Additive Cox proportional hazards models for next-generation sequencing data

Huda Mohammed Alshanbari

The University of Leeds

Department of Statistics

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

November 2017

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Acknowledgements

I would like to express my great thanks to my supervisors Dr Stuart Barber and Dr Arief Gusnanto for all the support, enthusiasm, guidance, helpful conversations and advice throughout my PhD study, and for their insights and comments on my thesis. Indeed, words cannot express my deep gratitude, it has been a great pleasure to work with them. Lots of thanks to all administrative staff of the School of Mathematics, and Tim Hainsworth, who has always been ready to help me with a computing problem.

This thesis is dedicated to the memory of my father, Mohammed Alshanbari, I miss him every day, may Allah forgive him and grant him hishighest paradise (Ameen). A great thank you to my beloved mother, for her emotional support, encouragement and spiritual support throughout my life. I am profoundly grateful to my husband Khalid Alrafaay and my daughter Saja who bring joy to my life. Throughout my PhD, I was financially supported by Princess Nourah Bint Abdulrahman University.

Abstract

Eighty-Nine Non-Small Cell Lung Cancer (NSCLC) patients experience chromosomal rearrangements called Copy Number Alteration (CNA), where the cells have abnormal number of copies in one or more regions in their genome, this genetic alteration are known to drive cancer development. An important aim of this thesis is to propose a way to combine the clinical covariate as fixed predictors with CNAs genomics windows as smoothing terms using the penalized additive Cox Proportional Hazard (PH) model. Most of the proposed prediction methods assume linearity of the CNAs genomic windows along with the clinical covariates. However, the continuous covariates can affect the hazard via more complicated nonlinear functional forms. Therefore, Cox PH model with continuous covariate are likely misspecified, because it is not fitting the correct functional form for the continuous covariates. Some reports of the work on combining the clinical covariates with high-dimensional genomic data in a clinical genomic prediction are based on standard Cox PH model. Most of them focus on applying variable selection to high-dimensional CNA genomic data.

Our main interest is to propose a variable selection procedure to select important nonlinear effects from CNAs genomic-windows. Two different approaches of feature selection are presented which are discrete and shrinkage. Discrete feature selection is based on penalized univariate variable selection, which identify the subset of the CNAs genomic-windows have the strongest effects on the survival time, while feature selection by shrinkage works by adding a second penalty to the penalized partial log-likelihood, that leads to penalizing the smoothing coefficients in the model, as a result some of the smoothing coefficient are being set to the zero.

For the NSCLC dataset, we find that the size of the tumor cells and spread cancer into the lymph nodes are significant factors that increase the hazard of the patients survival, and the estimate of the smooth log hazard ratio curves identify that some of the significant CNA genomic-windows contribute a higher or lower hazard of death to the survival of some significant CNA genomic-windows across the genome.

Contents

Acknowledgements **iii**

Abstract **iv**

Contents **v**

List of Figures **x**

List of Tables **xvi**

1 Introduction **1**

1.1 Objective of Thesis 2

1.2 Outline of Thesis 3

2 Estimation of Copy Number alteration in Lung Cancer Data **6**

2.1 Introduction 6

2.2 Description of Non-Small Cell Lung Cancer Dataset 6

2.2.1 Clinical Characteristics 6

2.2.2 Next-Generation Sequencing for Copy Number Alteration . . 8

2.3 Analysis of Copy Number Alteration. 9

2.3.1 GC Correction 10

2.3.2 A Smooth Segmentation Approach. 10

2.3.3 Genome-wide Normalization 12

2.3.4 Contamination Correction 13

2.4 Real Data Normalization 15
2.5 Conclusion 21

3 Survival Analysis 22
3.1 Introduction 22
3.2 Cox Proportional Hazards (PH) Models 23
3.3 Estimation of The Model Parameters 24
3.3.1 The Full Likelihood Function 24
3.3.2 The Partial Likelihood Function 24
3.4 Breslow Estimator 27
3.5 Checking the Proportional Hazard Assumption 29
3.6 Residual for Cox PH model 30
3.7 The Cox PH model for Clinical Characteristic Data 31
3.8 Conclusion 39

4 Generalized Additive Model 41
4.1 Introduction 41
4.2 Generalized Additive Model Overview 42
4.3 Spline Basis and Penalties 45
4.3.1 Radial Basis Functions 45
4.3.2 Cubic Splines 51
4.4 Penalized Iteratively Re-weighted Least Square Estimation (PIRLS) . 58
4.5 Selecting the Smoothing Parameter 62
4.6 Choosing the Number of Knots 63
4.6.1 Myopic algorithm 63
4.6.2 Full-search Algorithm 64
4.7 P-value 64
4.8 Logistic Regression with GAM for Clinical Characteristic of Lung Cancer Dataset 65
4.8.1 Modelling 65

4.9 Conclusion 68

5 Additive Cox PH Model 70

5.1 Introduction 70

5.2 Extending The Cox PH Model 72

5.3 The Smoothing Log Hazard Ratio 77

5.4 Simulating Survival Time from Standard Parametric Distributions . . 79

5.4.1 Generating Survival Times to Simulate Cox PH models 81

5.5 Simulation Study 82

5.5.1 Generating Survival Times to Simulate Additive Cox PH Models 82

5.5.2 Simulating Additive Cox Model for One Smooth Term 83

5.5.3 Simulating Additive Cox Model for Two Smoothing Terms . 85

5.5.4 Simulating Additive Cox Model with One Smoothing Term and One Categorical Variable 86

5.6 Test for Covariates Effects 88

5.6.1 Test for the Spline Effect Parameters 88

5.6.2 Simulation Under the Null Hypothesis of the Tests in the Penalized Additive Cox PH Model 92

5.7 Estimating the Smoothing Parameters and the Number of Knots . . . 113

5.8 Model Diagnostics 116

5.8.1 Estimation of $h_0(t)$ and $S(t|x)$ 116

5.8.2 Cox-Snell Residuals 116

5.9 Real Data Analysis 117

5.10 Conclusion 123

6 Variable Selection for Penalized Additive Cox PH Model 125

6.1 Introduction 125

6.2 Selecting Variables by Testing Individual Covariates 128

6.2.1 Univariate Selection Based on Standard Cox PH Model 128

6.2.2 Univariate Selection Based on Penalized Additive Cox PH Model129

6.3 Dependencies of Significant CNA genomic-windows 137
6.4 Forward Stepwise Selection 141
6.4.1 Cluster Analysis 154
6.5 Conclusion 160

7 Shrinkage Penalty Variable Selection 161

7.1 Introduction 161
7.2 Shrinkage Method 163
7.2.1 Double Penalty Approach 164
7.2.2 Shrinkage Approach 165
7.3 Simulation Study 167
7.3.1 Simulation Setting 167
7.3.2 Simulation Results 167
7.4 Choosing the Optimal Value of ϵ 171
7.4.1 Simulation Study 171
7.4.2 Simulation Study Using Real Data 174
7.5 Real Data Analysis 180
7.5.1 Fit the Shrinkage Penalized Additive Cox PH Model with One Smoothing Parameter 180
7.6 Forward Stepwise Selection Using the Shrinkage Approach 187
7.7 A Comparison between the Discrete Feature Selection and Shrinkage Feature Selection 197
7.8 Conclusion 203

8 Conclusion and Future Work 205

8.1 Conclusion 205
8.2 Future Work 207
8.2.1 Selecting the Optimal Values of the Smoothing Parameters . . 207

Bibliography 211

Appendices **228**

A Significant CNA genomics windows **229**

A.1 Significant CNA Genomics Windows in the Penalized Univariate Selection 229

A.2 The Common significant CNA Genomic-Windows for both Univariate Variable Selection and Penalized Univariate Variable Selection 233

List of Figures

2.1 Histogram of ratio and smoothed ratio 15

2.2 The fit of the mixture distribution to the smoothed ratio 16

2.3 Copy number ratio with smoothed-segmentation line along genome . 17

2.4 Copy number ratio with normalized and segmented DNAcopy 17

2.5 Copy number ratios for chromosome three 18

3.1 Plots of scaled Schoenfeld residuals 35

3.2 Plots of scaled Schoenfeld residuals 37

3.3 Plots of the martingale residuals versus Age with smooth curve the black solid line. 38

3.4 Plots of the estimated of Cox-Snell residual 38

3.5 Plots of the estimated cumulative hazard function 39

4.1 The plot of the smooth function of Age in model 20 68

5.1 The plot of the estimated log hazard ratio $\hat{f}(x)$ versus x. The gray lines are the estimated log hazard ratio $\hat{f}(x)$ for the 500 simulations, the black line indicates the true curve $f(x) = \sin(x)$, and the red dashed line is the average of $\hat{f}(x)$ for the 500 simulations. 84

5.2 The plot of the mean square error versus x, the dotted vertical lines indicates the locations of the knots. 85

5.3 Plots of the estimated two log hazard ratios 86

5.5 A histogram of the estimated parameter for the categorical term. . . . 87

5.4 Plot of the estimated log hazard ratio 88

5.6 Wald-type statistics result of $\hat{\boldsymbol{\beta}}_z$ for $\lambda = 0.001$ 94

5.7 Wald-type statistics result of $\hat{\boldsymbol{\beta}}_c$ for $\lambda = 0.001$ 95

5.8 The results of test statistics, degrees of freedom, and p-values when $\lambda = 0.001$ for both $\hat{\boldsymbol{\beta}}_z$, and $\hat{\boldsymbol{\beta}}_c$ 96

5.9 Histogram of the unconstrained estimated spline parameter $\hat{\boldsymbol{\beta}}_{zi}$ for $\lambda = 0.001$. 97

5.10 The Normal quantile-quantile plot of the unconstrained estimated spline parameter $\hat{\boldsymbol{\beta}}_{zi}$ for $\lambda = 0.001$. 97

5.11 Histogram of the constrained estimated parameter $\hat{\boldsymbol{\beta}}_{ci}$ for $\lambda = 0.001$. . 98

5.12 The Normal quantile-quantile plot of the constrained estimated parameter $\hat{\boldsymbol{\beta}}_{ci}$ for $\lambda = 0.001$. 98

5.13 Wald-type statistics result of $\hat{\boldsymbol{\beta}}_z$ for $\lambda = 0.01$ 99

5.14 Wald-type statistics result of $\hat{\boldsymbol{\beta}}_c$ for $\lambda = 0.01$ 100

5.15 The results of test statistics, degrees of freedom, and p-values when $\lambda = 0.01$ for both $\hat{\boldsymbol{\beta}}_z$, and $\hat{\boldsymbol{\beta}}_z$ 101

5.16 Histogram of the unconstrained estimated parameters $\hat{\boldsymbol{\beta}}_{zi}$ for $\lambda = 0.01$. 102

5.17 The Normal quantile-quantile plot of the unconstrained estimated parameters $\hat{\boldsymbol{\beta}}_{zi}$ for $\lambda = 0.01$. 102

5.18 Histogram of the constrained estimated parameters $\hat{\boldsymbol{\beta}}_{ci}$ for $\lambda = 0.01$. . 103

5.19 The Normal quantile-quantile plots of the constrained estimated parameters $\hat{\boldsymbol{\beta}}_{ci}$ for $\lambda = 0.01$. 103

5.20 Wald-type statistics result of $\hat{\boldsymbol{\beta}}_z$ for $\lambda = 0.01$ 104

5.21 Wald-type statistics result of $\hat{\boldsymbol{\beta}}_z$ for $\lambda = 0.01$ 105

5.22 Histogram of the unconstrained estimated parameters $\hat{\boldsymbol{\beta}}_{zi}$ for $\lambda = 0.1$. 106

5.23 The Normal quantile-quantile plot of the unconstrained estimated parameters $\hat{\boldsymbol{\beta}}_{zi}$ for $\lambda = 0.1$. 106

5.24 Histogram of the unconstrained estimated parameters $\hat{\boldsymbol{\beta}}_{zi}$ for $\lambda = 0.2$. 107

5.25 The Normal quantile-quantile plot of the unconstrained estimated parameters $\hat{\boldsymbol{\beta}}_{zi}$ for $\lambda = 0.2$. 107

5.26 Penalized score statistics results for $\lambda = 0$ 109

5.27 Penalized score statistics results for $\lambda = 0.01$ 110

5.28 Penalized score statistics results for $\lambda = 0.1$ 111

5.29 Penalized score statistics results for $\lambda = 0.5$ 112

5.30 Plots of the Cross-validated partial log-likelihood versus λ 119

5.31 The estimated log hazard ratio of $\hat{f}(Age)$ 121

5.32 Cumulative hazard of Cox-Snell residual from the penalized additive Cox model fit 122

5.33 Martingale residuals versus age 122

6.1 Correlation heat map all 1056 significant CNA genomic-windows. . . 137

6.2 Correlation heat map all 1056 significant CNA genomic-windows, the black box represents the highly-correlated block of Chromosome 7. . 138

6.3 The plot of $\log(se/s)$ versus the correlation r. The red solid line is the fitted loess line. 140

6.4 The plot of the estimated log hazard ratio for $\hat{f}(age)$ 147

6.5 The plot of the estimated log hazard ratio for each of the significant window of the CNA 148

6.6 The plot of the estimated log hazard ratio for each of the significant window of the CNA 149

6.7 The plot of the estimated log hazard ratio for each of the significant window of the CNA 150

6.8 The plot of the estimated log hazard ratio for each of the significant window of the CNA 151

6.9 The cumulative hazard of Cox-Snell residual from the fitted the penalized additive Cox PH model. 152

6.10 The plot of the estimated log hazard ratio for each of the not significant window of the CNA 153

6.11 The plot of the estimated log hazard ratio for y 154

6.12 The plots of the estimated log hazard ratio of the first 11 significant neighboring windows CNA. 155

6.13 The clusters of the log hazard ratio of the significant CNA genomic-windows. 157

6.14 The clusters of the log hazard ratio of the significant CNA genomic-windows. 158

6.15 The seven clusters of the log hazard ratio of the significant CNA genomic-windows. 159

7.1 Smoothing function estimates obtained applying the shrinkage approach for fixing ϵ and increasing λ. 169

7.2 Smoothing function estimates obtained applying the shrinkage approach for fixing λ and increasing ϵ. 170

7.3 Smoothing function estimates obtained applying the shrinkage approach for increasing λ and decreasing ϵ and vise versus in the shrinkage approach. 171

7.4 Five-fold CVPL for different values of λ and ϵ. 172

7.5 (a) Five-fold CVPL for different values of λ at the optimal value of $\epsilon_{opt} = 0$. (b) Five-fold CVPL for different values of ϵ at the optimal value of $\lambda_{opt} = 2$. 173

7.6 Plots of the estimated log hazard ratio at optimal values of $\lambda_{opt} = 2$ and different values of ϵ. 174

7.7 (a) Five-fold CVPL for different values of λ_1 at the optimal value pf $\epsilon_{1,opt}$. (b) Five-fold CVPL for different values of ϵ_1 at the optimal value of $\lambda_{1,opt} = 0.0102$. 177

7.8 Five-fold CVPL for different values of λ and ϵ. 177

7.9 (a) Five-fold CVPL for different values of λ_2 at the optimal value of $\epsilon_{2,opt}$. (b) Five-fold CVPL for different values of ϵ_2 at the optimal value of $\lambda_{2,opt} = 0.655$. 179

7.10 Five-fold CVPL for different values of λ_2 and ϵ_2. 179

7.11 The cross-validated partial log likelihood for the shrinkage approach. 181

7.12 Plots of the estimate log hazard ratio for age. 182

7.13 Plots of the estimate log hazard ratios using shrinkage approach. . . . 183

7.14 Plots of the estimate log hazard ratios using shrinkage approach. . . . 184

7.15 Plots of the estimate log hazard ratios using shrinkage approach. . . . 185

7.16 Plots of the estimate log hazard ratios using shrinkage approach. . . . 186

7.17 The estimated log hazard ratios for the $\hat{f}(Age)$ versus Age. 190

7.18 The plot of the estimated log hazard ratio for each of the significant window of the CNA using shrinkage approach 191

7.19 The plot of the estimated log hazard ratio for each of the significant window of the CNA using shrinkage approach 192

7.20 Cumulative hazard of Cox-Snell residual, the solid line, from the shrink additive Cox PH model fit, comparing to the identity line the red dashed line 193

7.21 Two clusters of the estimated log hazard ratio of the significant CNA genomic-windows. 195

7.22 The six clusters of the estimated log hazard ratio of the significant CNA genomic-windows. 196

7.23 The plot of the estimated log hazard ratio for each of the significant window of the CNA using both approaches 201

7.24 The plot of the estimated log hazard ratio for each of the significant window of the CNA using both approaches 202

8.1 Five-fold CVPL for different values of λ_1 and λ_2. 209

8.2 Five-fold CVPL for different values of ϵ_1 and ϵ_2. 209

8.3 Five-fold CVPL for different values of λ_1 and λ_2. 210

8.4 Five-fold CVPL for different values of ϵ_1 and ϵ_2. 210

List of Tables

2.1 Demographic and clinical characteristics of patients in Non-Small Cell Lung Cancer dataset. 7
2.2 Next-Generation Sequencing for CNA in lung cancer dataset 9
2.3 The output of the CNAnorm method 18
2.4 The total number of ploidy in CNA 19
2.5 The total number of ploidy in CNA 20

3.1 The -2 partial log likelihood on fitting 32 possible Cox PH model . . 32
3.2 Estimated values of the coefficients on fitting Cox PH model 33
3.3 The p-values from testing the PH assumption 34
3.4 The information of five hazards outliers. 34
3.5 The p-values from testing the PH assumption 36

4.1 The number of patients, median survival times, censored and uncensored observations for the response variables. 65
4.2 32 possible GAM models with the effective degree of freedom and AIC values for each model. 66
4.3 Estimated values of the coefficients on fitting GAM model 67

5.1 The values of -2 log partial likelihood and the number of parameters, AIC for each fitted additive Cox model. 118
5.2 Estimated values of the parameters on fitting additive Cox PH model. 120

6.1 Genes related to Non-Small Cell Lung Cancer. 131
6.2 Genes related to other types of the cancer. 132
6.3 Genes related to other types of the cancer. 133
6.4 Genes which are not related to any type of cancer. 133
6.5 Genes which are not related to any type of cancer. 134
6.6 Genes which are not related to any type of cancer. 135
6.7 Genes which are not related to any type of cancer. 136
6.8 41 significant CNA genomic windows 141
6.9 The result of including one significant CNA window in the model . . 143
6.10 The result of including two significant CNA windows in the model . . 144
6.11 The optimal values of the smoothing parameters in the multivariate penalized Cox model. 146
6.12 The seven clusters of the estimated log hazard ratio of the significant CNA genomic-windows. 158

7.1 The optimal values of the λ and ϵ for each univariate penalized additive Cox PH model. 175
7.2 The ordering of the significant CNA genomic-windows. 187
7.3 Optimal values of the smoothing parameters in the shrinkage approach 189
7.4 The effective degrees of freedom for each smoothing significant CNA genomic-windows 189
7.5 The six clusters of the estimated log hazard ratio of the significant CNA genomic-windows. 197
7.6 List of the common significant CNA genomic-windows 198
7.7 List of the common significant CNA genomic-windows 198
7.8 Estimated values of the parameters on fitting additive Cox PH model for both forward variable selection and shrinkage selection. 199

Chapter 1

Introduction

The third most common cancer in the United Kingdom is the Squamous Cell carcinoma which is subtypes of Non-Small Cell Lung Cancer (NSCLC). Non-Small Cell Lung Cancer (NSCLC) patients experience chromosomal rearrangements called Copy Number Alteration (CNA), which has a high degree of homology with regards to their chromosomal abnormalities (Zhang et al., 2009), this genetic alteration is known to drive cancer development. An important aim of this thesis is identifying CNAs of NSCLC to assess the severity of the chromosomes rearrangement, and to investigate the non-linear relationship between survival and CNAs. The Cox Proportional Hazards (PH) model, (Cox, 1972), has become a popular choice for modeling survival data, where the effect of the continuous covariates are assumed to be linear. To introduce possible nonlinear effects of the continuous covariates into a Cox model, the hazard can be expressed as an additive Cox model using smoothing methods Hastie and Tibshirani (1990b); Gray (1992). This flexibility would be a great help to further scientific understanding of the association between CNAs and survival. In this thesis we derive a new nonlinear technique to express the relationship between CNA genetics and survival, using the penalized additive Cox model for high-dimensional data.

The additive Cox PH model can not be directly applied in case of high-dimensional CNA data, and some CNAs are highly correlated. To handle this issue, we try to apply two different methods of feature variable selection: discrete and shrinkage.

1.1 Objective of Thesis

There are many problems related to including high-dimensional data in survival models. In this thesis, we will focus on three main problems, which can be summarized as follows:

- Developing penalized additive Cox PH models where the effect of the continuous covariates on the log hazard ratios are represented nonlinearly using radial basis functions as a smoothing term. Combining clinical characteristics, as fixed predictors, with CNA data as smoothing terms in the penalized additive Cox PH model, where each smoothing term having two constraints. The general linear constraints approach can be used to solve the constraints problem. We will discuses some aspect that are relevant to the penalized additive Cox PH model such as Cross-Validated partial log-likelihood (CVPL), the effective degrees of freedom (edf), and choosing the optimal number of knots.

- Modeling high-dimensional data is challenging as the number of covariates is generally much larger than the sample size. The problem is not only including high-dimensional data in the survival model, but it is also including the high-dimensional data as smoothing terms in the survival models. Each smoothing term can be express as a matrix where the number of rows is the number of observations, and the number of columns is the number of spline parameters, which is associated with the number of knots. Combining all these matrices together lead to very large smooth matrix for all CNA in the model, which can be computational demanding.

- Developing an approach that is able to select the important CNA data, and identifying which CNA to include in the penalized additive Cox PH model without incorporation a whole CNA in the model. This leads to simpler model for easier interpretation. Variable selection is tried to gain as much as possible information from CNA data. We introduce two variable selection methods: discrete and shrinkage.

