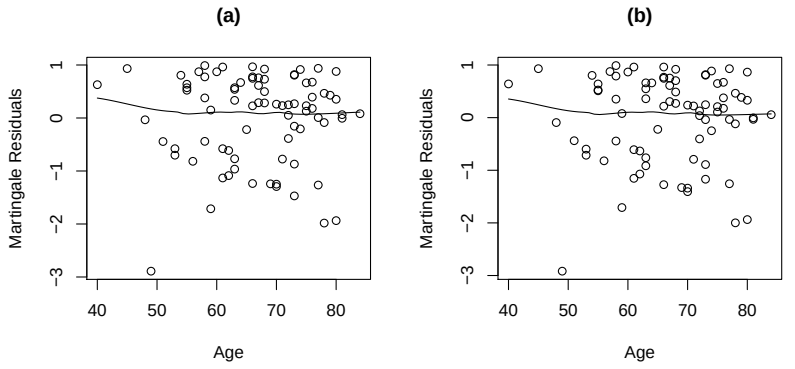


Figure 5.15: Martingale residuals based on smooth CNA (a) and DNAcopy CNA (b) against age with a smoothed curve using the Lowess method. This plot confirms that the age enters the proportional hazards model linearly.

5.9 Discussion

We have investigated how an extension of the Cox PH model for survival data is suitable to cope with high dimensional copy number alteration profiles, in addition to the clinical variables as (fixed) predictors. A key parameter in the model is θ which controls the amount of information in CNA profiles used in the model fitting, in terms of degrees of freedom of fit. We have also discussed computational methods to speed up the estimation of θ.

It is important to note that we have recoded the data so that the data used are absolute difference from copy number ratio one. We effectively assume that, given a genomic window, the hazard or risk of a patient with a total loss is the same as those who have copy number ratio two. This modelling assumption is reasonable since the Cox PH model assumes a linear predictor in the risk score R_i. A more relaxed assumption can be devised by assuming a smooth function. However, this is beyond the scope of the current research.

Our approach in this study is a genome-wide approach, in the sense that we take into account the all the CNA information in the genome. In this regard, the model is not embedded with a variable selection mechanism wich will be discussed later in Chapter 7. This is an interesting research challenge in its own right since the challenge in the context of CNA profile is much more complex: the correlation structure between genomic windows wich will be discussed later in Chapter 6.

It is worthwhile to notice that our proposed method in this chapter and in the next two chapters rely on Laplace approximation. It is well known in generalized linear mixed model (GLMM) that this method is called penalized quasi-likelihood (PQL) which has been introduced by Breslow & Clayton (1993). Pan & Huang (2014) argue that PQL method is the most commonly used method due to a convenient computation. However, this method may yield biased estimates of variance components, particularly for modeling correlated binary data and the true variance component is large Chen *et al.* (2015b).

Breslow & Lin (1995) shows that PQL estimators of regression coefficient and variance component were subject to bias when applied to correlated binary data and Poisson. Their numerical studies suggest that the biases are minimal for $0 \leq \theta \leq 0.25$

In our thesis, first we are not working with binary data nor a Poisson model. Also, it is clear from all results of simulation studies and real data analysis, that the variance components are very small (it is around 1×10^{-5}). Therefore, the results in this thesis are not suffering from the bias of PQL. Moreover, Our focus on this thesis is to incorporate the CNA (random effects) into the survival prediction and we used the most common method (PQL) to calculate the high dimensional integration in the marginal likelihood. In other words, our goal is not to compare the relative performance of different competitive estimation approach in estimating the parameter of a model. These different approaches are discussed in next two paragraphs.

In future research, one can use hierarchical generalized linear models (HL) introduced by Lee & Nelder (2001). However, Lee *et al.* (2006) claim that HL and PQL give the same estimator for the fixed effects β and random effects b but the variance component still suffer from bias especially in the cases where the response is Poisson or Binomial. The main difference between the two approaches is that the PQL approach estimate β and b by maximizing a penalized quasi-likelihood function, where the HL approach maximize the hierarchical likelihood function. To solve the weakness of PQL and HL, Sutradhar (2004) has proposed a generalized quasi-likelihood (GQL) approach that produces consistent as well as more efficient estimates as shown in Chowdhury & Sutradhar (2009). However, the GQL does not require any estimates for the random effects b and we actually are interested in the estimated of random effects b itself to know what is the effect of CNA in the hazard.

Another popular statistical method to incorporate the CNA in survival prediction is the numerical technique which is not in the scope of our thesis. This numerical techniques include Bayesian approach with sampling by Karim & Zeger (1992), MCEM algorithm by Booth & Hobert (1999), Gauss-Hermite quadrature (GHO) by Pan & Thompson (2003), and Quasi-Monte Carlo (QMC) by Pan & Thompson (2007). One common drawback in above literature is the time consuming. Newcombe *et al.* (2014) shows that the analysis of dataset with 20000 covariates were run about 28 hours, while it took less than 4 minuets with PQL.

Finally, our computational method and R package in this study can also be used for CNA profiles from array technology, provided that the (genome-wide) CNA profiles across individuals can be put in a matrix form. This means that CNA estimates across

individuals can be made into the same column in the data matrix, for each genomic region.

In summary, we investigated an extension of the standard Cox proportional hazard model to take into account cancer patients genome-wide copy number alteration (CNA) profiles. The genome-wide CNA profiles are considered as random predictors in the model in addition to the clinical variables as fixed predictors. The model enabled us to assess the significance of the fixed predictors, and to examine the genomic regions associated with the patients survival. Post-hoc analysis indicates that the model is suitable for the data and has a good fit. The model also enables us to estimate individual patient's survivor function and distinguish the survivor functions for different groups of patients at different risk.

Chapter 6

Extending Cox PH model : Taking dependences of CNA into account

6.1 Introduction

In Chapter 5, we considered an extension of the standard Cox proportional hazard model to take into account cancer patients' CNA profiles, in which CNAs are considered to be random predictors in the model, and the clinical variables as fixed predictors. Specifically, we assumed that the random effects b follow a normal distribution $b \sim N(0, D(\theta))$, with $D(\theta) = \theta I_q$ (I_q is the identity matrix of size q). The diagonal structure of $D(\theta)$ is the simplest possible structure for a variance-covariance matrix. This structure indicates independence between neighbouring genomic windows as a working assumption. However, CNAs generally have dependencies between neighbouring genomic windows, and have spatial characteristics which would have been ignored if we had used the method described in Chapter 5, or other methods described in Chapter 1 such as feature selection (e.g. Benjamini & Hochberg (1995)) or derived variables (e.g. Bair *et al.* (2006), Lee *et al.* (2013)). Huang *et al.* (2009) conclude that these methods can be adapted for survival analysis to model gene expression data; however, they are not a good match for CNA data as they ignore its spatial dependence structure.

In this chapter, we specifically address the spatial dependence structure of CNAs. In order to achieve this, we start in Section 6.2 by discussing other structures of

variance-covariance matrices of random effects. In Section 6.3, methods of imposing smoothness using first and second differences are presented. Section 6.4 discusses the mixture of normal and Cauchy distributions for first or second differences of random effects. Section 6.5 shows how to estimate the parameters on the model (fixed effects, random effects, and tuning parameters). Simulation studies are described and discussed in Section 6.6. Finally, the results of our lung cancer dataset are presented in Section 6.7.

6.2 Structures of variance-covariance matrices of random effects

In Chapter 5, the structure of the variance-covariance matrix was assumed to have a diagonal structure with equal variances and covariances of zero $D(\theta) = \theta I_q$. Thus,

$$D(\theta) = \theta \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}, \tag{6.1}$$

where $\theta = \sigma^2$. This structure is the simplest possible structure for a variance-covariance matrix which indicates independence between neighboring genomic windows. Therefore, in the next two subsections (6.2.1 and 6.2.2), we will discuss other, more complex structures for variance-covariance matrices to take into account the dependences between neighboring genomic windows.

6.2.1 Compound symmetry covariance matrix

The compound symmetry structure has the form

$$\Sigma(\theta) = \theta \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix} \tag{6.2}$$

In our study, we used only the first neighbouring structure, where we considered the first off-diagonal to have a correlation ρ and the remaining off-diagonal values to have 0 correlation. In other words, the structure of the variance-covariance matrix we consider is

$$\Sigma(\theta) = \theta \begin{bmatrix} 1 & \rho & \dots & 0 \\ \rho & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \rho \\ 0 & \dots & \rho & 1 \end{bmatrix} \tag{6.3}$$

This structure indicates dependence between the first neighbouring genomic windows; in other words, window 1 is correlated with window 2, and window 2 is correlated with window 3 and so on.

6.2.2 Inverse of covariance matrix

To allow for more neighbouring correlation we assume that the inverse of the variance-covariance matrix has the form

$$\Sigma^{-1}(\theta) = \frac{1}{\theta} \begin{bmatrix} 1 & \varrho & \dots & 0 \\ \varrho & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \varrho \\ 0 & \dots & \varrho & 1 \end{bmatrix} \tag{6.4}$$

Another way to deal with dependencies between neighbouring genomic windows and spatial characteristics of CNAs is by imposing smoothness in the assumption of random effects as will be describe in next sections.

6.3 Imposed smoothing

The assumption of the random effects b depends on the nature of the CNA. If we want to impose smoothness, we can do so by assuming that the first differences, which are

$$\Delta b_j = b_j - b_{j-1},$$

or that the second differences, which are

$$\Delta^2 b_j = b_j - 2b_{j-1} + b_{j-2},$$

to have a specific distribution (see Pawitan, 2013). We could use a normal distribution, but Huang *et al.* (2009) suggest that as we were dealing with CNAs, we need to allow for sudden jumps or large spatial changes. To have this flexibility, Huang *et al.* (2009) chose a heavy-tailed distribution, the Cauchy distribution, instead of a normal distribution, because normal distributions tend to convert jumps into gradual changes.

Therefore, in Section 6.3.1, we will discuss the implementation of a Cauchy distribution as a random effect in the Cox PH model. Then, in Sections 6.3.2 and 6.3.3, using a Cauchy distribution for the first- and second-order differences of b will be discussed.

6.3.1 Cauchy distribution

We extend the Cox PH model (by including CNA profiles as random predictors) to:

$$h_i(t|X, Z) = h_0(t) \exp\{X_i\beta + Z_i b\}. \tag{6.5}$$

We assume b to follow a multivariate Cauchy distribution with location 0 and scale $\Sigma(\theta)$; In other words, $b \sim Cauchy(0, \Sigma(\theta))$.

We considered the same approach in Chapter 6, which linked the likelihood approximation of (Breslow & Clayton (1993)) with the penalised likelihood concept to derive a generalisation of the model's estimation. The marginal (integrated) likelihood $L(h_0(t), \beta, \theta)$ for model (6.5) is

$$\begin{aligned} L(h_0(t), \beta, \theta) &= \int \prod_{i=1}^{n} h_i(t|b)^{\delta_i} S_i(t|b) p(b; D(\theta)) db \\ &= \int \prod_{i=1}^{n} [h_0(t) \exp(X_i\beta + Z_i b)]^{\delta_i} \exp[-H_0(t) \exp(X_i\beta + Z_i b)] p(b; \Sigma(\theta)) db, \end{aligned} \tag{6.6}$$

where the random effects b are integrated out. As b is restricted to follow a multivariate Cauchy distribution with a location 0 and a scale matrix $\Sigma(\theta)$, the probability density function (PDF) of this multivariate Cauchy distribution is

$$p(b; \Sigma(\theta))db = \Gamma(\frac{1+q}{2})\Gamma(\frac{1}{2})^{-1}\pi^{-\frac{q}{2}}|\Sigma(\theta)|^{-\frac{1}{2}}\left[1 + b'\Sigma(\theta)^{-1}b\right]^{-\frac{q+1}{2}}.$$

Unfortunately, the above integral equation, Eq. (6.6) is difficult to compute. Therefore, we follow Breslow & Clayton (1993) in their approach for the GLMM and use a Laplace approximation for the integral in (6.6).

First, instead of using the marginal likelihood $L(h_0(t), \beta, \theta)$ (Eq. (6.6)), we use the log of the marginal likelihood $\log(L(h_0(t), \beta, \theta)) = \ell(h_0(t), \beta, \theta)$, which is

$$\begin{aligned}\ell(h_0(t), \beta, \theta) = \int \sum_{i=1}^{n} \Big[\delta_i \big[\log(h_0(t)) + X_i\beta + Z_i b\big] - H_0(t)\exp(X_i\beta + Z_i b)\Big] \\ + \log\big(\Gamma(\frac{1+q}{2})\big) - \log\big(\Gamma(\frac{1}{2})\big) \\ - \frac{q}{2}\log(\pi) - \log\big(|\Sigma(\theta)|^{-\frac{1}{2}}\big) - \frac{q+1}{2}\log\big(1 + b'\Sigma(\theta)b\big)db.\end{aligned} \tag{6.7}$$

Then, we write Eq. (6.7) in the form:

$$e^{\ell(h_0(t),\beta,\theta)} \propto C|\Sigma(\theta)|^{-\frac{1}{2}} \int e^{-k(b)},$$

where C are constant terms that are not related to the parameters. Let k' and k'' denote the q vector and $q \times q$ dimensional matrix of the first- and second-order partial derivatives of k with respect to b. Ignoring the multiplicative constant C, the approximation returns

$$\ell(h_0(t), \beta, \theta) \approx -\frac{1}{2}\log|\Sigma(\theta)| - \frac{1}{2}\log|k''(\tilde{b})| - k(\tilde{b}),$$

where

$$k(\tilde{b}) = -\left[\left[\sum_{i=1}^{n}[\delta_i\big\{\log(h_0(t)) + X_i\beta + Z_i\tilde{b}\big\} - H_0(t)\exp(X_i\beta + Z_i\tilde{b})\right] - \frac{q+1}{2}\log\big(1 + \tilde{b}'\Sigma(\theta)^{-1}\tilde{b}\big)\right],$$

and $\tilde{b} = \tilde{b}(\beta, \theta)$ denotes the solution to the partial derivatives of $k(b)$ with respect to b. i.e $\tilde{b}$ satisfies

$$k'(\tilde{b}) = \frac{\partial k(\tilde{b})}{\partial \tilde{b}}.$$

The set of second partial derivative of $k(b)$ with respect to b denoted $k''(b)$ has the form

$$k''(\tilde{b}) = \frac{\partial^2 k(\tilde{b})}{\partial \tilde{b}\partial \tilde{b}'}$$

Therefore, the approximate marginal log likelihood using the Laplace approximation leads to:

$$\begin{aligned}\ell(h_0(t),\beta,\theta) \approx & -\frac{1}{2}\log|D(\theta)| \\ & -\frac{1}{2}\log\left|k''(\tilde{b})\right| \\ & +\sum_{i=1}^{n}\Big[\delta_i[\log(h_0(t))+X_i\beta+Z_i\tilde{b}] \\ & -H_0(t)\exp(X_i\beta+Z_i\tilde{b})\Big]-\frac{q+1}{2}\log\left(1+\tilde{b}'\Sigma(\theta)^{-1}\tilde{b}\right).\end{aligned} \tag{6.8}$$

if θ were known and b were considered a fixed-effects parameter, then the first two terms are ignored and β can be chosen to maximise the second two terms, which gives us a penalised log likelihood. Thus $(\hat{\beta},\hat{b}) = (\hat{\beta}(\theta),\hat{b}(\theta))$ where $\hat{b}(\theta) = \tilde{b}(\hat{\beta}(\theta))$, jointly maximise

$$\sum_{i=1}^{n}\Big[\delta_i[\log(h_0(t))+X_i\beta+Z_i\tilde{b}]-H_0(t)\exp(X_i\beta+Z_i\tilde{b})\Big]-\frac{q+1}{2}\log\left(1+\tilde{b}'\Sigma(\theta)^{-1}\tilde{b}\right). \tag{6.9}$$

Equation (6.9) is the full likelihood for a Cox model with b as another set of parameters and penalty terms. It turns out that the full likelihood can be maximised using penalised fixed-effect partial likelihood, as Cox showed in Cox *et al.* (1972):

$$l_p(\beta,\theta,b) = \sum_{i=1}^{n}\left[\delta_i(X_i\beta+Z_ib)-\delta_i\log\left(\sum_{j\in R(t_i)}\exp(X_j\beta+Z_jb)\right)\right]-\frac{q+1}{2}\log\left(1+b'\Sigma(\theta)^{-1}b\right), \tag{6.10}$$

where $\frac{q+1}{2}\log\left(1+b'\Sigma(\theta)^{-1}b\right)$ is the penalty term penalising extreme values of b.

6.3.2 Using a Cauchy distribution for the first differences of b

Here, we assumed first-order differences of b, i.e

$$\Delta b \equiv \begin{pmatrix} b_2-b_1 \\ b_3-b_2 \\ \vdots \\ b_q-b_{q-1}\end{pmatrix}$$

to follow a Cauchy distribution with location 0 and a scale θI_{q-1}, where $\theta = \sigma^2$. Assuming the first differences, Δb, have a Cauchy distribution with a location 0 and scale θI_{q-1} is equivalent to assuming b is Cauchy distribution with location 0 and an inverse scale matrix

$$\begin{aligned}\Sigma(\theta)^{-1} \equiv & \theta^{-1} R_1^{-1}, \\ \text{where } R_1^{-1} \equiv & \Delta'\Delta \\ = & \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix}\end{aligned} \tag{6.11}$$

((see Pawitan, 2013)).

6.3.3 Using a Cauchy distribution for the second differences of b

Here, we assumed second-order differences of b,

$$\Delta^2 b \equiv \begin{pmatrix} b_3 - 2b_2 + b_1 \\ b_4 - 2b_3 + b_2 \\ \vdots \\ b_q - 2b_{q-1} + b_{q-2} \end{pmatrix},$$

to follow a Cauchy distribution with location 0 and scale θI_{q-2}, where $\theta = \sigma^2$. Assuming the second differences, $\Delta^2 b$, have a Cauchy distribution with location 0 and a scale of θI_{q-2} is equivalent to assuming b is a Cauchy distribution with location 0 and an inverse scale matrix

$$\begin{aligned}\Sigma(\theta)^{-1} \equiv & \theta^{-1} R_2^{-1}, \\ \text{where } R_2^{-1} \equiv & (\Delta^2)'\Delta^2 \\ = & \begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & \cdots & 0 \\ -2 & 5 & -4 & 1 & 0 & \cdots & 0 \\ 1 & -4 & 6 & -4 & 1 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ 0 & \cdots & 1 & -4 & 6 & -4 & 1 \\ 0 & \cdots & 0 & 1 & -4 & 5 & -2 \\ 0 & \cdots & \cdots & 0 & 1 & -2 & 1 \end{pmatrix}\end{aligned} \tag{6.12}$$

((see Pawitan, 2013)).

6.4 Mixture of the products of normal and Cauchy distributions for first or second differences of b

In this section we introduce a novel algorithm based on a smooth extended Cox model (SCox) within a random effects model framework using penalised partial likelihood to model survival time using patients clinical characteristics as fixed effects and their CNA profiles as random effects. We assumed CNA coefficients b to be correlated random effects that follow a mixture of two distributions: normal as in Chapter 5 (to achieve shrinkage around the mean values), and Cauchy for the first- or second-order differences of b as in Section 6.3.2 and 6.3.3 (to gain smoothness). A similar idea was first discussed in Tibshirani *et al.* (2005), and later in Huang *et al.* (2009) in smoothed logistic regression .

to derive the penalised partial likelihood of the mixture model, we considered the same approach in Chapter 5 and Section 6.3.1, which linked the likelihood approximation of (Breslow & Clayton (1993)) with the penalised likelihood concept to derive a generalisation of the model estimation. The marginal (integrated) likelihood $L(h_0(t), \beta, \theta, w)$ is

$$\begin{aligned} L(h_0(t), \beta, \theta, w) &= \int \prod_{i=1}^{n} h_i(t|b)^{\delta_i} S_i(t|b) p(b; D(\theta), \Sigma(\theta)) db \\ &= \int \prod_{i=1}^{n} [h_0(t) \exp(X_i\beta + Z_i b)]^{\delta_i} \times \exp[-H_0(t) \exp(X_i\beta + Z_i b)] \\ &\times p(b; D(\theta), \Sigma(\theta)) db, \end{aligned} \tag{6.13}$$

where the unobserved (b) are integrated out.

Now b is restricted to follow a mixture of the product of a multivariate normal distribution (with a mean 0 and a variance-covariance matrix $D(\theta)$) and a Cauchy distribution (with a location 0 and a scale matrix of $\Sigma(\theta)$). The PDF of this product of

multivariate normal and Cauchy distributions is

$$\begin{aligned} PDF &= f_m(b; D(\theta), \Sigma(\theta)) \\ &= \frac{f_n^w(b; D(\theta)) \times f_c^{(1-w)}(b; \Sigma(\theta))}{Cons} \\ &= \frac{\left[|D(\theta)|^{-\frac{1}{2}} e^{-\frac{1}{2} b' D(\theta)^{-1} b}\right]^w \times \left[|\Sigma(\theta)|^{-\frac{1}{2}} \left[1 + b'\Sigma(\theta)^{-1} b\right]^{-\frac{q+1}{2}}\right]^{(1-w)}}{Cons}, \end{aligned} \tag{6.14}$$

where $Cons$ is the normalising constant which does not contain any parameters.

As before, the above integral equation, Eq. (6.13), is difficult to evaluate. Therefore, we follow the same steps explained in the previous section to find the Laplace approximation for the integral in (6.13).

First, The log of the marginal likelihood (Eq. (6.13)), $\ell(h_0(t), \beta, \theta, w)$ is

$$\begin{aligned} \ell(h_0(t), \beta, \theta, w) = \int \sum_{i=1}^{n} &\left[\delta_i \left[\log(h_0(t)) + X_i\beta + Z_i b\right] - H_0(t) \exp(X_i\beta + Z_i b)\right] \\ &+ w \log\left[|D(\theta)|^{-\frac{1}{2}} e^{-\frac{1}{2} b' D(\theta)^{-1} b}\right] \\ &+ (1-w) \log\left[|\Sigma(\theta)|^{-\frac{1}{2}} \left[1 + b'\Sigma(\theta)^{-1} b\right]^{-\frac{q+1}{2}}\right] - \log(Cons) db. \end{aligned} \tag{6.15}$$

Then, we write Eq. (6.15) in the form:

$$e^{\ell(h_0(t), \beta, \theta, w)} \propto C |D(\theta)|^{-\frac{1}{2}} |\Sigma(\theta)|^{-\frac{1}{2}} \int e^{-k(b)},$$

where C is a constant terms that are not related to the parameters. Let k' and k'' denote the q vector and $q \times q$ dimensional matrix of first- and second-order partial derivatives of k with respect to b. Ignoring the multiplicative constant C, the approximation returns

$$\ell(h_0(t), \beta, \theta, w) \approx -\frac{1}{2} \log(|D(\theta)||\Sigma(\theta)|) - \frac{1}{2} \log|k''(\tilde{b})| - k(\tilde{b}),$$

where

$$\begin{aligned} k(\tilde{b}) = -\Bigg\{ &\left[\sum_{i=1}^{n} [\delta_i \left[\log(h_0(t)) + X_i\beta + Z_i\tilde{b}\right] - H_0(t) \exp(X_i\beta + Z_i\tilde{b})\right] \\ &- \left(\frac{1}{2} \tilde{b}' D^{-1}(\theta) \tilde{b} + \frac{q+1}{2} \log\left(1 + \tilde{b}'\Sigma(\theta)^{-1}\tilde{b}\right)\right)\Bigg\}, \end{aligned}$$

and $\tilde{b} = \tilde{b}(\beta, \theta)$ denotes the solution to the partial derivatives of $k(b)$ with respect to b. i.e $\tilde{b}$ satisfies

$$k'(\tilde{b}) = \frac{\partial k(\tilde{b})}{\partial \tilde{b}}.$$

The set of second partial derivative of $k(b)$ with respect to b denoted $k''(b)$ has the form

$$k''(\tilde{b}) = \frac{\partial^2 k(\tilde{b})}{\partial \tilde{b} \partial \tilde{b}'}$$

Therefore, the approximate marginal log likelihood using the Laplace approximation leads to:

$$\begin{aligned}\ell(h_0(t), \beta, \theta, w) \approx & -\frac{1}{2}\log(|D(\theta)||\Sigma(\theta)| \\ & - \frac{1}{2}\log\left|k''(\tilde{b})\right| \\ & + \sum_{i=1}^{n}\left[\delta_i[\log(h_0(t)) + X_i\beta + Z_i\tilde{b}] - H_0(t)\exp(X_i\beta + Z_i\tilde{b})\right] \\ & - \left[w\big[\frac{1}{2}\tilde{b}'D(\theta)^{-1}\tilde{b}\big] + (1-w)\big[\frac{q+1}{2}\log\left(1 + \tilde{b}'\Sigma(\theta)^{-1}\tilde{b}\right)\big]\right].\end{aligned} \tag{6.16}$$

If both θ, w are known and b is considered a fixed-effects parameter, the first two terms are ignored and β is chosen to maximise the second two terms; this is a penalised log likelihood. Thus $(\hat{\beta}, \hat{b}) = (\hat{\beta}(\theta, w), \hat{b}(\theta, w))$, where $\hat{b}(\theta, w) = \tilde{b}(\hat{\beta}(\theta, w))$, jointly maximise

$$\begin{aligned}& \sum_{i=1}^{n}\left[\delta_i[\log(h_0(t)) + X_i\beta + Z_i\tilde{b}] - H_0(t)\exp(X_i\beta + Z_i\tilde{b})\right] \\ & - \left[w\big[\frac{1}{2}\tilde{b}'D(\theta)^{-1}\tilde{b}\big] + (1-w)\big[\frac{q+1}{2}\log\left(1 + \tilde{b}'\Sigma(\theta)^{-1}\tilde{b}\right)\big]\right].\end{aligned} \tag{6.17}$$

Equation (6.17) is the full likelihood for a Cox model with b as another set of parameters and penalty terms. It turns out that it can be maximised using penalised fixed effect partial likelihood, as Cox showed in Cox *et al.* (1972):

$$\begin{aligned}\ell_P(\beta, \theta, w, b) = & \sum_{i=1}^{n}\left[\delta_i(X_i\beta + Z_i b) - \delta_i\log\Big(\sum_{j \in R(t_i)}\exp^{(X_j\beta + Z_j b)}\Big)\right] \\ & - \left[w\big[\frac{1}{2}b'D(\theta)^{-1}b\big] + (1-w)\big[\frac{q+1}{2}\log\left(1 + b'\Sigma(\theta)^{-1}b\right)\big]\right],\end{aligned} \tag{6.18}$$

where $\left[w\left[\frac{1}{2}b'D(\theta)^{-1}b\right] + (1-w)\left[\frac{q+1}{2}\log\left(1+b'\Sigma(\theta)^{-1}b\right)\right]\right]$ is the penalty term penalising extreme value of b.

6.5 Parameter estimation

6.5.1 Estimation of β and b

1. To derive an estimation of β and b at fixed values for (θ, ρ) or (θ, ϱ) in Sections 6.2.1 and 6.2.2, we used the same partial differentiate of the log partial likelihood $l_P(\beta, b)$ as given by Eq. (5.6) from Chapter 5.

2. For the mixture of the products of a normal and a Cauchy distribution for first or second differences of b (see Section 6.4), we can derive an estimation of β and b at fixed θ, w by first partially differentiating the penalized partial log-likelihood $\ell_P(\beta, \theta, w, b)$ from Eq. (6.18) with respect to β and b. The resulting estimation equations for β and b, respectively, are

$$u(\beta) = \sum_{i=1}^{n} \delta_i \left[X_i - \frac{\sum_{j\in R(t_i)} X_j \exp^{(X_j\beta + Z_j b)}}{\sum_{j\in R(t_i)} \exp^{(X_j\beta+Z_j b)}} \right] \tag{6.19}$$

and

$$\begin{aligned} u(b) = &\sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j\in R(t_i)} Z_j \exp^{(X_j\beta + Z_j b)}}{\sum_{j\in R(t_i)} \exp^{(X_j\beta+Z_j b)}} \right] - \Big(w_n (D(\theta)^{-1} b) \\ &+ w_c \Big(\frac{(q+1)\Sigma(\theta)^{-1} b}{1 + b'\Sigma(\theta)^{-1} b} \Big) \Big). \end{aligned} \tag{6.20}$$

Estimates for $(\hat{\beta}(\theta, w), \hat{b}(\theta, w))$ can be found by alternating between solving (6.19) and (6.20) at a fixed value for (θ, w) using the Newton-Raphson algorithm. The inverse of the minus second partial derivative matrix H^{-1} (Hessian) can be used as an approximate covariance matrix. However, using this covariance matrix give a wide confidence interval as we can see in the simulation result. Gray (1992) suggested the estimate of the covariance to be $H^{-1} I_{PL} H^{-1}$, where I_{PL} is the standard Cox PH model information matrix, or the second derivative matrix of partial log likelihood with respect to β and b.

We calculate the confidence interval of the random effect as

$$\text{CI} = \hat{b} \pm 1.96 \times \sqrt{\text{var}(\hat{b})}.$$

Then, we compared the CI by using the previous two formulas of var$(\hat{b})$ with the CI by using bootstrap to see which formula was more appropriate. This comparison suggests the use of Gray's formula as we can see later in the simulation result.

6.5.2 Estimation of tuning parameters $K = (\theta, \rho, \varrho, w)$

To estimate the tuning parameters $K = (\theta, \rho)$, $K = (\theta, \varrho)$, or $K = (\theta, w)$, we used a cross-validation (CV) criterion, which is based on the unpenalised (standard) log partial likelihood ($\ell(\beta, K, b)$), as proposed by Verweij & Van Houwelingen (1993). We could have used different values of θ for the normal and Cauchy distributions, which would have given us one more tuning parameter. However, for simplicity we set θ to be the same for both distributions.

We used fivefold CV. For each K,we computed the cross-validated partial likelihood as

$$CV(K) = \sum_{m=1}^{5} l^{[s]}\big(\hat{\beta}_K^{[-s]}, \hat{b}_K^{[-s]}\big), \tag{6.21}$$

where $l^{[s]}\big(\hat{\beta}_K^{[-s]}, \hat{b}_K^{[-s]}\big)$ is the the unpenalised (standard) log partial likelihood for the sth validation set evaluated at $(\hat{\beta}_K^{[-s]}, \hat{b}_K^{[-s]})$ (the coefficient estimates from the sth training sets). We selected a value for K that maximised $CV(K)$ over a fine grid of values.

We can compute standard errors for the CV curve at each tuning parameter value for K. We computed the sample standard deviation of $CV_1(K), \ldots, CV_5(K)$, and finally we used

$$SE(K) = SD(K)/\sqrt{5} \tag{6.22}$$

for the standard error of $CV(K)$. We chose a value for K where the error is within one standard error. In other words, we take the model whose error is within one standard error of the minimal error.

6.6 Simulation results

We generated CNA , Z, with a dimension of 85×100 from a multivariate normal distribution $Z \sim N(0, \tau)$, where τ is called the *variance-covariance matrix* of the data (the CNAs) as follows:

$$\tau = 1 \times \begin{bmatrix} 1 & .5 & 0 & \dots & 0 & 0 \\ .5 & 1 & .5 & & & 0 \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & & & \ddots & \ddots & .5 \\ 0 & 0 & & 0 & .5 & 1 \end{bmatrix}. \tag{6.23}$$

Then, to make the CNA (Z) have a signal, we did the following:

1. We separated the patients in our study based on their survival time. The **first** group was comprised of the patients whose survival time exceeded the median of the group we studied, while the **second** group were the patients whose survival time was less than the median.

2. In the first 10 $(1 - 10)$ columns of Z,we added 6 to the patients' CNAs in the **first** group, and we added 2 to the patients' CNAs in the **second** one. Therefore, the first group gains more CNA than the second group in the first 10 columns.

3. In the second 10 columns $(11 - 20)$, we inverted those processes.

4. For the rest of the windows $(21 - 100)$, we added 4 for all patients.Therefore, booth groups have the same expected CNA.

6.6.1 Simulation study: compound symmetry covariance matrix (first neighboring structure)

In Figure 6.1, we can see the effects of choosing different values for ρ when $\theta = 0.001$ on the estimation of random effects b. From the left panel to the right panel, $\rho = 0, 0.5,$ and 0.9, respectively. When $\rho = 0$, the structure of the variance-covariance matrix is the diagonal structure explained in the previous chapter (Chapter 5). For all

setting, the first 20 values reveal a signal as we anticipated. We can see the effect of including ρ in the estimation of b for neighborhood values.

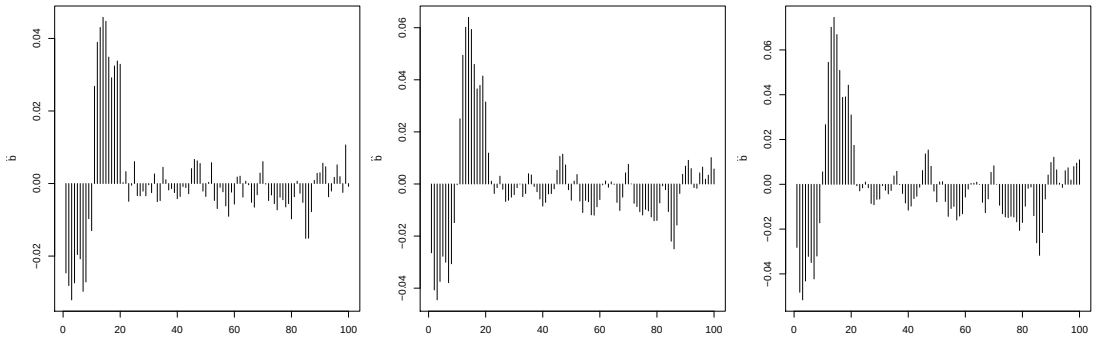


Figure 6.1: Estimation of random effects b when $\theta = 0.001$ ($\rho = (0, 0.5, \text{ and } 0.9)$ from left to right, respectively)

In order to estimate the optimal tuning parameters for $K = (\theta, \rho)$, we used the CV criterion as explained in Section 6.5.2. Figure 6.2 shows CV for different values of θ and ρ.

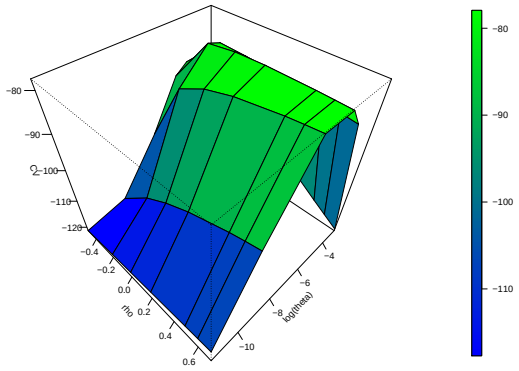


Figure 6.2: Five-fold CV (E.q (6.21)) for different values of θ and ρ

The estimation of the random effects b, based on the optimal tuning parameters $K = (\theta = 0.005, \rho = 0.2)$, can be seen in Figure 6.3.

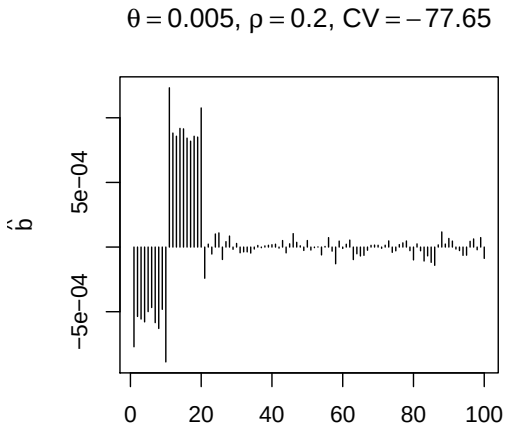


Figure 6.3: Estimation of the random effects b in compound symmetry covariance matrix (first neighboring structure) model based on optimal tuning parameters

6.6.2 Simulation study: inverse of the covariance matrix

Figure 6.4 shows the effects of choosing different values for ϱ when $\theta = 0.001$ on the estimation of random effects b. From left to right, $\varrho = 0, -0.3,$ and -0.4 respectively. When $\varrho = 0$, the structure of the covariance matrix is the diagonal structure explained in the previous chapter (Chapter5).

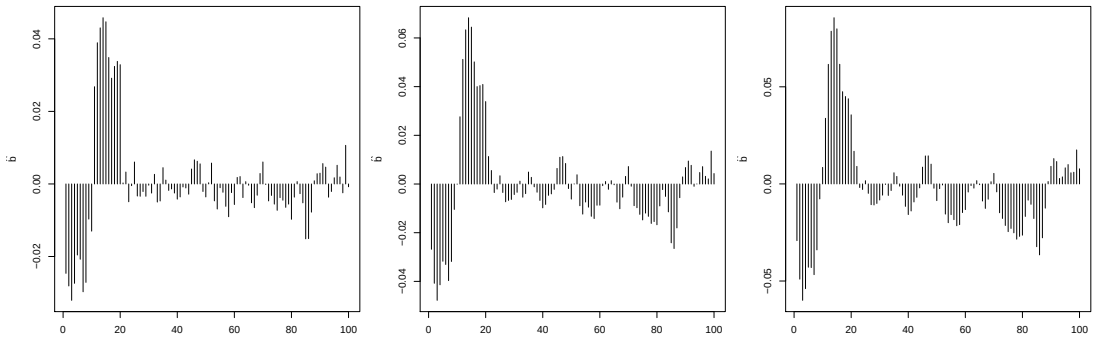


Figure 6.4: Estimation of random effects b when $\theta = 0.001$ ($\varrho = (0, -0.3, and - 0.4)$ from left to right, respectively)

In order to estimate the optimal tuning parameters for $K = (\theta, \varrho)$, we again used the CV criterion as explained in Section 6.5.2. Figure 6.8 shows CV for different values of θ and ϱ.

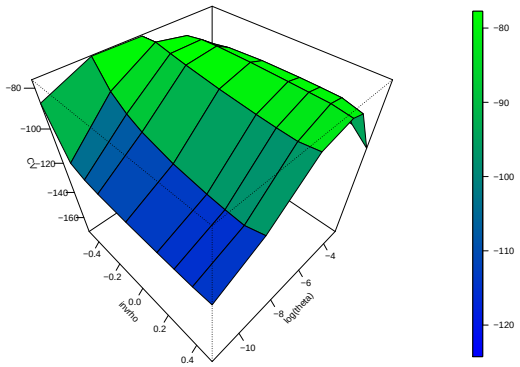


Figure 6.5: Five-fold CV (E.q (6.21)) for different values of θ and ϱ

Estimations of the random effects b, based on the optimal tuning parameters $K = (\theta = 0.001, \varrho = -0.4)$, are shown in Figure 6.9.

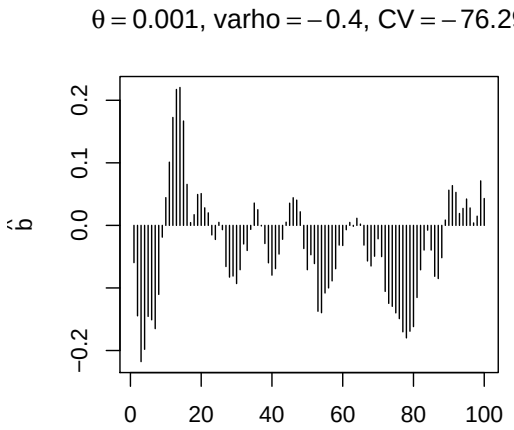


Figure 6.6: Estimation of the random effects b in Inverse of the covariance matrix model based on optimal tuning parameters

6.6.3 Simulation study: mixture of the products of normal and Cauchy distributions for first or second differences of b

Figure 6.7 shows the effects of choosing different weights of w when $\theta = 0.001$ on the estimation of random effects b can be seen. In this figure, from left to right, $w = 1, 0.5,$ and 0, respectively; the first differences are presented in the top panels, while the bottom panels show the second differences. When $w = 1$, the method was reduced to only a normal distribution, as explained in the previous chapter (Chapter5), and when $w = 0$ the method was reduced to a Cauchy distribution for first or second differences. Choosing a first or second differences of b to be Cauchy alone, as can be seen on right panels of Figure 6.7, smoothed out the random effects. In other words, the random effects estimates b are over smoothed.

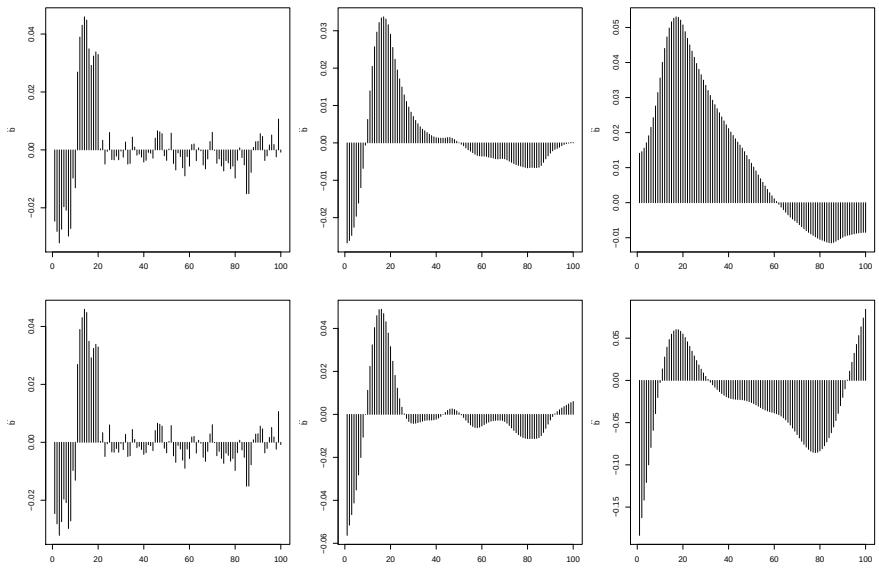


Figure 6.7: Top panels show estimations of random effects b with a Cauchy distribution for the first differences, and bottom panels show estimations for the second differences; in both rows $\theta = 0.001$, and from left to right $w = (0, 0.5, and1)$, respectively.

In order to estimate the optimal tuning parameters for $K = (\theta, w)$, we also used the CV criterion as explained in Section 6.5.2. Figure 6.8 shows CV for different values of θ and w.

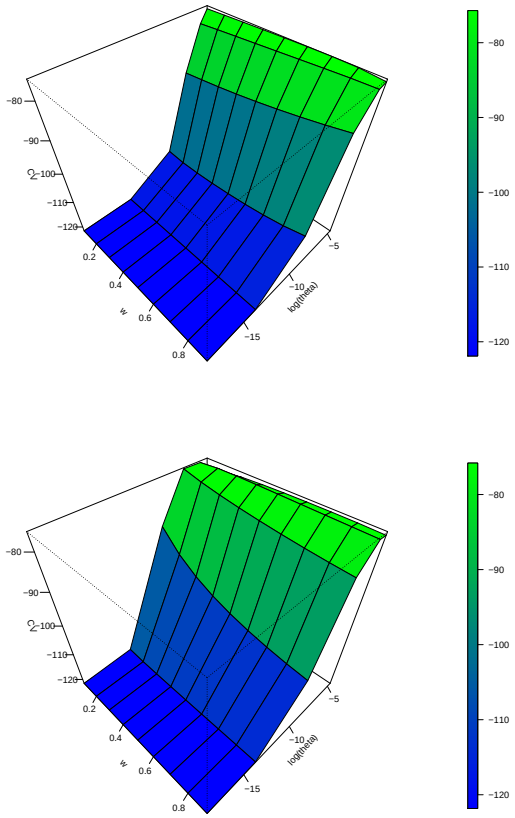


Figure 6.8: Five-fold CV (E.q (6.21)) for different values of θ and w for first and second differences (top and bottom, respectively)

The estimation of the random effects b, based on the optimal tuning parameters $K = (w = 0.7, \theta = 0.01)$ for the first differences and $K = (w = 0.3, \theta = 0.005)$ for the second differences, can be seen in Figure 6.9 (left and right panels, respectively).

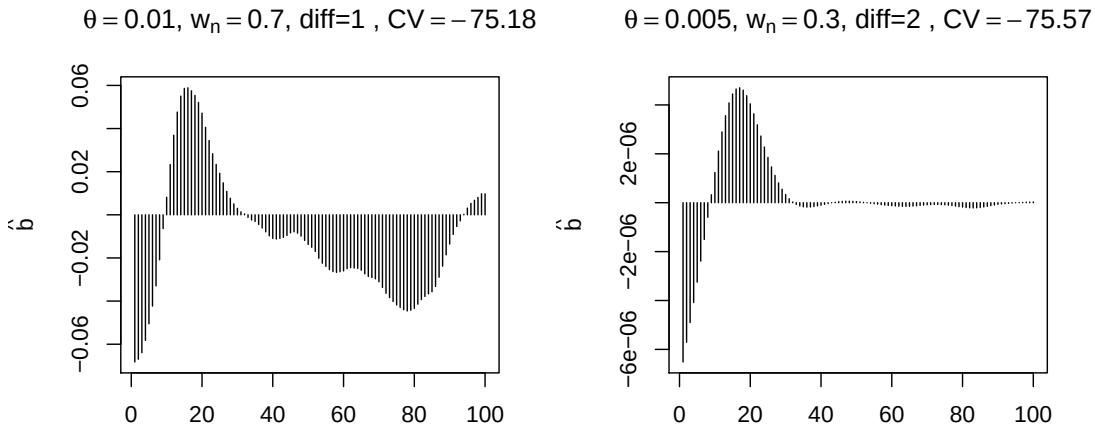


Figure 6.9: Estimation of the random effects b based on the optimal tuning parameters for the first and second differences (left and right, respectively)

Finally, we gathered all estimations of the random effects b in Figure 6.10. We can see that the best model based on CV (greater is better) is the mixture of normal and Cauchy distributions for the first differences (CV=-75.18). The second-best model is the the mixture of normal and Cauchy distributions for the second differences (CV=-75.57). The third-best model is the normal distribution with an inverse of covariance matrix (CV=-76.26). The fourth-best model is the normal distribution with a compound symmetry covariance matrix (first neighbouring structure) (CV=-77.65). Finally, the worst model is the normal distribution with a diagonal structure of covariance matrix (CV=-77.83).

As we have imposed a correlation between the first neighbouring windows in the simulation setting, we expect the model with a diagonal structure of covariance matrix (presented in Chapter 5) to be the worst, because the diagonal structure indicates independence. For CNAs, we believe that the mixture of normal and Cauchy distributions for second-order differences works better; in this simulation, however, the first-order differences work better than the second-order differences. This is because we only have two jumps in this simulation, and we imposed a correlation between the first neighbouring windows. This is not the case for the CNAs; there we have many jumps, and we have strong serial correlations.

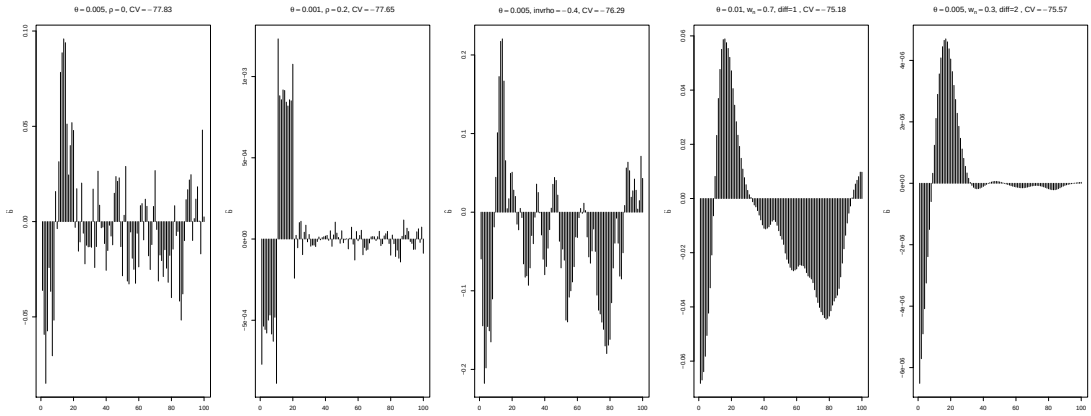


Figure 6.10: Estimation of the random effects b based on the optimal tuning parameters for all methods

6.6.4 Simulation study: confidence interval of the random effects

Figure 6.11 shows the estimation of the random effects b for one simulation along with the confidence interval CI. We use three different way to calculate the CI of b as explained in Section 6.5. First, we use the $\text{var}(\hat{b})$ to be H^{-1}. Second, we use the $\text{var}(\hat{b})$ to be $H^{-1}I_{PL}H^{-1}$. Third, we calculate the CI of b by using bootstrap. For bootstrap, we resample the data with replacement, and the size of the resample must be equal to the size of the original data set. We repeat this routine 1000 times.