1.2 Outline of Thesis

Chapter 2 begins with the NSCLC dataset description for the clinical characteristics, and CNA data. The recent method of detecting CNA from next-generation sequences data is reviewed in Chapter 2. This method is applied for each patient individually to estimate CNA as a ratio of the tumour sample to the normal sample from the patients, the result of this method is a matrix with dimension 85×13253, where 85 is the number of patients and 13253 is the number of CNA variables.

In Chapter 3, the background of modeling survival analysis based on the standard Cox proportional hazard model when the number of covariates is less than the sample size ($p < n$) is presented with an application using the clinical characteristics only. However, standard Cox PH model in the case of the high-dimensional data setting where the number of covariates is much larger than the sample size ($p > n$) is not discussed in this thesis, because it assumes the linearity of the covariates, while we are interested in non-linear form of the covariates.

The aim of Chapter 4, is to provide an overview of Generalized Additive Models (GAM) based on penalized likelihood framework and discussing some aspects that are relevant to this thesis such as Generalized Cross Validation (GCV), the effective degrees of freedom (edf), and choosing the optimal number of knots. Logistic regression with GAM for the clinical characteristics is presented.

In Chapter 5, we present a novel statistical model based on the penalized additive Cox PH model where the smoothing term is the radial basis function. This extension of Cox PH model allows us to include the clinical characteristics as fixed effect and CNA variables as smoothing terms. The method of estimating the model parameters for both the fixed effect and smooth spline effect based on maximizing the penalized partial log likelihood is presented. The estimate of the smooth log hazard ratio for the continuous covariate is presented. We generalize the idea of Bender et al. (2005) to generate survival data from the additive Cox model. Simulations studies are described and discussed to assess the proposed model. The test statistics for testing the smooth spline effect in the model using the penalized version of the score, Wald and likelihood ratio test statistics are discussed, simulation studies under the null hypothesis is discussed. Cross validation partial log likelihood (CVPL) which is adapted to select the optimal smoothing parameters and optimal number of knots in the penalized additive Cox PH model is presented. Since the model diagnosis is an important part of the model process, Breslow's estimator of the baseline cumulative hazard rate, the estimate of the survival function, Cox-Snell residuals, and Martingale residuals are presented. Finally, results and evaluation of NSCLC clinical data only are presented.

In Chapter 6,we produce variable selection methods based on discrete feature selection, in which only subsets of the CNA variables are selected and included in the multivariate penalized additive Cox PH model. This feature selection can be done discretely, by considering a strategy to assess whether CNA feature variables should be included in the model. We generalized univariate variable selection in Bøvelstad et al. (2007) to select the significant CNA variables by performing penalized univariate variable selection. We compared our penalized univariate selection method with univariate variable selection in Bøvelstad et al. (2007). As a result, the penalized univariate selection identify more significant CNA variables than the univariate selection method. Forward stepwise selection is used to include this significant CNA variables in the multivariate penalized additive Cox PH model. Three different Clustering techniques are used to identify the similar log hazard ratio shapes of the significant CNA variables across the genome.

In Chapter 7, we propose the double shrinkage penalty approach for variable selection. The idea of the shrinkage approach is to add a second penalty to the penalized partial log-likelihood, that leads to penalizing the smoothing spline coefficients in the model, as a result some of the smoothing coefficient are being set to the zero. Simulation studies are described and discussed to assess the proposal of the double shrinkage approach to the penalized additive Cox PH model.

In Chapter 8, a summary of our findings and the final conclusion is included with some suggestions for further work.

Chapter 2

Estimation of Copy Number alteration in Lung Cancer Data

2.1 Introduction

In this chapter, a description of the Non-Small Cell Lung Cancer (NSCLC) dataset, which includes the clinical characteristics and cancer genomes data, is introduced in Section 2.2. The method of estimating the Copy Number Alterations (CNA) based on the approach that described in Gusnanto et al. (2012) is presented in Section 2.3. The results obtained after the analysis for one patients (LS168) are shown in Section 2.4.

2.2 Description of Non-Small Cell Lung Cancer Dataset

2.2.1 Clinical Characteristics

The dataset that we were working with contained information for 89 patients with Non-Small Cell Lung Cancer (NSCLC), in particular Squamous Cell Carcinoma lung cancer, who had been seen in the Department of Thoracic Surgery at Leeds Teaching Hospital in Leeds, United Kingdom, between 1994 and 2003. For each of these patients, Next-Generation Sequencing data (NGS) and clinical characteristics data are available. The demographical and clinical characteristics of patients are described in Belvedere et al. (2012) which is summarized in Table 2.1.

Covariate			# of patients
Gender		Male	63
		Female	26
Status		Censored	23
		Uncensored	66
	T1	The tumor is smaller than 2 cm	23
Tumour Size	T2	The tumour is between 3 -7cm	59
	T3	The tumour is larger than 7cm	7
	N0	There are no lymph nodes	47
Nodal Status	N1	There are cancer cells in 1 or 2 nearby lymph nodes	35
	N2	There are cancer cells in 3 to 6 nearby lymph nodes	7
	G1	Low-grade	2
	G2	Intermediate-grade	46
Grade	G3	High-grade	37
	GX	The grade cannot be assessed	4

Table 2.1: Demographic and clinical characteristics of patients in Non-Small Cell Lung Cancer dataset.

The response variable is the survival time (in days, range 34-4565 days), the median survival time is 680 days. Survival status is either uncensored or censored, uncensored when the information is available, and censored when the information on the time-to-event is not available due to the event does not occurring before the study ended. The covariate includes age of the patients at the time of surgery (numerical variable), the range of age is between 39 and 84. Gender variable indicates the gender of the patients whether male or female. Stage T of cancer is an ordinal variable which indicates the size of the tumor cells, which can be level 1,2, or 3, with 1 for smallest size and 3 for the largest size . Stage N is also an ordinal variable which describes whether the cancer has spread to the nearby lymph nodes, there are 3 levels, 0, 1, and 2. Stage-N0 means that there is no cancer in the lymph nodes. Stage-N1 means there are cancer cells in lymph nodes within the hilum lymph nodes. Stage-N2 means the cancer has spread to the lymph nodes. Grade is an ordinal variable which describes how abnormal the cancer cells are when compared to normal cells. Tumors cells can be graded as 1, 2, 3, or 4, the lower the grade the slower the growth rate of the cancer, and grade GX means that the grade can not be assessed. There were 4 patients having grade GX, so we remove them from the data, the number of patients is reduced from 89 to 85.

2.2.2 Next-Generation Sequencing for Copy Number Alteration

NSCLC patients experience chromosomal rearrangements called Copy Number Alterations (CNA), where the cells have abnormal number of copies in one or more regions in their genome. The Next-Generation Sequencing (NGS) data from clinical samples come directly from patients. DNA libraries were obtained and sequenced by using ILLumina GAII sequencer, the preparation of DNA can be found in Belvedere et al. (2012). DNA sequencing refers to the methods that are used to determine the orders of nucleotide bases in a DNA molecule, namely adenine (A), guanine (G), cytosine (C) and thymine (T). The normal sample was created from a pool of 20 normal British individuals. The effect of pooling the data from different individuals is to reduce the variability due to individual specific variation, so the total number of normal read is 4,386,893.

The DNA was collected from the 89 patients; small DNA fragments (called *reads*) from each patient are mapped to a human normal genome. The genome was split into windows with average 300 tumour reads per window. To calculate the Copy Number Alteration (CNA), we count the number of reads that fall into fixed-size, 200 kb window size, non-overlapping genomic 'windows'. The information from each window is recorded as rows in Table 2.2. The number of reads from the patient in the "Test" column were compared to the number of reads from the normal samples in the "Normal" column. The Guanine-Cytosine (GC) column contains the percentage of nitrogenous bases in a read that are either Guanine or Cytosine in each window, and the right most column is the ratio of test reads to normal reads. The total number of windows is 15490, which was consistent across all patients, whilst the total number of test reads were different for each window in the samples. For example, for one patient the total number of test reads are 1,820,403, for more information see Tables 2.4 and 2.5 in Section 2.4.

Chromosome	Window starting position	Test	Normal	GC	Ratio
1	1	5	6	42.96	0.84
1	200001	1	1	39.87	1.00
1	400001	6	11	45.06	0.54
⋮	⋮	⋮	⋮	⋮	⋮
10	18600001	131	283	40.82	0.46
10	18800001	99	318	39.95	0.31
10	19000001	93	310	37.70	0.30
⋮	⋮	⋮	⋮	⋮	⋮
22	28200001	102	319	43.29	0.32
22	28400001	107	292	41.96	0.36
22	28600001	101	328	38.01	0.31

Table 2.2: The number of reads for one patient (Test) and normal human genome (Norm), guanine-cytosine (GC) percentage and test/norm read (Ratio).

2.3 Analysis of Copy Number Alteration.

As mentioned earlier, this analysis based on CNAnorm method developed by Gusnanto et al. (2012), we also use the same notations and arguments and the notation applies to a generic patient. To calculate the CNA, we compute the ratio of test reads to normal reads as follows: the observed number of reads from a tumour sample is denoted by x_{jk} for each chromosome $j = 1, \ldots, h$, and window $k = 1, \ldots, n_j$, where n_j is the number of windows in chromosome j and $h = 22$, and the observed number of reads in a normal sample is denoted by y_{jk}. In order to identify the CNAs in the tumour genome that either gain or loss, we estimate CNAs as the observed ratio of the tumour sample to the normal sample in each window:

$$\hat{\rho}_{jk} = r_{jk} = \frac{x_{jk}}{y_{jk}}. \tag{2.1}$$

In the normal sample there are two copies of (outosomole) chromosomes, while in the tumor sample may have zero, one, two or more duplication. Therefore, the ratio can takes any value in $G = \{0, 0.5, 1, 1.5, 2, \dots\}$ which corresponds to tumor copy number $P = \{0, 1, 2, 3\dots\}$. However, the ratio r_{jk} does not take into account the different number of reads in each genome, the recorded number of reads is different (number of test/normal reads), the size of tumour and normal genome is different, and the contamination of the tumour sample with a normal sample. All these problems make the estimate $\hat{\rho}_{jk}$ not belong to G, and the CNAs corresponding to normal genomic region will not centered to a ratio 1. Estimation of CNA requires several steps which are described in more details as follows.

2.3.1 GC Correction

Chen et al. (2013) showed that in the NGS data, the GC is bias, and this can be explained by the low coverage of reads in the GC-poor or GC-rich regions of a genome. However, the ratio r_{jk} is influenced by the GC content in the genome windows (Boeva et al., 2011). The aim of this step is to remove the dependency of the ratio r_{jk} on GC content. In order to achieve that, a quadratic local regression model can be used. Gusnanto et al. (2012) expressed the GC-correction ratio as follows

$$r_{jk}^{\text{norm}} = \frac{\kappa}{A_{jk}} r_{jk}, \tag{2.2}$$

where κ represents the median of r_{jk}, and the estimated Loess pointwise mean of $r_{jk}$is denoted by A_{jk}. For simplicity, the superscript norm is dropped, so the notation r_{jk} in the following steps have been GC-corrected.

2.3.2 A Smooth Segmentation Approach.

This step is based on the smooth segmentation approach by Huang et al. (2007), which is necessary in case of the small number of reads in each window. This smooth segmentation approach based on a linear model under the assumption that the second-order differences of the random effect parameters follows a Cauchy distribution, so estimates of the random effects are the segmented ratios $\tilde{r}_{jk}$. Let $b_1, \dots, b_n$, be the fixed genomic position and n is the number of genomic position, with $b_1 < b_2 < \dots < b_n$, and $r_1, \dots, r_n$ are the observed ratio. The model is

$$r_i = f(b_i) + \epsilon_i, \quad i = 1, \dots, n \tag{2.3}$$

where the errors $\epsilon_1, \dots, \epsilon_n$ are independent and identically distributed (IID) t-distribution with location at 0, unknown dispersion σ^2 and k degrees of freedom. To estimate $f(b_i)$ based on observations r_{jk}, Huang et al. (2007) assumed B-splines of degree zero with the observed b_i as knots, and the smoothness of f can be expressed by the second-order differences $a_i = \nabla^2 f_i = f_i - 2f_{i-1} + f_{i-2}$ which are IID Cauchy distributed to allow smooth transition. Using maximum likelihood approach, the log-likelihood can be expressed as

$$\ell(f, \sigma^2, \sigma_f^2) = \log(p(r|f)) + \log(p(f)). \tag{2.4}$$

The first term on the right hand side of equation (2.4) is obtained from the t-density with k degrees of freedom, which can be expressed as:

$$\begin{aligned}\log(p(r|f)) &= \log\Gamma(k/2 + 1/2) - \log\Gamma(k/2) + \frac{k}{2}\log(k)\\ &\quad - \frac{1}{2}\log(\pi\sigma^2) - \frac{k+1}{2}\log\left\{k + \frac{(r-f)^2}{\sigma^2}\right\},\end{aligned} \tag{2.5}$$

The second term in equation (2.4) is the random effects which comes from Cauchy model, with located at 0 and has scale factor σ_f^2 and can be written a

$$\log(p(f)) \equiv \ell(a) = -(n-2)\log(\pi\sigma_f) - \sum_{i=1}^{n-2}\log\left(1 + \frac{a^2}{\sigma_f^2}\right). \tag{2.6}$$

Taking the first derivative of $\log(p(r, f))$ with respect to the random effect f we obtained $\frac{1}{\sigma^2}W(r - f)$, where $W = \text{diag}(\frac{k+1}{k+(r_i-f_i)^2/\sigma^2})$, and then taking the first derivative of $\log(p(f))$ with respect to a, we have $-\frac{2a}{a^2+\sigma_f^2}$, which can be written in vector form $-\frac{1}{\sigma_f^2}D^{-1}a$, where $a = \nabla^2 f$, ∇^2 is the $(n-2) \times n$ second order differences matrix, and $D^{-1} = \mathrm{diag}[2/(1+a^2/\sigma_f^2)]$. Therefore, combining the above results, the first derivative of $\ell(f, \sigma^2, \sigma_f^2)$ with respect to f is

$$S(f) = \frac{1}{\sigma^2}W(r-f) - \frac{1}{\sigma_f^2}(\nabla^2)^T D^{-1}(\nabla^2)f. \tag{2.7}$$

In order to solve $S(f) = 0$, Huang et al. (2007) proposed an Iterative Re-weighted Least Square (IRLS) method. Given the initial value of f, we can compute the matrices W and D, and solve the updating equation $f = \Big(W + \lambda(\nabla^2)^T D^{-1}(\nabla^2)\Big)^{-1} Wr$, where $\lambda = \sigma/\sigma_f$. The estimation of the dispersion parameter σ can be calculated as $\hat{\sigma} = \mathrm{median}\{|r_i - \hat{f}_i|\}$. The degree of freedom in the model is $df \equiv df(\lambda) = \mathrm{trace}\{(W + \lambda(\nabla^2)^T D^{-1}(\nabla^2))^{-1}\boldsymbol{W}\}$. The estimation of λ can be done by minimizing the AIC criterion, $AIC(\lambda) = -2\sum \log(p(r|\hat{f})) + 2df$. The estimate of σ_f is $\hat{\sigma}/\lambda$. The estimate of the random effect $\hat{f}_i$ is the segmented ratio $\tilde{r}_{jk}$.

2.3.3 Genome-wide Normalization

The aim of this step is to correct the location of the distribution of the copy number ratio, to achieve this we have to estimate δ_{CNA} from the segmented ratio $\tilde{r}_{jk}$, where δ_{CNA} represents the genome-wide normalization coefficient. In the previous step the segmented ratio $\tilde{r}_{jk}$ displays a multi-modal distribution, each mode indicated the position of the CNA in G, and the corresponding copy number in P. However, the modes of the segmented ratio distribution $\tilde{r}_{jk}$ are not centered on the expected CNA in G, to correct that we need to estimate δ_{CNA} from the segmented ratio $\tilde{r}_{jk}$. The mixture normal distributions can be used to fit the distribution of smooth segmented ratio $\tilde{r}_{jk}$.

$$p(\tilde{r}_{jk}) = \sum_{m=1}^{M} \pi_m N(\tilde{r}_{jk}; \mu_m, \sigma_m^2). \tag{2.8}$$

where μ_m are the mean, σ_m^2 are the variance for each normal distribution, and π_m are the unknown mixing proportions, such that $\sum_{m=1}^{M} \pi_m = 1$,$0 \leq \pi_m \leq 1$ for $m = 1, \ldots, M$, and M is the number of component. Each value of the means μ_m corresponds to the value in G that indicates the ratio of tumour to normal copy number, and the corresponding tumour copy number P.

Gusnanto et al. (2012) estimated the mixture component in equation (2.8) using the standard expectation-maximization (EM) algorithm. Akaike's Information Criterion (AIC) is used to select the number of components M in the model. After estimating $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_M)$, the relationship between $\hat{\boldsymbol{\mu}}$ and the corresponding tumour copy number in P is described by a linear regression model.

The component in the mixture model in equation (2.8) that corresponds to the normal ploidy, which has ratio 1 in G is identified as the most common component $\nu = \text{argmax}_m \hat{\pi}_m$, this indicates that the ν^{th} components is assigned to have $\nu - 1$ copy number, because the first component is corresponding to zero copy number.

The genome-wide normalization coefficient was estimated as $\hat{\delta}_{\text{CNA}} = \frac{1}{\hat{\mu}_\nu}$, as a result the estimation of $\hat{\delta}_{\text{CNA}}$ identified the mixture component that corresponds to normal ratio, so then shifts the whole distribution multiplicatively which makes the normal ratio center at one. The estimation of crude CNA can be computed as follows

$$\hat{\rho}_{jk}^{a} = \tilde{r}_{jk}\hat{\delta}_{\text{CNA}}, \tag{2.9}$$

where $\hat{\rho}_{jk}^{a}$ is the estimates of CNA where the contamination is still presents, for the purpose of finding the estimate of CNA that makes the estimate comparable between samples, we need to characterized any contamination in order to make the suitable correction.

2.3.4 Contamination Correction

In the case of no contamination, the smoothed ratio $\tilde{r}_{jk}$ takes any value G, while if the contamination is present the smooth ratio $\tilde{r}_{jk}$ shrunk toward ratio one (see Gusnanto et al. (2012) supplementary material).

To estimate the contamination, Gusnanto et al. (2012) assumed that the contamination shrinks the CNA linearly towards a ratio of one. For example, it $\rho_{jk} = 2$, then the CNAs will shrink to $1 < \rho_{jk} < 2$, while if $\rho_{jk} = 0.5$, the CNA will shrink between 0.5 and 1. However, from the previous step the normal copy number is centered at one, so we can assume that the estimation of CNAs has occurred from a shrinkage of the non-contaminated $\hat{\rho}_{jk}$ around ratio one.

$$\hat{\rho}^{a}_{jk} = 1 + (\hat{\rho}_{jk} - 1) \times (1 - \hat{\psi}), \tag{2.10}$$

where $\hat{\psi}$ is the estimate of contamination proportions, which is between zero and 1. If $\hat{\psi}$ is equal to zero, the estimate of crude CNA will be equal to the ratio $\hat{\rho}_{jk}$. The estimation of $\hat{\psi}$ can be done by investigating how the estimate in $\hat{\boldsymbol{\mu}}$ have been shrunk towards $\hat{\boldsymbol{\mu}}_{v}$ that corresponds to the copy number two. To do this they normalized the estimates $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_M)$ into $\hat{\boldsymbol{\mu}}^{c} \equiv \{\hat{\mu}^{c}_{m}\} = \{\hat{\mu}_m \hat{\delta}_{\text{CNA}}\}$ for $m = 1, \ldots, M$. The estimate of $\hat{\psi}$ is (Gusnanto et al., 2012)

$$\hat{\psi} = \frac{1}{M-1} \sum_{m,\ m \neq \nu} \left(1 - \frac{|\mu^{c}_{m} - \mu^{c}_{\nu}|}{\mu^{c}_{\nu}} \frac{1}{0.5 \times |P_m - P_\nu|} \right), \tag{2.11}$$

where P_m is the set of the copy number without P_ν. From equation (2.10) the estimate of CNA can be written as

$$\hat{\rho}_{jk} = 1 + (\hat{\rho}^{a}_{jk} - 1) \times \frac{1}{(1 - \hat{\psi})}. \tag{2.12}$$

The estimation of CNA by $\hat{\rho}_{jk}$ now considers the different read depths, genome size and contamination. This can make the estimate comparable between different pairs of samples.

2.4 Real Data Normalization

The normalization steps have already been applied to the 85 patients by using R package CNAnorm as described by Gusnanto et al. (2012). Tables (2.4) and (2.5) present the ploidy number, the number of test reads and the estimated contamination for the 85 patients. There are only 4 patients with estimated ploidy equal to 4 (tetraploid), and 81 patients with estimated with ploidy equal to 2 (diploid), more invigorations is needed regrading whether or not it would be useful to include the number of ploidy and the contamination information of Tables (2.4) and (2.5) as a additional variables in the survival models.

The result of only one of the patients (LS168) out of the 89 patients will be presented as an illustrative example. Figure 2.1 illustrates the impact of the smoothing on the distribution of copy number ratio. The left panel shows the ratios r_{jk}, which do not clearly identify the multi-modality in the distribution of ratio r_{jk} in the genome. The right panel is the smoothed ratios $\tilde{r}_{jk}$, which clearly identify the multi-modality in the distribution of the smoothed ratio $\tilde{r}_{jk}$.