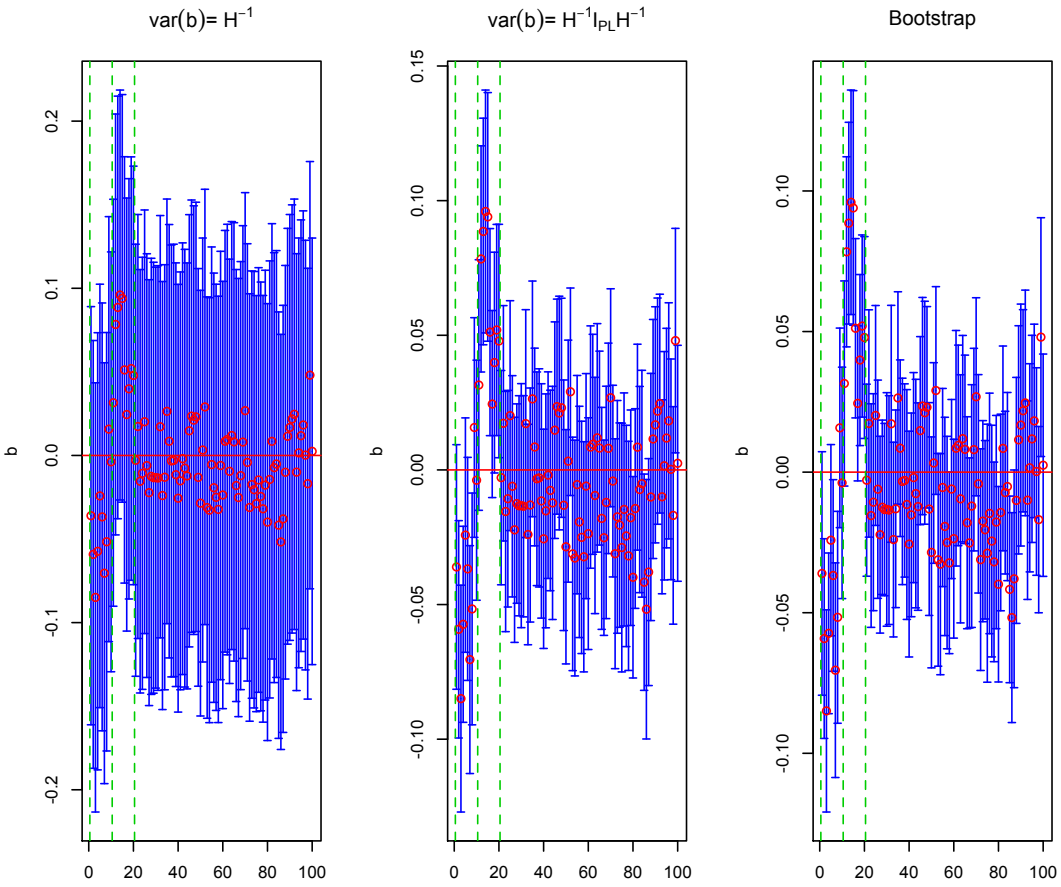


Figure 6.11: The estimation of random effects b along with CI. In the left panel, $\text{var}(\hat{b}) = H^{-1}$, middle panel $\text{var}(\hat{b}) = H^{-1}I_{PL}H^{-1}$, and right panel used bootstrap. The green dotted lines indicates windows which have a signal $(1:10, 11:20)$

It is clear from looking to Figure 6.11, that the CI by using Gray's formula, $\text{var}(\hat{b}) = H^{-1}I_{PL}H^{-1}$, is similar to the CI by using bootstrap.

6.7 Real data

In this section, we will only discuss the method where the random effects were assumed to be a mixture of the products of normal and Cauchy distributions for second differences of b (SCox). The results based on a compound symmetry covariance matrix

(Coxrho) and an inverse of covariance matrix (Coxinv) will be presented in Appendix B.

6.7.1 Model fit: Estimation of tuning parameters $K = (\theta, w)$

An important parameter to be estimated from the SCox PH model is $K = (\theta, w)$. These parameters are important in the interpretation; as θ goes to zero (in limit terms), the estimates of the random effects will be zero, and no information on CNAs is taken into account in the model. Also, w controls the smoothness of the model. We will be usually interested in models for more than one amount of regularisation. It is possible to solve a two-dimensional grid of θ and w; however, we found this to be computationally impractical, and to do a poor job of model selection (similar to Simon et al.'s (Simon *et al.* (2011)) finding). Instead, we fixed the mixing parameter w and computed solutions for a path of θ. We set the mixing parameter w to be 0.5 to give equal weight to the normal and Cauchy distributions.

Figure 6.12 shows the cross validation partial likelihood (CVPL) (θ) when $w = 0.5$. To estimate θ, we use the principle of CV with the one standard error rule (Eq. (6.22)). By applying the one standard error rule, the figure gives a value of $\log(\theta) = -10.105$, corresponding to $\theta = 4.09 \times 10^{-5}$.

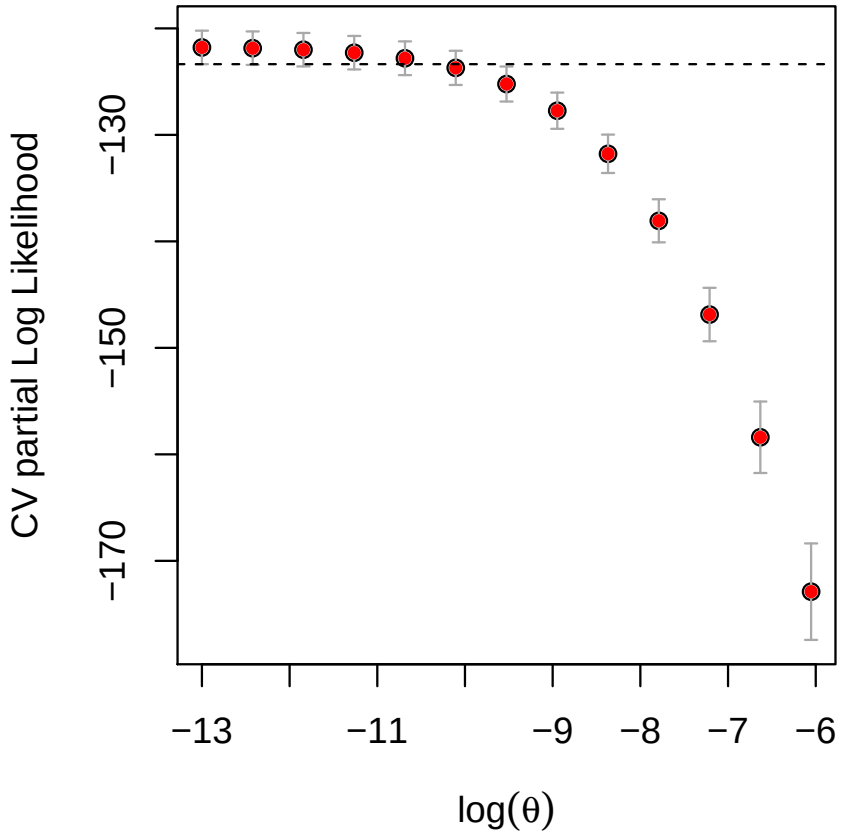


Figure 6.12: CVPL (θ); the horizontal dotted line indicates one standard error (Eq. (6.22)) of CVPL(θ)

6.7.2 Model fit: fixed effects

Using the optimal value for θ, the estimates of the fixed effects and their inferences can be seen in Table 6.1. For comparison, we estimate the fixed effects under both conditions with and without the CNA profiles in the model. The table indicates that the variables Age, Stage-T, and Stage-N are statistically significant (p-value < 0.05). The estimates indicate that, all else being equal, the hazard ratio increases by about six percent ($e^{0.055} \approx 1.06$) with each one-year increase in Age-at-operation. The positive estimates of Stage-T3 indicate that larger tumour size is associated with a significant increase in the hazard (relative to Stage-T1 as the baseline). Similarly, the estimates of Stage-N2 indicate that a wider spread of cancer cells to the lymph nodes increases the

hazard significantly (relative to Stage-N0 as the baseline).

Table 6.1: Summary of fixed predictors

Predictor	Estimate	Exp	Std.Error	$z$values	p-value
(Without CNA profiles)					
Age	0.0551	1.06	0.0164	3.37	0.0008
StageT2	0.1818	1.20	0.3215	0.57	0.5700
StageT3	1.7623	5.83	0.6392	2.76	0.0058
StageN1	0.3616	1.44	0.3019	1.20	0.2300
StageN2	1.3653	3.92	0.4824	2.83	0.0047
(With CNA profiles)					
Age	0.0571	1.06	0.0164	3.48	0.0005
StageT2	0.1858	1.20	0.3215	0.58	0.5633
StageT3	1.9010	6.69	0.6392	2.97	0.0029
StageN1	0.3451	1.41	0.3019	1.14	0.2529
StageN2	1.3245	3.76	0.4824	2.74	0.0060

6.7.3 Model fit: random effects

The random effects estimates b of the full SCox PH model, using CNA profile from smooth segmentation, are presented in Figure 6.13. The magnitude of the estimates is relatively small (compared to the fixed effects estimates, for example). This is due to the shrinkage effect on the estimation of random effects: 80 observations were used to estimate almost 14,000 variables. Positive estimates of random effects indicate that the relevant windows are associated with an increase of the hazard ratio, while negative estimates of random effects indicate the opposite.

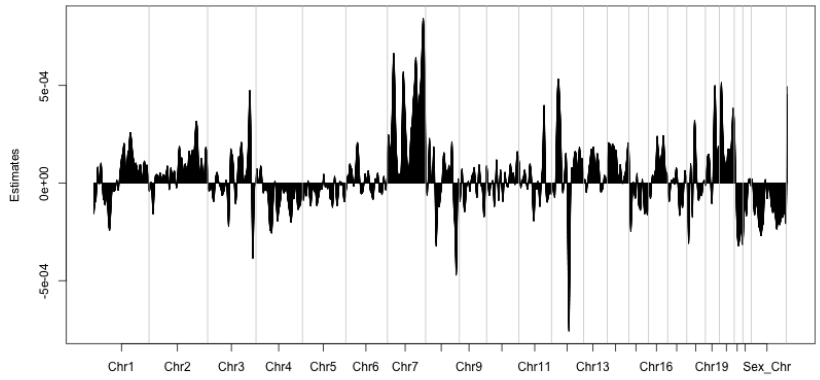


Figure 6.13: Random effects estimate b in the SCox model, using CNA profiles. Genomic windows with missing values (e.g. in the centromere regions) were excluded from analysis, hence these are not plotted. A more detailed view of the random effects estimates in each chromosome is presented in the next figure.

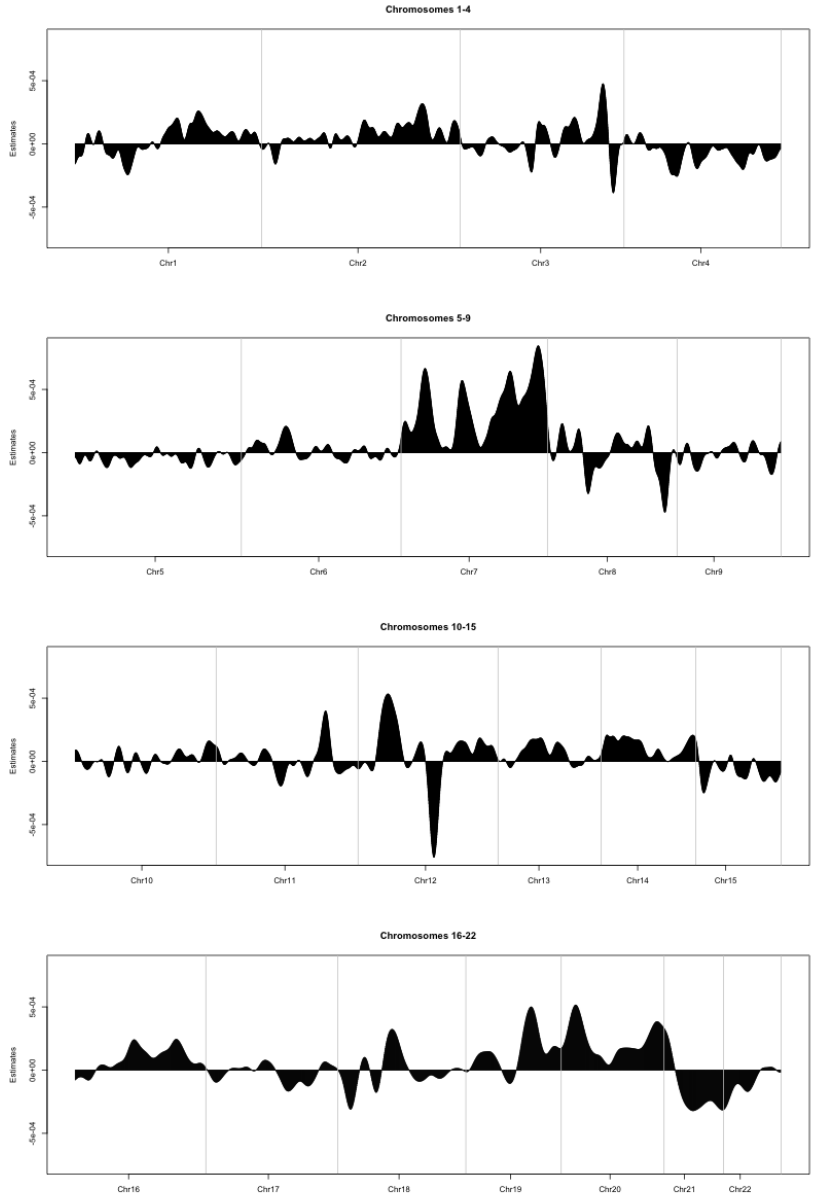


Figure 6.14: Detailed views of the random effects estimates b in each chromosome, using CNA profiles from smooth segmentation

Figure 6.15 shows the random effects estimates b of the full SCox PH model along with the significant windows, (red dotted lines), in term of CI. In other words, the windows whose CI does not include zero (1273 out of 13968).

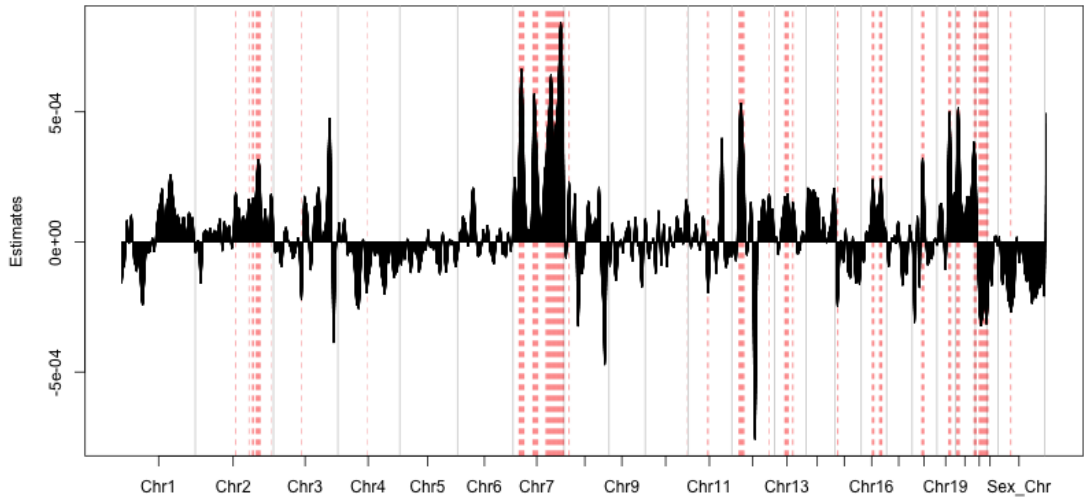


Figure 6.15: Random effects estimate b in the SCox model, using CNA profiles along with the significant windows, (red dotted lines), in term of CI. In other words, the windows whose CI does not include zero (1273 out of 13968).

6.7.4 Cumulative hazard rate and estimates of survival functions

To show that SCox PH modelling with CNA profiles is able to distinguish individuals at different levels of risk, we estimate the survivor functions for three individuals in the lung cancer dataset. These individuals correspond to low, medium, and high levels of risk based on their risk scores R_i, which equate to the 10th, 50th, and 90th percentiles of the distribution of R_i in the dataset. Figure 6.16 shows the estimated survivor functions for the three individuals using smooth-segmented profiles as random predictors. The figures indicate that the median survival times for low, medium, and high risk individuals are approximately 7.6 years, 2.46 years, and 7.3 months, respectively.

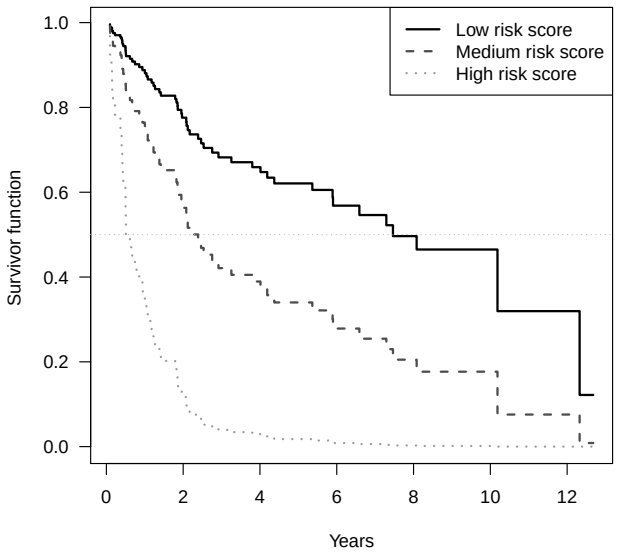


Figure 6.16: Estimated survival functions from SCox PH model for three individuals in the 10th (low risk), 50th (medium risk), and 90th (high risk) percentiles of risk set R_i, based on smooth-segmented profiles as random predictors

6.7.5 Model diagnostics

As part of model diagnostics, we plotted the cumulative hazard of the Cox-Snell residuals from the model fitting based on smoothed CNAs as shown in Figure 6.17 (solid black line). The figure indicates that the cumulative hazard line is very close to the identity line, which suggests that the SCox PH model is suitable and has a reasonably good fit for the CNA profile data. The cumulative hazard line near the top right corner of the figure is slightly jagged, as expected, due to rare events (deaths) near the end of the survival time distribution.

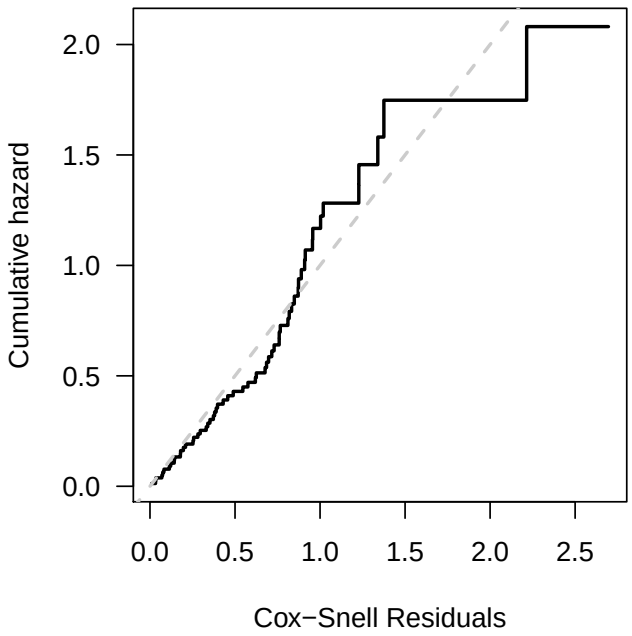


Figure 6.17: Cumulative hazard of Cox-Snell residuals (solid black line) from the SCox PH model fit, in comparison to the identity line (dashed line), based on smooth-segmented CNA profiles

6.8 Discussion

In this chapter, we have investigated how to deal with dependencies between neighbouring genomic windows and spatial characteristics of CNAs which would have been ignored if we had used the methods described in the previous chapter. We introduced three novel algorithms (Coxrho, Coxinv, and SCox) within a random effects model framework using penalised partial likelihood to model survival time using the patients' clinical characteristics as fixed effects and their CNA profiles as random effects. A key parameter in the model is the tuning parameters ($K = (\theta, \rho)$, $K = (\theta, \varrho)$, or $K = (\theta, w)$), which controls the amount of information in the CNA profiles used in the model fitting.

For the tuning parameter estimates K, we used five-fold CV partial likelihood. The tuning parameters have been recovered in the simulation study because we have imposed a strong signal in the simulated CNAs. However, for our real data, we used

the principle of the one standard error rule for the CV curve at each tuning parameter K by choosing values for K within one standard error.

Our approach in this study was a genome-wide approach, in the sense that we took into account all of the CNA information in the genome. In this regard, the model was not embedded with a variable selection mechanism, which will be discussed in the next chapter. We could use the idea of confidence interval (CI) to do a variable selection; however we found some genomics windows fall on the border of CI which make it harder to choose the significant windows.

Finally, our computational method and R package in this study is available, and can also be used for CNA profiles from array technology, provided that the (genome-wide) CNA profiles across individuals can be put into matrix form. This means that CNA estimates across individuals can be made into the same column in the data matrix, for each genomic region.

To sum up, In modelling cancer patients survival, we described three different estimation procedures using the Cox proportional hazard model to take into account cancer patients genome-wide CNAs. Unlike the extended Cox method described in the previous chapter, the new methods deal with dependencies between neighbouring genomic windows and their spatial characteristics.

The genome-wide CNA profiles are considered random predictors in the model, and the clinical variables as fixed predictors. We have three different scenarios for the distribution of CNAs:

1. Normal with mean zero and a compound symmetry covariance matrix (Coxrho), as described in Section 6.2.1;

2. Normal with mean zero and an inverse covariance matrix (Coxinv), as described in Section 6.2.2; and

3. Correlated random effects that follow a mixture of two distributions, normal and Cauchy, for the first or second differences as described in Section 6.3.

These models enabled us to assess the significance of the fixed predictors, and to examine the genomic regions associated with the patients survival. The models also enabled us to estimate individual patients' survivor functions, and distinguish the survivor functions for different groups of patients at different risk levels.

Chapter 7

Extending Cox PH model : Sparse solution

7.1 Introduction

In the previous chapter we addressed dependencies between neighbouring genomic windows and a particular spatial characteristic of CNAs. However, models based on the methods explained in the previous chapter are not embedded within a variable selection mechanism. They employ all predictors regardless of their relevance, which makes it difficult to interpret the results they produce.

In this chapter, we introduce a novel algorithm based on a sparse-smoothed Cox (SSCox) model within a random effects-model framework, using penalised partial likelihood to model survival time using patients' clinical characteristics as fixed effects and CNA profiles as random effects. We assume CNA coefficients to be correlated random effects that followed a mixture product of three distributions: normal (to achieve shrinkage around the mean values (Chapter 5)), Cauchy (for the second-order differences, to gain smoothness (Chapter 6)), and Laplace (to achieve sparsity).

This chapter presents a full gradient algorithm for maximising the penalised partial likelihood. We generalised Goeman (2010)'s idea, which follows the gradient of the likelihood from a given starting value of b and uses the full gradient at each step. Furthermore, the algorithm can automatically switch to a Newton-Raphson when it gets close to the optimum values to avoid the tendency of gradient-ascent algorithms of slow convergence.

The organization of this chapter is as follows. Section 7.2 discusses the extension of Cox PH model to include the copy number alteration as random effects. In Section 7.3 and 7.4, the estimation of the unknown parameters of the model is discussed. Simulation studies and comparison with previous methods are described and discussed in Section 7.5. Finally, the results and evaluation of the lung cancer dataset are found in Section 7.6.