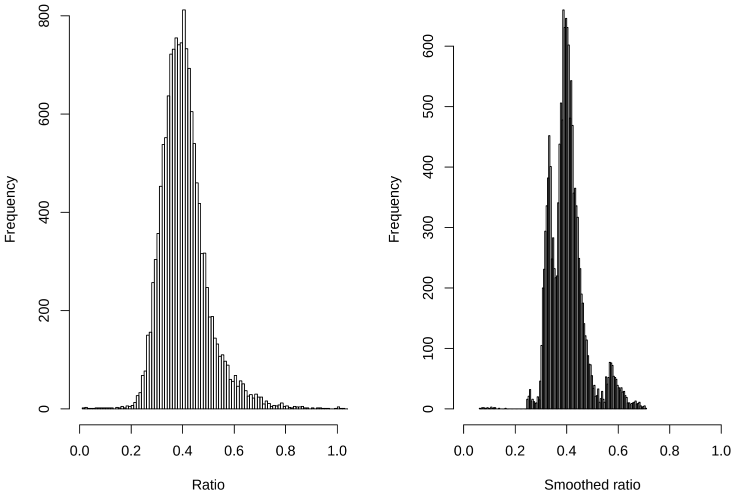


Figure 2.1: The left panel is the histogram of ratio as reported by CNAnorm. The right panel is the smoothed ratio across the genome of patient LS168.

We fit the mixture model in equation (2.8) to the distribution of the smoothed ratio $\tilde{r}_{jk}$, as seen in Figure 2.2. The optimal number of mixture components is $M = 7$ as determined by minimizing AIC, with estimates of the means $\hat{\mu}_m = (0.33, 0.36, 0.39, 0.42, 0.46, 0.50, 0.58)$. The estimated proportion of the common mixture component is $\hat{\pi}_3 = 0.28$ which illustrates that the third component is the most common one ($\nu = 3$) suggestion that the tumour genome is diploid, so $\hat{\mu}_3 = 0.39$ and $\delta_{\text{CNA}} = 2.56$. After scaling for the diploid component to have ratio one, the scaled estimates $\hat{\mu}^c$ were $0.84, 0.92, 0.99, 1.07, 1.17, 1.40$ and 1.48. This gives the estimate of the tumour contamination $\hat{\psi} = 71.13\%$.

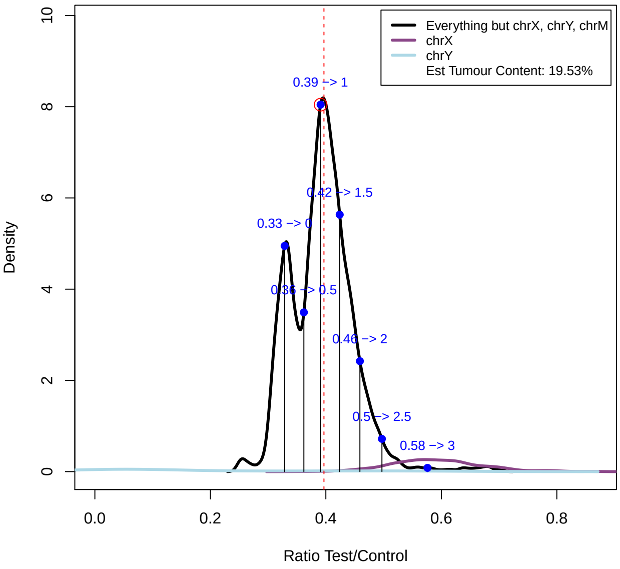


Figure 2.2: The fit of the mixture distribution to the smoothed ratio, the black vertical solid lines indicate the mean of the distribution. The estimated tumour content is shown in the legend of patient LS168.

Figure 2.3 shows the normalized copy number ratio with estimates of copy number, as a smoothed signal lines across the genome. Segmented ratio using DNACopy is presented in Figure 2.4,the vertical solid lines separate the chromosomes, Grey dots represent normalized reads per window, solid thick lines are the normalized and segmented DNAcopy, and the horizontal dashed lines are the median copy number. The solid horizontal line across the entire figure is the mixture model closest to the median, the green triangles are points outside the graph.

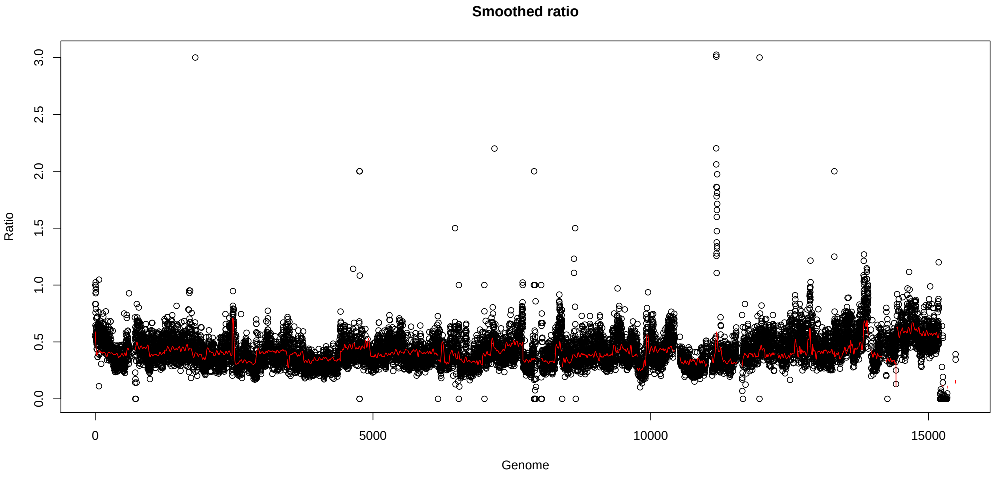


Figure 2.3: The copy number ratio with estimates of copy number, as a smoothed signal, lines across the genome of patient LS168.

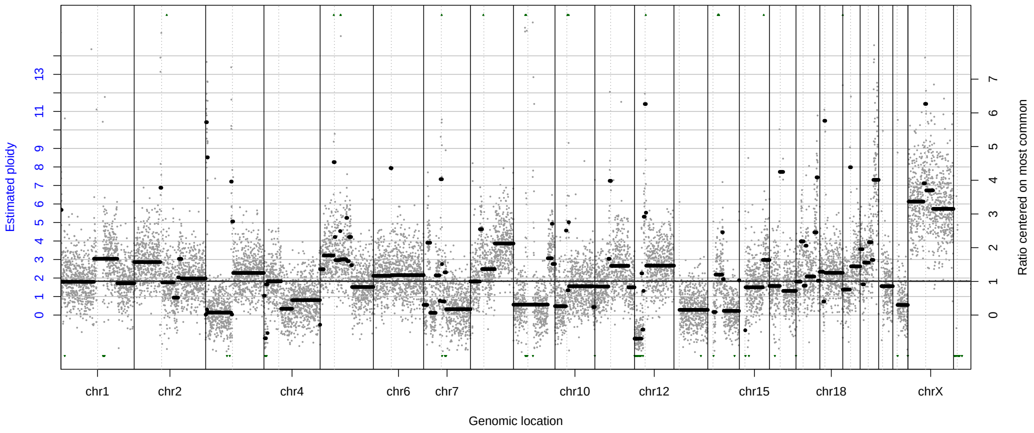


Figure 2.4: The normalized copy number ratio with estimates of copy number, the solid thick black lines are the normalized and segmented DNAcopy of patient LS168.

Figure 2.5 shows the output for chromosome three before and after normalization.

There are 15,490 genomic-windows per patients, with some missing data due to centromeres of the genome, each window that contain one missing value we impute it by the mean of that windows, as result, the estimate of CNA is 13253 genomics-windows for each patients. Finally, the result of the normalization steps for patient LS168 are presented in Table 2.3.

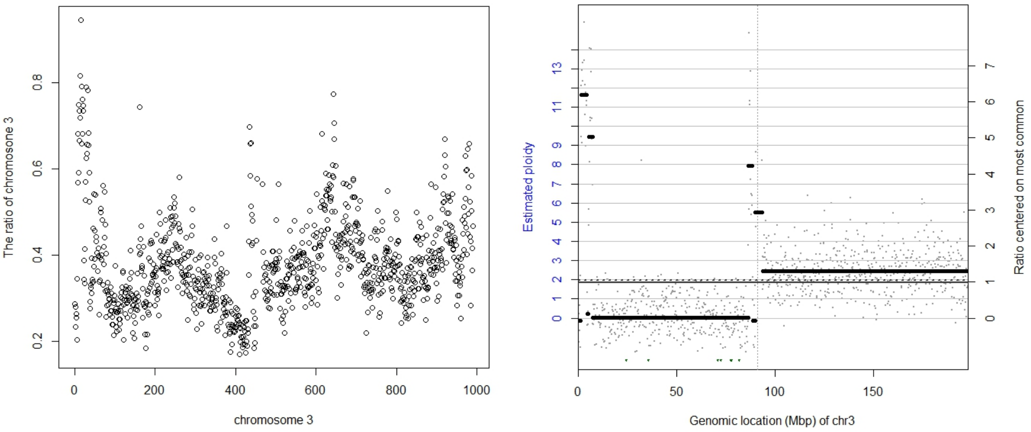


Figure 2.5: Copy number ratios for chromosome three before (left) and after (right) normalization, the black solid line is the estimate of CNA of patient LS168.

#	Chr	Pos	Ratio	Ratio.n	Ratio.s.n	SegMean	SegMean.n
1	chr1	1	0.83	4.80	0.58	-0.82	2.99
2	chr1	200001	1	8.19	0.57	-0.82	2.99
3	chr1	400001	0.54	1.40	0.56	-0.82	2.99
4	chr1	600001	0.56	2.04	0.55	-0.82	2.99
5	chr1	800001	0.64	0.98	0.54	-0.8	2.99
6	chr1	1000001	0.75	1.77	0.53	-0.82	2.99
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
14410	chr22	50200001	0.52	0.39	0.31	-1.58	0.39
14411	chr22	50400001	0.42	0.00	0.31	-1.58	0.39
14412	chr22	50600001	0.54	0.32	0.31	-1.58	0.39
14413	chr22	50800001	0.56	0.55	0.31	-1.58	0.39
14414	chr22	51000001	0.48	0.26	0.31	-1.58	0.39
14415	chr22	51200001	0.25	-0.65	0.31	-1.58	0.39

Table 2.3: The output of the CNAnorm method for patients LS168.

Patient	Number of Ploidy	Number of Test	Contamination
LS 168	2	1820403	71.13
LS169	2	2017296	49.79
LS170	2	1877104	86.49
LS171	2	3141892	72.15
LS172	2	1926593	74.60
LS173	2	3217576	79.82
LS174	2	2434153	87.27
LS182	2	275944	78.43
LS187	2	99541	57.81
LS188	2	680068	75.91
LS189	2	613568	89.99
LS192	2	934719	79.11
LS193	2	394880	92.33
LS194	2	1054930	42.40
LS195	2	528435	31.17
LS197	2	1011972	86.25
LS199	2	1159264	60.02
LS200	2	888820	75.85
LS202	2	496917	79.72
LS203	4	1145927	34.10
LS204	2	427771	63.65
LS206	2	275195	86.10
LS238	2	1084707	59.92
LS243	2	1211597	38.18
LS244	2	1065651	81.69
LS245	2	112859	51.36
LS246	2	1069590	78.57
LS249	2	679410	75.21
LS251	2	616019	82.30
LS254	2	1062136	81.60
LS255	2	700787	77.24
LS256	2	771187	84.89
LS257	2	1294166	72.62
LS258	2	1147664	86.23
LS259	2	976499	70.42
LS260	2	983376	93.65
LS264	2	389219	83.34
LS265	2	98367	76.81
LS266	2	541334	61.38
LS270	2	124166	67.80
LS272	2	961265	89.13
LS273	2	1071641	90.43
LS274	2	300947	83.62
LS277	2	1873814	80.92

Table 2.4: The number of ploidy, number of test read and the estimate of contamination for 89 patients.

Patient	Number of Ploidy	Number of Test	Contamination
LS281	2	1051811	91.60
LS282	4	1061886	92.93
LS283	2	1602892	91.72
LS286	2	2312598	87.55
LS287	2	1224020	78.23
LS289	2	468956	82.94
LS290	2	1545591	63.78
LS291	2	2202839	58.04
LS292	2	778032	81.40
LS293	2	964380	54.49
LS294	2	1162422	83.74
LS295	2	16563	82.84
LS296	2	1586895	56.23
LS297	2	587496	88.02
LS299	2	8212772	69.65
LS300	4	1455529	88.91
LS302	2	325607	78.61
LS303	4	1523161	88.12
LS304	2	1278349	71.04
LS306	2	1514125	85.42
LS307	2	1394593	85.40
LS352	2	1001953	82.16
LS353	2	1500424	63.00
LS354	2	850667	85.70
LS355	2	1991428	60.83
LS357	2	271962	80.06
LS359	2	1624446	74.17
LS362	2	1563365	73.43
LS364	2	1862728	61.62
LS366	2	1316441	41.40
LS367	2	1107864	72.92
LS369	2	507639	39.17
LS370	2	306645	88.29
LS375	2	1252654	65.56
LS376	2	878661	78.75
LS379	2	552071	66.54
LS382	2	1554592	78.33
LS383	2	1429903	36.83
LS384	2	41823	86.07
LS387	2	2143304	66.07
LS388	2	749250	75.13

Table 2.5: The number of ploidy number of test read and the estimate of contamination for 89 patients.

2.5 Conclusion

We have applied the method to estimate CNA from patients genomic sample using the method explained in Gusnanto et al. (2012). The observed ratio (test/normal read) is not necessarily take the expected value of the normal genome, this is because the random error, different number of read, and different size, and the contamination. The random error is solved by using the smoothing method. The other problems is solved by acknowledging the multi-modality in the distribution of the segmented ratio, and correcting the location of the distribution to the corresponding different copy number. The estimation of CNA by $\hat{\rho}_{jk}$ now considers the different read depths, genome size and contamination, which makes the estimate comparable between different pairs of samples. The normalization methods was applied to the 89 patients genomic sample.

Chapter 3

Survival Analysis

3.1 Introduction

The most frequently model used in survival analysis is the Proportional Hazard (PH) model introduced by Cox (1972), which is used to explore the relationship between the survival time of patients and a set of covariates. This chapter deals with the background of the standard setting for the Cox PH model, where the number of covariates is less than the sample size, $n < p$. However, in the case of high-dimensional data where the number of covariates is larger than the sample size, $p \gg n$, various techniques have been proposed in the literature, which are not discussed in this thesis. Some examples can be found in Van Houwelingen et al. (2006); Simon et al. (2011); Bøvelstad et al. (2007, 2009).

This chapter is organized as follows. In Section 3.3 we introduce the method of estimating the regression coefficients by using partial log likelihood maximization. The Breslow estimator for estimating the baseline hazard and the corresponding survival function estimator are introduced in Section 3.4. Section 3.5 shows the method of testing the proportional hazard assumption based on a scaled Schoenfeld residual (Schoenfeld, 1982). This chapter describes the application of the standard Cox PH model using clinical characteristics only. Later, we shall explore the use of CNA data in standard Cox PH models as smooth function of CNA via additive model.

3.2 Cox Proportional Hazards (PH) Models

Cox's PH model was invented by Cox (1972), and has proven to be a popular mathematical model for the analysis of survival data, because it allows the survival probability to depend not only on time, but also on the set of covariates. Let δ_i for $i = 1, \dots, n$ be the event indicator for the i^{th} patient, where δ_i is equal to 1 if the survival time t_i is uncensored, and to zero if t_i is censored. Let $\boldsymbol{X}$ be a matrix of size $n \times p$, where the number of rows of $\boldsymbol{X}$ correspond to the number of patients, and the columns of $\boldsymbol{X}$ correspond to the clinical characteristics as continuous and categorical covariates. The i^{th} row of $\boldsymbol{X}$ is $\boldsymbol{x}_i$, which is a p vector of the covariates for the i^{th} patient. Let $h_0(t)$ be the baseline hazard function, which is describes the hazard for a hypothetical individual with all covariates being equal to zero. The hazard function for the i^{th} individual can be written as

$$h_i(t|\boldsymbol{X}) = h_0(t)\varphi(\boldsymbol{X}_i),$$

where $\varphi(\boldsymbol{X}_i)$ is a relative hazard function, that is the hazard at time t for an individual whose vector of covariate variable is $\boldsymbol{X}_i$, relative to the hazard for an individual with all covariates being equal to zero. According to the specification of the Cox model (Cox, 1972) the relative hazard function must be a non-negative function, and $h_i(t)$ is positive. A common model for $\varphi(\boldsymbol{X}_i)$ is $\exp(\boldsymbol{X}_i\boldsymbol{\beta})$ resulting in the standard proportional hazard model written as

$$h_i(t) = h_0(t)\exp(\boldsymbol{X}_i\boldsymbol{\beta}) \tag{3.1}$$

The phrase proportional hazard refers to the ratio of hazard functions corresponding to any two patients i and j, where $i \neq j$ does not depending on time t:

$$\frac{h(t|\boldsymbol{X}_i)}{h(t|\boldsymbol{X}_j)} = \frac{h_0(t)\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{h_0(t)\exp(\boldsymbol{X}_j\boldsymbol{\beta})} = \exp[(\boldsymbol{X}_i - \boldsymbol{X}_j)\boldsymbol{\beta}]. \tag{3.2}$$

Equation (3.1) has two components that need to be estimated, the first component is the unknown coefficients of the covariates variables; $\boldsymbol{\beta} = [\beta_1, \dots, \beta_p]^T$, and the second component is the baseline hazard function $h_0(t)$. The estimation of the model parameters will be discussed further in the following section.

3.3 Estimation of The Model Parameters

3.3.1 The Full Likelihood Function

The full likelihood function of the proportional hazard model can be written as

$$\begin{aligned} L(\boldsymbol{\beta}, h_0(t)) &= \prod_{i=1}^{n} [f(t_i|\boldsymbol{X}_i)]^{\delta_i} [S(t_i|\boldsymbol{X}_i)]^{1-\delta_i}, \\ &= \prod_{i=1}^{n} [h(t_i|\boldsymbol{X}_i)]^{\delta_i} S(t_i|\boldsymbol{X}_i), \end{aligned} \tag{3.3}$$

where $h(t_i|\boldsymbol{X}_i)$ is the hazard function and $S(t_i|\boldsymbol{X}_i)$ is the survival function. Replacing $h(t_i|\boldsymbol{X}_i)$ by $h_0(t_i)\exp(\boldsymbol{X}_i\boldsymbol{\beta})$, and $S(t_i|\boldsymbol{X}_i)$ by $\exp\{-H_0(t_i)\exp(\boldsymbol{X}_i\boldsymbol{\beta})\}$, the full likelihood becomes

$$L(\boldsymbol{\beta}, h_0(t)) = \prod_{i=1}^{n} \Big\{ h_0(t_i)\exp(\boldsymbol{X}_i\boldsymbol{\beta}) \Big\}^{\delta_i} \exp\Big\{ -H_0(t_i)\exp(\boldsymbol{X}_i\boldsymbol{\beta}) \Big\}. \tag{3.4}$$

The full maximum likelihood requires that we maximize (3.4) with respect to the unknown parameter $\boldsymbol{\beta}$ and the unspecified baseline hazard function $h_0(t_i)$ which is difficult without specifying the form for the baseline hazard. Cox (1972) used the conditional argument to derive the partial likelihood function, which is presented in the following subsection.

3.3.2 The Partial Likelihood Function

An important feature of the PH model is that the estimated regression parameter $\boldsymbol{\beta}$ can be obtained without specifying the baseline hazard $h_0(t)$. Cox (1972) derived a partial likelihood function for the i^{th} patient for a PH model. The derivation for the partial likelihood function can be found in Chapter 3 of Collett (2003). We assume that only one individual dies at each death time, so there are no ties in the data and there is no assumption about the form of the baseline hazard. Let $t_1 < t_2, \cdots < t_n$ denote the ordered observed follow-up times, δ_i is the event indicator, which is 1 if the individual is dead and zero otherwise, and $\boldsymbol{x}_i$ is a vector of covariates for the i^{th} individual who dies at t_i or censored. The risk set at time t, denoted by $R(t) = \{j : t_j \geq t\}$, which is the set of all individuals who are still alive and uncensored at a time just prior to t.