7.2 SSCox PH model

similar to Chapter 5, and 6, we incorporated genome-wide CNA profiles into the original Cox PH model (Cox *et al.* (1972)). To recap the notations, let δ_i as the event indicator for the i-th patient, $i = 1, 2, \ldots, n$, where $\delta_i = 1$ if the survival time of the i-th patient, t_i, is uncensored, and $\delta_i = 0$ if their survival time, t_i, is censored. We defined X to be a matrix of size $n \times p$, where the columns of X corresponded to the different pieces of clinical information being used as fixed predictors, and the rows of X corresponded to different patients. We designated the rows of X as X_i, which is a p vector of fixed predictors for the i-th patient. The matrix Z is of size $n \times q$, with $n \ll q$, where q is the number of genomic regions in the CNA profiles (in our lung cancer cohort, $n = 80$ and $q = 13968$). We also assigned $h_0(t)$ to be the baseline hazard function, which indicates the baseline hazard rate for all of the patients in the group across time, and does not depend on any predictor.

Then we extended the Cox PH model to include the CNA profiles as random predictors,

$$h_i(t|X) = h_0(t) \exp\{X_i\beta + Z_i b\}, \tag{7.1}$$

where b is a $q-$vector of random effects that we assumed to follow a mixture model that combined shrinkage, smoothness and sparseness. This mixture model is a mixture of the product of a multivariate normal distribution, Cauchy distribution for the second deferences of b (as expained in Chapter 6), and Laplace distribution. In other words, The PDF of this product of multivariate normal, Cauchy, and Laplace distributions is

$$f(b) = \frac{f_n^{w_n}(b; D(\theta_1)) \times f_c^{w_c}(b; \Sigma(\theta_2)^{-1}) \times f_l^{w_l}(b, \sqrt{\theta_3})}{C},$$

where C is the normalising constant which does not contain any parameters. The details of these distributions explain in next paragraphs.

First, to shrink b_i so that it varied little around its mean value, we assumed b to be a normal distribution with mean zero, and covariance variance matrix $D(\theta_1)$, where $D(\theta_1) \equiv \theta_1 I_q$ (I_q is an identity matrix of size q), and $\theta_1 = \sigma_1^2$ (as explained in chapter 5) .

Secondly, Huang *et al.* (2009) suggest that as we were dealing with CNAs, we need to allow for sudden jumps or large spatial changes. To ensure this flexibility, Huang *et al.* (2009) chose a heavy-tailed distribution, the Cauchy distribution, instead of a normal distribution, because normal distributions tend to convert jumps into gradual changes. Therefore (as explained in Chapter 6), to achieve this smoothness, we assumed second-order differences of b,

$$\Delta^2 b \equiv \begin{pmatrix} b_3 - 2b_2 + b_1 \\ b_4 - 2b_3 + b_2 \\ \vdots \\ b_q - 2b_{q-1} + b_{q-2} \end{pmatrix}$$

to follow a Cauchy distribution with a location of zero and a scale of $\theta_2 I_{n-2}$, where $\theta_2 = \sigma_2^2$. In this case, $\Delta^2 b$ Cauchy with a location of zero and a scale of $\theta_2 I_{n-2}$ is equivalent to b Cauchy with a location of zero and an inverse scale matrix of

$$\begin{aligned} \Sigma(\theta)^{-1} \equiv & \theta_2^{-1} R_2^{-1}, \\ \text{where } R_2^{-1} \equiv & (\Delta^2)' \Delta^2 \\ = & \begin{pmatrix} 1 & -2 & 1 & \cdots & \cdots & \cdots & 0 \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ 0 & & & & 1 & -2 & 1 \end{pmatrix} \end{aligned} \tag{7.2}$$

((see Pawitan, 2013)).

Finally, to achieve the sparseness necessary for variable selection, we assume b_i to follow a Laplace distribution with a location of zero and a scale factor of $\sqrt{\theta_3}$, where $\theta_3 = \sigma_3^2$.

For simplicity, we set $\theta_1 = \theta_2 = \theta_3$ so that we would only have to look for one tuning parameter θ instead of three different values for θs.

The addition of the random predictors Z also extend the log partial likelihood for β to include the random effect b, as

$$\begin{aligned} l_p(\beta, \theta, b) = \sum_{i=1}^{n} \Bigg[\delta_i (X_i\beta + Z_i b) - \delta_i \log\Bigg(\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b) \Bigg) \Bigg] \\ - \Bigg[w_n \big[\frac{1}{2} b' D(\theta)^{-1} b \big] \\ + (w_c) \big[\frac{q+1}{2} \log\big(1 + b'\Sigma(\theta)^{-1} b\big) \big] \\ + \frac{(w_l)}{\sqrt{\theta}} \big[\sum_{k=1}^{q} |b_i| \big] \Bigg], \end{aligned} \tag{7.3}$$

where $\Bigg[w_n \big[\frac{1}{2} b' D(\theta)^{-1} b \big] + (w_c) \big[\frac{q+1}{2} \log\big(1 + b'\Sigma(\theta)^{-1} b\big) \big] + \frac{(w_l)}{\sqrt{\theta}} \big[\sum_{k=1}^{q} |b_k| \big] \Bigg]$ is the part of the partial log likelihood that corresponds to the mixture of normal, second-order difference Cauchy, and Laplace assumptions for b, and $R(t_i)$ is a set of patients who are at risk at time t_i (the risk set).

The estimation of the model parameters β and b, and the tuning parameter $K = (\theta, w_n, w_c, w_l)$, is done by first estimating β and b at fixed K, as described in Section 7.3. The estimation of the tuning parameter K is done via five-fold cross-validation partial likelihood, introduced by Verweij & Van Houwelingen (1993), as described in Section 7.4. In practice, we alternate the two estimation steps to obtain all of the estimates.

7.3 Estimation of β and b for fixed tuning parameters $K = (\theta, w_n, w_c, w_l)$

Estimation of fixed effects β

We derived an estimation of β at a fixed value for K by first partially differentiating the log partial likelihood $l_p(\beta, b)$ from Eq. (7.3) with respect to β. The resulting estimating

equation for β is:

$$\sum_{i=1}^{n} \delta_i \left[X_i - \frac{\sum_{j \in R(t_i)} X_j \exp(X_j\beta + Z_j b)}{\sum_{j \in R(t_i)} \exp(X_j\beta + Z_j b)} \right]. \tag{7.4}$$

Estimation of random effects b

To derive an estimation of the random effects b, we could use a standard convex optimizer such as as Broyden-Fletcher-Goldfarb-Shanno (BFGS) method (Head & Zerner (1985)), conjugate gradient (Møller (1993)), and Nelder-Mead (Lagarias *et al.* (1998)) or any other standard convex optimizer. However all of these methods suffer from two general issues. First, they use an approximate solution. Second, they are computationally slow as can bee seen in Table 7.1. In other words, It take too much time to estimate the random effect. This is become more harm when we include these estimation in the cross validation method.

Alternatively, we derive an estimation of b by using **Gradient Ascent**, as proposed by Goeman (2010), with some modifications. Goeman (2010) chose to use only the log partial likelihood because only used a Laplace distribution whereas we included the penalty of the normal and Cauchy distributions in the log partial likelihood. Although we combined the log partial likelihood with the penalty of both distributions, this combination is still a highly regular function: concave and everywhere at least twice differentiable. As a result, one case of our method is the same version as Goeman (2010)'s method, used when the weights of the normal and Cauchy distributions are equal to zero.

To understand why gradient ascent is the method of choice for our estimation, it is worthwhile to look more closely into the penalized log partial likelihood function (7.3) that is to be optimized. We rewrote the target function of Eq. (7.3) as a sum of two terms:

$$l_p(\beta, b) = l_{\text{pnc}} - \frac{(w_l)}{\sqrt{\theta}} \Big[\sum_{k=1}^{q} |b_k| \Big] \tag{7.5}$$

1. The first term, l_{pnc}, which corresponds to the first three terms on the right hand side of Eq. (7.3), is the log partial likelihood plus the penalty of the normal and Cauchy parts. In the models we are interested in, this is a highly regular function: concave and everywhere at least twice differentiable.

2. The second term, the penalty of the Laplace part, $P(b) = \frac{(w_l)}{\sqrt{\theta}}\left[\sum_{k=1}^{q} |b_k|\right]$, is less well behaved: it is concave and continuous, but is only differentiable at points with $b_k \neq 0$ for all k.

The penalized log partial likelihood (7.5) is not differentiable everywhere because of the lack of differentiability of the Laplace penalty function. Therefore, The Newton-Raphson algorithm can not be applied directly. An alternative approach is the gradient ascent algorithm. A gradient ascent algorithm for the optimization of one coefficient just calculates the derivative at that point and takes a step in that direction. When the derivative is zero, it estimates.

Even though the gradient ascent is robust in its simplicity, when a basic gradient ascent algorithm is applied on the lasso problem, it is not successful. To solve that, Goeman (2010) defined a directional derivative as

$$l'_p(b; v) = \lim_{t \to 0} \frac{1}{t}\{l_p(b + tv) - l_p(b)\}$$

for every point b and every direction $v \in \Re^q$. The gradient can then be defined for every b as the scaled direction of steepest ascent. The algorithm follows the gradient in the direction v_{opt} which maximizes $l'_p(b; v)$ among all v such that $||v|| = 1$

The gradient $g(b)$, which is a vector of length q, can be calculated from the unpenalized log partial likelihood gradient with the penalty of normal and Cauchy parts, $(d(b) = \partial l_{\text{pnc}}/\partial(b) = (d(b_1), \ldots, d(b_q))')$, as

$$g(b_k) = \begin{cases} d(b_k) - \frac{(w_l)}{\sqrt{\theta}} \text{ sign } (b_k) & \text{if } b_k \neq 0 \\ d(b_k) - \frac{(w_l)}{\sqrt{\theta}} \text{ sign } (d(b_k)) & \text{if } b_k = 0 \text{ and } |d(b_k)| > \frac{(w_l)}{\sqrt{\theta}} \\ 0 & \text{otherwise,} \end{cases}$$

Where $k = (1, 2, \ldots, q)$ and

$$\text{sign } (b_k) = \begin{cases} 1 & \text{if } b_k > 0 \\ 0 & \text{if } b_k = 0 \\ -1 & \text{if } b_k < 0 \end{cases}$$

This gradient is discontinuous at every point where the penalized log likelihood l_p is not differentiable, i.e at every point with $b_k = 0$ for some k.

Then gradient ascent algorithm is constructed that updates the coefficients b with step size t until convergence:

$$b_{\text{new}} = b_{\text{old}} + tg(b_{\text{old}})$$

The gradient ascent algorithm from Goeman (2010) also includes Newton-Raphson steps to include the fast optimization properties of the Newton-Raphson when near the optimum. Let t_{opt} be the optimum and t_{edge} the borders of the sub domain, where a sub domain is a space that does not include any zero. Where,

$$t_{\text{edge}} = \min_k \left\{ -\frac{b_k}{g(b_k)} : \text{sign}(b_k) = -\text{sign}\{g(b_k)\} \neq 0 \right\},$$

and

$$t_{\text{opt}} = -\frac{l'_p(b; g(b))}{l''_p(b; g(b))}.$$

$l'_p(b; g(b))$ and $l''_p(b; g(b))$ is the directional first and second derivative, respectively, for every b and $g(b)$

$$l'_p(b; g(b)) = g(b).g(b)/||g(b)||$$
$$l''_p(b; g(b)) = g(b)' \frac{\partial^2 l_{\text{pnc}}}{\partial b \partial b'} g(b).$$

Calculating the full $q \times q$ Hessian matrix of l_{pnc} to calculate the directional second derivative is hardly and ever necessary because the direction $g(b)$ of interest, which is the direction of the gradient, will have many zeros.

Then the algorithm is shown below

Algorithm 1 Gradient ascent algorithm for penalized partial log likelihood of Sparse smoothed Cox with NewtonRaphson.

1: **Start** with initial values $b^{(0)}$

2: *For steps* $s = 0, 1, 2, \ldots$: iterate

$$b^{(s+1)} = \begin{cases} b^{(s)} + t_{\text{edge}} g(b^{(s)}) & \text{if } t_{\text{opt}} \geq t_{\text{edge}} \\ b_{NR}^{s+1} & \text{if } t_{\text{opt}} < t_{\text{edge}} \text{ and } \text{sign}(b_{NR}^{(s+1)}) = \text{sign}(b_+^s) \\ b^{(s)} + t_{\text{opt}} g(b^{(s)}) & \text{otherwise,} \end{cases}$$

3: **End if** convergence occurs when $g(b) = 0$.

where b_+ indicates the set of active variables and b_{NR} is a NewtonRaphson step:

$$b^{(s+1)} = b^{(s)} - \{\frac{\partial^2 l_{\text{pnc}}}{\partial b \partial b'}\}^{-1} g(b^{(s)}).$$

7.4 Estimating the tuning parameters $K = (\theta, w_n, w_c, w_l)$ by cross-validation

We use a cross-validation criterion, which was based on the unpenalised log partial likelihood (standard Cox), proposed by Verweij & Van Houwelingen (1993), to estimate the tuning parameters $K = (\theta, w_n, w_c, w_l)$. We used fivefold cross-validation as explained in Chapter 6, Section 6.5.2

To choose the path of θ for the set of weights (w_n, w_c, w_l), we started from a minimum value of θ, namely $\theta_{\text{min}} = \max_i \frac{w_l}{|d_i(0)|}$, which yielded the estimate $\hat{b} = 0$. We then set the maximum value of θ to be $\theta_{\text{max}} = \theta_{\text{min}}/\epsilon$, and computed solutions over a grid of v between θ_{min} and θ_{max}, where $\theta_j = \theta_{\text{min}}(\theta_{\text{max}}/\theta_{\text{min}})^{j/v}$ for $j = 0, \ldots, v$.

7.5 Numerical study

We compared our proposed method, the SSCox PH model, with a sparse Cox PLS having an L_1 penalty (SPLS-L1) and a sparse Cox PLS with an HL penalty (SPLS-HL) (presented in Lee *et al.* (2013)). All comparisons are based on 100 simulation datasets. To get a high-dimension setting, we had to set the sample size for n to be less than the number of predictors q. Therefore, we chose the values of $n = 100$ and $p = 200$, with $\%30$ censoring rate.

7.5.1 Simulation setting

We first conducted simulations to assess the performance of the SSCox PH model. We followed the simulation setting of Bøvelstad *et al.* (2007), Nygård *et al.* (2008), and Lee *et al.* (2013). The steps taken to produce this simulation are explained in the next paragraph.

First, the covariate matrix Z was generated from a multivariate normal distribution with a zero mean vector and the 200×200 covariance matrix Σ. Here we assumed

$\Sigma = diag\{\Sigma_c\}_{c=1,....,10}$, with a 20×20 matrix Σ_c. Σ_c has diagonal elements $\sigma_c^2 = 1$ and off-diagonal elements $\rho\sigma_c^2$ for all $c = 1, ..., 10$. In this simulation, we set ρ equal to 0.9, because it is close to our real data and Lee *et al.* (2013) argued that their methods work better with higher correlation.

For $i = 1, ..., n$, we assumed

$$\eta_i = \sum_{j=1}^{40} z_{ij} b_j,$$

where z_{ij} is the (i, j) element of Z. $b_j = \exp(-\alpha(j-1))$ and $b_{j+20} = -b_j$ for $j = 1, ..., 20$. This setting indicated that only the first 40 covariates are associated with survival time among 200 covariates.

The regression parameters were exponentially decaying, and the speed of the decay was controlled by the parameter α. We used a slow decay, where $\alpha = 0.0141$, such that $\exp(-49\alpha) = 0.5$.

Then, we generated the survival time T_i from a Weibull distribution with a hazard rate of $h_0(t_i) = 5t_i^4$, and the censoring time C_i was taken to be uniform $(0, 3)$ distributed, which gave censoring rates of approximately 35%.

7.5.2 Simulation results: one simulation and computational time comparison

Figure 7.1 shows the estimation of the random effects b for one random simulation, with the optimal θ based on the extended Cox PH model presented in Chapter 5 ($w_n = 1, w_c = 0, w_l = 0$), ($w_n = 0, w_c = 0, w_l = 1$), ($w_n = 0.5, w_c = 0, w_l = 0.5$) and SSCox model ($w_n = 0.4, w_c = 0.2, w_l = 0.4$).

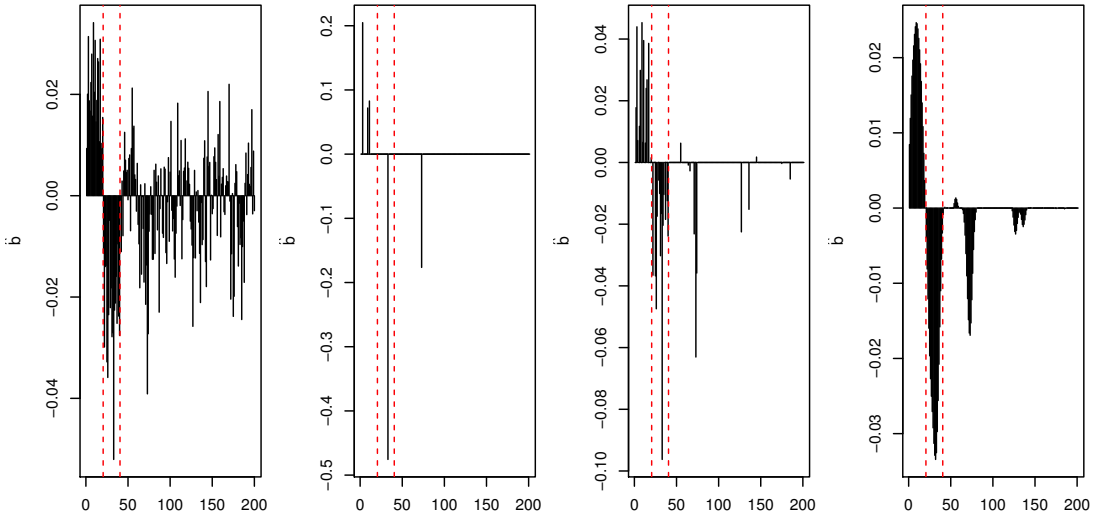


Figure 7.1: Estimation of the random effects b based on the extended Cox PH model presented in Chapter 5 ($w_n = 1, w_c = 0, w_l = 0$), ($w_n = 0, w_c = 0, w_l = 1$), ($w_n = 0.5, w_c = 0, w_l = 0.5$) and SSCox model ($w_n = 0.4, w_c = 0.2, w_l = 0.4$), from left to right

We can see that by setting the wight of Laplace and Cauchy to be equal to zero ($w_n = 1, w_c = 0, w_l = 0$), our proposed method SSCox reduces to the extended Cox PH introduced in Chapter 5 (Ridge). As we can see from the left panel of Figure 7.1, none of the estimation of random effects b has been set to be equal to zero because the ridge penalty (L_2-penalty) is used. Moreover, our setting in the simulation indicated that only the first 40 covariates are associated with survival time among 200 covariates. The first 40 values reveal a signal as we anticipated, but the rest covariates (41-200) are not equal to zero.

In the second left panel of Figure 7.1, we set the wight of Normal and Cauchy to be equal to zero ($w_n = 0, w_c = 0, w_l = 1$); this setting reduces our proposed method (SSCox) to the Lasso solution (Tibshirani *et al.* (1997)). It is clear that Lasso tends to identify only one of the correlated features which are associated with the outcome; this is a well known potential problem with the Lasso penalty (L_1-penalty).

In third left panel of Figure 7.1, we set the wights of Normal and Laplace to be equal to 0.5 ($w_n = 0.5, w_c = 0, w_l = 0.5$); this setting reduces our proposed method

(SSCox) to the the elastic net (EN) proposed by Zou & Hastie (2005) for linear regression and Engler & Li (2007) for survival analysis. This method adds a ridge-type penalty to the Lasso which improves Lasso's ability to identify sets of correlated genes associated with outcome. However, as we can see from the figure that EN method did not address the spatial dependence structure of CNAs. Therefore, we can see from the right panel of Figure 7.1 the effect of imposing smoothness by using Cauchy on the second differences of the random effects ($w_n = 0.4, w_c = 0.2, w_l = 0.4$).

As we mentioned before that in order to derive an estimation of the random effects b, we could use a standard convex optimizer such as as BFGS method. However, it is computationally slow as can bee seen in Table 7.1. Alternatively, we derive an estimation of b by using gradient Ascent algorithm with and without switching to a Newton-Raphson algorithm. Table 7.1 shows computation time (in seconds) in our proposed method SSCox by using standard convex optimizer (BFGS), gradient ascent, and gradient ascent with Newton-Raphson algorithm for one random simulation. All convergence criteria and other settings were set equal in the three algorithms.

Table 7.1: Computation time comparison between the standard convex optimizer (BFGS), and full gradient approach of SSCox with and without without switching to a NewtonRaphson algorithm for one random simulation , time is calculated by seconds

Method	BFGS	SSCox (without NR)	SSCox(NR)
Lasso	203	11	4
Elastic net	251	15	6
SSCox	441	71	33

Table 7.1 shows that the full gradient algorithm can be much quicker than the standard convex optimizer (BFGS). This is become more obvious when we include these estimation in the cross validation method.

7.5.3 Simulation results: comparative study

For each data set, we evaluated the following methods:

- SPLS-L1: sparse Cox PLS with L1 penalty Lee *et al.* (2013)
- SPLS-HL: sparse Cox PLS with HL penalty Lee *et al.* (2013)

- Our proposed method (SSCox)

We evaluated these methods with respect to variable selection and prediction.

1. As a measure of variable selection, following Chun & Keleş (2010), we calculated the average of the sensitivity and specificity, defined by

   - sensitivity: the proportion of nonzero estimates among the true nonzero elements of b, and
   - specificity: the proportion of zero estimates among the true zero elements of b.

2. As a measure of the methods' prediction power, following Nygård *et al.* (2008), we computed

$$-2pl = -2l_{p_{\text{Cox}}^{(test)}}\left(\hat{b}^{train}\right), \tag{7.6}$$

   where $l_{p_{\text{Cox}}^{(test)}}\left(\hat{b}^{train}\right)$ is the log partial likelihood of standard Cox PH for independent test data evaluated at $\hat{b}^{train}$, which is the estimate of b based on training set. For comparison purposes, we calculated the difference between the $-2PL$ of each method and the $-2PL$ of the true model.

3. We also looked at the criteria for prediction performance, because Lee *et al.* (2013) argued that Eq. (7.6) might assess it fragmentarily. In the simulation setting, we actually know the true failure time $T_i^{(test)}$ of the test dataset, and we can predict that the median survival time $\hat{T_i^m}^{(test)}$ follows,

$$\hat{T_i^m}^{(test)} = \hat{H_0}^{-1}\big[-\log(1/2)\exp\big(-Z_i^{test}\hat{b}^{train}\big)\big],$$

   where Z_i^{test} are the covariates of the i-th individual in the test data, and $\hat{H}_0(t)$ is the estimated cumulative base line hazard function for the test data evaluated at $\hat{b}^{train}$.

   With $T_i^{(test)}$ and $\hat{T_i^m}^{(test)}$, we computed the sum of squared prediction error (SSPE) and sum of absolute prediction error (SAPE) as the prediction measures,

$$SSPE = \sum_{i=1}^{n}\left(T_i^{(test)} - \hat{T_i^m}^{(test)}\right)^2$$

and

$$SAPE = \sum_{i=1}^{n} |T_i^{(test)} - \hat{T_i^m}^{(test)}|.$$

The prediction performance was better when $-2pl$, SSPE, and SAPE were smaller.

Before making the comparison between the competitive methods, it is worth to take a look to the estimation of random effects b based on each method. Figure 7.2 shows the estimation of the random effects b for one simulation with the optimal θ also based extended Cox PH model presented in Chapter 5 ($w_n = 1, w_c = 0, w_l = 0$), SSCox ($w_n = 0.4, w_c = 0.2, w_l = 0.4$), SPLS-L1 method, and SPLS-HL method.