The probability that the i^{th} individual with covariate vector $\boldsymbol{x}_i$ dies at t_i, conditioned on that only one death in $R(t_i)$ occurring at time t_i can be express as

$$\begin{aligned}
&\Pr(\text{ individual with covariates } \boldsymbol{x}_i \text{ dies at } t_i \text{ |one death at } t_i \text{ })\\
&= \frac{\Pr(\text{ individual with covariates } \boldsymbol{x}_i \text{ dies at } t_i \text{ })}{\Pr(\text{one death at } t_i \text{ })}\\
&= \frac{\Pr(\text{ individual with covariates } \boldsymbol{x}_i \text{ dies at } t_i \text{ })}{\sum_{j\in R(t_i)} \Pr(\text{ individual with covariates } \boldsymbol{x}_j \text{ dies at } t_{(i)} \text{ })}.
\end{aligned} \tag{3.5}$$

If the probabilities of death at time t_i are replaced by the probabilities of death in the interval $(t_i, t_i + \Delta t)$, expression (3.5) becomes

$$\frac{\Pr(\text{ individual with covariates } \boldsymbol{x}_i \text{ dies in } (t_i, t_i + \Delta t) \text{ })}{\sum_{j\in R(t_i)} \Pr(\text{ individual with covariates } \boldsymbol{x}_j \text{ dies in } (t_i, t_i + \Delta t) \text{ })}, \tag{3.6}$$

Dividing the numerator and denominator in expression (3.6) by Δt, and taking the limit as $\Delta t \to 0$, we obtain the ratio of the corresponding hazards of death at time t_i:

$$\begin{aligned}
&\lim_{\Delta t \to 0} \frac{\Pr(\text{ individual with covariates } \boldsymbol{x}_i \text{ dies on } (t_i, t_i + \Delta t) \text{ })/\Delta t}{\sum_{j\in R(t_i)} \Pr(\text{ individual with covariates } \boldsymbol{x}_j \text{ dies on } (t_i, t_i + \Delta t) \text{ })/\Delta t},\\
&= \frac{\text{hazard of death at time } t_i \text{ for individual with variable } \boldsymbol{x}_i}{\sum_{j\in R(t_i)}\{\text{hazard of death at time } t_i \text{ for individual with variable } \boldsymbol{x}_j\}}\\
&= \frac{h(t_i|\boldsymbol{X}_i)}{\sum_{j\in R(t_{(i)})} h(t_i|\boldsymbol{X}_j)}\\
&= \frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{\sum_{j\in R(t_{(i)})} \exp(\boldsymbol{X}_j\boldsymbol{\beta})}.
\end{aligned}$$

The product of this conditional probability is the partial likelihood function for all observations,

$$L_{pl}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{\sum_{j\in R(t_i)} \exp(\boldsymbol{X}_j\boldsymbol{\beta})} \right)^{\delta_i},$$

The partial log-likelihood is

$$\ell_{pl}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left(\boldsymbol{X}_i\boldsymbol{\beta} - \log \sum_{j\in R(t_i)} \exp(\boldsymbol{X}_j\boldsymbol{\beta}) \right). \tag{3.7}$$

The Newton-Raphson algorithm is used to maximize the partial log-likelihood function (3.7). Let $U_{pl}(\boldsymbol{\beta})$ be the $p \times 1$ vector of the first derivative of the partial log-likelihood function (3.7) with respect to $\boldsymbol{\beta}$

$$U_{pl}(\boldsymbol{\beta}) = \frac{\partial \ell_{pl}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \delta_i \Big(\boldsymbol{X}_i - \frac{\sum_{j\in R_{(t_i)}} \boldsymbol{X}_j \exp(\boldsymbol{X}_j\boldsymbol{\beta})}{\sum_{j\in R_{(t_i)}} \exp(\boldsymbol{X}_j\boldsymbol{\beta})} \Big).$$

The negative of the matrix of the second derivative of the partial log likelihood is the information matrix, which is denoted as $I_{pl}(\boldsymbol{\beta})_{p\times p}$ and is given by

$$I_{pl}(\boldsymbol{\beta}) = -\Big[\frac{\partial^2 \ell_{pl}(\boldsymbol{\beta})}{\partial\beta_l \partial\beta_m}\Big] = \sum_{i=1}^{n} \delta_i \left(\frac{\Big(\sum_{j\in R_{(t_i)}} X_{jl} \exp(\boldsymbol{X}_j\boldsymbol{\beta})\Big)\Big(\sum_{j\in R_{(t_i)}} X_{jm} \exp(\boldsymbol{X}_j\boldsymbol{\beta})\Big)}{\Big(\sum_{j\in R_{(t_i)}} \exp(\boldsymbol{X}_j\boldsymbol{\beta})\Big)^2} \right.$$
$$\left. - \frac{\Big(\sum_{j\in R_{(t_i)}} X_{jl} \exp(\boldsymbol{X}_j\boldsymbol{\beta})\Big)\Big(\sum_{j\in R_{(t_i)}} X_{jm}X_{jl} \exp(\boldsymbol{X}_j\boldsymbol{\beta})\Big)}{\Big(\sum_{j\in R_{(t_i)}} \exp(\boldsymbol{X}_j\boldsymbol{\beta})\Big)^2} \right)$$

An estimate of the vector of the $\boldsymbol{\beta}$ parameters at the $(s+1)^{\text{th}}$ iteration is

$$\hat{\boldsymbol{\beta}}_{s+1} = \hat{\boldsymbol{\beta}}_s + I_{pl}(\hat{\boldsymbol{\beta}}_s)^{-1} U_{pl}(\hat{\boldsymbol{\beta}}_s). \tag{3.8}$$

The estimator of the variance of the estimated parameters is the inverse of the information matrix evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, is given by $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = I_{pl}^{-1}(\hat{\boldsymbol{\beta}})$, and the square root of the diagonal of the estimated variance is the standard error of $\hat{\boldsymbol{\beta}}$, denoted $\mathrm{se}(\hat{\boldsymbol{\beta}})$.

As $n \to \infty$, the $\hat{\beta}$ will converge to the true value of $\boldsymbol{\beta}$, and $\hat{\boldsymbol{\beta}}$ approximately follows a normal distribution $N(\boldsymbol{\beta}, I_p^{-1}(\boldsymbol{\beta}))$. The null hypothesis $H_0 : \boldsymbol{\beta} = 0$ can be tested using the Wald statistics, $z = (\frac{\hat{\boldsymbol{\beta}}}{\text{se}(\hat{\boldsymbol{\beta}})})^2$, which have asymptotically a χ_p^2 distribution under the null hypothesis and p is the dimension of $\boldsymbol{\beta}$. Subset selection with model selection criteria, such as the Akaike Information Criterion $AIC = -2\log \ell_{pl}(\hat{\beta}) + 2p$, can be used to identify relevant variables and choose the best model, where $\ell_{pl}(\hat{\beta})$ is the partial log likelihood for the estimated parameters. There is built-in function in R package for Cox PH model in survival package, but we wrote R code to carry out the Newton-Raphson algorithm in case of standard Cox PH model, in order to extend this later on when we include the CNA genomic-windows as smoothing terms in the model, in Chapters 5, 6 and 7.

3.4 Breslow Estimator

In this section we focus on estimating the baseline hazard, $h_0(t)$, which dropped out of the partial likelihood. Breslow (1974) proposed a method that is based on the full likelihood function by fixing $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ in equation (3.4). This method estimates the baseline hazard, so the estimated survival function can be obtained by knowing the estimate of the cumulative hazard function and the estimate of the regression parameters. Estimating hazard function can be found in Chapter 8 of Klein and Moeschberger (1997). The full likelihood with $\boldsymbol{\beta}$ equal to $\hat{\boldsymbol{\beta}}$ is a function of $h_0(t)$ only:

$$\begin{aligned} L(h_0|\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}) &= \prod_{i=1}^{n} [h_0(t_i)]^{\delta_i} [\exp(\boldsymbol{X}_i \hat{\boldsymbol{\beta}})]^{\delta_i} \exp[-H_0(t_i)\exp(\boldsymbol{X}_i\hat{\boldsymbol{\beta}})] \\ &= \left\{ \prod_{i=1}^{n} [h_0(t_i)]^{\delta_i} [\exp(\boldsymbol{X}_i \hat{\boldsymbol{\beta}})]^{\delta_i} \right\} \left\{ \prod_{i=1}^{n} \exp[-H_0(t_i)\exp(\boldsymbol{X}_i\hat{\boldsymbol{\beta}})] \right\} \\ &= \left\{ \prod_{i=1}^{n} [h_0(t_i)]^{\delta_i} [\exp(\boldsymbol{X}_i \hat{\boldsymbol{\beta}})]^{\delta_i} \right\} \left\{ \exp\left[\sum_{i=1}^{n} -H_0(t_i)\exp(\boldsymbol{X}_i\hat{\boldsymbol{\beta}}) \right] \right\}. \end{aligned} \tag{3.9}$$

Equation (3.9) has two components including $h_0(t_i)$. In the first component $[h_0(t_i)]^{\delta_i}$ when $\delta_i = 1$, $h_0(t_i)$ will be a large value, which leads to a large value in the likelihood. The second component is $\exp(-H_0(t_i)) = \exp(-\int_0^{t_i} h_0(u)du)$, so to maximize (3.9), $H_0(t_i)$ should have a small value. This can be obtained when $\hat{h}_0(t) = 0$ for all $t \notin \{t_{(1)}, \cdots, t_{(m)}\}$, where m is the number of deaths among n, so $H_0(t_i) = \sum_{t_j \leq t_i} h_0(t_{(j)})$. After some simplification, Klein and Moeschberger (1997), show that equation (3.9) can be written as

$$L(h_0(t_j)) = \left\{ \prod_{j=1}^{m} h_0(t_j) \exp(\boldsymbol{X}_j \hat{\boldsymbol{\beta}}) \right\} \exp\left\{ -\sum_{j=1}^{m} h_0(t_j) \sum_{j \in R(t_i)} \exp(\boldsymbol{X}_j \hat{\boldsymbol{\beta}}) \right\}. \quad (3.10)$$

The log likelihood is

$$\ell(h_0(t_j) = \sum_{j=1}^{m} \{\log h_0(t_j) + \boldsymbol{X}_j \hat{\boldsymbol{\beta}}\} - \sum_{j=1}^{m} h_0(t_j) \sum_{i \in R(t_j)} \exp(\boldsymbol{X}_i \hat{\boldsymbol{\beta}})\}. \quad (3.11)$$

Differentiating (3.11) with respect to $h_0(t_j)$ and setting the derivative to zero gives the maximum likelihood estimator for $h_0(t_j)$.

$$\widehat{h}_0(t_j) = \frac{1}{\sum_{i \in R(t_j)} \exp(\boldsymbol{X}_i \hat{\boldsymbol{\beta}})}. \quad (3.12)$$

The estimate of the cumulative baseline hazard function of $H_0(t)$ is

$$\widehat{H}_0(t) = \sum_{t_j \leq t} \frac{1}{\sum_{i \in R(t_j)} \exp(\boldsymbol{X}_i \hat{\boldsymbol{\beta}})}, \quad (3.13)$$

and the corresponding baseline survival function is

$$\widehat{S}_0(t|x_i) = \exp(-\hat{H}_0(t)). \quad (3.14)$$

3.5 Checking the Proportional Hazard Assumption

The proportionality of the hazard assumption is the main assumption of the Cox model. The PH assumption means that the hazard ratio between two sets of covariates is constant over time, because the common baseline hazard function cancels out in the ratio of the two hazards. Various methods have been proposed to test the proportional hazard assumption. This section presents the method of testing the proportional hazard assumption based on a scaled Schoenfeld residual (Schoenfeld, 1982), which is a measure of the difference between the observed and expected value of the covariate at each time (Therneau and Grambsch, 2000). If the proportional hazard assumption is reasonable, the Schoenfeld residual will be independent of time. This approach considers one covariate at a time, the result of which is that one p-value per covariate is included in the fitted Cox regression model. The derivation for the Schoenfeld residuals can be found in Chapter 4 of Collett (2003), which involves taking the first derivative of the partial log likelihood for the i-th covariate,

$$\tilde{r}_i(\hat{\boldsymbol{\beta}}) = \delta_i\Big(\boldsymbol{X}_i - \frac{\sum_{j\in R_{(t_i)}} \boldsymbol{X}_j \exp(\boldsymbol{X}_j\hat{\boldsymbol{\beta}})}{\sum_{j\in R_{(t_i)}} \exp(\boldsymbol{X}_j\hat{\boldsymbol{\beta}})}\Big).$$

Schoenfeld residuals are known as partial score residuals, because their sum is equals the partial log likelihood score function whose solution is the estimated model parameters, $\sum_i \tilde{r}_i(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. These residuals are plotted against time to validate the proportional hazard assumption, and if the residual falls randomly around a horizontal line centered at zero, the proportional hazard assumption is thought to hold. Otherwise, the proportional hazard assumption does not hold. We carried out the test as follows:

1. Performed the Cox proportional hazard model and obtained the Schoenfeld residual for each covariate.

2. Ranked death times as suggested by Kleinbaum and Klein (2005).

3. Tested the correlation, ρ, between the Schoenfeld residual and ranked death

times:

The null hypothesis is that the correlation, ρ, between the Schoenfeld residual and the ranked death times is zero, $H_0 : \rho = 0$. Rejection of the null hypothesis leads to the conclusion that the proportional hazard assumption is violated. An alternative was proposed by Grambsch and Therneau (1994), in which the scaling of the residuals is by done by multiplying it with the an estimate of their variance, and that approach is used in the survival package in R (Therneau, 2015; Therneau and Grambsch, 2000). Let the vector of Schoenfeld residuals for the i^{th} individual be $\tilde{\boldsymbol{r}}_i = (\tilde{r}_{1i}, \tilde{r}_{2i}, \ldots, \tilde{r}_{pi})^T$. The scaled Schoenfeld residual r^*_{ki}, is defined as

$$\boldsymbol{r}^*_i = m \ \mathrm{Var}(\hat{\boldsymbol{\beta}})\tilde{r}_i. \tag{3.15}$$

where m is the number of deaths among n individuals, and $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ is the variance covariance matrix of the estimated parameters in the Cox PH regression model.

3.6 Residual for Cox PH model

Several types of residuals have been proposed in survival analysis which can be found in may articles and books (Klein and Moeschberger, 1997; Grambsch and Therneau, 1994; Therneau and Grambsch, 2000). In this chapter we briefly introduce Cox-Snell and Martingale residuals, Cox-Snell residual is defined as (Cox and Snell, 1968)

$$r_{Ci} = \widehat{H}_0(t_i) \exp(\boldsymbol{X}_i\hat{\boldsymbol{\beta}}) \tag{3.16}$$

where $\widehat{H}_0(t_i)$ is the estimate of the baseline cumulative hazard function at time t_i. The Nelson-Aalen estimator used in practice which is given by

$$\widehat{H}_0(t_i) = -\log \hat{S}_0(t_i) = \sum_{i:t_i \leq t_j} \frac{\delta_i}{\sum_{j \in R(t_i)} \exp(\boldsymbol{X}_i\boldsymbol{\beta})}. \tag{3.17}$$

The Cox-Snell residual can be also expressed as

$$r_{Ci} = \widehat{H}_i(t) = -\log \hat{S}_i(t_i) \tag{3.18}$$

where $\widehat{S}_i(t_i) = \exp(-\widehat{H}_i(t))^{\exp(\boldsymbol{X}_i\hat{\boldsymbol{\beta}})}$. The Cox-Snell residual can be compared graphically to the cumulative hazard function of an exponential distribution with mean equal to one. The plot of the cumulative hazard estimator of the residual by the Nelson-Aalen versus the Cox-Snell should be a straight line passing through 0 with unit slope.

The Martingales residuals are defined for the i^{th} individual as:

$$r_{Mi} = \delta_i - r_{Ci} = \delta_i - \hat{H}_i(t_i). \tag{3.19}$$

where r_{Ci} is the Cox-Snell residual, and $\hat{H}_i(t_i)$ is the cumulative hazard. The residual r_{Mi} can be viewed as the difference between the observed number of deaths for the i^{th} patient between 0 and t_i and the expected numbers of death in that interval. r_{Mi}'s have mean 0, and the range of Martingale residuals is between $-\infty$ and 1. Therefore, Martingale value near to 1 represent patients died sooner and large negative values mean that the patient lived too long or it were censored. Martingale residuals are very useful as they can be used for determining the functional form of a covariate to be included in the model, this can be obtain by plotting the Martingale residual versus the continuous covariate, and then the points can be fitted using nonparametric LOESS smoother method, the fitted smooth lines should be linear to satisfy the Cox proportional hazards model assumptions.

3.7 The Cox PH model for Clinical Characteristic Data

We now analyzes the clinical (non-CNA) part of our data set using the Cox PH model. The standard Cox PH model is only applicable to the situation where the number of covariates is less than the sample size. The predictors in the model in the clinical characteristic are presented in Table 2.1. The clinical variables we consider are age, gender, tumor grade, tumor stage, and nodes status, female, G_1, T_1 and N_0 are a part of the baseline hazard. The response variable is the survival time (in days). Table 3.1 shows the 32 possible model combinations for 5 variables with AIC value for each model.

Model	Variables in the model	$-2\log\hat{\ell}_{pl}$	q	AIC
1)	intercept			478.567
2)	Age	473.504	1	475.504
3)	Gender	477.982	1	479.982
4)	Stage N	475.810	2	479.810
5)	Stage T	479.509	2	478.509
6)	Grade	477.087	2	481.087
7)	Age+Gender	473.249	2	477.249
8)	Age+Stage T	466.405	3	478.405
9)	Age+Stage N	467.225	3	473.225
10)	Age+Grade	471.734	3	477.734
11)	Gender+Stage T	472.312	3	478.312
12)	Gender+Stage N	475.530	3	481.530
13)	Gender+Grade	476.384	3	482.384
14)	Stage T+ Stage N	471.842	4	479.842
15)	Stage T+Grade	473.098	4	481.098
16)	Stage N+Grade	474.321	4	482.321
17)	Age+Gender+Stage T	464.680	4	472.680
18)	Age+Gender+Stage N	467.225	4	475.225
19)	Age+Gender+Grade	471.399	4	479.399
20)	Age+Stage T+Stage N	459.463	5	469.463
21)	Age+Stage T+Grade	464.639	5	474.639
22)	Age+Stage N+Grade	464.958	5	474.958
23)	Gender+Stage T+Stage N	470.289	5	480.289
24)	Gender+Stage T+Grade	470.982	5	480.982
25)	Stage T+Stage N+Grade	470.360	6	482.360
26)	Gender+Stage N+Grade	473.942	5	483.942
27)	Gender+Stage T+Stage N+Grade	468.871	6	482.871
28)	Age+Gender+Stage T+Stage N	458.940	6	470.940
29)	Age+Gender+Stage N+Grade	457.859	7	471.859
31)	Age+Gender+StageN+Grade	464.932	6	476.932
32)	Age+Gender+StageT+StageN+Grade	457.287	8	473.287

Table 3.1: The values of -2 log likelihood, the number of parameters q, and the AIC for each possible standard Cox PH model.

The best model is the model that has the smallest AIC value, which is the model that contains Age, Stage-T and Stage-N, this is also confirmed by the stepwise selection procedure. Table 3.2 shows the estimated parameters for the best Cox PH model that includes Age, Stage-T and Stage-N. However, the p-values of both Stage-N1 and Stage-T2 are not significant. One way to solve this is to combine the levels of Stage-T1 with Stage-T2, and level Stage-N1 with Stage-N2. This will give a model with significant p-values as shown in Table 3.2.

Parameter	Estimate	exp(Estimate)	Standard error	p-value
Age	0.05	1.05	0.03	0.02
Stage-N1	0.34	1.41	0.28	0.22
Stage-N2	1.33	3.80	0.48	0.01
Stage-T2	0.15	1.16	0.30	0.62
Stage-T3	1.80	6.05	0.57	0.00
(combining level of stages)				
Age	0.05	1.05	0.01	0.00
Stage-N2	1.17	3.23	0.45	0.01
Stage-T3	1.83	6.21	0.53	0.00

Table 3.2: The estimated coefficients, exponential estimated coefficients, and standard errors of estimated coefficients and the p-value for the fitting a proportional hazard model that includes Age, Stage-T and Stage-N.

The model for the ith patients in the first part in the Table 3.2 can be expressed as

$$\begin{aligned} h_i(t) = h_0(t)\exp\{&0.05\,\mathrm{Age}_i + 0.34\,\mathrm{StageN1}_i + 1.34\,\mathrm{StageN2}_i + 0.15\,\mathrm{StageT2}_i \\ &+ 1.80\,\mathrm{StageT3}_i)\} \end{aligned} \tag{3.20}$$

The interpretation of the model is that the positive estimate of Stage-N2 means that the wider spread of cancer cells near the lymph nodes will increase the hazard level, the hazard for patients having Stage-N1 and Stage-N2 are 1.41 and 3.80 times that for those of who are in Stage-N0. Similarly, the positive estimate of Stage-T3 indicates that large tumor size increases the hazard risk, so those patients who are in Stage-T2 and at Stage-T3 have a hazard about 1.16 and 6.05 times that of those who are in Stage-T1. For all patients every year of age increases hazard by 5%.

The test and graphical diagnostics for the PH model based on the Schoenfeld residual are applied to the best model, which gives the result shown in Table 3.3 and Figure 3.1.

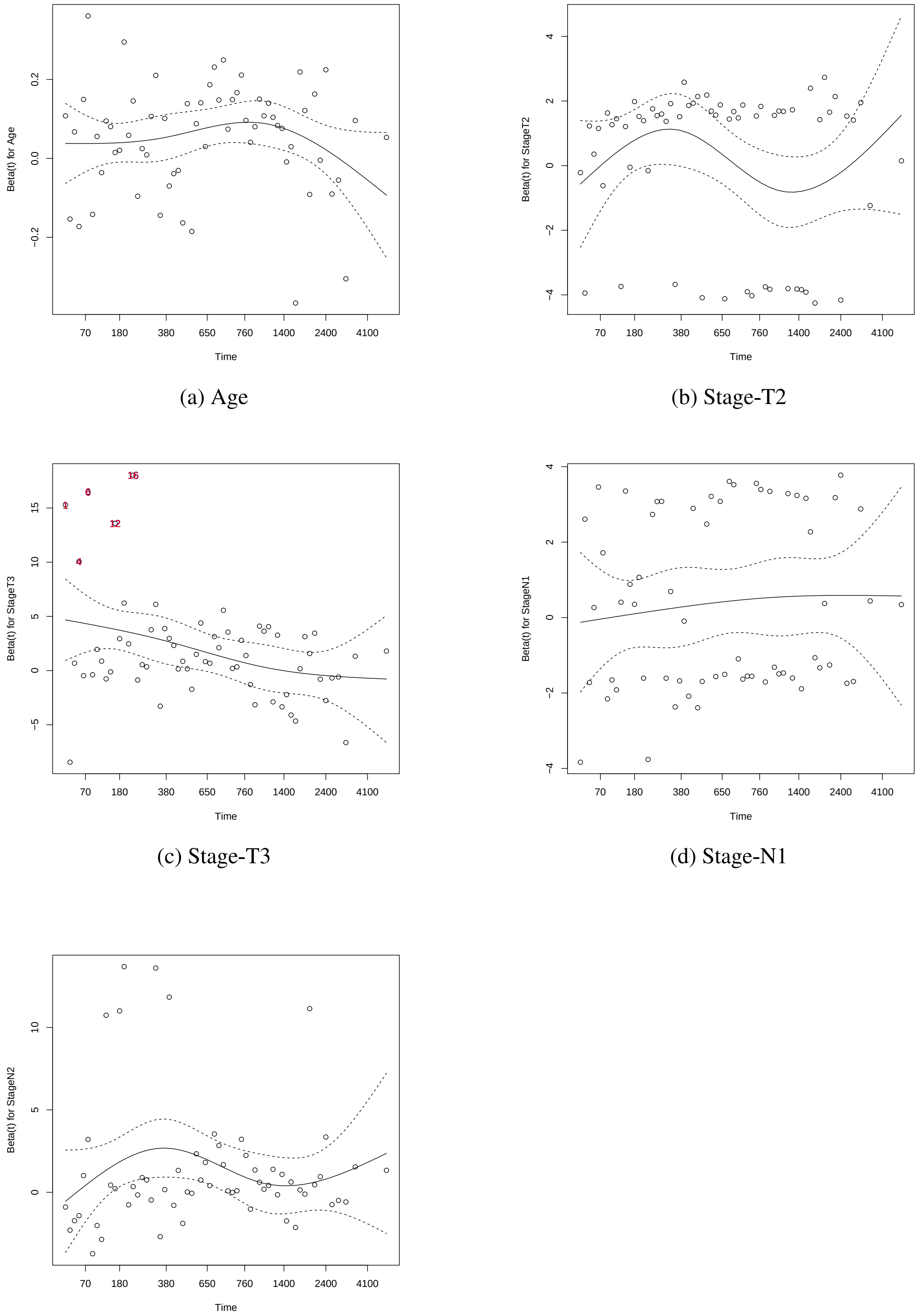
Parameter	ρ	χ^2	p-value
Age	-0.0215	0.0356	0.85043
Stage-T2	-0.1126	0.7898	0.37415
Stage-T3	-0.3600	8.9400	0.00279
Stage-N1	0.0952	0.5835	0.44493
Stage-N2	-0.0387	0.0941	0.75899
GLOBAL		9.7663	0.08213

Table 3.3: The p-values from testing the Cox PH assumption for the model that contains Age, Stage-T and Stage-N.