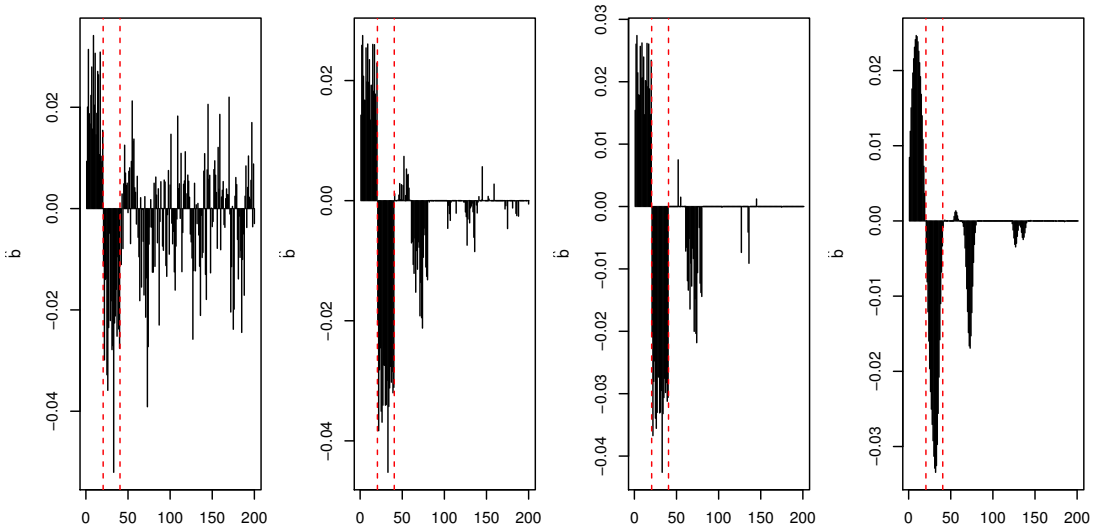


Figure 7.2: Estimation of the random effects b based on the extended Cox PH model presented in Chapter 5 ($w_n = 1, w_c = 0, w_l = 0$), SPLS-L1 method, SPLS-HL method, and SSCox model ($w_n = 0.4, w_c = 0.2, w_l = 0.4$), from left to right

It is clear from looking to Figure 7.2 that the extended Cox PH model (left panel) is the only method that does not do feature selection. Our proposed method (SScox) is different from SPLS-L1 and, SPLS-HL in the smoothing imposed in the random effects b.

We summarise the simulation result from the exponential slow decay model with $\rho = 0.9$ in Table 7.2 and Figure 7.3.

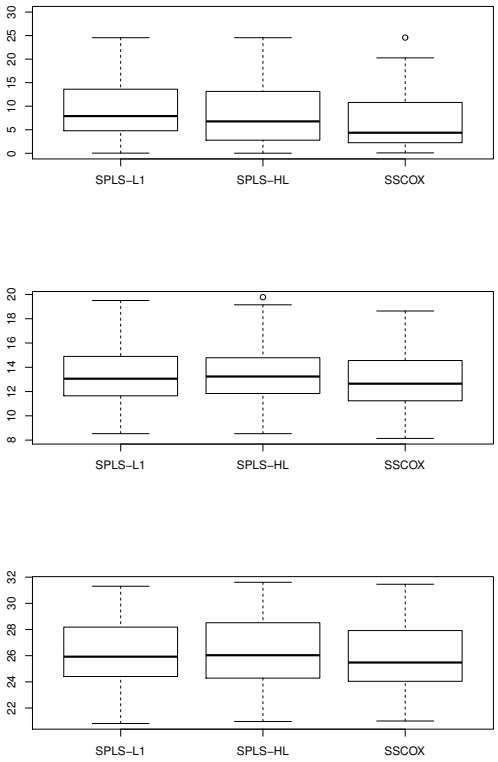


Figure 7.3: Box plots of the absolute difference of the -2 unpenalised likelihood from Eq. (7.6) of a method and the true model (top); the SSPE (middle); and the SAPE (bottom)

As can be seen from Figure 7.3, the SSCox model generally performed better in making predictions than SPLS-L1 or SPLS-HL. The SSCox has the smallest -2pl, MSPE, and MAPE. For the MSPE, the means of 100 simulations are $13.28, 13.42,$ and 12.73 for SPLS-L1, SPLS-HL, and SSCox, respectively, while the MAPE values were $26.09, 26.21,$ and 25.66.

In terms of sensitivity and specificity, we know that ordinary Cox PLS models and Cox models with ridge penalties always have a specificity equal to zero. However, as all methods which were compared here have the idea of sparseness, we can see that the specificity was not equal to zero. Table 7.2 shows that all methods had a large degree of sensitivity, and the largest amount of specificity for SSCox. In fact, SSCox had a

sensitivity equal to 0.976, because the smoothness imposed in our method interacted with the zero line when the parameters changed their signs.

Table 7.2: Performance measures for variable selection

Method	Sensitivity	Specificity
SPLS-L1	0.998	0.273
SPLS-HL	0.985	0.575
SSCox	0.976	0.592

7.6 Real data analysis

7.6.1 Model fit: estimating the tuning parameters K

An important parameter to be estimated from the SSCox PH model is the tuning parameter $K = (\theta, w_n, w_c, w_l)$. These parameters are important in the interpretation; as θ moves towards zero (in limit terms), the estimates of the random effects will be closer to zero, and no information in CNAs was taken into account in the model. Also w_l controls the sparseness of the model, while w_c controls the smoothness of the model. We would usually be interested in models for more than one amount of regularisation. One could solve a 3 dimensional grid of θ, w_n, and w_c; however, we found this to be computationally impractical, and to do a poor job of model selection similar to Simon *et al.* (2011) findings. Instead, we fixed the mixing parameter and computed solutions for a path of θ values (as regulated the degree of sparsity). We began the path with θ_{min}, set sufficiently small so that $b = 0$, and increased θ until we were near the unregularised solution.

Figure 7.4 shows the CVPL(θ) when $w_l = 0.5, w_n = 0.3,$ and $w_c = 0.2$. To estimate θ, we used the principle of cross validated partial likelihood (CVPL) with one standard error rule (Eq. (6.22)). By applying the one standard error rule, the figure gives a $\log(\theta) = -6.075$, corresponding to $\theta = 0.0023$.

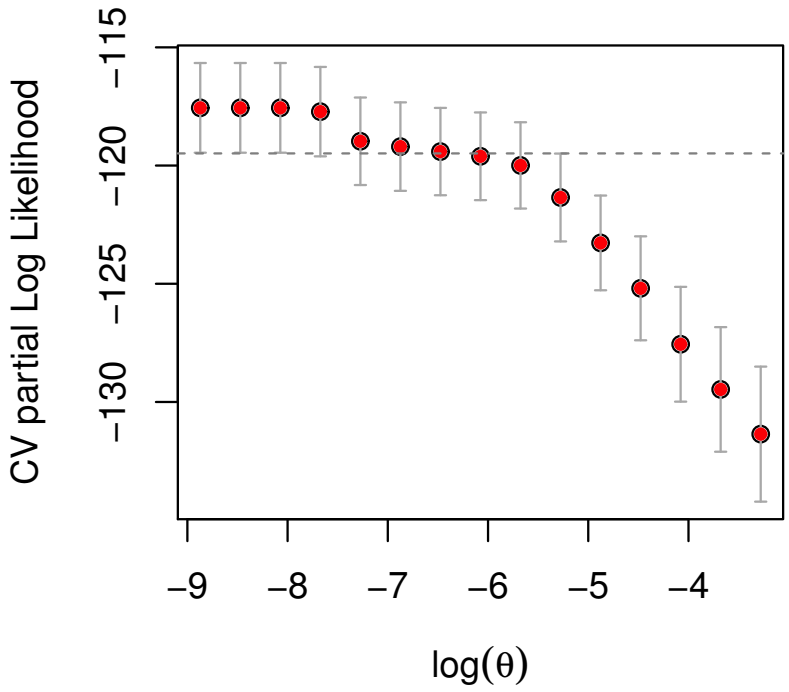


Figure 7.4: Cross validated partial likelihood (CVPL(θ)); the horizontal dotted line indicates one standard error (Eq. (6.22)) of CVPL(θ)

7.6.2 Model fit: fixed predictors

Using this value of θ, the estimates of the fixed effects and their inferences can be seen in Table 7.3. For comparison, we estimated the fixed effects under both conditions with and without the CNA profiles in the model. The table shows that the variables of Age, Stage-T, and Stage-N are statistically significant ($p < 0.05$). The estimates indicate that the hazard ratio increases by about six percent ($e^{0.055} \approx 1.06$) as age-at-operation increases by one year (all else being equal). The positive estimates of Stage-T3 indicate that larger tumour size is associated with a significant increase in the hazard level (relative to Stage-T1 as the baseline). Similarly, the estimates of Stage-N2 indicate that a wider spread of cancer cells to the lymph nodes significantly increases

the hazard level (relative to the Stage-N0 as the baseline).

Table 7.3: Summary of fixed predictors

Predictor	Estimate	Exp	Std.Error	$z$values	p-value
(Without CNA profiles)					
Age	0.0551	1.06	0.0164	3.37	0.0008
StageT2	0.1818	1.20	0.3215	0.57	0.5700
StageT3	1.7623	5.83	0.6392	2.76	0.0058
StageN1	0.3616	1.44	0.3019	1.20	0.2300
StageN2	1.3653	3.92	0.4824	2.83	0.0047
(With CNA profiles)					
Age	0.0551	1.06	0.0164	3.37	0.0008
StageT2	0.1817	1.20	0.3215	0.57	0.5679
StageT3	1.7622	5.83	0.6392	2.76	0.0058
StageN1	0.3616	1.43	0.3019	1.20	0.2301
StageN2	1.3653	3.92	0.4824	2.83	0.0047

7.6.3 Model fit: random effects

The random effect estimates b of the full Cox PH model, using CNA profiles from smooth segmentation, are presented in Figure 7.5. The magnitude of the estimates is relatively small (compared to the fixed effects estimates, for example). This is due to the effect of shrinkage on the estimation of random effects: 80 observations were used to estimate almost 14,000 variables.

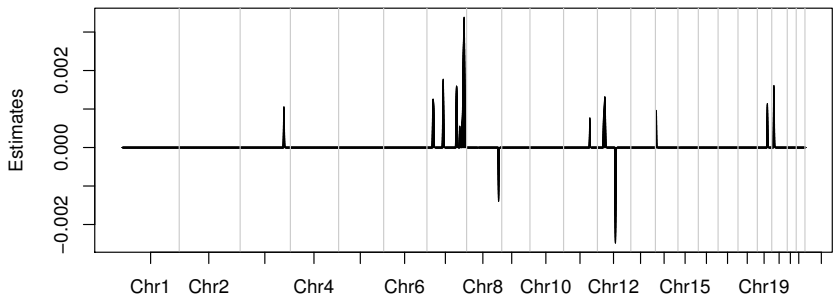


Figure 7.5: Random effects estimates b in the full model, using CNA profiles. Genomic windows with missing values (for example in the centromere regions) were excluded from analysis, and hence not plotted. A more detailed view of the random effects estimates in each chromosome is presented in Figure 7.6 .

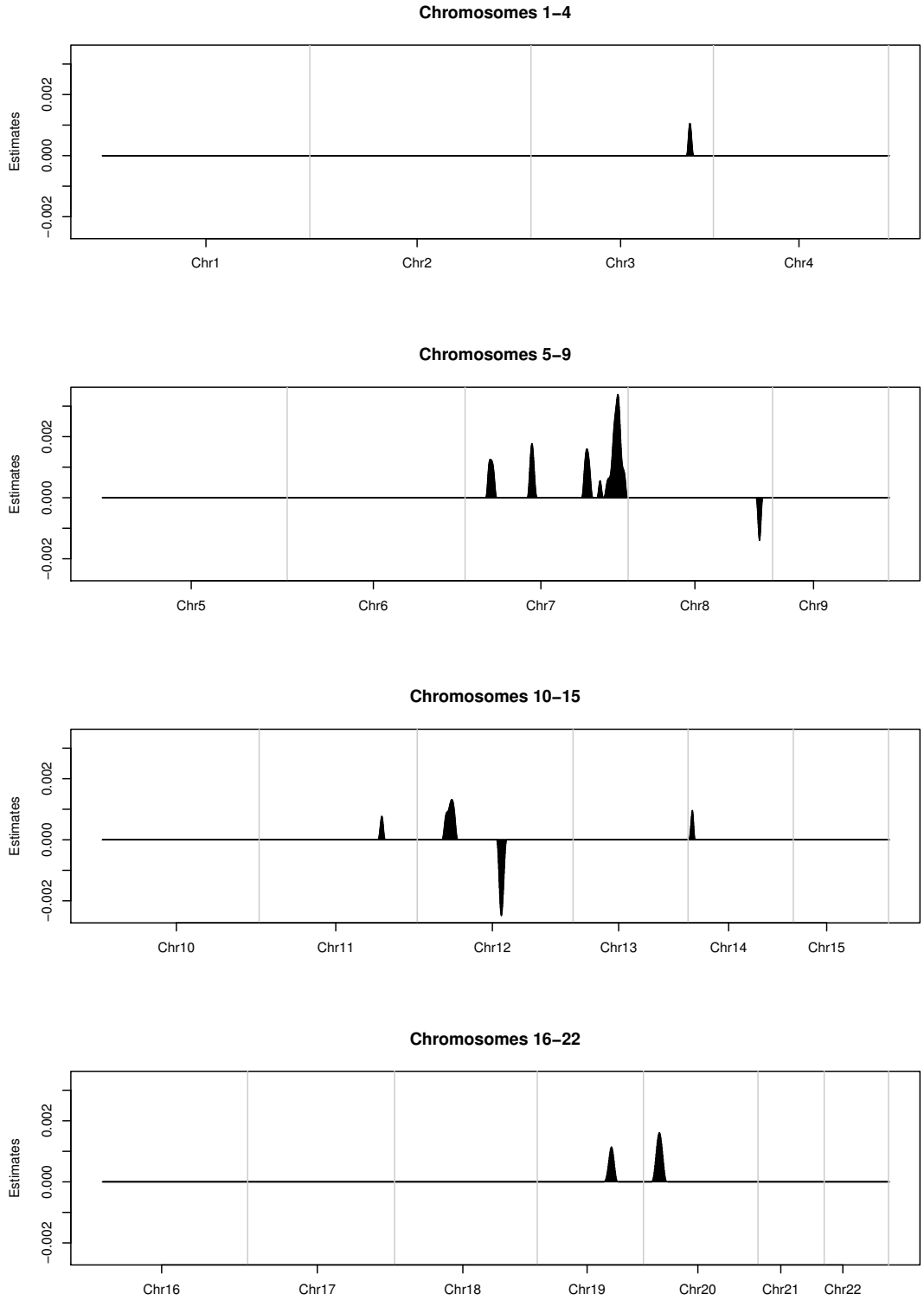


Figure 7.6: Estimates of the random effects $\hat{b}$ in the full model. Genomic windows with missing values were also removed from this figure.

Positive estimates of random effects indicated that the relevant windows are associated with increases in hazard levels, while negative estimates of random effects indicated the opposite. In this regard, Pelosi *et al.* (2007), Antoniou *et al.* (2013), and Flacco *et al.* (2015) found that TERC copy number in chromosome 3 gain in early-stage non-small-cell lung cancer (NSCLC). Moreover, there are many studies that show the involvement of other chromosomes in NSCLC. For example, researchers such as Lee *et al.* (1987), Buckingham *et al.* (2007), and Kitada & Yamasaki (2008) have demonstrated the involvement of chromosome 7 in NSCLS. These genes and more relative genes, which are summarised in Table 7.4, are associated with increases in hazard levels in NSCLC.

On the other hand, negative estimates of random effects indicated that the relevant windows are associated with decreases in hazard levels. The negative estimates appeared in chromosome 8 and 12. There are many studies shows the relation of genes in chromosome 8 in 12 with tumour suppressor (protective). For example, Schemionek *et al.* (2016) found that MTSS1 in chromosome 8 decreased clonogenic capacity and motility of murine myeloid progenitor cells and reduced tumor growth. Also, Yue *et al.* (2012) showed that ZHX2 overexpression significantly reduced the growth of tumor in mice. Li *et al.* (2015) showed that NDUFB9 was a suppressor of breast cancer cell proliferation, migration and invasion, and Cheng *et al.* (2016) found that FAM84B significantly reduced vitro cell growth, migration and invasion. In chromosome 12, KRAS, YEATS4, FRS2, SLCO1B3, and LGR5 have tumour-suppressive activity as Zhang *et al.* (2001), Pikor *et al.* (2013), Valencia *et al.* (2011), Wu *et al.* (2014), and Lee *et al.* (2008) , respectively, showed in there studies.

Table 7.4: Genes related to NSCLC

Gene	Chromosome	Reference
TERC	3	Flacco *et al.* (2015), Antoniou *et al.* (2013) Pelosi *et al.* (2007)
SKIL	3	Pelosi *et al.* (2007)
ECT2L	3	Justilien *et al.* (2011)
TNFSF10	3	Lee *et al.* (2014)
TWIST1	7	Avasarala *et al.* (2015)
IL6	7	Zhou *et al.* (2015)
MACC1	7	Wang *et al.* (2015b)
HOXA1	7	Xiao *et al.* (2014)
HOXA4	7	Kang (2013)
HOXA5	7	Wang *et al.* (2015a)
FSCN1	7	Luo *et al.* (2015)
HOXA11	7	Hwang *et al.* (2013)
GUSB	7	Zhang *et al.* (2010)
CAV1	7	Tian *et al.* (2015)
CAV2	7	Kettunen *et al.* (2004)
MET	7	Domblides *et al.* (2015)
CFTR	7	Son *et al.* (2011)
PPP1R3	7	Huang *et al.* (2013)
CALU	7	Turacli *et al.* (2015)
MEST	7	HIROFUMINAKANISHI *et al.* (2004)
VIPR2	7	Moody *et al.* (2000)
XRCC2	7	Butkiewicz *et al.* (2011), Sullivan *et al.* (2014)
CDK5	7	Lockwood *et al.* (2008), Choi *et al.* (2009)
ARHGEF5	7	He *et al.* (2013)
CASP2	7	Muppani *et al.* (2011)
BRAF	7	Sereno *et al.* (2015)
EPHB6	7	Bulk *et al.* (2012)
AKR1B10	7	Kang *et al.* (2011)
EGFR	7	Boukakis *et al.* (2010)
MMP1	11	Bi *et al.* (2015)
MMP3	11	Zhao *et al.* (2015)
MMP7	11	Lopez-Ayllon *et al.* (2015)
MMP8	11	Bi *et al.* (2015)
MMP10	11	Zhang *et al.* (2014)
MMP12	11	Tian *et al.* (2015)
CADM1	11	Jang *et al.* (2015), Gyhorffy *et al.* (2013)
CEACAM1	19	Fiori *et al.* (2012)
CEACAM3	19	Beauchemin & Arabzadeh (2013)
CEACAM5	19	Chen *et al.* (2015c)
CEACAM6	19	Han *et al.* (2014)
CEACAM7	19	Beauchemin & Arabzadeh (2013)
TGFB1	19	Vizoso *et al.* (2015)
AKT2	19	Chen *et al.* (2015a)
LGALS4	19	Selamat *et al.* (2012)
PCNA	20	Huang *et al.* (2015a)
BMP2	20	Tan & Chen (2014)
BMP7	20	Lazar *et al.* (2013)

Table 7.5 shows some genes with nonzero regression coefficients that we found in our analysis, though we did not locate any studies showing a relationship between these genes and lung cancer. There is, however, evidence that these genes are related to other types of cancer.

Table 7.5: Genes with nonzero regression coefficients that are related to cancers other than lung cancer

Gene	Chromosome	Related Cancer	Reference
PDCD10	3	Prostate	Fu *et al.* (2016)
ABCB5	7	Breast and skin	Lal *et al.* (2016)
SBDS	7	Leukemia	Aalbers *et al.* (2013)
TES	7	Breast and colorectal	Li *et al.* (2016)
ST7	7	Breast and prostate	Hooi *et al.* (2006)
POT1	7	Breast	Motevalli *et al.* (2014)
FLNC	7	Breast and prostate	
SMO	7	Pancreatic	Guo *et al.* (2013)
TRIM24	7	Breast and liver	Chambon *et al.* (2011)
CUL1	7	Breast	Bai *et al.* (2013)
ING3	7	Liver and prostate	Almami *et al.* (2016)
YAP1	11	Breast and liver	Yu *et al.* (2013)
BIRC2	11	Cervical and leukemia	Mak *et al.* (2014)
BIRC3	11	Breast and pancreatic	Gan *et al.* (2016)
ARHGEF1	19	Breast and Colorectal	Huang *et al.* (2015b)
CD79A	19	Leukemia	Palanca-Wessels *et al.* (2015)
ZFP36	19	Liver and prostate	Zhu *et al.* (2015)
RASSF2	20	Cervical and breast	Perez-Janices *et al.* (2015)

Table 7.6 shows some genes with nonzero regression coefficients that we found in our analyses. In contrast to those shown in Table 7.5 above, though, we found no prior studies showing evidence of any relationship between these genes and any type of cancer.

Table 7.6: Genes with nonzero regression coefficients but no known relationship to any type of cancer

genes							
(Chromosome 3)							
ZBBX	WDR49	SERPINI2	GOLIM4	LRC31	ACTR3	MYNN	SAMD7
SEC62	PHC3	SIC7A14	SIC2A2	TINK	TMEM212	PLD1	NCEH1
NLGN1	NCEH1	NLGN1	SPATA16				
(Chromosome 7)							
TMEM196	RPL21P75	ITGBB	SP8	CDCA7L	RAPGEF5	STEAPIB	STK31
TOMM7	KLHL7	NUPL2	GPNMB	TRA2A	CCDC126	MPP6	DFNA5
OSBL3	CYCS	NPVF	SNX10	SKAP2	CBX3	PRR15	ZNF736
ASL	TPST1	CRCP	KCTD7	TYW1	IMMP2l	LRRN3	DOCK4
ZNF277	CAPZA2	IFRD1	LSMEM1	TMEM168	GBR85	FOXP2	ASZ1
CTTNBP	MDF1C	LSM8	TFEC	SPAM1	GBR37	CCDC136	GRM8
ZNF800	GCC!	ARF5	PAX4	ZC3HC1	SND1	LRRC4	RBM28
OPNISW	IMPDH1	HILPDA	STRIP2	METTL28	FAM71F2	KCD	ATP6VIF
IRF5	TNP03	TSPAN33	AHCY12	SMKR1	NRF1	KIHDC10	TMEM209
SSMEM1	CEP41	COPG2	MKLN!	TSGA13	CPA2	EXOC4	CNOT4
TBXAS1	NOBOX	GALNT15	DGKI	PDIA4	OR2FI	WDR91	SIC13A4
SSPO	ABCF2	CHRM2	KDM7A	AOC1	STRA8	PTN	KLF14
MGAM2	ABCBB	AGBL3	TRPV6	TRIM24	TMUB1	TMEM213	ATG98
CNTNAP2	OR9A2	TC26	NUP205	TCAF2	PIP	MGAM	FASTIK
(Chromosome 8)							
SNTB1	ATAD2	DERLI	FBX03	TBC1D31	FAM83A	SQLE	TMEM65
C8	WPYHV1	FER1LS	FAM9IAI	KLHL38	ANXA13	TRMT12	TATDN1
(Chromosome 11)							
CNTN5	ARHGAP42	TMEM133	PGR	TRPC6	CEP126	ANGPTL5	C11
TMEM123	DCUN1D5	DYNC2H1	DDT1				
(Chromosome 12)							
PDE3A	SPX	TBCID15	TPH2	TRHDE	IAPP	PYROXD1	GYS2
LPHB	ABCC9	ST8SIAI	C2CD5	ETNK1	LYRM5	BCAT1	C12
LRMB	RASSF8	CASC1	TSPAN11	LMNTD1	TIPR2	HLHE41	ARNTL2
SSPN	FGFR10P2	SMCO2	MANCS4	PPFIBP1	ASUN	TM7SF3	CCDC91
MED21	STK38L	REP15	FAR2	ERGC2	OVCH1	TMTC1	IPO8
CAPRIN2	SYT10	METTL20	PKP2	DDX11	BICD1	OVOS2	FAM60A
AMN1	KIAA1551	FGD4	LLPH	TMPIM4	IRAK3	HEIB	GRIP1
CAND1	DYRK2	IFNG	IL26	IL22	MDM1	RAP18	NUP107
CPM	SLC35E3	MDM2	CPSF6	RAP1B	LYZ	BEST3	CCT2
RAP3IP	LRRC10	KCNMP4	CNOT2	MYRF1	PTPRB	PTPRR	TSPAN8
RAB21	TRHDE	ZFC3H1	THAP2	TMEM19			
(Chromosome 19)							
GRAMD1A	SCN1B	HPN	FXYD3	LG14	FXYD1	FXYD7	FXUD5
FAM187B	LSR	USF2	HAMP	MAG	CD22	FFAR1	FFAR3
GPR42	KRTDAP	DMKN	SBSN	TMEM147	ATP4A	PSENEN	LIN37
HSPB6	PROSER3	ARHGAP33	PRODH2	NPHS1	TYROBP	APLP1	UPK1A
LRFN3	FFAR2	THAPS	HAUS5	KMT2B	GAPDHS	ETV2	COX6B1
KIRREL2	RBM42	SDHAF1	WDR62	ZBTB32	U2AFIL4	IGFLR1	ALKBH6
NFKBID	HCST	TBCB	COX7A1	ZNF565	CLIP3	ZNF146	
ZNF420	HKR1	ZNF850	CAPNS1	OVOL3	POLR21	ZFP14	SIPA1L3
DPF1	SPINT2	PPPIR14A	KCNK6	WDR78	C19	ACTN4	ECH1
MAP4K1	CAPN12	EIF3K	HNRNPL	RYR1	IFNL3	TTC9B	PLD3
BLVRB	PRX	SHKBP1	EGLN2	GRIK5	MAP3K10	CYP2A6	BCKDHA
DEDD2	RPS19	B3GNT8	AXL	MIA	ITRKC	CCDC79	TMEM91
B9D2	ERICH4	EXOC5	POU2F2				
(Chromosome 20)							
PRNP	PRND	SLC23A2	TMEM230	CDS2	GPCPD1	CRLS1	PROKR2
C20	GHGB	LRRN4	FERMT1	MCM8	TRMT6	TMX4	HAO1
PICBL	HA01	PICB1	PLCB4	PAK7	LAMP5	ANKEF1	SNAP25
MKKS	SLX41P	BTBD3					

Finally, Figure 7.7 presents the random effects estimates $\hat{b}$ paths for the SSCox PH model for our lung cancer dataset.