The Pearson product-moment correlation between the scaled Schoenfeld residual and time for each covariates in the best model in ρ column, the test statistic in the χ^2 column, and $p-$value in the third column. The Global test for all the covariates is 0.082 which indicates no evidence for non proportionality. However, Stage-T3 shows non proportionality as indicate by the p-value in Table 3.3. This is because Stage-T3 represent the size of the tumour larger than 7 cm which is the most severe case. The proportional hazards assumption states that the hazard of any variables must be constant. That is, hazard of Stage-T3 should not fluctuate with time. Figure 3.1 (c) shows the plot of scaled Schoenfeld residuals against time for Stage-T3, There are five outliers hazard associated with Stage-T3 between 34 and 199 days. Table 3.4 represents the informations of five hazards outliers.

#	survival	survival	Age	Stage-T2	Stage-T3	Stage-N1	Stage-N2
#	time	status					
1	34	1	66	0	1	0	0
4	54	1	45	0	1	1	0
6	81	1	80	0	1	1	0
12	179	1	55.	0	1	1	0
16	199	1	65	0	1	0	0

Table 3.4: The information of five hazards outliers.



(a) Age

(b) Stage-T2

(c) Stage-T3

(d) Stage-N1

(e) Stage-N2

Figure 3.1: Plots of scaled Schoenfeld residuals against time for each covariate in the best model. The solid line is a smoothing-spline fit to the plot, the dashed line is the ± 2 standard error band around the fit.

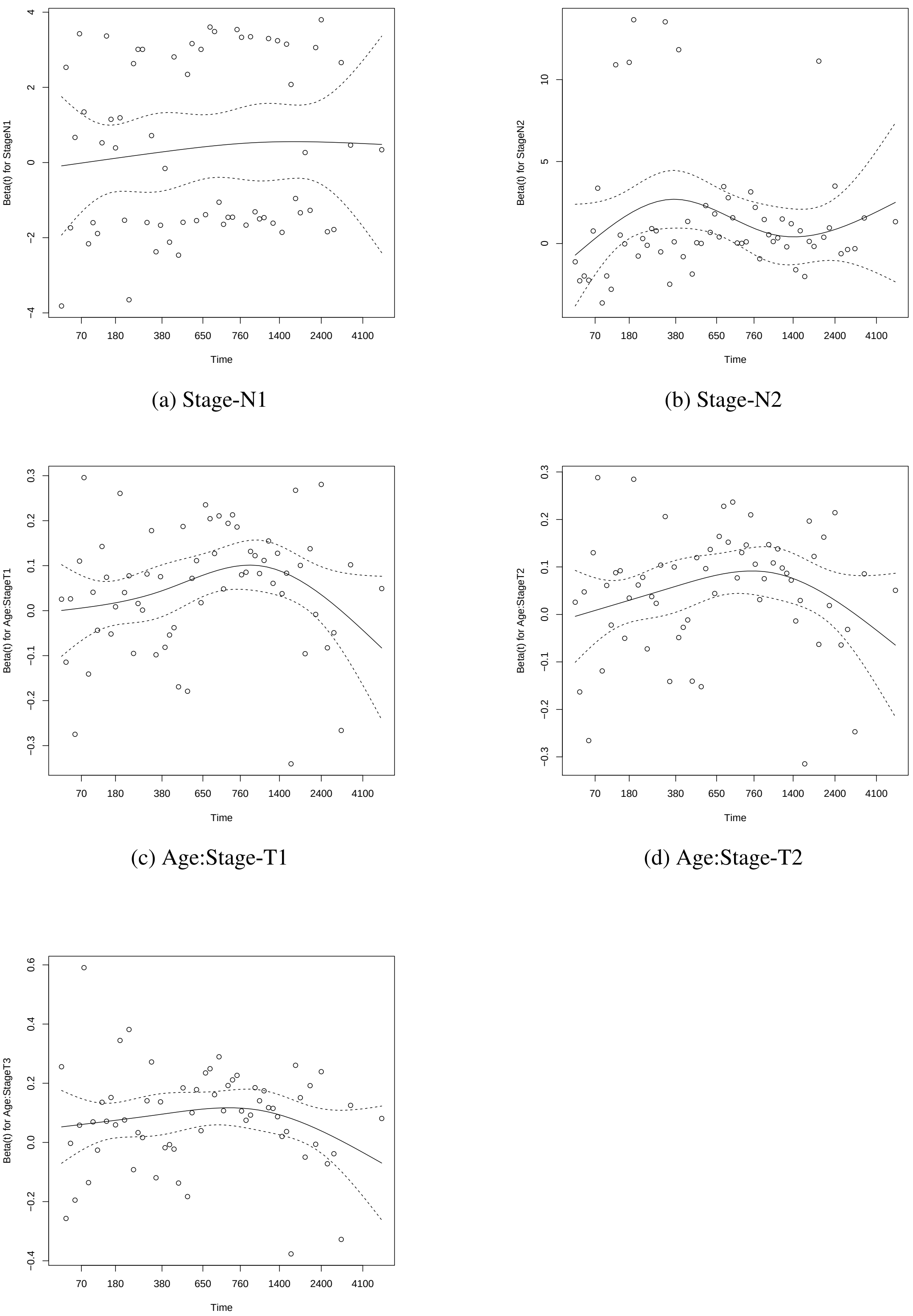
Schemper (1992) discussed the consequences of violated PH assumptions for Cox's proportional hazards model. One possible consequences of violated PH assumptions is that the relative risk for covariates with hazard ratios increasing over time is overestimated and for covariates whose hazard ratios are non-constant over time, the power of corresponding tests decreases because of suboptimal weights for combining the information provided by the risk sets of times where failures occur. To solve this problem we built up a model that contained an interaction Age with Stage-T, and Stage-N, the result of test and graphical diagnostics for the PH model based on the Schoenfeld residual are shown in Table 3.5 and Figure 3.2.

Parameter	ρ	χ^2	p-value
StageN1	0.0863	0.4708	0.493
StageN2	-0.0258	0.0423	0.837
Age:StageT1	0.0952	0.6719	0.412
Age:StageT2	0.0587	0.2458	0.620
Age:StageT3	-0.0456	0.1574	0.692
GLOBAL		6.1530	0.292

Table 3.5: The p-values from testing the Cox PH assumption for the model that contains Age:Stage-T and Stage-N.

The Global test for all the covariates is 0.292 which indicates no evidence for non proportionality. This result can be added to the previous investigation which allows us to believe that the Stage-T3 can be satisfied with the PH assumption.

Martingale and Cox-Snell residuals for a best model fitted are computed. Figure 3.3 shows the plot of the Martingale residuals against continuous covariates with a nonparametric LOESS-smoother, which is a common approach used to assess the functional form of a covariate. The resulting Cox-Snell residualsl plot is appear in Figure 3.5.



(a) Stage-N1

(b) Stage-N2

(c) Age:Stage-T1

(d) Age:Stage-T2

(e) Age:Stage-T3

Figure 3.2: Plots of scaled Schoenfeld residuals against time for each covariate in the best model. The solid line is a smoothing-spline fit to the plot, the dashed line is the ± 2 standard error band around the fit.

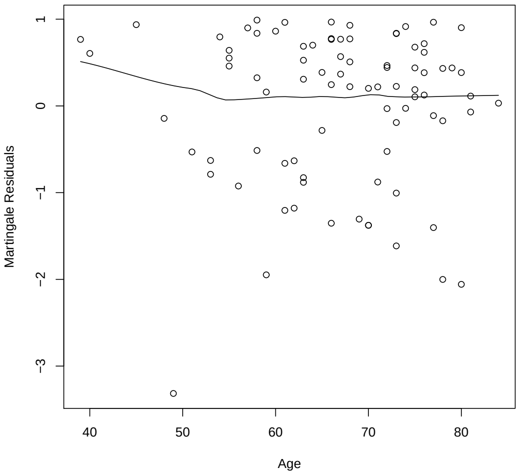


Figure 3.3: Plots of the martingale residuals versus Age with smooth curve the black solid line.

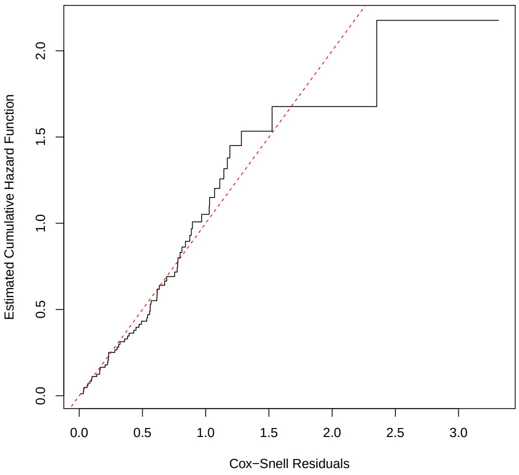


Figure 3.4: Plots of the estimated cumulative hazard function versus the of Cox-Snell residual for the best model, the red solid line is the identity line.

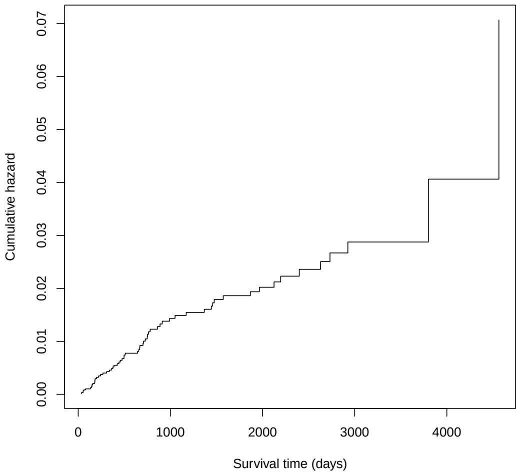


Figure 3.5: Plots of the estimated cumulative hazard function versus survival time (in days).

3.8 Conclusion

In this chapter, we introduced the standard Cox PH model when the number of covariates is less than the sample size. The model parameters are estimated by maximizing the partial log likelihood function, the estimated parameters have an asymptotic normal distribution with mean equal to the estimated parameters and covariance matrix given by the inverse observed Fisher information matrix at the estimated parameters. The estimate baseline hazard function can be obtained using the Breslow estimator, and survival function can be obtained plugging in the Breslow estimator and the estimated parameters. The Schoenfeld residuals are used to testing the PH assumption.

The clinical data of NSCLC were included in the Cox PH model, we fit 32 possible models, each model containing one or more variables. The variable selection is performed by AIC to choose the best model, the best model includes Age, and Stage-T and Stage-N shown in Table 3.2. The results show that the positive estimate of Stage-N2 and Stage-T leads to an increase the hazard level, a 5% increase in hazard level for every 1-year older in age. The PH assumption is satisfied for Age and Stage N_2 but violated for Stage T_3.

Chapter 4

Generalized Additive Model

4.1 Introduction

The Generalized Additive Model (GAM), which was first proposed by Hastie and Tibshirani (1990b), is a Generalized Linear Model (GLM) McCullagh and Nelder (1989) where the linear predictor includes smooth functions of some or all of the covariates. GAM assumes that the mean of the response variable depends on an additive predictor through a link function, and the response probability distribution is a member of the exponential family of distributions. Various methodologies in the literature for fitting a GAM have been developed. The first method was the back-fitting algorithm proposed by Hastie and Tibshirani (1990b), which is based on an iterative procedure of smoothing partial residuals in order to estimate each smooth model component. The generalized smoothing spline approach Wahba (1990); Green and Silverman (1994) is another method of fitting GAM via a penalized regression smoothing spline approach, which is considered in this chapter. The theory of penalized regression smoothing is explained in Wood (2006). The penalized regression spline approach consists of three main steps:

1. Representation the smooth term with a penalized regression spline.

2. Estimating the model coefficients by using penalized log likelihood maximiza-

tion.

3. Estimating the smoothing parameters by minimization of the Generalized Cross Validation (GCV) score.

Complexities arise from the number of predictors, p, being larger than the sample size n; Marra and Wood (2011) presented the variable selection methods of GAM.

Most of the definitions and methods here are drawn from Wood (2006). The layout of this chapter is as follows: an overview of the generalized additive model is briefly presented in Section 4.2. The representation of the smoothing term in GAM using penalized regression splines is presented in Section 4.3. The purpose of this section is to present how the basis functions and the roughness penalty are constructed, for simplicity, the one smoothing term is presented, which can be extended to more than one smoothing terms. Fitting the GAM model is discussed in Section 4.4. The method of estimating the smoothing parameter is also presented in Section 4.5. Choosing the optimal number of knots by minimizing GCV is introduced in Section 4.6. Testing the hypothesis that each smoothing term in the model is equal to zero is introduced in Section 4.7. Finally, a logistic regression with the GAM smoothing term for the clinical characteristic of the NSCLC dataset is presented in Section 4.8.

4.2 Generalized Additive Model Overview

The response variable y_i, $i = 1, \ldots, n$, is an independent observation from a distribution belonging to an exponential family, and $x_{1i}, x_{2i}, \ldots, x_{qi}$ are the covariates. Generalized additive model proposes that the mean of the response variable y_i is linked to an additive effect of the covariate variables via a known link function. The GAM can be expressed as:

$$g(\mu_i) = X_i^L \boldsymbol{\zeta} + \sum_{j=1}^{p} f_j(x_j), \tag{4.1}$$

where $g(\mu_i)$ is a known link function, which describes how the mean, $\mu_i = E(y_i)$ depends on the additive predictor $\eta_i = g(\mu_i)$.

The parametric model matrix of size $n \times q$ for any parametric component such as the intercept, or categorical covariates is denoted by $\boldsymbol{X}^L$, where the i^{th} row of the parametric matrix $\boldsymbol{X}^L$ is $\boldsymbol{X}_i^L$, and the corresponding unknown vector of parametric parameters is $\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_q]^T$, $f_j(x_{j_i})$ are unknown smooth functions of the j^{th} covariates x_{j_i} that may be a vector value. Various penalized regression smoothers can be used to represent the smooth functions $f_j(x_{j_i})$, such as cubic regression spline, cubic B-spline, truncated polynomial spline, or radial basis function, for representing a single covariate.

The main challenge is how to estimate $f_j(x_j)$. The general idea is to determine a basis function b_j, so that the j^{th} smoothing function can be presented as

$$f_j(x_j) = \sum_{k=1}^{q_j} b_{jk}(x_j)\beta_{jk},$$

where x_j is a vector of the j^{th} covariate, β_{jk} are the coefficients of the j^{th} smoothing function that need to estimated, and q_j is a number of knots or a number of basis function, so estimating f_j is equivalent to estimating the coefficients β_{jk}. The j^{th} smoothing term can be expressed in vector-matrix notation as

$$\boldsymbol{f}_j = \boldsymbol{X}_j\boldsymbol{\beta}_j,$$

where $\boldsymbol{f}_j$ is the vector of the j^{th} smooth term with $\boldsymbol{f}_{j_i} = f_j(x_{j_i})$, $\boldsymbol{X}_j$ is the j^{th} smooth matrix of size $n \times q_j$, where the i^{th} row of the the j^{th} smooth matrix $\boldsymbol{X}_j$ is $\boldsymbol{X}_{ji} = \{b_{j1}(x_{j_i}), \ldots, b_{jq_j}(x_{j_i})\}$, and the vector of the coefficients for the j^{th} smoothing term is $\boldsymbol{\beta}_j = [\beta_{j1}, \beta_{j2}, \ldots, \beta_{jq_j}]^T$, Therefore, We can write each smooth function in the model in terms of its introduced smooth matrix.

The model (4.1) is not identifiable. To solve this problem each smooth function is subject to centering constraints so that the sum of $\boldsymbol{f}_j$ is equal to zero, more details can be found in (Wood, 2006) Section 4.2. Having centered the model matrices for the smoothing functions, model (4.1) can be expressed as

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta}, \tag{4.2}$$

where $\boldsymbol{\beta}^T = [\boldsymbol{\zeta}^T, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T]$, and $\boldsymbol{X} = [\boldsymbol{X}^L, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p]$ is the model matrix, and $\boldsymbol{X}_i$ is the i^{th} row of the model matrix $\boldsymbol{X}$.

Model (4.2) can be represented as a GLM, and the parameters $\boldsymbol{\beta}$ can be estimated by standard likelihood maximization, which can be obtained using the Iterative Reweighted Least Square (IRLS) method. However, if q_j is large, then the model will be overfitted, so introducing a penalty for each smoothing term can solve the problem of smoothing, so the model parameter can be estimated by penalized likelihood maximization. The penalty can be presented in terms of the integrated square second derivative of the smoothing function. The penalized likelihood can be written as

$$\ell_{pen}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j=1}^{p}\lambda_j \int [f_j''(x_j)]^2 dx_j, \tag{4.3}$$

where $\ell(\boldsymbol{\beta})$ is the log likelihood of the model, and $\lambda_j > 0$ are the smoothing parameters that control the smoothness, the fraction of 1/2 is included for the convenience of the derivative representation. The integration is over the range of x_j, and $\int [f_j''(x_j)]^2 dx_j$ is a measure of the total change in the function $f_j''(x_j)$, over the range of x_j, so $\lambda_j \int [f''(x)]^2 dx_j$ encourages $f_j(x_j)$ to be smooth. However, larger value of λ_j, will lead to a smoother $f_j(x_j)$. The penalty $\sum_{j=1}^{p}\lambda_j \int [f_j''(x_j)]^2 dx_j$ can be expressed as a quadratic form in $\boldsymbol{\beta}$. Note that the penalty form depends on the basis function that we choose.

$$\begin{aligned}
\int [f_j''(x_j)]^2 dx_j &= \int \left[\frac{\partial^2 f_j(x_j)}{\partial x_j^2}\right] dx_j = \int \left[\frac{\partial^2 \sum_{k=1}^{q_j} b_{jk}(x_j)\beta_{jk}}{\partial x_j^2}\right] dx_j \\
&= \int \left[\boldsymbol{\beta}_j^T \boldsymbol{b}_j''(x_j)\right]^2 dx_j = \int \boldsymbol{\beta}_j^T \boldsymbol{b}_j''(x_j)\boldsymbol{b}_j''(x_j)^T \boldsymbol{\beta}_j dx_j \\
&= \boldsymbol{\beta}_j^T \left[\int \boldsymbol{b}_j''(x_j)\boldsymbol{b}_j''(x_j)^T dx_j\right]\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^T \boldsymbol{S}_j \boldsymbol{\beta}_j,
\end{aligned}$$

where $\boldsymbol{b}_j''(x_j)$ is a vector involving the second derivatives of the basis function for the j^{th} smoothing term with respect to the j^{th} covariate x_j, which follows that

$$\sum_{j=1}^{p} \lambda_j \int [f_j''(x_j)]^2 dx_j = \sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_j^T \boldsymbol{S}_j \boldsymbol{\beta}_j.$$

The penalized likelihood can be expressed as

$$\ell_{pen}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2} \sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_j^T S_j \boldsymbol{\beta}_j, \tag{4.4}$$

where $\boldsymbol{S}_j$ is the penalty matrix with known elements. For given values of the smoothing parameters λ_j, and defining the block diagonal matrix $\boldsymbol{S} = \sum_j \lambda_j \boldsymbol{S}_j$, the penalized log likelihood function can be written as;

$$\ell_{pen}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}. \tag{4.5}$$

The model parameter vector $\boldsymbol{\beta}$ can be estimated by maximizing the penalized likelihood, which can be done using the Penalized Iteratively Re-weighted Least Square (PIRLS) method. However, before coming to the details of the model parameters estimation, we would like to explain how the smoothing term, and the penalty are constructed. Two basis functions are discussed here: the radial basis function and cubic splines, which are used in this thesis. The radial basis function is mainly used in Chapter2 5, 6 and 7, while the cubic spline is used in this chapter.