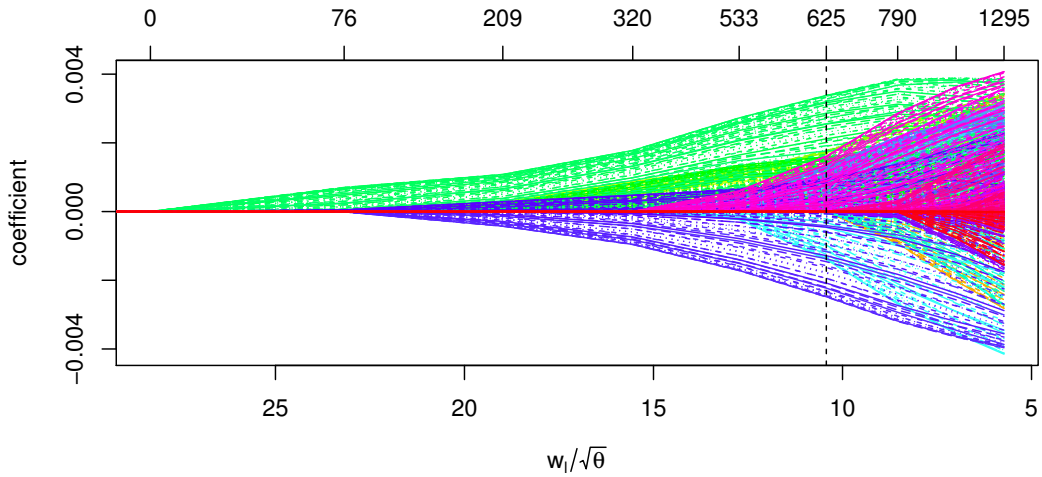


Figure 7.7: Random effects estimates $\hat{b}$ paths for the SSCox PH model for our lung cancer dataset

Each curve in Figure 7.7 corresponds to a variable (CNAs' Window). It shows the path of its coefficient against the the penalty $(w_l/\sqrt{\theta})$ of the whole coefficient vector. The top x-axis indicates the number of nonzero coefficients (active) at the current penalty $((w_l/\sqrt{\theta}))$. The vertical dotted line indicates the optimal θ that we have chosen for our lung cancer data set. In our R package users may also wish to annotate the curves; this can be done by setting (**label = TRUE**) in the plot command. We do not plot the annotation here as we have 1295 variables.

7.6.4 Cumulative hazard rate and estimates of survival function

To show that the Cox PH modelling with CNA profiles is able to distinguish individuals at different levels of risk, we estimated the survivor functions for three individuals in the lung cancer dataset. These three individuals had low, medium, and high levels of risk, based on their risk scores R_i, which corresponded respectively with the 10th, 50th, and 90th percentiles of the distribution of R_i in our dataset.

Figure 7.8 shows the estimated survivor functions for these three individuals using smooth-segmented CNA profiles as random predictors. The figure shows that the

median survival times for the low-, medium-, and high-risk individuals were approximately 7.5 years, 2.5 years, and 8 months, respectively.

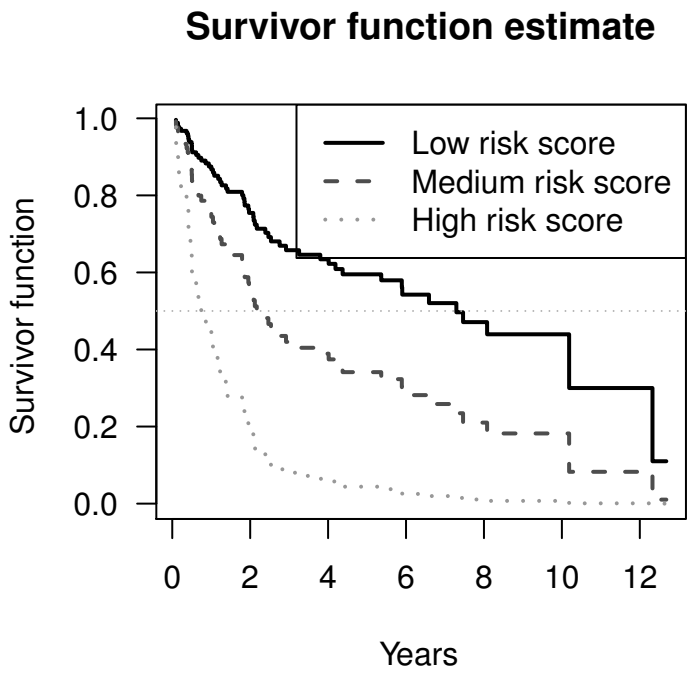


Figure 7.8: Estimated survival functions from the extended Cox PH model for three individuals in the 10th, 50th, and 90th percentiles of risk set R_i , representing low-, medium-, and high-risk individuals respectively. The horizontal dotted line marks the 50% survival probability level.

7.6.5 Model diagnostics

As part of model diagnostics, we plotted the cumulative hazard of the Cox-Snell residuals from the model fitting as shown in Figure 7.9 (solid black line). As can be seen in that figure, the cumulative hazard line is very close to the identity line, which suggests that the extended Cox PH model is suitable and has a reasonably good fit for our CNA profile data. The cumulative hazard line near the top right corner of the figure is slightly jagged, as expected, due to rare events (deaths) near the far end of the survival time distribution.

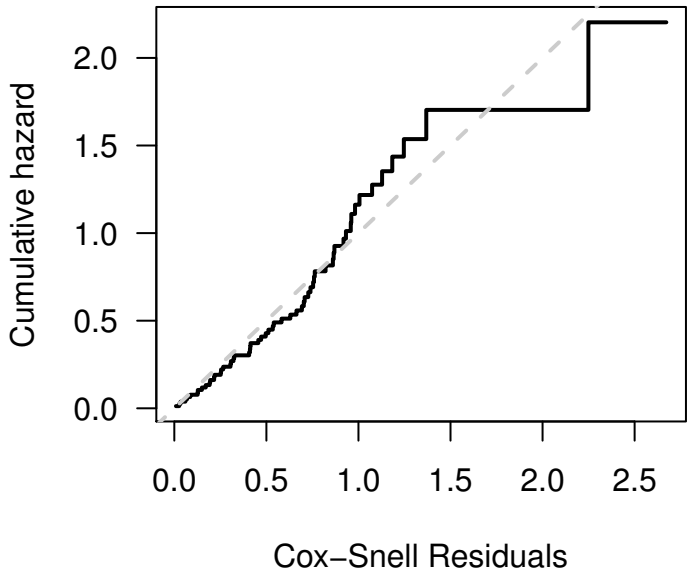


Figure 7.9: Cumulative hazard of Cox-Snell residuals (solid black line) from the Cox PH model fit, compared to the identity line (grey dashed line), based on CNA profiles

7.6.6 The assessment of the prediction performance of SSCox PH by comparing with Cox PH model (fixed effects only)

Concordance Statistics (C-statistic) are estimated to assess the prediction performance of any model using the approach proposed in Uno *et al.* (2011). A C-statistic is a measure of the concordance between an estimated risk score and the survival times (Harrell *et al.* (1996)). Let R be the the risk score calculated from a model (e.g., $R_i = X_i\hat{\beta} + Z_i\hat{b}$ for SSCox PH, and $R_i = X_i\hat{\beta}$ for Cox PH with fixed effects only.), then C-statistic is

$$Pr(R_1 > R_2 | T2 > T1),$$

which captures how well the ordering of the survival times matches the ordering of the estimated risk scores. The estimate C-statistics for Cox PH model with fixed effects only (Age+StageT+StageN) is 0.67, while the estimate C-statistic for SSCox PH is

0.72. This result shows that by including the CNA into the model (SSCox PH), we obtain a better prediction performance.

Another measure to asses the prediction performance of SSCox is receiver operating characteristic (ROC curve) and area under the curve (AUC). In survival analysis, prognostic ROC curve is a graphical approach to show the discriminative capacity of the marker: a receiver operating characteristic (ROC) curve by plotting 1 minus the survival in the high-risk group versus 1 minus the survival in the low risk group. Also, AUC represents the probability that a patient in the low- risk group has a longer lifetime than a patient in the high-risk group. To calculate the Prognostic ROC curve for the two models (Cox PH with fixed effects only and SSCox PH), we estimate the survival curve for each patient

$$\hat{S}_i = \{\hat{S_0(t)}\}^{\exp\{R_i\}}.$$

Then we divided the data into a high and low risk groups based on the risk score R_i's. There are more than one way to choose the cut off point to split the group and the result would be comparable. To be more specific, we chose the median of risk score R_i's to be the cut off point so the that the number of events is equal in the two groups. Finally, we compare the the two groups with model-based average survival function where the average is taken pointwise over time.

Figure 7.10 shows prognostic ROC curve for the two models (fixed effects only and fixed with CNA) which indicates that by including the CNA, we get better prediction accuracy. The AUC for the model with fix only is 0.70, while the AUC for the model with fix and CNA is 0.73, which also suggest that by adding CNA, we will get better prediction accuracy.

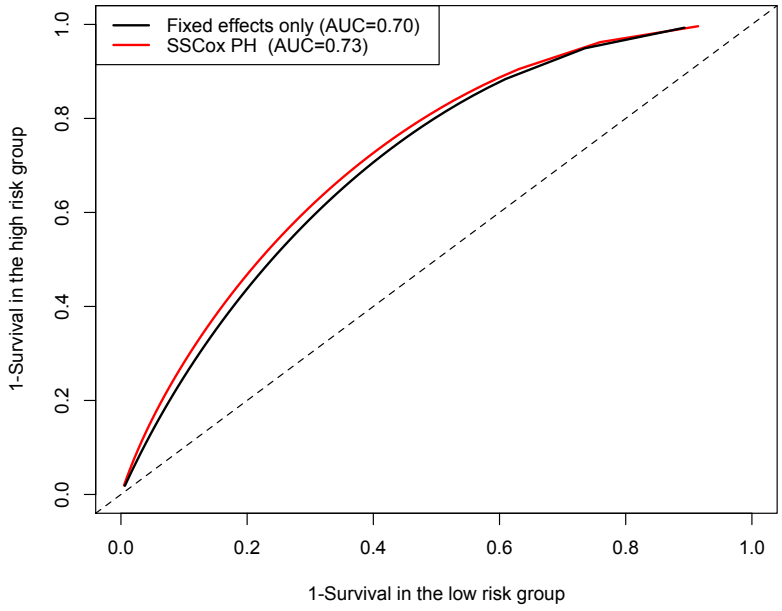


Figure 7.10: Prognostic ROC curve for the two models (fixed effects only and fixed with CNA (SSCox PH)).

7.7 Discussion

In this chapter, we have investigated how the SSCox PH model for survival data is suitable for coping with the high dimensional CNA profiles, in addition to the clinical variables as (fixed) predictors. The SSCox model allows for sparse variable selection, smoothness, and dimension reduction at the same time. A key parameter in the model is $K = (\theta, w_n, w_c, w_l)$, which controlled the amount of information from the CNA profiles that was used in the model fitting.

We have also compared our proposed method with two other methods, SPLS-LI and SPLS-HL, recently presented by Lee *et al.* (2013). Our proposed method, the SSCox model, was found to perform better than the SPLS-L1 or SPLS-HL in making predictions, as it has the lowest -2pl, MSPE, and MAPE.

Also, we compared the SSCox PH model with Cox PH model (Fixed effects only) and we found that by adding CNA, we will get better prediction accuracy. Not only

that but also we identified some relative genes associated with NSCLC. Some of these genes have been found in previous studies as shown in Table 7.4. Also we found new genes with no prior studies showing evidence of any relationship between these genes and any type of cancer. These genes need to be investigated more in a clinical trial.

Even though we used different assumption in the random effects in Chapter $5-7$, the median survival times for the low-, medium-, and high-risk individuals were similar in these different models; it was approximately 7.5 years, 2.5 years, and 8 months, respectively. This finding are consistent with clinicians views as shown in Molina *et al.* (2008).

Finally, our computational method and R package in this study can also be used for CNA profiles from an array of technology, provided that the (genome-wide) CNA profiles across individuals can be put into matrix form. This means that CNA estimates across individuals can be entered into the same column in the data matrix, for each genomic region.

To sum up, we described a new sparse-smoothed estimation procedure in Cox proportional hazard modelling that takes into account cancer patients genome-wide CNA profiles. Unlike the standard Cox PLS and Cox with ridge penalty methods, our new method automatically selects relevant variables without sacrificing prediction performance. Not only that, we also imposed smoothness to deal with the spatial structure of CNAs.

Genome-wide CNA profiles are considered as random predictors in our model, in addition to using the clinical variables as fixed predictors. We assumed CNA coefficients to be correlated random effects that followed a mixture of three distributions: normal, Cauchy (for second-order differences, to achieve smoothness), and Laplace (to achieve sparsity).

Chapter 8

Conclusion and further work

The objective of our thesis was to establish or develop statistical methods that are able to include the CNA (ultra-high dimensional data) in survival analysis. There are already established approaches to include high dimensional data in survival prediction. These approach can be classified into two classes: feature selection and derived variables (feature extraction). Even though these two approaches have been widely used to deal with high dimensional data, they suffer from major drawbacks. For the feature selection method, it is easy to implement, but it will select highly correlated features which may lead to a poor performance. On the other hand, derived variable methods do not automatically lead to selection of relevant variables. This is because they construct latent variables that are linear combination of all original covariates, so performance is expected to be reduced if a large number of covariates are in fact unrelated. Both approaches can be adapted for survival analysis to model gene expression data; however, they are not a good match for CNA data as they ignore its spatial dependence structure and do not accommodate serial correlation that exists in the CNA.

We addressed our objective by achieving the following :

- In Chapter 2, we prepared and estimated the CNA of our lung cancer data set. We end up with a matrix with dimension 89×13968, where 89 is the number of the patients and 13968 is the number of genomic windows.

- CNA can be estimated as the ratio of a tumour sample to a normal sample. Therefore, in chapter 3 we investigated the approximations of the distribution of the ratio of two Poisson random variables. In other words, if $X \sim Pois(\lambda_x)$ and

$Y \sim Pois(\lambda_y)$ with X and Y are independent; then approximations of the distribution $Z = \frac{X}{Y}$, conditional on $Y \neq 0$ is examined.

- In Chapter 4, we reviewed some concept of survival analysis, and we applied non-parametric and semi-parametric methods in the clinical part (fixed effects only) of lung cancer data set.

- In Chapter 5, we investigated an extension of the standard Cox proportional hazard model to take into account cancer patients genome-wide CNA profiles. The genome-wide CNA profiles are considered as random predictors in the model in addition to the clinical variables as fixed predictors. the random effects are assumed to be normal distribution with mean zero and diagonal structure covariance matrix which has equal variances and covariances of zero.

- In Chapter 6, we described three different estimation procedures using the Cox proportional hazard model to take into account CNAs. Unlike the extended Cox method described in the Chapter 5, these methods deal with dependencies between neighboring genomic windows and their spatial characteristics. The genome-wide CNA profiles are considered as random predictors in the model, and the clinical variables as fixed predictors. We have three different scenarios for the distribution of the random effects:

  1. Normal with mean zero and a compound symmetry covariance matrix (Coxrho).
  2. Normal with mean zero and an inverse covariance matrix (Coxinv).
  3. Correlated random effects that follow a mixture of two distributions, normal and Cauchy, for the first or second differences (SCox).

- In Chapter 7, we described a new sparse smoothed estimation procedure in Cox proportional hazard model (SSCox) to take into account cancer patients CNAs. We assumed CNA coefficients to be correlated random effect that follow a mixture of three distribution:Normal, Cauchy for the second differences to achieve smoothness, and Laplace to achieve sparsity.

In epidemiological research, it is not always the case that the response variable (outcome) is time to event data. The response can be normal or nonnormal distribution

such as the binary (Y is 1 with probability p and 0 with probability $1-p$) and the Poisson. For instance, the binary distribution is useful when the outcome of an experiment is category with 2 levels such as being classified into two different clinical groups . As a working example, suppose that we want to model tumour histological subtypes based on the patients' clinical data and their CNA profiles.

In order to deal with these different type of response, for further research, it could be interesting to adapt the idea of including the CNA as a random predictor in the generalized linear model. The only part will change is the likelihood part, instead of using the log partial likelihood, we will use the log likelihood of the model either normal, Poisson or logistic. I have started writing the code in R for the generalized linear model, however, this is beyond the scope of this thesis.

Also, In future research, one could penalize the variance of the random effects instead of penalizing the random effects itself. This setting indicates that if the variance is zero, its corresponding random effects is no longer random and is actually a constant, which can be absorbed by fixed effects. This idea was discussed in Pan & Huang (2014) for GLMM and it could be generalized to be used in survival analysis.

Appendix A

Number of ploidy for the 89 patents along with the estimated contamination and the number of reads

Patient	number of ploidy	method	contamination	number of reads
LS168	2	Mixture	71.13	1820403
LS169	2	Mixture	49.79	2017296
LS170	2	Mixture	86.49	1877104
LS171	2	Mixture	72.15	3141892
LS172	2	Mixture	74.6	1926593
LS173	2	Mixture	79.82	3217576
LS174	2	Mixture	87.27	2434153
LS182	2	Mixture	78.43	275944
LS187	2	Mixture	57.81	99541
LS188	2	Mixture	75.91	680068
LS189	2	Mixture	89.99	613568
LS192	2	Mixture	79.11	934719
LS193	2	Mixture	92.331	394880
LS194	2	Density	42.4	1054930
LS195	2	Mixture	31.17	528435
LS197	2	Mixture	86.25	1011972
LS199	2	Mixture	60.02	1159264

LS200	2	Mixture	75.85	888820
LS202	2	Mixture	79.72	496917
LS203	4	Density	34.1	1145927
LS204	2	Density	63.65	427771
LS206	2	Density	86.1	275195
LS238	2	Mixture	59.92	1084707
LS243	2	Mixture	38.18	1211597
LS244	2	Mixture	81.69	1065651
LS245	2	Mixture	51.36	1128509
LS246	2	Mixture	78.57	1069590
LS249	2	Mixture	75.21	679410
LS251	2	Mixture	82.3	616019
LS254	2	Mixture	81.6	1062136
LS255	2	Mixture	77.24	700787
LS256	2	Mixture	84.89	771187
LS257	2	Mixture	72.62	1294166
LS258	2	Mixture	86.23	1147664
LS259	2	Mixture,density	70.42	976499
LS260	2	Mixture	93.652	983376
LS262	2	Density	71.7	1008864
LS264	2	Mixture	83.34	389219
LS265	2	Mixture	76.81	98367
LS266	2	Mixture,density	61.38	541334
LS270	4	Density	67.8	1687135
LS272	2	Mixture	89.13	961265
LS273	4	Mixture	90.43	1071641
LS274	2	Mixture	83.62	300947
LS277	2	Mixture	80.92	1873814
LS281	2	Mixture	91.6	1051811
LS282	4	Mixture	92.533	1061886
LS283	2	Mixture	91.72	1602892
LS286	2	Mixture	87.55	2312598
LS287	2	Mixture	78.22	1224020
LS289	2	Mixture	82.94	468956
LS290	2	Mixture	63.78	1545591
LS291	2	Mixture	58.04	2202839
LS292	2	Mixture	81.4	778032
LS293	2	Mixture	54.49	964380

LS294	2	Mixture	83.74	1162422
LS295	2	Mixture	82.84	16563
LS296	2	Mixture	56.23	1586895
LS297	2	Mixture	88.02	587496
LS298	2	Density	68.7	1193645
LS299	2	Mixture	69.65	821277
LS300	4	Mixture	88.91	1455529
LS302	2	Mixture	78.61	325607
LS303	4	Mixture	88.12	1523161
LS304	2	density	71.04	1278349
LS306	2	Mixture	85.42	1514125
LS307	2	Mixture	85.4	1394593
LS352	2	Mixture	82.16	1001953
LS353	2	Density	63	1500452
LS354	2	Mixture	85.7	850667
LS355	2	Mixture	60.83	1991428
LS357	2	Mixture	80.08	271962
LS359	2	Mixture	74.17	1624446
LS360	2	Mixture	73.43	1479857
LS362	2	Mixture	73.43	1563365
LS364	2	Mixture	61.62	1862728
LS366	2	Density	41.4	131644
LS367	2	Mixture	72.92	1107864
LS369	2	Mixture	39.17	507639
LS370	2	Mixture	88.29	306645
LS375	2	Mixture	65.56	1252654
LS376	2	Mixture	78.75	878661
LS378	2	Mixture	83.8	718320
LS379	2	Mixture	66.54	552071
LS382	2	Mixture	78.33	1554592
LS383	2	Mixture	36.83	1429903
LS384	2	Mixture	86	418423
LS387	2	Mixture	66.1	2143304
LS388	4	Mixture	75.12	749250

Table A.1: Number of ploidy for the 89 patents along with the estimated contamination and the number of reads

Appendix B

Additional Figures for Chapter 6

Figure B.1, shows the random effect estimates b based on Coxrho, while Figure B.1 shows the random effect estimates b based on Coxinv.

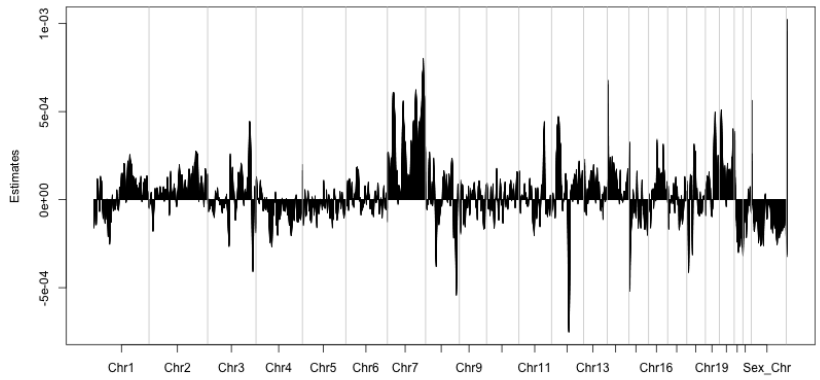


Figure B.1: Random effects estimate b in the Coxrho model, using CNA profiles. Genomic windows with missing values (e.g. in the centromere regions) were excluded from analysis, hence these are not plotted. A more detailed view of the random effects estimates in each chromosome is presented in the next figure.

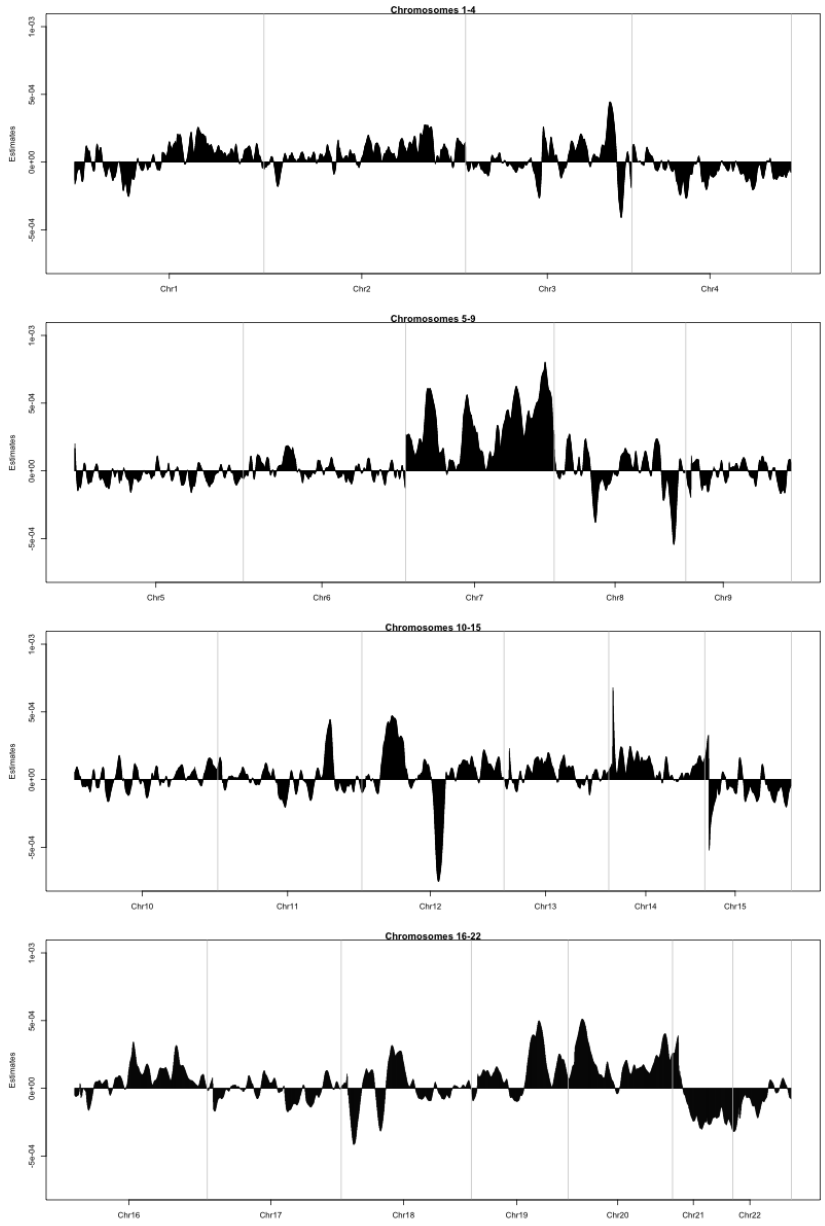


Figure B.2: Detailed views of the random effects estimates b in the Coxrho model, using CNA profiles from smooth segmentation

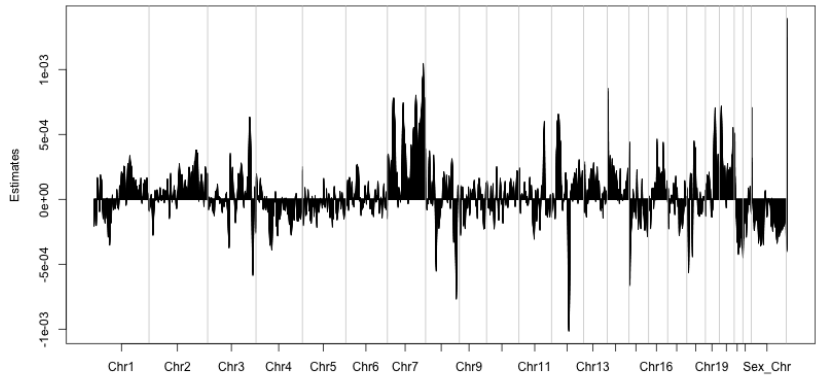


Figure B.3: Random effects estimate b in the Coxinv model, using CNA profiles. Genomic windows with missing values (e.g. in the centromere regions) were excluded from analysis, hence these are not plotted. A more detailed view of the random effects estimates in each chromosome is presented in the next figure.

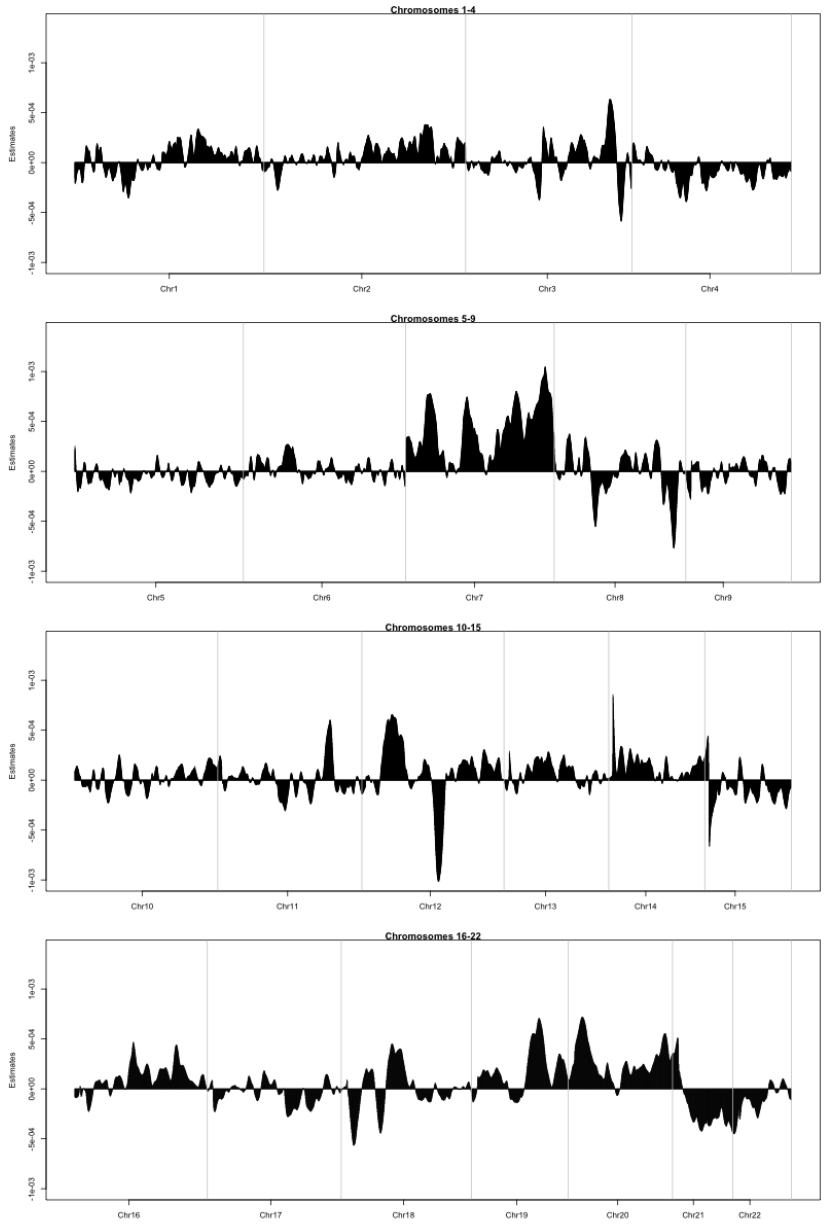


Figure B.4: Detailed views of the random effects estimates b in the Coxinv model, using CNA profiles from smooth segmentation

References

AALBERS, A.M., CALADO, R.T., YOUNG, N.S., ZWAAN, C.M., KAJIGAYA, S., BARUCHEL, A., GELEIJNS, K., HAAS, V., KASPERS, G.J., REINHARDT, D. *et al.* (2013). Absence of SBDS mutations in sporadic paediatric acute myeloid leukaemia. *British Journal of Haematology*, **160**, 559–561. 152

ABYZOV, A., URBAN, A.E., SNYDER, M. & GERSTEIN, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, **21**, 974–984. 16

ALBERTS, B. (2008). *Molecular Biology of the Cell with CD*. Garland. 3

ALMAMI, A., HEGAZY, S.A., NABBI, A., ALSHALALFA, M., SALMAN, A., ABOU-OUF, H., RIABOWOL, K. & BISMAR, T.A. (2016). ING3 is associated with increased cell invasion and lethal outcome in ERG-negative prostate cancer patients. *Tumor Biology*, 1–8. 152

ANTONIADIS, A., FRYZLEWICZ, P. & LETUÉ, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics*, **37**, 531–552. 2

ANTONIOU, K., SAMARA, K., LASITHIOTAKI, I., MARGARITOPOULOS, G., SOUFLA, G., LAMBIRI, I., GIANNARAKIS, I., DROSITIS, I., SPANDIDOS, D. & SIAFAKAS, N. (2013). Differential telomerase expression in idiopathic pulmonary fibrosis and non-small cell lung cancer. *Oncology Reports*, **30**, 2617–2624. 150, 151

AVASARALA, S., VAN SCOYK, M., RATHINAM, M.K.K., ZERAYESUS, S., ZHAO, X., ZHANG, W., PERGANDE, M.R., BORGIA, J.A., DEGREGORI, J., PORT, J.D. *et al.* (2015). PRMT1 is a novel regulator of epithelial-mesenchymal-transition in

non-small cell lung cancer. *Journal of Biological Chemistry*, **290**, 13479–13489. 151

BAGLAMA, J. & REICHEL, L. (????). irlba: Fast truncated svd, pca and symmetric eigendecomposition for large dense and sparse matrices, 2015. *R package version*, **2**. 79

BAI, J., YONG, H., CHEN, F., MEI, P., LIU, H., LI, C., PAN, Z., WU, Y. & ZHENG, J. (2013). Cullin1 is a novel marker of poor prognosis and a potential therapeutic target in human breast cancer. *Annals of Oncology*, **24**, 2016–2022. 152

BAIR, E., HASTIE, T., PAUL, D. & TIBSHIRANI, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, **101**. 101

BEAUCHEMIN, N. & ARABZADEH, A. (2013). Carcinoembryonic antigen-related cell adhesion molecules (CEACAMs) in cancer progression and metastasis. *Cancer and Metastasis Reviews*, **32**, 643–671. 151

BELVEDERE, O., BERRI, S., CHALKLEY, R., CONWAY, C., BARBONE, F., PISA, F., MACLENNAN, K., DALY, C., ALSOP, M., MORGAN, J. *et al.* (2012). A computational index derived from whole-genome copy number analysis is a novel tool for prognosis in early stage lung squamous cell carcinoma. *Genomics*, **99**, 18–24. 8, 17, 18

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300. 5, 101

BI, H.X., SHI, H.B., ZHANG, T. & CUI, G. (2015). PRDM14 promotes the migration of human non-small cell lung cancer through extracellular matrix degradation in vitro. *Chinese Medical Journal*, **128**, 373. 151

BOEVA, V., ZINOVYEV, A., BLEAKLEY, K., VERT, J.P., JANOUEIX-LEROSEY, I., DELATTRE, O. & BARILLOT, E. (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, **27**, 268–269. 16, 20

BOOTH, J.G. & HOBERT, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**, 265–285. 99

BOUKAKIS, G., PATRINOU-GEORGOULA, M., LEKARAKOU, M., VALAVANIS, C. & GUIALIS, A. (2010). Deregulated expression of hnRNP A/B proteins in human non-small cell lung cancer: parallel assessment of protein and mRNA levels in paired tumour/non-tumour tissues. *BMC Cancer*, **10**, 1. 151

BØVELSTAD, H.M., NYGÅRD, S., STØRVOLD, H.L., ALDRIN, M., BORGAN, Ø., FRIGESSI, A. & LINGJÆRDE, O.C. (2007). Predicting survival from microarray data comparative study. *Bioinformatics*, **23**, 2080–2087. 5, 15, 138

BRESLOW, N. (1974). Covariance analysis of censored survival data. *Biometrics*, **30**, 89. 64, 84

BRESLOW, N.E. & CLAYTON, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, **88**, 9–25. 75, 98, 104, 105, 108

BRESLOW, N.E. & LIN, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 81–91. 98

BUCKINGHAM, L.E., COON, J.S., MORRISON, L.E., JACOBSON, K.K., JEWELL, S.S., KAISER, K.A., MAUER, A.M., MUZZAFAR, T., POLOWY, C., BASU, S. *et al.* (2007). The prognostic value of chromosome 7 polysomy in non-small cell lung cancer patients treated with gefitinib. *Journal of Thoracic Oncology*, **2**, 414–422. 150

BULK, E., YU, J., HASCHER, A., KOSCHMIEDER, S., WIEWRODT, R., KRUG, U., TIMMERMANN, B., MARRA, A., HILLEJAN, L., WIEBE, K. *et al.* (2012). Mutations of the EPHB6 receptor tyrosine kinase induce a pro-metastatic phenotype in non-small cell lung cancer. *PloS One*, **7**, 44591–44599. 151

BUTKIEWICZ, D., RUSIN, M., SIKORA, B., LACH, A. & CHOAZY, M. (2011). An association between DNA repair gene polymorphisms and survival in patients with

resected non-small cell lung cancer. *Molecular Biology Reports*, **38**, 5231–5241. 151

CHAMBON, M., ORSETTI, B., BERTHE, M.L., BASCOUL-MOLLEVI, C., RODRIGUEZ, C., DUONG, V., GLEIZES, M., THENOT, S., BIBEAU, F., THEILLET, C. *et al.* (2011). Prognostic significance of TRIM24/TIF-1α gene expression in breast cancer. *The American Journal of Pathology*, **178**, 1461–1469. 152

CHEN, B., TAN, Z., GAO, J., WU, W., LIU, L., JIN, W., CAO, Y., ZHAO, S., ZHANG, W., QIU, Z. *et al.* (2015a). Hyperphosphorylation of ribosomal protein S6 predicts unfavorable clinical survival in non-small cell lung cancer. *Journal of Experimental & Clinical Cancer Research*, **34**, 1. 151

CHEN, P.N. (2002). Asymptotic refinement of the Berry-Esseen constant. *Unpublished Manuscript*. 35

CHEN, Y., FEI, Y. & PAN, J. (2015b). Quasi-monte carlo estimation in generalized linear mixed model with correlated random effects. *Open Access Library Journal*, **2**, 1. 98

CHEN, Y.C., LIU, T., YU, C.H., CHIANG, T.Y. & HWANG, C.C. (2013). Effects of GC bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PloS One*, **8**, e62856. 20

CHEN, Y.M., LAI, C.H., CHANG, H.C., CHAO, T.Y., TSENG, C.C., FANG, W.F., WANG, C.C., CHUNG, Y.H., HUANG, K.T., CHEN, H.C. *et al.* (2015c). Baseline, Trend, and Normalization of Carcinoembryonic Antigen as Prognostic Factors in Epidermal Growth Factor Receptor-Mutant Nonsmall Cell Lung Cancer Patients Treated with First-Line Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitors. *Medicine*, **94**, e2239. 151

CHENG, C., CUI, H., ZHANG, L., JIA, Z., SONG, B., WANG, F., LI, Y., LIU, J., KONG, P., SHI, R. *et al.* (2016). Genomic analyses reveal FAM84B and the NOTCH pathway are associated with the progression of esophageal squamous cell carcinoma. *GigaScience*, **5**, 1. 150

CHIU, C.G., NAKAMURA, Y., CHONG, K.K., HUANG, S.K., KAWAS, N.P., TRICHE, T., ELASHOFF, D., KIYOHARA, E., IRIE, R.F., MORTON, D.L. *et al.* (2014). Genome-wide characterization of circulating tumor cells identifies novel prognostic genomic alterations in systemic melanoma metastasis. *Clinical Chemistry*, **60**, 873–885. 73

CHOI, H.S., LEE, Y., PARK, K.H., SUNG, J.S., LEE, J.E., SHIN, E.S., RYU, J.S. & KIM, Y.H. (2009). Single-nucleotide polymorphisms in the promoter of the CDK5 gene and lung cancer risk in a Korean population. *Journal of Human Genetics*, **54**, 298–303. 151

CHOWDHURY, M.R.I. & SUTRADHAR, B. (2009). Generalized quasi-likelihood versus hierarchical likelihood inferences in generalized linear mixed models for count data. *Sankhyā: The Indian Journal of Statistics, Series B (2008-)*, 55–78. 99

CHUN, H. & KELEŞ, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 3–25. 142

COLLETT, D. (2003). *Modelling survival data in medical research*. CRC press. 46

COX, D.R. (1975). Partial likelihood. *Biometrika*, **62**, 269–276. 60, 63

COX, D.R. & SNELL, E.J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 248–275. 67, 69

COX, D.R. *et al.* (1972). Regression models and life tables. *Journal of the Royal Statistical Society. Series B*, **34**, 187–220. 1, 60, 61, 74, 77, 106, 110, 132

DOMBLIDES, C., CORTOT, A. & WISLEZ, M. (2015). MET exon 14 mutation, new target in lung sarcomatoid carcinoma. *Bulletin du Cancer*, **102**, 966. 151

DUAN, J., ZHANG, J.G., DENG, H.W. & WANG, Y.P. (2013). Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PloS One*, **8**, e59128. 16

ENGLER, D.A. & LI, Y. (2007). Survival analysis with large dimensional covariates: an application in microarray studies. *Harvard University Biostatistics Working Paper Series*. 6, 141

FAN, J. & LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, 74–99. 2

FAN, J., FENG, Y., WU, Y. *et al.* (2010). High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, 70–86, Institute of Mathematical Statistics. 2

FELLER, W. (2008). *An Introduction to Probability Theory and its Applications*, vol. 2. John Wiley & Sons. 36

FIORI, V., MAGNANI, M. & CIANFRIGLIA, M. (2012). The expression and modulation of CEACAM1 and tumor cell transformation. *Annali Dell'Istituto Superiore Di SanitÃ*, **48**, 161–171. 151

FLACCO, A., LUDOVINI, V., BIANCONI, F., RAGUSA, M., BELLEZZA, G., TOFANETTI, F.R., PISTOLA, L., SIGGILLINO, A., VANNUCCI, J., CAGINI, L. *et al.* (2015). MYC and Human Telomerase gene (TERC) Copy Number Gain in Early-stage Non–small Cell lung Cancer. *American Journal of Clinical Oncology*, **38**, 152–158. 150, 151

FREEMAN, J.L., PERRY, G.H., FEUK, L., REDON, R., MCCARROLL, S.A., ALTSHULER, D.M., ABURATANI, H., JONES, K.W., TYLER-SMITH, C., HURLES, M.E. *et al.* (2006). Copy number variation: new insights in genome diversity. *Genome Research*, **16**, 949–961. 16

FU, X., ZHANG, W., SU, Y., LU, L., WANG, D. & WANG, H. (2016). MicroRNA-103 suppresses tumor cell proliferation by targeting PDCD10 in prostate cancer. *The Prostate*. 152

GAN, H., LIU, H., ZHANG, H., LI, Y., XU, X., XU, X. & XU, J. (2016). SHh-Gli1 signaling pathway promotes cell survival by mediating baculoviral IAP repeat-containing 3 (BIRC3) gene in pancreatic cancer cells. *Tumor Biology*, 1–8. 152

GATZA, M.L., SILVA, G.O., PARKER, J.S., FAN, C. & PEROU, C.M. (2014). An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nature Genetics*, **46**, 1051–1059. 73

GOEMAN, J.J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, **52**, 70–84. 11, 14, 131, 135, 136, 137

GRAY, R.J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, **87**, 942–951. 82, 111

GREEN, P.J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, 245–259. 77

GREENWOOD, M. *et al.* (1926). A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer.*. 52

GUO, J., GAO, J., LI, Z., GONG, Y., MAN, X., JIN, J. & WU, H. (2013). Adenovirus vector-mediated Gli1 siRNA induces growth inhibition and apoptosis in human pancreatic cancer with Smo-dependent or Smo-independent Hh pathway activation in vitro and in vivo. *Cancer Letters*, **339**, 185–194. 152

GUSNANTO, A., WOOD, H.M., PAWITAN, Y., RABBITTS, P. & BERRI, S. (2012). Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics*, **28**, 40–47. 12, 16, 17, 18, 20, 22, 23

GUSNANTO, A., TAYLOR, C.C., NAFISAH, I., WOOD, H.M., RABBITTS, P. & BERRI, S. (2014). Estimating optimal window size for analysis of low-coverage next-generation sequence data. *Bioinformatics*, **30**, 1823–1829. 18

GYHORFFY, B., SUROWIAK, P., BUDCZIES, J. & LANCZKY, A. (2013). Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PloS One*, **8**, e82241. 151

HAAN, J.C., LABOTS, M., RAUSCH, C., KOOPMAN, M., TOL, J., MEKENKAMP, L.J., VAN DE WIEL, M.A., ISRAELI, D., VAN ESSEN, H.F., VAN GRIEKEN, N.C. *et al.* (2014). Genomic landscape of metastatic colorectal cancer. *Nature Communications*, **5**. 73

HAN, H.S., SON, S.M., YUN, J., JO, Y.N. & LEE, O.J. (2014). MicroRNA-29a suppresses the growth, migration, and invasion of lung adenocarcinoma cells by targeting carcinoembryonic antigen-related cell adhesion molecule 6. *FEBS Letters*, **588**, 3744–3750. 151

HARRELL, F.E., LEE, K.L. & MARK, D.B. (1996). Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361–387. 156

HE, P., WU, W., WANG, H., LIAO, K., ZHANG, W., XIONG, G., WU, F., MENG, G. & YANG, K. (2013). Co-expression of Rho guanine nucleotide exchange factor 5 and Src associates with poor prognosis of patients with resected non-small cell lung cancer. *Oncology Reports*, **30**, 2864–2870. 151

HEAD, J.D. & ZERNER, M.C. (1985). A Broyden, Fletcher, Goldfarb, Shanno optimization procedure for molecular geometries. *Chemical Physics Letters*, **122**, 264–270. 42, 135

HENSON, J., TISCHLER, G. & NING, Z. (2012). Next-generation sequencing and large genome assemblies. *Pharmacogenomics*, **13**, 901–915. 3

HINKLEY, D.V. (1969). On the ratio of two correlated normal random variables. *Biometrika*, **56**, 635–639. 36, 38

HIROFUMINAKANISHI, T.S., KATOH, M., WATANABE, A., IGISHI, T., MASAHIROKODANI, S., NAKAMOTO, M. & SHIGEOKA, Y. (2004). Loss of imprinting of PEG1/MEST in lung cancer cell lines. *Oncology Reports*, **12**, 1273–1278. 151

HOOI, C.F., BLANCHER, C., QIU, W., REVET, I., WILLIAMS, L., CIAVARELLA, M., ANDERSON, R., THOMPSON, E., CONNOR, A., PHILLIPS, W. *et al.* (2006). ST7-mediated suppression of tumorigenicity of prostate cancer cells is characterized by remodeling of the extracellular matrix. *Oncogene*, **25**, 3924–3933. 152

HUANG, J., GUSNANTO, A., O'SULLIVAN, K., STAAF, J., BORG, Å. & PAWITAN, Y. (2007). Robust smooth segmentation approach for array CGH data analysis. *Bioinformatics*, **23**, 2463–2469. 19, 20, 21