4.3 Spline Basis and Penalties

4.3.1 Radial Basis Functions

The primary references for this section are Ruppert et al. (2003), Lancaster and Šalkauskas (1986), Wood and Augustin (2002), and Wood (2006). Radial basis function is defined in Ruppert et al. (2003) as

$$f(x_i) = \alpha_1 + \alpha_2 x_i + \sum_{k=1}^{n_k} \alpha_{1k} |x_i - x_k^*|^3, \quad \text{subject to} \quad \sum_{k=1}^{n_k} \alpha_{1k} = \sum_{k=1}^{n_k} \alpha_{1k} x_k{}^* = 0, \quad (4.6)$$

where x_k^* for $k = 1, \ldots, n_k$ are the knots in the range of x_i, and n_k is the number of the knots. Ruppert (2002) suggests that the location and number of knots are fixed in advance, we used the evenly spaced knots through the range of observed x_i. The constraints are imposed on the parameters of the third part of $f(x_i)$. These constraints come from the natural cubic spline constraints, which means that the smooth function is linear in the tails on the boundary knots by conditioning $f''(x_1^*) = f''(x_{n_k}^*) = 0$, more information is provided in the derivation of the penalty term in equation (4.12). The radial basis function can be written as matrix-vector notation, such that $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{\beta}_c$ where $\boldsymbol{X}$ is a $n \times (n_k + 2)$ smooth matrix, such that $\boldsymbol{X} = [\boldsymbol{1}, \breve{\boldsymbol{X}}]$, and the i^{th} row of the matrix $\breve{\boldsymbol{X}}$ is $\breve{\boldsymbol{X}}_i = [x_i, |x_i - x_1^*|^3, |x_i - x_2^*|^3, \ldots, |x_i - x_{n_k}^*|^3]$, which depends only on the distance between the observation and the knots. The smooth matrix $\boldsymbol{X}$ can be constructed as

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & |x_1 - x_1^*|^3 & |x_1 - x_2^*|^3 & \ldots & |x_1 - x_{n_k}^*|^3 \\ 1 & x_2 & |x_2 - x_1^*|^3 & |x_2 - x_2^*|^3 & \ldots & |x_2 - x_{n_k}^*|^3 \\ \vdots & \vdots & \vdots & \vdots & & \\ 1 & x_n & |x_n - x_1^*|^3 & |x_n - x_2^*|^3 & \ldots & |x_n - x_{n_k}^*|^3 \end{bmatrix}, \qquad (4.7)$$

and let $\boldsymbol{\beta}_c = [\alpha_1, \alpha_2, \alpha_{11}, \ldots, \alpha_{1n_k}]^T$ be the $n_k + 2$ vector of unknown constraint parameters, which is subject to the constraints, $\boldsymbol{C}\boldsymbol{\beta}_c = \boldsymbol{0}$, where $\boldsymbol{C}$ is a $2 \times (n_k + 2)$

matrix. These constraints can be expressed as;

$$\boldsymbol{C}\boldsymbol{\beta}_c = \begin{bmatrix} 0 & 0 & 1 & 1 & \cdots & 1 \\ 0 & 0 & x_1^* & x_2^* & \cdots & x_{n_k}^* \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_{11} \\ \vdots \\ \alpha_{1n_k} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{4.8}$$

The radial basis function involves $n_k + 2$ parameters and two linear constraints. The general linear constraints approach can be used to solve the constraints problem, which can be found in Wood (2006) Section 1.8.1. The main idea is to re-write the model in terms of n_k unconstrained parameters. QR decomposition can only be applied to the constraints matrix $\boldsymbol{C}$ if the number of rows is greater than the number of columns, therefore we consider the $\boldsymbol{C}^T$ instead of $\boldsymbol{C}$. Let $\boldsymbol{C}^T$ be

$$\boldsymbol{C}^T = \boldsymbol{Q} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{bmatrix}, \tag{4.9}$$

where $\boldsymbol{Q}$ is $(n_k+2) \times (n_k+2)$ square and orthogonal matrix such that $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}_{n_k+2}$ and $\boldsymbol{R}$ is a 2×2 upper triangular matrix. $\boldsymbol{Q}$ can be partitioned as $\boldsymbol{Q} = [\boldsymbol{D} : \boldsymbol{Z}]$, where $\boldsymbol{Z}$ is a semi-orthogonal matrix of size $(n_k + 2) \times (n_k)$, where $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{I}$, such that

$$\boldsymbol{C}^T = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{Z} \end{bmatrix} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{bmatrix}.$$

The aim is to find the orthogonal matrix $\boldsymbol{Q}$, such that

$$\boldsymbol{C}\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{R}^T & 0 \end{bmatrix},$$
$$\boldsymbol{C} \begin{bmatrix} \boldsymbol{D} : \boldsymbol{Z} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}^T & 0 \end{bmatrix}.$$

This means that $\boldsymbol{CZ} = \mathbf{0}$ and $\boldsymbol{CD} = \boldsymbol{R}^T$. Let

$$\boldsymbol{\beta}_c = \boldsymbol{Z}\boldsymbol{\beta}_z, \tag{4.10}$$

where $\boldsymbol{\beta}_z$ is an n_k vector of unknown parameters, then $\boldsymbol{C}\boldsymbol{\beta}_c = \mathbf{0}$ for any $\boldsymbol{\beta}_z$ since

$$\boldsymbol{C}\boldsymbol{\beta}_c = \begin{bmatrix} \boldsymbol{R}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{D}^T \\ \boldsymbol{Z}^T \end{bmatrix} \boldsymbol{Z}\boldsymbol{\beta}_z = \begin{bmatrix} \boldsymbol{R}^T & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \boldsymbol{I} \end{bmatrix} \boldsymbol{\beta}_z = \mathbf{0}.$$

The radial basis function can be expressed as

$$\boldsymbol{f} = \boldsymbol{X}\boldsymbol{Z}\boldsymbol{\beta}_z. \tag{4.11}$$

Derivation of the Penalty Term

The radial basis function is defined in equation (4.6), which can be rewritten as;

$$f(x_i) = \alpha_1 + \alpha_2 x_i + \sum_{k=1}^{n_k} \alpha_{1k} g_k(x_i),$$

where $g_k(x_i) = |x_i - x_k^*|^3$; this may be rewritten as

$$g_k(x_i) = \text{sign}(x_i - x_k^*)(x_i - x_k^*)^3.$$

The first and second derivatives of $f(x_i)$ with respect to x_i are

$$f'(x_i) = \alpha_2 + 3\sum_{k=1}^{n_k} \alpha_{1k}\text{sign}(x_i - x_k^*)(x_i - x_k^*)^2,$$
$$f''(x_i) = 6\sum_{k=1}^{n_k} \alpha_{1k}\text{sign}(x_i - x_k^*)(x_i - x_k^*).$$

The natural spline constraints require $f''(x_1^*) = f''(x_{n_k}^*) = 0$, therefore

$$-6\sum_{k=1}^{n_k}\alpha_{1k}(x_1^* - x_k^*) = 0 \quad \text{and} \quad 6\sum_{k=1}^{n_k}\alpha_{1k}(x_{n_k}^* - x_k^*) = 0. \tag{4.12}$$

Solving these two simultaneous equations gives $\sum_{k=1}^{n_k}\alpha_{1k} = \sum_{k=1}^{n_k}\alpha_{1k}x_k^* = 0$. Therefore, this is equation (4.6)

$$f(x_i) = \alpha_1 + \alpha_2 x_i + \sum_{k=1}^{n_k}\alpha_{1k}|x_i - x_k^*|^3; \qquad \sum_{k=1}^{n_k}\alpha_{1k} = \sum_{k=1}^{n_k}\alpha_{1k}x_k^* = 0$$

Since f is a natural cubic spline, $f''(x_i) = 0$ for all $x_i < x_1^*$ and for all $x_i > x_{n_k}^*$. Therefore, for any value $a > \max(|x_1^*|, |x_{n_k}^*|)$,

$$\int_{-\infty}^{\infty}[f''(x)]^2 dx = \int_{-a}^{a}[f''(x)]^2 dx.$$

For any twice differentiable function $h(x)$ and for any $s \in [-m, m]$, where s is a single knot, we have

$$\begin{aligned}
\int_{-a}^{a} h''(x_i)(\text{sign}(x_i - s)\ (x_i - s)^3)'' dx &= 6\int_{-a}^{a} h''(x_i)(\text{sign}(x_i - s)(x_i - s))dx_i \\
&= 6\int_{-a}^{s} h''(x_i)\text{sign}(x_i - s)(x_i - s)dx_i \\
&+ 6\int_{s}^{a} h''(x_i)\text{sign}(x_i - s)(x_i - s)dx_i \\
&= -6\int_{-a}^{s} h''(x_i)(x_i - s)dx_i + 6\int_{s}^{a} h''(x_i)(x_i - s)dx_i \\
&= -6\left([h'(x_i)(x_i - s)]_{-a}^{s} - \int_{-a}^{s} h'(x_i)dx_i\right) \\
&+ 6\left([h'(x_i)(x_i - s)]_{s}^{a} - \int_{s}^{a} h'(x_i)dx_i\right) \\
&= 6[h(m)(m - s) - h(-m)(-m - s) + 2h(s)].
\end{aligned}$$

In order to compute $[f''(x)]^2$ we let

$$\begin{aligned}\int_{-m}^{m} h''(x_i)f''(x_i)dx &= \int_{-m}^{m} h''(x_i)6\sum_{k=1}^{n_k}\alpha_{1k}\text{sign}(x_i - s)(x_i - s)dx_i,\\ &= 6\sum_{k=1}^{n_k}\alpha_{1k}\int_{-m}^{m} h''(x_i)(\text{sign}(x_i - s)(x_i - s))dx.\end{aligned}$$

Using the above result, and setting $s = x_k^*$ gives

$$\begin{aligned}\int_{-m}^{m} h''(x_i)f''(x_i)dx &= 6\sum_{k=1}^{n_k}\alpha_{1k}\left(6[\underbrace{(m-s)h'(m)}_{=0} - \underbrace{(-m-s)h'(-m)}_{=0} + \underbrace{2h(s)}_{\text{setting} s=x_k^*}\ ]\right)\\ &= 6\sum_{k=1}^{n_k}\alpha_{1k}2h(x_k^*)\\ &= 12\sum_{k=1}^{n_k}\alpha_{1k}h(x_k^*).\end{aligned}$$

Setting $h(x_i) = f(x_i)$ gives;

$$\begin{aligned}\int_{-\infty}^{\infty}[f''(x_i)]^2dx &= 12\sum_{k=1}^{n_k}\alpha_{1k}f(x_k^*)\\ &= 12\sum_{k=1}^{n_k}\alpha_{1k}\Big[\alpha_1 + \alpha_2 x_k^* + \sum_{j=1}^{n_k}\alpha_{jk}|x_k^* - x_j^*|^3\Big]\\ &= 12\sum_{k=1}^{n_k}\underbrace{\alpha_{1k}(\alpha_1 + \alpha_2 x_k^*)}_{=0} + 12\sum_{k=1}^{n_k}\alpha_{1k}\sum_{j=1}^{n_k}\alpha_{1j}|x_k^* - x_j^*|^3\\ &= 12\sum_{k=1}^{n_k}\sum_{j=1}^{n_k}\alpha_{1k}\alpha_{1j}|x_k^* - x_j^*|^3\\ &= \boldsymbol{\beta}_1^T\boldsymbol{S}\boldsymbol{\beta}_1,\end{aligned}$$

where $\boldsymbol{\beta}_1 = [\alpha_{11},\ldots,\alpha_{1n_k}]^T$ is the corresponding parameters, and $\boldsymbol{S}$ is the $n_k \times n_k$ matrix with (k, j) element being $12|x_k^* - x_j^*|^3$. Commonly the penalty can be expressed in a quadratic form of the full parameters vector as follows

$$\boldsymbol{K} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times n_k} \\ \mathbf{0}_{n_k\times 2} & \boldsymbol{S}_{n_k\times n_k} \end{bmatrix}_{(n_k+2)\times(n_k+2)}, \tag{4.13}$$

where $\boldsymbol{K}$ is a square penalty matrix of size $(n_k + 2) \times (n_k + 2)$, while the first two rows and columns of $\boldsymbol{K}$ are zero. The penalty matrix can be expressed in a quadratic form of the full parameters vector as

$$\int \left[f''(x)\right]^2 dx = \boldsymbol{\beta}_z^T \boldsymbol{Z}^T \boldsymbol{K} \boldsymbol{Z} \boldsymbol{\beta}_z. \tag{4.14}$$

4.3.2 Cubic Splines

Cubic splines are constructed by using sections of cubic polynomials joined at the knots, so that they are continuous up to and including the second derivative. The derivation for the cubic spline function can be found in Lancaster and Šalkauskas (1986) and Poirier (1973). The aim in this section is to construct a cubic function $f(x)$, where the number of knots is less than the number of data points, or the location of the knots does not correspond with y. However, the case of the number of knots being equal to the data points is discussed in Green and Silverman (1994). Let a set of knots $\{x_j^*\}, \;\; j = 1, \ldots, n_k$ be given which satisfy $x_1^* \le x_2^* \le \cdots \le x_{n_k}^*$, where x_1^* and $x_{n_k}^*$ are the minimum and maximum value of the data point respectively. On each interval $[x_j^*, x_{j+1}^*]$ for $j = 1, \ldots, n_k - 1$, the cubic piecewise function $f_j(x)$ is defined. The result of connecting all these piecewise functions is a curve $y = f(x)$, whose first and second derivatives are both continuous on the interval $[x_1^*, x_{n_k}^*]$. Most of the definitions and methods here are drawn from Lancaster and Šalkauskas (1986). The spline $f(x)$ is created by considering a piecewise cubic function $f_j(x)$ for each interval $[x_j^*, x_{j+1}^*]$, which can be defined as

$$f(x) = \begin{cases} f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3, & x_1^* \leq x \leq x_2^*, \\ \vdots & \vdots \\ f_{n_k-1}(x) = a_{n_k-1} + b_{n_k-1} x + c_{n_k-1} x^2 + d_{n_k-1} x^3, & x_{n_k-1}^* \leq x \leq x_{n_k}^*. \end{cases}$$

There are $n_k - 1$ segments of the piecewise cubic function, where each segment has 4 unknown coefficients. Requiring continuity and smoothness at the knots will force the the cubic spline to interpolate at the knots. The strategy is to first construct the linear spline to interpolate $f_j''(x_i)$ and then integrate that twice to obtain $f_j(x_i)$. Since $f(x_i)$ is a set of cubic polynomials with a continuous second derivative, $f_j''(x_i)$ is the linear interpolating spline over (x_j^*, z_j^*). Let $f_j''(x_j^*) = z_j^*$, where z_j^* is the value of the second derivative of the spline at the knot. The natural splines condition on the boundary knot is used, so this implies $f_1''(x_1^*) = z_1^* = 0, \;\; f_{n_k-1}''(x_{n_k}^*) = z_{n_k}^* = 0$, although the other z_j^* are unknown and need to be estimated; let

$$f_j''(x_i) = f_j''(x_j^*) + m_j(x_i - x_j^*), \tag{4.15}$$

where the slope is $m_j = \frac{z_{j+1}^* - z_j^*}{x_{j+1}^* - x_j^*} = \frac{z_{j+1}^* - z_j^*}{h_j}$ and $h_j = x_{j+1}^* - x_j^*$, then

$$f_j''(x_i) = \frac{(x_{j+1}^* - x_i) z_j^*}{h_j} + \frac{z_{j+1}^*(x_i - x_j^*)}{h_j}. \tag{4.16}$$

Integrate equation (4.16) twice to obtain $f_j(x_i)$:

$$f_j'(x_i) = -z_j^* \frac{(x_{j+1}^* - x_i)^2}{2h_j} + z_{j+1}^* \frac{(x_{j+1}^* - x_i)^2}{2h_j} + E_j, \tag{4.17}$$

$$f_j(x_i) = z_j^* \frac{(x_{j+1}^* - x_i)^3}{6h_j} + z_{j+1}^* \frac{(x_i - x_j^*)^3}{6h_j} + E_j x_i + G_j. \tag{4.18}$$

The values of E_j and G_j can be obtained by using the interpolation condition. The values of the spline at the knots are $f_j(x_j^*) = y_j^*$ and $f_j(x_{j+1}^*) = y_{j+1}^*$. This leads to writing E_j and G_j in terms of the spline at the knots. The cubic spline for each segment is defined as

$$\begin{aligned} f_j(x_i) = z_j^* \frac{(x_{j+1}^* - x_i)^3}{6h_j} + z_{j+1}^* \frac{(x_i - x_j^*)^3}{6h_j} + \left(\frac{y_{j+1}^*}{h_j} - \frac{y_j^*}{h_j} - \frac{h_j}{6}(z_{j+1}^* - z_j^*) \right) x_i \\ + \frac{x_{j+1}^* y_j^*}{h_j} - \frac{x_j^* y_{j+1}^*}{h_j} - \frac{h_j}{6} x_{j+1}^* z_j^* + \frac{h_j}{6} x_j^* z_{j+1}^*; \end{aligned}$$

simplifying, we obtain

$$\begin{aligned} f_j(x_i) &= y_j^* \frac{(x_{j+1}^* - x_i)}{h_j} + y_{j+1}^* \frac{(x_i - x_j^*)}{h_j} \\ &+ \left(\frac{(x_{j+1}^* - x_i)^3}{6h_j} - \frac{h_j(x_{j+1}^* - x_i)}{6} \right) z_j^* \\ &+ \left(\frac{(x_i - x_j^*)^3}{6h_j} - \frac{h_j(x_i - x_j^*)}{6} \right) z_{j+1}^*. \quad x_i \in [x_j^*, x_{j+1}^*]. \end{aligned} \tag{4.19}$$

Summing over j leads to the cubic spline basis parameterized in terms of its value, and values of it is first and second derivative at the knots. In order to make the spline smooth at the knots, constraints must be applied. The interpolation condition $f_j(x_j^*) = y_j^*$ and $f_j(x_{j+1}^*) = y_{j+1}^*$ guarantee that $f(x_i)$ is continuous, and $f_j'(x_{j+1}^*) = f_{j+1}'(x_{j+1}^*)$, $f_j''(x_{j+1}^*) = f_{j+1}''(x_{j+1}^*)$ guarantee the smoothness . Imposing the condition $f_j'(x_{j+1}^*) = f_{j+1}'(x_{j+1}^*)$ yields

$$\frac{y_j^*}{h_j} - \left(\frac{1}{h_j} + \frac{1}{h_{j+1}} \right) y_{j+1}^* + \frac{y_{j+2}^*}{h_{j+1}} = \frac{z_j^* h_j}{6} + \left(\frac{h_j}{3} + \frac{h_{j+1}}{3} \right) z_{j+1}^* + \frac{z_{j+2}^* h_{j+1}}{6}.$$

The smoothing condition can be expressed in a vector-matrix notation as

$$\boldsymbol{D}\boldsymbol{y}^* = \boldsymbol{B}\boldsymbol{z}^*, \tag{4.20}$$

where $\boldsymbol{z}^* = (z_2, \cdots, z_{n_k-1})^T$ is the vector of the second derivative at the knots (note $z_1^* = z_{n_k}^* = 0$), and $\boldsymbol{y}^* = (y_1^*, \cdots, y_{n_k}^*)^T$ is the vector of the value of the spline at the knots. The two tri-diagonal matrices $\boldsymbol{D}$ and $\boldsymbol{B}$ are defined in terms of the spacing between successive knots by known $h_j = x_{j+1}^* - x_j^*$, for $j = 1, \cdots, n_k - 1$. The matrix

$\boldsymbol{D}$ is defined as a $(n_k - 2) \times (n_k)$ matrix;

$$\boldsymbol{D} = \begin{bmatrix} \frac{1}{h_1} & -\left(\frac{1}{h_1} + \frac{1}{h_2}\right) & \frac{1}{h_2} & \cdots & 0 & 0 \\ 0 & \frac{1}{h_2} & -\left(\frac{1}{h_2} + \frac{1}{h_3}\right) & \frac{1}{h_3} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{h_{n_k-1}} & -\left(\frac{1}{h_{n_k-2}} + \frac{1}{h_{n_k-1}}\right) & \frac{1}{h_{n_k-1}} \end{bmatrix}_{(n_k-2)\times(n_k)} .$$

The $(n_k - 2) \times (n_k - 2)$ symmetric matrix $\boldsymbol{B}$ is defined to be ;

$$\boldsymbol{B} = \begin{bmatrix} \frac{1}{3}(h_1 + h_2) & \frac{1}{6h_1} & 0 & \cdots & 0 & 0 \\ \frac{1}{6h_2} & \frac{1}{3}(h_2 + h_3) & \frac{1}{6h_3} & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \frac{1}{6h_{n_k-2}} & \frac{1}{3}(h_{n_k-2} + h_{n_k-1}) \end{bmatrix}_{(n_k-2)\times(n_k-2)} .$$

Equation (4.20) can be written as $\boldsymbol{F}\boldsymbol{y}^* = \boldsymbol{z}^*$, where

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{B}^{-1}\boldsymbol{D} \\ \boldsymbol{0} \end{bmatrix}$$

Therefore, equation (4.19) can be rewritten in terms of the unknown parameter $\boldsymbol{y}^*$. Using the notation in (Wood, 2006), section 4.1.2, the cubic spline can written as

$$f_j(x_i) = a_j^-(x_i)y_j^* + a_j^+(x_i)y_{j+1}^* + c_j^-(x_i)\boldsymbol{F}_j y^* + c_j^+(x_i)\boldsymbol{F}_{j+1}y^* \text{ if } x_j^* \le x_i \le x_{j+1}^*, \tag{4.21}$$

where,

$$\begin{aligned}
a_j^-(x_i) &= \frac{(x_{j+1}^* - x_i)}{h_j},\\
a_j^+(x_i) &= \frac{(x_i - x_j^*)}{h_j},\\
c_j^-(x_i) &= \left(\frac{(x_{j+1}^* - x_i)^3}{6h_j} - \frac{h_j(x_{j+1}^* - x_i)}{6}\right),\\
c_j^+(x_i) &= \left(\frac{(x_i - x_j^*)^3}{6h_j} - \frac{h_j(x_i - x_j^*)}{6}\right).
\end{aligned}$$

This can be written in a matrix form which maps $\boldsymbol{y}^*$ to the evaluated spline. The model matrix $\ddot{\boldsymbol{X}}$ is defined as the sum of two matrices, where $\ddot{\boldsymbol{X}} = \boldsymbol{A} + \boldsymbol{R}$, such that

$$\boldsymbol{A} = \begin{bmatrix}
a_1^-(x_1) & a_1^+(x_1) & 0 & 0 & 0\\
\vdots & \vdots & \ddots & \vdots & \vdots\\
a_1^-(x_{n_1}) & a_1^+(x_{n_1}) & 0 & 0 & 0\\
0 & a_2^-(x_{n_1+1}) & a_2^+(x_{n_1+1}) & 0 & 0\\
\vdots & \vdots & \ddots & \vdots & \vdots\\
0 & a_2^-(x_{n_2}) & a_2^+(x_{n_2}) & 0 & 0\\
\vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & 0 & a_{n_k-1}^+(x_{n_1+n_2+\cdots+1}) & a_{n_k-1}^+(x_{n_1+n_2+\cdots+1})\\
\vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & 0 & a_{n_k-1}^+(x_n) & a_{n_k-1}^+(x_n)
\end{bmatrix}_{n\times(n_k)},$$

where n_1 is the number of observations in the first segment, n_2 is the number of observations in the second segment and x_n is the number of observations in the n_k segment.

Then

$$\boldsymbol{R} = \begin{bmatrix} c_1^-(x_1)F_1 + c_1^+(x_1)F_2 & \cdots & c_1^-(x_{n_1})F_1 + c_1^+(x_{n_1})F_2 \\ c_2^-(x_{n_1+1})F_2 + c_2^+(x_{n_1+1})F_3 & \cdots & c_2^-(x_{n2})F_2 + c_2^+(x_{n2})F_3 \\ \vdots & \ddots & \vdots \\ c_{n_k-1}^-(x_{n_1+\cdots+1})F_{n_k-1} + c_{n_k-1}^+(x_{n_1+\cdots+1})F_{n_k} & \cdots & c_{n_k-1}^-(x_n)F_{n_k-1} + c_{n_k-1}^+(x_n)F_{n_k} \end{bmatrix}_{n\times(n_k)} .$$

Derivation of the Penalty Term

The second derivative of the spline is defined in equation 4.16 as

$$\begin{aligned} f_j''(x_i) &= \frac{(x_{j+1}^* - x_i)}{h_j} z_j^* + \frac{(x_i - x_j^*)}{h_j} z_{j+1}^*, \\ &= \sum_{j=2}^{nk-2} z_j^* d_j(x), \end{aligned}$$

where

$$d_j(x_i) = \begin{cases} \frac{x_i - x_j^*}{h_{j-1}} & x_{j-1}^* \le x_i \le x_j^* \\ \frac{x_{j-1}^* - x_i}{h_j} & x_j^* \le x_i \le x_{j+1}^* \\ 0 & x_i^* \ge x_{j+1}^* \end{cases} .$$

we get

$$\int [f''(x)]^2 dx = \boldsymbol{z^{*T}} \int \boldsymbol{d}(x)\boldsymbol{d}(x)^T dx \boldsymbol{z^*}.$$

$\int \boldsymbol{d}(x)\boldsymbol{d}(x)^T$ is tri-diagonal and the diagonal elements are

$$\begin{aligned}
\int_{x_{j-1}^*}^{x_{j+1}^*} (d_j(x_i))^2 dx_i &= \int_{x_{j-1}^*}^{x_j^*} (d_j(x_i))^2 dx_i + \int_{x_j^*}^{x_{j+1}^*} (d_j(x_i))^2 dx_i \\
&= \int_{x_{j-1}^*}^{x_j^*} \frac{(x_i - x_{j-1}^*)^2}{h_{j-1}^2} dx_i + \int_{x_j^*}^{x_{j+1}^*} \frac{(x_{j+1}^* - x_i)^2}{h_j^2} dx_i \\
&= \Big[\frac{(x_i - x_{j-1}^*)^3}{3h_{j-1}^2}\Big]_{x_{j-1}^*}^{x_j^*} - \Big[\frac{(x_{j+1} - x_i^*)^3}{3h_j^2}\Big]_{x_j^*}^{x_{j+1}^*} \\
&= \frac{h_{j-1}}{3} + \frac{h_j}{3}.
\end{aligned}$$

The off-diagonal elements are

$$\begin{aligned}
\int_{x_{j-1}^*}^{x_j^*} d_j(x_i) d_{j-1}(x_i) dx_i &= \int_{x_{j-1}^*}^{x_j^*} \frac{(x_i - x_{j-1}^*)}{h_{j-1}} . \frac{(x_j^* - x_i)}{h_{j-1}} dx_i \\
&= \frac{1}{h_{j-1}^2} \int_{x_{j-1}^*}^{x_j^*} (x_i - x_{j-1}^*)(x_j^* - x_i) dx_i \\
&= \frac{1}{h_{j-1}^2} \Big[\underbrace{(x_i - x_{j-1}^*)\frac{(x_j^* - x_i)^2}{2}\Big]_{x_{j-1}^*}^{x_j^*}}_{=0} - \int_{x_{j-1}^*}^{x_j^*} \frac{-(x_j - x_i^*)^2}{2} dx_i \Big] \\
&= \frac{1}{h_{j-1}^2} \Big[\frac{(x_j^* - x_i)^3}{6}\Big]_{x_{j-1}^*}^{x_j^*}, \\
&= \frac{h_{j-1}}{6}.
\end{aligned}$$

The penalty matrix is

$$\begin{aligned}
\int \big[f''(x_i)\big]^2 dx_i &= \boldsymbol{z}^{*T} \boldsymbol{B} \boldsymbol{z}^* \\
&= \boldsymbol{y}^{*T} \boldsymbol{D}^T \underbrace{\boldsymbol{B}^{-1} \boldsymbol{B}}_{I} \boldsymbol{B}^{-1} \boldsymbol{D} \boldsymbol{y}^* \\
&= \boldsymbol{y}^{*T} \underbrace{\boldsymbol{D}^T \boldsymbol{B}^{-1} \boldsymbol{D}}_{(n_k)\times(n_k)} \boldsymbol{y}^*
\end{aligned}$$

4.4 Penalized Iteratively Re-weighted Least Square Estimation (PIRLS)

The probability density function of y_i given θ_i is defined as

$$\begin{aligned} f^*(y_i|\theta_i) &= \exp\left\{\frac{y_i\theta_i - b_i(\theta_i)}{a_i(\phi)} + c_i(y_i,\phi)\right\}, \\ &= \exp\left\{\frac{w_i\left(y_i\theta_i - b_i(\theta_i)\right)}{\phi} + c_i(y_i,\phi)\right\}, \end{aligned}$$

where θ_i is the parameter of interest, which is also called the canonical parameter. The functions $b_i(\theta_i)$, $c_i(y_i,\phi)$ and $a_i(\phi)$ are arbitrary functions, and $a_i(\phi) = \phi/w_i$ where w_i is a known constant, which is often equal to 1, and ϕ is dispersion parameter. The parameter θ_i is determined by μ_i, which has the properties

$$\mu_i = E(y_i) = b_i'(\theta_i), \tag{4.22}$$

$$\text{Var}(y_i) = a(\phi)b_i''(\theta_i) = \phi V(\mu_i), \tag{4.23}$$

where $b_i'(\theta_i)$ and $b_i''(\theta_i)$ are the first and second derivatives of $b_i(\theta_i)$. The canonical link function is defined since we let $\eta_i = \theta_i = \boldsymbol{X}_i\boldsymbol{\beta}$, so

$$\begin{aligned} g(\mu_i) &= g(b'(\theta_i)) = \eta_i \\ b'(\theta_i) &= g^{-1}(\eta_i), \\ \mu_i &= g^{-1}(\eta_i), \\ \mu_i &= g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta}). \end{aligned}$$

The likelihood of $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f^*(y_i|\theta_i),$$

the log-likelihood of $\boldsymbol{\beta}$ is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log[f^*(y_i|\theta_i)].$$

The penalized log likelihood function is defined in equation 4.5 as

$$\ell_{pen}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{S}\boldsymbol{\beta}, \tag{4.24}$$

Given the values of the smoothing parameters λ_j, the parameters $\boldsymbol{\beta}$ can be estimated by maximizing the penalized log likelihood. To maximize $\ell_{pen}(\boldsymbol{\beta})$ we partially differentiate $\ell_{pen}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_j$ and set the result to zero, such that:

$$\frac{\partial \ell_{pen}(\boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} - [\boldsymbol{S}\boldsymbol{\beta}]_j = 0 \quad \text{for all } j = 1, \ldots, p,$$

where $[.]_j$ denotes the jth element of a vector. The derivative of $\ell(\boldsymbol{\beta})$ with respect to β_j is

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\left(y_i\frac{\partial \theta_i}{\partial \beta_j} - b_i'(\theta_i)\frac{\partial \theta_i}{\partial \beta_j}\right), \qquad \text{for all } j = 1, \ldots, p.$$

Using the chain rule, we obtain

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i}.\frac{\partial \mu_i}{\partial \beta_j}.$$

Since $\mu = b'(\theta)$ for any distribution belonging to the exponential family, we have

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \implies \frac{\partial \theta_i}{\partial \mu_j} = \frac{1}{b''(\theta_i)}.$$

Therefore,

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\left(\frac{y_i - b'(\theta)}{b''(\theta_i)}\frac{\partial \mu_i}{\partial \beta_j}\right). \tag{4.25}$$

We substitute 4.22, and 4.23 into (4.25) to obtain

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \right).$$

The estimate is required to satisfy

$$\begin{aligned} \frac{\partial l_{pen}(\boldsymbol{\beta})}{\partial \beta_j} &= 0 \qquad (4.26) \\ \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \right) - [\boldsymbol{S}\boldsymbol{\beta}]_j &= 0, \end{aligned}$$

Define $\mathcal{S}$ to be

$$\mathcal{S} = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \right) + \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}.$$

If the variance $V(\mu_i)$ is fixed, then the solution of (4.26) is equivalent to minimizing $\mathcal{S}$ with respect to $\boldsymbol{\mu} = (\mu_i, \dots, \mu_n)^T$ which can be solved by PIRLS algorithm. Given a starting value $\boldsymbol{\mu}^0$, and then compute $V = \text{diag}(V(\mu_1^{[0]}), \dots, V(\mu_n^{[0]}))$ to minimize the $\mathcal{S}$:

$$\mathcal{S} = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \right) + \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta} = \|\sqrt{\boldsymbol{V}^{-1}}[\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})]\|^2 + \boldsymbol{\beta}^T S \boldsymbol{\beta}.$$

$\boldsymbol{\mu}$ is replaced by its first order Taylor expanison, we can write

$$\begin{aligned} \boldsymbol{\mu}(\boldsymbol{\beta}^{[k]}) \approx & \boldsymbol{\mu}(\boldsymbol{\beta})^{[k]} + \frac{\partial \mu_i}{\partial \beta_j} (\boldsymbol{\beta} - \boldsymbol{\beta}^{[\boldsymbol{k}]}) \\ \approx & \boldsymbol{\mu}^{[k]} + J_{ij}(\boldsymbol{\beta} - \boldsymbol{\beta}^{[k]}), \end{aligned}$$

where $\boldsymbol{J}$ is the Jacobian matrix with elements $J_{ij} = \frac{\partial \mu_i}{\partial \beta_j}|_{\hat{\boldsymbol{\beta}}^{[\boldsymbol{k}]}}$, such that;

$$\|\sqrt{\boldsymbol{V}^{[k]^{-1}}}[\boldsymbol{y} - \boldsymbol{\mu}^{[k]} - \boldsymbol{J}[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{[\boldsymbol{k}]}]]\|^2 + \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta},$$

since we have,

$$g(\mu_i) = \boldsymbol{X}_i\beta \Longrightarrow g'(\mu_i)\frac{\partial \mu_i}{\partial \beta_j} = X_{ij} \Longrightarrow J_{ij} = \frac{\partial \mu_i}{\partial \beta_j}|_{\hat{\beta}^{[k]}} = \frac{X_{ij}}{g'(\mu_i^{[k]})}.$$

Defining the diagonal matrix $\boldsymbol{G}^{[\boldsymbol{k}]} = \text{diag}(g'(\mu_1^{[k]}), \dots, g'(\mu_n^{[k]}))$ leads to $\boldsymbol{J} = (\boldsymbol{G}^{[\boldsymbol{k}]})^{-1}\boldsymbol{X}$. So we can finally write

$$\mathcal{S} \approx \|\sqrt{\boldsymbol{V}^{[k]^{-1}}}(\boldsymbol{G}^{[k]})^{-1}\left[\boldsymbol{G}^{[k]}(\boldsymbol{y} - \boldsymbol{\mu}^{[k]}) + \boldsymbol{X}\hat{\boldsymbol{\beta}}^{[\boldsymbol{k}]} - \boldsymbol{X}\boldsymbol{\beta}\right]\|^2 + \boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta},$$

where $\boldsymbol{\eta}^{[k]} = \boldsymbol{X}\hat{\boldsymbol{\beta}^{[\boldsymbol{k}]}}$. The pseudo-data is $\boldsymbol{z}^{[k]} = \boldsymbol{G}^{[k]}(\boldsymbol{y} - \boldsymbol{\mu}^{[k]}) + \boldsymbol{\eta}^{[k]}$, and $\boldsymbol{W}^{[k]}$ is a diagonal matrix, such that $\boldsymbol{W}_{ii}^{[k]} = \frac{1}{V(\mu_i^{[k]})g'(\mu_i^{[k]})^2}$. The penalized problem can be written as:

$$\|\sqrt{\boldsymbol{W}^{[k]}}(\boldsymbol{z}^{[k]} - \boldsymbol{X}\boldsymbol{\beta})\|^2 + \boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}. \tag{4.27}$$

The maximum penalized likelihood estimates $\hat{\boldsymbol{\beta}}$ can be estimated by the iteration procedure until convergence, as follows:

1. Given the current $\boldsymbol{\mu}^{[k]}$, calculate the diagonal weight matrix $w_i^{[k]}$ and the pseudo-data $\boldsymbol{z}^{[k]}$.
2. Minimize $\|\sqrt{\boldsymbol{W}^{[k]}}(\boldsymbol{z}^{[k]} - \boldsymbol{X}\boldsymbol{\beta})\|^2 + \boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}$ with respect to $\boldsymbol{\beta}$ to obtain $\hat{\boldsymbol{\beta}}^{[k+1]}$ and evaluate $\boldsymbol{\eta}^{[k+1]} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{[\boldsymbol{k}+\boldsymbol{1}]}$.
3. Compute the fitted values $\mu_i^{[k+1]} = g^{-1}(\eta_i^{[k+1]})$.
4. Increase k.

The expression $\|\sqrt{\boldsymbol{W}^{[k]}}(\boldsymbol{z}^{[k]} - \boldsymbol{X}\boldsymbol{\beta})\|^2 + \boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta}$ can be minimized by differentiating it with respect to $\boldsymbol{\beta}$ and then setting it to zero, such that

$$\hat{\boldsymbol{\beta}}^{[\boldsymbol{k}+\boldsymbol{1}]} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}^{[\boldsymbol{k}]}\boldsymbol{X} + \boldsymbol{S}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}^{[\boldsymbol{k}]}\boldsymbol{z}^{[\boldsymbol{k}]}.$$

The influence matrix is $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X} + \boldsymbol{S})^{-\boldsymbol{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}$ and the effective degree of freedom is defined as the trace of the influence matrix, tr$(\boldsymbol{H})$, which controls the smoothness of the curve. Using large values of smoothing parameters, the result of the model fitting will have few degrees of freedom, because the penalization reduces the models degree of freedom. The residual variance for the GAM model is estimated in a similar manner to the residual variance in linear regression, $\hat{\sigma}^2 = \frac{\|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{y}\|^2}{n-\text{tr}(\boldsymbol{H})}$. The estimation for the scale parameter can be obtained using the Pearson-like estimator, as $\hat{\phi} = \frac{\sum_i V(\hat{\mu}_i)^{-1}(y_i-\hat{\mu}_i)^2}{n-\text{tr}(\boldsymbol{H})}$. The variance-covariance matrix for the estimators $\hat{\boldsymbol{\beta}}$ is given by;

$$\boldsymbol{V}_e = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}+\boldsymbol{S})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X}+\boldsymbol{S})^{\boldsymbol{-1}}\phi, \tag{4.28}$$

and therefore, $\hat{\boldsymbol{\beta}} \underset{ap}{\sim} N(E(\hat{\beta}), \boldsymbol{V}_e)$.

4.5 Selecting the Smoothing Parameter

Penalized likelihood maximization estimates the model parameters $\boldsymbol{\beta}$ for given values of the smoothing parametersλ. Wood (2006) suggested two methods for estimating the smoothing parameters λ, based on either the known or unknown scale parameter ϕ. For known scale parameter ϕ, the smoothing parameters can be estimated by minimizing Mallow's C_p (Mallows, 1973), or using the Un-Biased Risk Estimator (UBRE) (Craven and Wahba, 1979), $\nu_u(\lambda) = \|\boldsymbol{y}-\boldsymbol{H}\boldsymbol{y}\|^2/n - \sigma^2 + 2\text{tr}(\boldsymbol{H})\sigma^2/n$. In the case of an unknown scale parameter, the estimation of the smoothing parameters λ is obtained by generalized cross validation. The Leave One Out Cross Validation (LOOCV) criterion was introduced by Wahba et al. (1979), which is based on a method of minimizing the average mean square error in predicting a new observation y using the fitting model by leaving one observation y_i out and fitting the model for the remaining data. This process continues for all data points in turn. The LOOCV score is given by;

$$LOOCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}^{[-i]})^2, \tag{4.29}$$

where $\widehat{f}^{[-i]}$ is the prediction of f_i obtained by fitting a model to all data except the i^{th} value. It is ineffectual to calculate LOOCV score by leaving out one variable at a time, and fitting the n models, but fortunately it can be shown that

$$LOOCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}_i)^2/(1 - \boldsymbol{H}_{ii})^2, \tag{4.30}$$

where $\widehat{f}$ is the estimate from fitting to all the data, and $\boldsymbol{H}$ is the corresponding influence matrix. In practice the weights, $(1 - \boldsymbol{H}_{ii})$, are often replaced by the mean weight, $tr(\boldsymbol{I} - \boldsymbol{H})/n$, which gives the generalized cross validation score

$$GCV(\lambda) = \frac{n\sum_{i=1}^{n}(y_i - \widehat{f}_i)^2}{[tr(\boldsymbol{I} - \boldsymbol{H})]^2}. \tag{4.31}$$

4.6 Choosing the Number of Knots

In this section two algorithms for selection of the knots based on GCV are discussed. The primary reference for this section can be found in Ruppert et al. (2003) Chapter 5.

4.6.1 Myopic algorithm

The idea of a myopic algorithm is based on a sequence of trial values of knots, taking into account that the knots are equally spaced and the number of knots has to be less than $n - p - 1$, where n is the sample size and p is the number of covariates. This is performed by taking two values of knots from a sequence, fitting two GAMs with optimal smoothing parameter value λ for each model separately, and then computing the GCV for both models. This continues for the whole sequence and is stopped when there is no improvement in GCV score.

4.6.2 Full-search Algorithm

The idea of the full-search algorithm is more general than the myopic algorithm, which considers all the values of the sequence of knots, computes GCV by fitting the GAM with optimal smooth parameter value of λ separately for each values of the knots, and selects the knots that minimize the GCV score. Therefore, the myopic algorithm takes less computational time than the full-search algorithm.

4.7 P-value

In this section we are interested on testing a subset $\boldsymbol{\beta}_j$ of $\boldsymbol{\beta}$ is identically zero, which means there is no effect of the j^{th} smoothing term. To test the hypotheses regarding nonlinearity of the j^{th} smoothing term, the null hypothesis is that $\boldsymbol{f}_j = \mathbf{0}$. Let $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}_j}$ be the covariance matrix of $\hat{\boldsymbol{\beta}}_j$, extracted from $\boldsymbol{V}_e$. Under the null hypothesis, $\boldsymbol{\beta}_j = \mathbf{0}$, we have $\hat{\boldsymbol{\beta}}_j \underset{ap}{\sim} N(\mathbf{0}, \boldsymbol{V}_{\hat{\boldsymbol{\beta}}_j})$. However, $\boldsymbol{V}_{\hat{\boldsymbol{\beta}}_j}$ is not a full rank so it is not invertible. Let $u = \text{rank}(\boldsymbol{V}_{\hat{\boldsymbol{\beta}}_j})$ and $\boldsymbol{V}^{u-}_{\hat{\boldsymbol{\beta}}_j}$ is the rank u pseudo-inverse of the covariance matrix, then the Wald test statistics is performed using the result under the null hypothesis.

$$T_j = \hat{\boldsymbol{\beta}}_j^T \boldsymbol{V}^{u-}_{\hat{\boldsymbol{\beta}}_j} \hat{\boldsymbol{\beta}}_j \underset{ap}{\sim} \chi^2_u \tag{4.32}$$

In the case of the covariance matrix $\boldsymbol{V}_e$ containing an unknown scale parameter, ϕ, the p-value can be calculated from the following test

$$\frac{\hat{\boldsymbol{\beta}}_j^T \boldsymbol{V}^{u-}_{\hat{\boldsymbol{\beta}}_j} \hat{\boldsymbol{\beta}}_j / u}{\hat{\phi}/(n - edf)} \sim F_{u,edf}, \tag{4.33}$$

where edf is equal to the effective degrees of freedom for the model.