HUANG, J., SALIM, A., LEI, K., O'SULLIVAN, K. & PAWITAN, Y. (2009). Classification of array CGH data using smoothed logistic regression model. *Statistics in Medicine*, **28**, 3798–3810. 10, 101, 104, 108, 133

HUANG, J., SUN, C., WANG, S., HE, Q. & LI, D. (2015a). microRNA miR-10b inhibition reduces cell proliferation and promotes apoptosis in non-small cell lung cancer (NSCLC) cells. *Molecular BioSystems*, **11**, 2051–2059. 151

HUANG, O., WU, D., XIE, F., LIN, L., WANG, X., JIANG, M., LI, Y., CHEN, W., SHEN, K. & HU, X. (2015b). Targeting rho guanine nucleotide exchange factor ARHGEF5/TIM with auto-inhibitory peptides in human breast cancer. *Amino Acids*, **47**, 1239–1246. 152

HUANG, W., JIN, Y., YUAN, Y., BAI, C., WU, Y., ZHU, H. & LU, S. (2013). Validation and target gene screening of hsa-miR-205 in lung squamous cell carcinoma. *Chinese Medical Journal*, **127**, 272–278. 151

HWANG, J.A., LEE, B.B., KIM, Y., PARK, S.E., HEO, K., HONG, S.H., KIM, Y.H., HAN, J., SHIM, Y.M., LEE, Y.S. *et al.* (2013). HOXA11 hypermethylation is associated with progression of non-small cell lung cancer. *Oncotarget*, **4**, 2317. 151

JANG, S.M., AN, J.H., KIM, C.H., KIM, J.W. & CHOI, K.H. (2015). Transcription factor FOXA2-centered transcriptional regulation network in non-small cell lung cancer. *Biochemical and Biophysical Research Communications*, **463**, 961–967. 151

JUSTILIEN, V., JAMEISON, L., DER, C.J., ROSSMAN, K.L. & FIELDS, A.P. (2011). Oncogenic activity of Ect2 is regulated through protein kinase cι-mediated phosphorylation. *Journal of Biological Chemistry*, **286**, 8149–8157. 151

KANG, J.U. (2013). Characterization of amplification patterns and target genes on the short arm of chromosome 7 in early-stage lung adenocarcinoma. *Molecular Medicine Reports*, **8**, 1373–1378. 151

KANG, M., LEE, E., YOON, S., JO, J., LEE, J., KIM, H., CHOI, Y., KIM, K., SHIM, Y., KIM, J. *et al.* (2011). AKR1B10 is associated with smoking and smoking-related non-small-cell lung cancer. *Journal of International Medical Research*, **39**, 78–85. 151

KARIM, M.R. & ZEGER, S.L. (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics*, 631–644. 99

KETTUNEN, E., ANTTILA, S., SEPPÄNEN, J.K., KARJALAINEN, A., EDGREN, H., LINDSTRÖM, I., SALOVAARA, R., NISSÉN, A.M., SALO, J., MATTSON, K. *et al.* (2004). Differentially expressed genes in nonsmall cell lung cancer: expression profiling of cancer-related genes in squamous cell lung cancer. *Cancer Genetics and Cytogenetics*, **149**, 98–106. 151

KITADA, K. & YAMASAKI, T. (2008). The complicated copy number alterations in chromosome 7 of a lung cancer cell line is explained by a model based on repeated breakage-fusion-bridge cycles. *Cancer Genetics and Cytogenetics*, **185**, 11–19. 150

KLAMBAUER, G., SCHWARZBAUER, K., MAYR, A., CLEVERT, D.A., MITTERECKER, A., BODENHOFER, U. & HOCHREITER, S. (2012). cn. MOPS: mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research*, gks003. 16

LAGARIAS, J.C., REEDS, J.A., WRIGHT, M.H. & WRIGHT, P.E. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on Optimization*, **9**, 112–147. 135

LAL, S., SUTIMAN, N., OOI, L., WONG, Z., WONG, N., ANG, P. & CHOWBAY, B. (2016). Pharmacogenetics of ABCB5, ABCC5 and RLIP76 and doxorubicin pharmacokinetics in Asian breast cancer patients. *The Pharmacogenomics Journal*. 152

LAZAR, V., SUO, C., OREAR, C., VAN DEN OORD, J., BALOGH, Z., GUEGAN, J., JOB, B., MEURICE, G., RIPOCHE, H., CALZA, S. *et al.* (2013). Integrated molecular portrait of non-small cell lung cancers. *BMC Medical Genomics*, **6**, 1. 151

LEE, D., LEE, W., LEE, Y. & PAWITAN, Y. (2011). Sparse partial least-squares regression and its applications to high-throughput data analysis. *Chemometrics and Intelligent Laboratory Systems*, **109**, 1–8. 10

LEE, D., LEE, Y., PAWITAN, Y. & LEE, W. (2013). Sparse partial least-squares regression for high-throughput survival data analysis. *Statistics in Medicine*, **32**, 5340–5352. 7, 15, 101, 138, 139, 141, 142, 158

LEE, J.C., LEE, W.H., MIN, Y.J., CHA, H.J., HAN, M.W., CHANG, H.W., KIM, S.A., CHOI, S.H., KIM, S.W. & KIM, S.Y. (2014). Development of TRAIL resistance by radiation-induced hypermethylation of DR4 CpG island in recurrent laryngeal squamous cell carcinoma. *International Journal of Radiation Oncology, Biology, Physics*, **88**, 1203–1211. 151

LEE, J.S., PATHAK, S., HOPWOOD, V., TOMASOVIC, B., MULLINS, T.D., BAKER, F.L., SPITZER, G. & NEIDHART, J.A. (1987). Involvement of chromosome 7 in primary lung tumor and nonmalignant normal lung tissue. *Cancer Research*, **47**, 6349–6352. 150

LEE, W., BELKHIRI, A., LOCKHART, A.C., MERCHANT, N., GLAESER, H., HARRIS, E.I., WASHINGTON, M.K., BRUNT, E.M., ZAIKA, A., KIM, R.B. *et al.* (2008). Overexpression of OATP1B3 confers apoptotic resistance in colon cancer. *Cancer Research*, **68**, 10315–10323. 150

LEE, Y. & NELDER, J.A. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 987–1006. 99

LEE, Y., NELDER, J.A. & PAWITAN, Y. (2006). *Generalized linear models with random effects: unified analysis via H-likelihood*. CRC Press. 99

LESCH, S.M. & JESKE, D.R. (2009). Some suggestions for teaching about normal approximations to poisson and binomial distribution functions. *The American Statistician*, **63**, 274–277. 34

LI, H., HUANG, K., GAO, L., WANG, L., NIU, Y., LIU, H., WANG, Z., WANG, L., WANG, G. & WANG, J. (2016). TES inhibits colorectal cancer progression through activation of p38. *Oncotarget*. 152

LI, L.D., SUN, H.F., LIU, X.X., GAO, S.P., JIANG, H.L., HU, X. & JIN, W. (2015). Down-Regulation of NDUFB9 Promotes Breast cancer Cell Proliferation, Metastasis by Mediating Mitochondrial Metabolism. *PloS One*, **10**, e0144441. 150

LOCKWOOD, W., CHARI, R., COE, B., GIRARD, L., MACAULAY, C., LAM, S., GAZDAR, A., MINNA, J. & LAM, W. (2008). DNA amplification is a ubiquitous mechanism of oncogene activation in lung and other cancers. *Oncogene*, **27**, 4615–4624. 151

LOPEZ-AYLLON, B.D., DE CASTRO-CARPENO, J., RODRIGUEZ, C., PERNA, O., DE CACERES, I.I., BELDA-INIESTA, C., PERONA, R. & SASTRE, L. (2015). Biomarkers of erlotinib response in non-small cell lung cancer tumors that do not harbor the more common epidermal growth factor receptor mutations. *International Journal of Clinical and Experimental Pathology*, **8**, 2888. 151

LU, T.X., FAN, L., WANG, L., WU, J.Z., MIAO, K.R., LIANG, J.H., GONG, Q.X., WANG, Z., YOUNG, K.H., XU, W. *et al.* (2015). MYC or BCL2 copy number aberration is a strong predictor of outcome in patients with diffuse large B-cell lymphoma. *Oncotarget*, **6**, 18374. 73

LUO, A., YIN, Y., LI, X., XU, H., MEI, Q. & FENG, D. (2015). The clinical significance of FSCN1 in non-small cell lung cancer. *Biomedicine and Pharmacotherapy*, **73**, 75–79. 151

MAGI, A., BENELLI, M., YOON, S., ROVIELLO, F. & TORRICELLI, F. (2011). Detecting common copy number variants in high-throughput sequencing data by using jointSLM algorithm. *Nucleic Acids Research*, **39**, e65–e65. 16

MAK, P.Y., MAK, D.H., RUVOLO, V., JACAMO, R., KORNBLAU, S.M., KANTARJIAN, H., ANDREEFF, M. & CARTER, B.Z. (2014). Apoptosis repressor with caspase recruitment domain modulates second mitochondrial-derived activator of caspases mimetic-induced cell death through BIRC2/MAP3K14 signalling in acute myeloid leukaemia. *British Journal of Haematology*, **167**, 376–384. 152

MAKABE, H. & MORIMURA, H. (1955). A normal approximation to Poisson distribution. *Rep. Stat. Appl. Res., JUSE*, **4**, 37–46. 34

MAMPAEY, E., FIEUW, A., VAN LAETHEM, T., FERDINANDE, L., CLAES, K., CEELEN, W., VAN NIEUWENHOVE, Y., PATTYN, P., DE MAN, M., DE RUYCK, K. *et al.* (2015). Focus on 16p13. 3 Locus in Colon Cancer. *PloS One*, **10**, e0131421. 73

MANTEL, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, **58**, 690–700. 52

MASSY, W.F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, **60**, 234–256. 7

MILLER, C.A., HAMPTON, O., COARFA, C. & MILOSAVLJEVIC, A. (2011). ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PloS One*, **6**, e16327. 16

MOESCHBERGER, M.L. & KLEIN, J. (2003). *Survival analysis: Techniques for censored and truncated data: Statistics for Biology and Health*. Springer. 46

MOLINA, J.R., YANG, P., CASSIVI, S.D., SCHILD, S.E. & ADJEI, A.A. (2008). Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. In *Mayo Clinic Proceedings*, vol. 83, 584–594, Elsevier. 159

MØLLER, M.F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, **6**, 525–533. 135

MOODY, T., WALTERS, J., CASIBANG, M., ZIA, F. & GOZES, Y. (2000). VPAC1 receptors and lung cancer. *Annals of the New York Academy of Sciences*, **921**, 26–32. 151

MOTEVALLI, A., YASAEI, H., VIRMOUNI, S.A., SLIJEPCEVIC, P. & ROBERTS, T. (2014). The effect of chemotherapeutic agents on telomere length maintenance in breast cancer cell lines. *Breast Cancer Research and Treatment*, **145**, 581–591. 152

MUPPANI, N., NYMAN, U. & JOSEPH, B. (2011). TAp73alpha protects small cell lung carcinoma cells from caspase-2 induced mitochondrial mediated apoptotic cell death. *Oncotarget*, **2**, 1145–1154. 151

National Institutes of Health (2016). National human genome research institute. http://www.genome.gov/10000002/education. 3

NEWCOMBE, P., ALI, H.R., BLOWS, F., PROVENZANO, E., PHAROAH, P., CALDAS, C. & RICHARDSON, S. (2014). Weibull regression with bayesian variable selection to identify prognostic tumour markers of breast cancer survival. *Statistical methods in medical research*, 0962280214548748. 99

NGUYEN, D.V. & ROCKE, D.M. (2002). Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, **18**, 1625–1632. 7

NYGÅRD, S., BORGAN, Ø., LINGJÆRDE, O.C. & STØRVOLD, H.L. (2008). Partial least squares Cox regression for genome-wide data. *Lifetime Data Analysis*, **14**, 179–195. 15, 138, 142

OLSHEN, A.B., VENKATRAMAN, E., LUCITO, R. & WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572. 19

PALANCA-WESSELS, M.C.A., CZUCZMAN, M., SALLES, G., ASSOULINE, S., SEHN, L.H., FLINN, I., PATEL, M.R., SANGHA, R., HAGENBEEK, A., ADVANI, R. *et al.* (2015). Safety and activity of the anti-CD79B antibody–drug conjugate polatuzumab vedotin in relapsed or refractory b-cell non-hodgkin lymphoma and chronic lymphocytic leukaemia: a phase 1 study. *The Lancet Oncology*, **16**, 704–715. 152

PAN, J. & HUANG, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing*, **24**, 725–738. 98, 162

PAN, J. & THOMPSON, R. (2003). Gauss-hermite quadrature approximation for estimation in generalised linear mixed models. *Computational Statistics*, **18**, 57–78. 99

PAN, J. & THOMPSON, R. (2007). Quasi-monte carlo estimation in generalized linear mixed models. *Computational Statistics & Data Analysis*, **51**, 5765–5775. 99

PARK, P.J., TIAN, L. & KOHANE, I.S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**, 120–127. 7

PAWITAN, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press. 79

PAWITAN, Y. (2013). *In all Likelihood: Statistical Modelling and Inference using Likelihood*. OUP Oxford. 104, 107, 133

PAWITAN, Y., BJÖHLE, J., WEDREN, S., HUMPHREYS, K., SKOOG, L., HUANG, F., AMLER, L., SHAW, P., HALL, P. & BERGH, J. (2004). Gene expression profiling for prognosis using Cox regression. *Statistics in Medicine*, **23**, 1767–1780. 84, 89

PELOSI, G., DEL CURTO, B., TRUBIA, M., NICHOLSON, A.G., MANZOTTI, M., VERONESI, G., SPAGGIARI, L., MAISONNEUVE, P., PASINI, F., TERZI, A. *et al.* (2007). 3q26 amplification and polysomy of chromosome 3 in squamous cell lesions of the lung: a fluorescence in situ hybridization study. *Clinical CancerResearch*, **13**, 1995–2004. 150, 151

PEREZ-JANICES, N., BLANCO-LUQUIN, I., TORREA, N., LIECHTENSTEIN, T., ESCORS, D., CORDOBA, A., VICENTE-GARCIA, F., JAUREGUI, I., DE LA CRUZ, S., ILLARRAMENDI, J.J. *et al.* (2015). Differential involvement of RASSF2 hypermethylation in breast cancer subtypes and their prognosis. *Oncotarget*, **6**, 23944. 152

PETO, R. & PETO, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, 185–207. 49

PIKOR, L.A., LOCKWOOD, W.W., THU, K.L., VUCIC, E.A., CHARI, R., GAZDAR, A.F., LAM, S. & LAM, W.L. (2013). YEATS4 is a novel oncogene amplified in non–small cell lung cancer that regulates the p53 pathway. *Cancer Research*, **73**, 7301–7312. 150

REDON, R., ISHIKAWA, S., FITCH, K.R., FEUK, L., PERRY, G.H., ANDREWS, T.D., FIEGLER, H., SHAPERO, M.H., CARSON, A.R., CHEN, W. *et al.* (2006). Global variation in copy number in the human genome. *Nature*, **444**, 444–454. 16

RIPATTI, S. & PALMGREN, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, **56**, 1016–1022. 75, 82, 83, 86

SCHEMIONEK, M., HERRMANN, O., REHER, M.M., CHATAIN, N., SCHUBERT, C., COSTA, I., HAENZELMANN, S., GUSMAO, E., KINTSLER, S., BRAUNSCHWEIG, T. *et al.* (2016). MTSS1 is a critical epigenetically regulated tumor suppressor in CML. *Leukemia*, **30**, 823–832. 150

SELAMAT, S.A., CHUNG, B.S., GIRARD, L., ZHANG, W., ZHANG, Y., CAMPAN, M., SIEGMUND, K.D., KOSS, M.N., HAGEN, J.A., LAM, W.L. *et al.* (2012). Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Research*, **22**, 1197–1211. 151

SERENO, M., MORENO, V., RUBIO, J.M., GOMEZ-RAPOSO, C., SANCHEZ, S.G., JUSDADO, R.H., FALAGAN, S., TEBAR, F.Z. & SAENZ, E.C. (2015). A significant response to sorafenib in a woman with advanced lung adenocarcinoma and a BRAF non-V600 mutation. *Anti-cancer Drugs*, **26**, 1004–1007. 151

SIMON, N., FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R. *et al.* (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, **39**, 1–13. 122, 145

SON, J.W., KIM, Y.J., CHO, H.M., LEE, S.Y., LEE, S.M., KANG, J.K., LEE, J.U., LEE, Y.M., KWON, S.J., CHOI, E. *et al.* (2011). Promoter hypermethylation of the CFTR gene and clinical/pathological features associated with non-small cell lung cancer. *Respirology*, **16**, 1203–1209. 151

SULLIVAN, I., SALAZAR, J., MAJEM, M., PALLARES, C., DEL RIO, E., PAEZ, D., BAIGET, M. & BARNADAS, A. (2014). Pharmacogenetics of the DNA repair pathways in advanced non-small cell lung cancer patients treated with platinum-based chemotherapy. *Cancer Letters*, **353**, 160–166. 151

SUTRADHAR, B.C. (2004). On exact quasilikelihood inference in generalized linear mixed models. *Sankhyā: The Indian Journal of Statistics*, 263–291. 99

TAN, X. & CHEN, M. (2014). MYLK and MYL9 expression in non-small cell lung cancer identified by bioinformatics analysis of public expression data. *Tumor Biology*, **35**, 12189–12200. 151

THERNEAU, T.M., GRAMBSCH, P.M. & FLEMING, T.R. (1990). Martingale-based residuals for survival models. *Biometrika*, **77**, 147–160. 69

THERNEAU, T.M., GRAMBSCH, P.M. & PANKRATZ, V.S. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, **12**, 156–175. 79

TIAN, Z.Q., LI, Z.H., WEN, S.W., ZHANG, Y.F., LI, Y., CHENG, J.G. & WANG, G.Y. (2015). Identification of commonly dysregulated genes in non-small-cell lung cancer by integrated analysis of microarray data and qRT-PCR validation. *Lung*, **193**, 583–592. 151

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. & KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 91–108. 108

TIBSHIRANI, R. *et al.* (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385–395. 2, 6, 140

TURACLI, I.D., OZKAN, A.C. & EKMEKCI, A. (2015). The comparison between dual inhibition of mTOR with MAPK and PI3K signaling pathways in KRAS mutant NSCLC cell lines. *Tumor Biology*, **36**, 9339–9345. 151

UNO, H., CAI, T., PENCINA, M.J., D'AGOSTINO, R.B. & WEI, L. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, **30**, 1105–1117. 156

VALENCIA, T., JOSEPH, A., KACHROO, N., DARBY, S., MEAKIN, S. & GNANAPRAGASAM, V.J. (2011). Role and expression of FRS2 and FRS3 in prostate cancer. *BMC Cancer*, **11**, 484–492. 150

VAN HOUWELINGEN, H.C., BRUINSMA, T., HART, A.A., VAN'T VEER, L.J. & WESSELS, L.F. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*, **25**, 3201–3216. 83

VERWEIJ, P.J. & VAN HOUWELINGEN, H.C. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, **12**, 2305–2314. 112, 134, 138

VERWEIJ, P.J. & VAN HOUWELINGEN, H.C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine*, **13**, 2427–2436. 6

VIZOSO, M., PUIG, M., CARMONA, F.J., MAQUEDA, M., VELÁSQUEZ, A., GÓMEZ, A., LABERNADIE, A., LUGO, R., GABASA, M., RIGAT-BRUGAROLAS, L.G. *et al.* (2015). Aberrant DNA methylation in non-small cell lung cancer-associated fibroblasts. *Carcinogenesis*, **36**, 1453–1463. 151

WANG, Y., XU, L. & JIANG, L. (2015a). miR-1271 promotes non-small-cell lung cancer cell proliferation and invasion via targeting HOXA5. *Biochemical and Biophysical Research Communications*, **458**, 714–719. 151

WANG, Z., CAI, M., WENG, Y., ZHANG, F., MENG, D., SONG, J., ZHOU, H. & XIE, Z. (2015b). Circulating MACC1 as a novel diagnostic and prognostic biomarker for nonsmall cell lung cancer. *Journal of Cancer Research and Clinical Oncology*, **141**, 1353–1361. 151

WOLD, S., JONSSON, J., SJÖRSTRÖM, M., SANDBERG, M. & RÄNNAR, S. (1993). DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Analytica Chimica Acta*, **277**, 239–253. 7

WU, C., QIU, S., LU, L., ZOU, J., LI, W.F., WANG, O., ZHAO, H., WANG, H., TANG, J., CHEN, L. *et al.* (2014). RSPO2–LGR5 signaling has tumour-suppressive activity in colorectal cancer. *Nature Communications*, **5**. 150

XIAO, F., BAI, Y., CHEN, Z., LI, Y., LUO, L., HUANG, J., YANG, J., LIAO, H. & GUO, L. (2014). Downregulation of HOXA1 gene affects small cell lung cancer cell survival and chemoresistance under the regulation of miR-100. *European Journal of Cancer*, **50**, 1541–1554. 151

XIE, C. & TAMMI, M.T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Boinformatics*, **10**, 80. 16

YU, S.J., HU, J.Y., KUANG, X.Y., LUO, J.M., HOU, Y.F., DI, G.H., WU, J., SHEN, Z.Z., SONG, H.Y. & SHAO, Z.M. (2013). MicroRNA-200a promotes anoikis resistance and metastasis by targeting YAP1 in human breast cancer. *Clinical Cancer Research*, **19**, 1389–1399. 152

YUE, X., ZHANG, Z., LIANG, X., GAO, L., ZHANG, X., ZHAO, D., LIU, X., MA, H., GUO, M., SPEAR, B.T. *et al.* (2012). Zinc fingers and homeoboxes 2 inhibits hepatocellular carcinoma cell proliferation and represses expression of Cyclins A and E. *Gastroenterology*, **142**, 1559–1570. 150

ZHANG, F., WANG, Z.M., LIU, H.Y., BAI, Y., WEI, S., LI, Y., WANG, M., CHEN, J. & ZHOU, Q.H. (2010). Application of RT-PCR in formalin-fixed and paraffin-embedded lung cancer tissues. *Acta Pharmacologica Sinica*, **31**, 111–117. 151

ZHANG, H.H. & LU, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika*, **94**, 691–703. 2

ZHANG, J.J., ZHU, Y., XIE, K.L., PENG, Y.P., TAO, J.Q., TANG, J., LI, Z., XU, Z.K., DAI, C.C., QIAN, Z.Y. *et al.* (2014). Yin Yang-1 suppresses invasion and

metastasis of pancreatic ductal adenocarcinoma by downregulating MMP10 in a MUC4/ErbB2/p38/MEF2C-dependent mechanism. *Molecular Cancer*, **13**, 1. 151

ZHANG, Z., WANG, Y., VIKIS, H.G., JOHNSON, L., LIU, G., LI, J., ANDERSON, M.W., SILLS, R.C., HONG, H., DEVEREUX, T.R. *et al.* (2001). Wildtype Kras2 can inhibit lung carcinogenesis in mice. *Nature Genetics*, **29**, 25–33. 150

ZHAO, S.D. & LI, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, **105**, 397–411. 2

ZHAO, Z.G., JIN, J.Y., ZHANG, A.M., ZHANG, L.P., WANG, X.X., SUN, J.G. & CHEN, Z.T. (2015). MicroRNA profile of tumorigenic cells during carcinogenesis of lung adenocarcinoma. *Journal of Cellular Biochemistry*, **116**, 458–466. 151

ZHOU, W., YIN, M., CUI, H., WANG, N., ZHAO, L., YUAN, L., YANG, X., DING, X., MEN, F., MA, X. *et al.* (2015). Identification of potential therapeutic target genes and mechanisms in non-small-cell lung carcinoma in non-smoking women based on bioinformatics analysis. *Eur. Rev. Med. Pharmacol. Sci*, **19**, 3375–3384. 151

ZHU, J.G., YUAN, D.B., CHEN, W.H., HAN, Z.D., LIANG, Y.X., CHEN, G., FU, X., LIANG, Y.K., CHEN, G.X., SUN, Z.L. *et al.* (2015). Prognostic value of ZFP36 and SOCS3 expressions in human prostate cancer. *Clinical and Translational Oncology*, 1–10. 152

ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429. 6

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320. 6, 141