4.8 Logistic Regression with GAM for Clinical Characteristic of Lung Cancer Dataset

In medical research, the common statistical model for binary data is the logistic model. This section represents the logistic regression model using the GAM, with one smoother for continuous covariates and other covariates in their original form, since they are categorical. The link function is logit. In this model the response variable y_i is 0 or 1, with $y_i = 1$ when the survival time is less than the median, based on the uncensored observations. Alternatively, $y_i = 0$ if the survival time is greater then the median, which is also based on the uncensored observations. Censored observations were not taken into account in this GAM setting. Table 4.1 shows the binary response variables, the number of patients that are involved, the median survival times (in days) and the number of censored/uncensored observations.

Number of patients	Median survival	y_i	Status	
			Censored	Uncensored
85	645	$y_i = 1$ if $t < 645$	0	32
		$y_i = 0$ if $t > 645$	22	31

Table 4.1: The number of patients, median survival times, censored and uncensored observations for the response variables.

4.8.1 Modelling

We fit 32 generalized additive logistic models to the binary response variable, with smooth terms for Age with 5 knots, and a categorical variable for the rest. The models were fitted using the $\boldsymbol{mgcv}$ package function in $\boldsymbol{R}$ Wood (2006), where the smooth term is the cubic regression spline, which is described in Section 4.3.2. The AIC values are used for variable selection, where AIC is defined as $-2\ell(\hat{\boldsymbol{\beta}}) + 2edf$, where $\ell(\hat{\boldsymbol{\beta}})$ is the log likelihood of the estimated parameter, $\hat{\boldsymbol{\beta}}$ are the maximum penalized likelihood estimates and edf is the effective degrees of freedom for the model. The UBRE score is also used for the smoothing parameters selection for known scale parameters. The AIC values and the degree of freedom for the 32 possible models are given in Table 4.2.

#	parameters in the model	edf	AIC
1	intercept	1.00	89.32
2	f(Age)	2.54	91.17
3	Gender	2.00	91.28
4	Stage T	3.00	81.17
5	Stage N	3.00	89.88
6	Grade	3.00	90.50
7	f(Age)+Gender	3.40	93.14
8	f(Age)+Stage T	5.59	80.81
9	f(Age)+Stage N	4.51	91.72
10	f(Age)+Grade	4.00	92.38
11	Gender+Stage T	4.00	81.92
12	Gender+Stage N	4.00	91.87
13	Gender+Grade	4.00	92.50
14	Stage T+Stage N	5.00	81.47
15	Stage T+ Grade	5.00	79.69
16	Stage N+Grade	5.00	91.81
17	f(Age)+Gender+Stage T	6.37	82.10
18	f(Age)+Gender+Stage N	5.55	93.70
19	f(Age)+Gender+Grade	5.00	94.38
20	f(Age)+Stage T+Stage N	8.69	77.75
21	f(Age)+Stage T+Grade	6.84	80.74
22	f(Age)+Stage N+Grade	6.04	93.81
23	Gender+Stage T+Stage N	6.00	82.51
24	Gender+Stage T+Grade	6.00	81.12
25	Stage T+Stage N+Grade	7.00	81.34
26	Gender+ Stage N+ Grade	6.00	93.81
27	Gender+Stage T+Stage N+Grade	8.00	82.81
28	f(Age)+Gender+Stage T+ Stage N	9.71	78.73
29	f(Age)+Gender+Stage T+Stage N	7.78	82.27
30	f(Age)+Stage T+Stage N+Grade	10.53	80.54
31	f(Age)+Gender+Stage N+Grade	7.13	95.80
32	f(Age)+Gender++Stage T+Stage N+Grade	11.59	81.73

Table 4.2: 32 possible GAM models with the effective degree of freedom and AIC values for each model.

The best model that has the smallest AIC value is the model that contains Age, as a smooth function, the optimal value of the smoothing parameter is $\lambda_{opt} = 0.010$ with corresponding minimum UBRE score 0.2341, Stage-T and Stage-N. Table 4.3 shows the estimated parameters for the best model that includes Age, as smoothing term, Stage-T and Stage-N.

Parameter	Estimate	Standard error	p-value
Intercept	-2.404	0.8449	0.004
$f(Age)$			0.233
Stage-T2	1.703	0.764	0.025
Stage-T3	4.094	2388	0.999
Stage-N1	0.212	0.658	0.747
Stage-N2	3.989	2.349	0.089
Intercept	-2.295	2.295	-0.317
Age	0.013	0.031	0.655
StageT2	1.291	0.675	0.055
StageT3	17.982	1069.241	0.986
StageN2	0.034	0.614	0.955
StageN3	2.016	1.208	0.095

Table 4.3: The estimated coefficients, and standard errors of estimated coefficients and the p-value for the fitting a logistic generalized additive model that includes Age, Stage-T and Stage-N.

However, the p-values of both Stage-N1, Stage-N3, Stage-T3, and $f(Age)$ are not significant. The p-values for including a non-linear smoothing terms $f(Age)$ is equal to 0.233, over a linear age effect is equal to 0.655, which means there no effect of Age. Figure 4.1 shows the estimate of the smooth function of Age, where the vertical axis is on the scale of the additive predictor of the model. The horizontal axis indicates the Age, the solid line represents the predictor values, the two dashed lines represent the 95% confidence interval, and the small lines along the horizontal axis are the “rug”, showing the values of the covariate of Age for each patient. The highest risk age is between 55 and 65. Somewhat surprisingly, the risk seems to be same for all patients older than age 68.

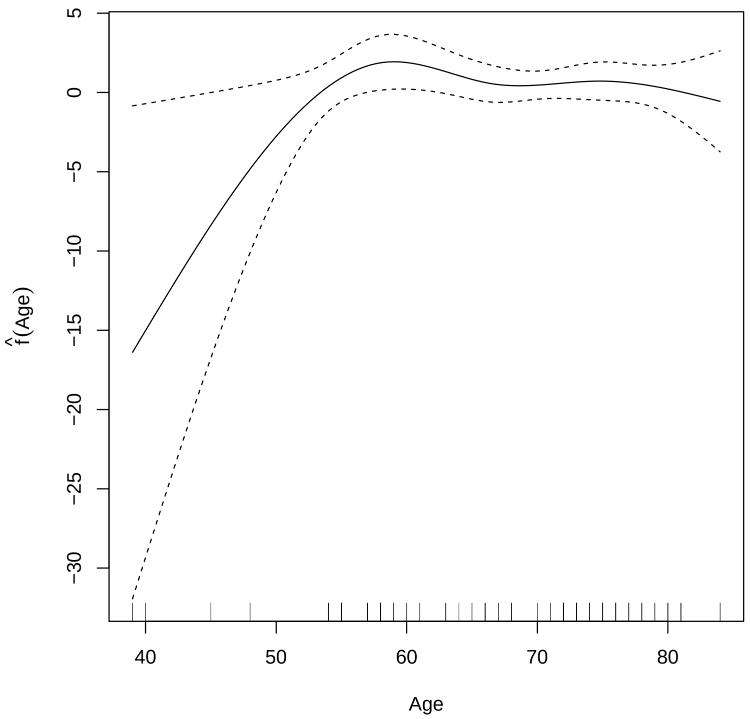


Figure 4.1: The plot of the smooth function of Age in model 20, the solid line represents the predictor values, the two dashed lines represent the 95% confidence interval, and the small lines along the horizontal axis are the “rug”, showing the values of the covariate of Age for each patient.

4.9 Conclusion

GAM produced a flexible statistical method for nonlinear relationships between independent and dependent variables in the exponential family form. In this chapter, we presented how GAM is constructed using the penalized regression spline and how the estimated parameters are obtained by PIRLS. The method of estimating the smoothing parameters using either UBRE or GCV is included. Two algorithms for selecting the optimal number of knots are discussed. The generalized additive logistic model is commonly used in medical research, which can be used to identify and characterize the effect of clinical characteristics on a binary response variable. In the logistic GAM example, the PIRLS was used to predict the unspecified smoothing function, the cubic spline smoother, with 5 knots, was employed to the continuous independent variable. The UBRE score was used to select the optimal smoothing parameter. The resulting curves of Age shows that patients aged 55 to 65 years have higher risk, while patients age 68 or older have slightly lower risk.

Chapter 5

Additive Cox PH Model

5.1 Introduction

The Cox proportional hazard model (PH) forces the log hazard ratio to be linear in the predictors. To express non-linear effects in the Cox model, the hazard can be expressed in terms of additive predictors. The purpose of this method is to examine the flexibility of a survival model that does not have to be linear in the predictors. Using smoothing terms in the Cox PH model allows nonlinear predictor effects to be detected and modeled. In the literature, cubic B-splines is most commonly used to represent the smoothing term in the additive Cox PH model. However, the spline parameters for the cubic B-spline and cubic regression spline are unconstrained.

O'Sullivan (1988); Gray (1992, 1994); Tsujitani et al. (2012); Tsujitani and Tanaka (2013); Nan et al. (2005); Cadarso-Surez et al. (2010); Meira-Machado et al. (2013); Wang et al. (2017a) studied nonparametric estimation of relative risk using a cubic B-spline representation of the smoothing term in the additive Cox PH model, with the parameter estimates obtained by maximizing the penalized partial log likelihood function using Newton-Raphson procedure. Sleeper and Harrington (1990) used the cubic B-spline in the Cox PH model with a small number of knots, and suggested 5 knots or less to clarify the effect of the covariate in survival. The coefficients were estimated in a manner similar to Cox regression coefficients using partial likelihood without penalty term.

Hastie and Tibshirani (1990a) and O'Sullivan (1988) presented a cubic B-spline with knots at all data points in the Cox PH model, as a result using the full information matrix could be time-consuming in the process of estimating the model parameters. Because of this, Hastie and Tibshirani (1990a) used the banded approximation to the information matrix instead of the full information matrix, which is implemented by setting off-diagonal elements of the information matrix to zero. Inference on GAM estimates is not well developed in the case of the Cox PH model because the standard GAM does not take into account censored observations.

Sleeper and Harrington (1990) and Gray (1992) noticed that the estimated B-spline functions gives poor estimates around the boundary knots. This chapter introduces a new smoothing term to the additive Cox PH model, using a radial basis function to build the survival model, aiming to capture nonlinear patterns of CNA genomic-windows, and age in the NSCLC data, and model their relationship with the survival time of patients. The radial basis function is used because it satisfies the natural cubic spline conditions, which are the second derivative at the boundary knots is equal to zero, this means the radial basis function has a better ending behavior at the boundary knots compared to the cubic B-spline. This point is discussed more in Section 5.5.2. By identifying nonlinear predictors and estimating their effects, we can estimate the current risk and thus determine which variable corresponds to lower or higher risk, that could affect the survival time, and which variables could increase the risk of death. The new smoothing term in the additive Cox PH model allows us to study both constrained and unconstrained smooth spline coefficients, which gives more flexibility in the model. In this thesis the number of knots is smaller than the number of observations, and the full information matrix is used in the Newton-Raphson procedure.

This chapter is organized as follows: Section 5.2 describes the statistical methodology for fitting the penalized additive Cox PH model using the penalized partial likelihood approach. The estimation of the log hazard ratio is present in Section 5.3. To investigate the stability of our method we perform a simulation study in Section 5.5.

Generating survival times to simulate the additive Cox model is introduced in Section 5.5.1. The test statistics for the spline effect are given in Section 5.6.1. Section 5.7 discusses the method of choosing the optimal values of the smoothing parameters and the optimal number of knots, using Cross-Validation Partial Log-likelihood (CVPL) method. Model diagnostics are present in Section 5.8. Finally, Section 5.9 shows the main result of modeling the clinical characteristic of the NSCLC.

5.2 Extending The Cox PH Model

The standard Cox model as defined in equation (3.1) has the form

$$h_i(t|\boldsymbol{X}) = h_0(t)\exp\{\eta_i\} = h_0(t)\exp\{\boldsymbol{X}_i\boldsymbol{\beta}\},$$

where $\eta_i = \boldsymbol{X}_i\boldsymbol{\beta}$ is the linear predictor, $\boldsymbol{X}$ is a parametric model matrix of size $n \times p$, where the i^{th} row of this matrix is denoted by $\boldsymbol{X}_i$, which is p vector of the covariates of the i^{th} patient, and $\boldsymbol{\beta}$ is the p vector of unknown parameters.

Replacing the linear predictor by the additive predictor allows us to include the genome-wide CNA profile and age as smoothing terms. Incorporating genome-wide CNA profiles into the Cox PH model as smoothing terms will make the parameters inestimable. The main challenges lies in the dimension of the smoothing matrix of all the CNA profiles. For the j^{th} CNA genomic-windows variable, the size of the j^{th} smooth matrix is $n \times q_j$, where q_j is the number of the basis function that represent the j^{th} smoothing term. Combine all these matrices together lead to very large smooth matrix for all CNA genomic-windows in the model, which can be computationally demanding.

Variable selection is very important in this case, which tries to gain as much information as possible from CNA data, without incorporating the large number of CNA genomic-windows variables in the model. As a result of this, we obtain simpler models that identify the important predictors in the CNA genomic-windows. Variable selection is discussed in Chapters 6 and 7.

The extension of the Cox PH model formed by including the CNA genome-windows profiles as smoothing terms can be expressed as

$$h_i(t) = h_0(t)\exp\{\boldsymbol{X}_i^L\boldsymbol{\zeta} + \sum_{j=1}^{p} f_j(x_j)\}, \tag{5.1}$$

where $h_0(t)$ is the baseline hazard function, which is the hazard function for a patient with values of all covariates equal to zero. We define $\boldsymbol{X}^L$ to be the parametric model matrix of size $n \times q$, that includes the continuous, categorical, or both continuous and categorical variables. We denote the rows of the parametric model matrix $\boldsymbol{X}^L$ as $\boldsymbol{X}_i^L$, which is the q vector of the fixed-effect covariates for the i^{th} patient. Here $\boldsymbol{X}^L$ is the same as $\boldsymbol{X}$ in equation (3.1), and $\boldsymbol{\zeta}$ is a q-vector of the parametric model parameters (the fixed effect parameters).

The f_j are the unknown smooth function of the covariate x_{j_i}. To estimate the model, we need to specify a basis function for each smoothing term. For simplicity, we assume the same basis function for each smoothing term, with the number of knots being equal for each smoothing term. In this chapter, the smoothing term is the radial basis function with 5 equally spaced knots, which is described in Section 4.3.1. The radial basis function is defined in equation (4.6), removing the intercept from the radial basis function, because it is absorbed in the baseline hazard, the radial basis function can be expressed as

$$f_j(x_{j_i}) = \alpha_{j0}x_{j_i} + \sum_{k=1}^{n_k} \alpha_{jk}|x_{j_i} - x_{jk}^*|^3, \quad j = 1, \dots, p, \quad k = 1, \dots, n_k, \tag{5.2}$$

subject to

$$\sum_{k=1}^{n_k} \alpha_{jk} = \sum_{k=1}^{n_k} \alpha_{jk}x_{jk}^* = 0. \tag{5.3}$$

where x_{jk}^* for $k = 1, \dots, n_k$ are the knots in the range of x_{j_i} for $i = 1, \dots, n$ and n_k is the number of the knots. Equation (5.2) can be written as matrix-vector notation, such that $\boldsymbol{f}_j = \boldsymbol{X}_j\boldsymbol{\beta}_{cj}$, where $\boldsymbol{f}_j$ is the$n$- vector of the j^{th} smoothing term with $\boldsymbol{f}_{j_i} = f_j(x_{j_i})$, $\boldsymbol{X}_j$ is the j^{th} smoothing matrix of size $n \times (n_k + 1)$, the i^{th} row of the j^{th} smooth matrix $\boldsymbol{X}_j$ is $\boldsymbol{X}_{ji} = [x_{ji}, |x_{ji} - x^*_{j1}|^3, |x_{ji} - x^*_{j2}|^3, \ldots, |x_{ji} - x^*_{jn_k}|^3]$, which depends only on the distance between the observation and the knots. Note, this matrix is the same as equation (4.7) except the first column which is removed, and $\boldsymbol{\beta}_{cj}$ is $(n_k + 1)$ vector of the j^{th} constrained spline effect parameter, such that $\boldsymbol{\beta}_{cj} = [\alpha_{j0}, \alpha_{j1}, \ldots, \alpha_{jn_k}]^T$. Each spline effect parameter has two constraints, which can be expressed as $\boldsymbol{C}_j\boldsymbol{\beta}_{cj} = \boldsymbol{0}$, where $\boldsymbol{C}_j$ is $2 \times (n_k + 1)$ matrix. These constraints can be expressed as;

$$\boldsymbol{C}_j\boldsymbol{\beta}_{cj} = \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 0 & x^*_{j1} & x^*_{j2} & \cdots & x^*_{jn_k} \end{bmatrix} \begin{bmatrix} \alpha_{j0} \\ \alpha_{j1} \\ \vdots \\ \alpha_{jn_k} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{5.4}$$

where $\boldsymbol{C}_j$ is as the same as equation (4.8), where the first column is removed. Using the general linear constraints approach as discussed in Section 4.3.1, the constrained spline effect parameter can be expressed as $\boldsymbol{\beta}_{cj} = \boldsymbol{Z}_j\boldsymbol{\beta}_{zj}$, where $\boldsymbol{Z}_j$ is a semi-orthogonal matrix of size $(n_k+1)\times(n_k-1)$, $\boldsymbol{Z}_j^T\boldsymbol{Z}_j = \boldsymbol{I}$, and $\boldsymbol{\beta}_{zj}$ is a vector of n_k-1 unconstrained spline effect parameters for the j^{th} smoothing term. Therefore, the j^{th} smoothing term, which can be expressed as $\boldsymbol{f}_j = \boldsymbol{X}_j\boldsymbol{Z}_j\boldsymbol{\beta}_{zj}$, so we can write each smoothing term in the model in terms of its introduced unconstrained spline effect parameters.

The additive Cox PH model can be written as

$$h_i(t) = h_0(t)\exp\{\boldsymbol{X}_i\boldsymbol{\beta}\}, \tag{5.5}$$

where $\boldsymbol{X}_i$ is the i^{th} row of the matrix $\boldsymbol{X} = [\boldsymbol{X}^L, \boldsymbol{X}_1\boldsymbol{Z}_1, \boldsymbol{X}_2\boldsymbol{Z}_2, \ldots, \boldsymbol{X}_p\boldsymbol{Z}_p]$. The model matrix $\boldsymbol{X}$ includes the columns of the parametric model; $\boldsymbol{X}^L$, and columns that represent the spline basis $\boldsymbol{X}_jZ_j$ for $j = 1, \ldots, p$. The full set of parameters in the model is denoted by $\boldsymbol{\beta}$, which contains $\boldsymbol{\zeta}$ and all the unconstrained spline effect parameters vector $\boldsymbol{\beta}_{zj}$, so that $\boldsymbol{\beta}^T = [\boldsymbol{\zeta}^T, \boldsymbol{\beta}_{z1}^T, \boldsymbol{\beta}_{z2}^T, \ldots, \boldsymbol{\beta}_{zp}^T]$.

The estimation of the parameters $\boldsymbol{\beta}$ can be obtained using partial likelihood method. The partial likelihood is

$$L_{pl}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp[\boldsymbol{X}_i \boldsymbol{\beta}]}{\sum_{j \in R(t_i)} \exp[\boldsymbol{X}_j \boldsymbol{\beta}]} \right)^{\delta_i}, \tag{5.6}$$

where δ_i is an indicator variable, with $\delta_i = 1$ when the individual is dead and zero if the individual is censored. The unpenalized partial log-likelihood function $\ell_{pl}(\boldsymbol{\beta})$ can be written as

$$\ell_{pl}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \boldsymbol{X}_i \boldsymbol{\beta} - \sum_{i=1}^{n} \delta_i \log \left(\sum_{j \in R(t_i)} \exp(\boldsymbol{X}_j \boldsymbol{\beta}) \right). \tag{5.7}$$

The first derivative of the partial log likelihood with respect to $\boldsymbol{\beta}$ gives the score vector

$$U_{pl}(\boldsymbol{\beta}) = \frac{\partial \ell_{pl}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \delta_i \left(\boldsymbol{X}_i - \frac{\sum_{j \in R_{(t_i)}} \boldsymbol{X}_j \exp(\boldsymbol{X}_j \boldsymbol{\beta})}{\sum_{j \in R_{(t_i)}} \exp(\boldsymbol{X}_j \boldsymbol{\beta})} \right),$$

The Fisher information matrix is calculated as the negative second derivative of the partial log likelihood with respect to $\boldsymbol{\beta}$:

$$\begin{aligned} I_{pl}(\boldsymbol{\beta}) = -\Big[\frac{\partial^2 \ell_{pl}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_l \partial \boldsymbol{\beta}_m}\Big] = \sum_{i=1}^{n} \delta_i \Bigg(& \frac{\Big(\sum_{j \in R_{(t_i)}} X_{jl} \exp(\boldsymbol{X}_j \boldsymbol{\beta})\Big)\Big(\sum_{j \in R_{(t_i)}} X_{jm} \exp(\boldsymbol{X}_j \boldsymbol{\beta})\Big)}{\Big(\sum_{j \in R_{(t_i)}} \exp(\boldsymbol{X}_j \boldsymbol{\beta})\Big)^2} \\ & - \frac{\Big(\sum_{j \in R_{(t_i)}} X_{jl} \exp(\boldsymbol{X}_j \boldsymbol{\beta})\Big)\Big(\sum_{j \in R_{(t_i)}} X_{jm} X_{jl} \exp(\boldsymbol{X}_j \boldsymbol{\beta})\Big)}{\Big(\sum_{j \in R_{(t_i)}} \exp(\boldsymbol{X}_j \boldsymbol{\beta})\Big)^2} \Bigg) \end{aligned}$$

The penalized partial log-likelihood can be used to solve the smoothing problems, which can be written as the difference between the partial log-likelihood $\ell_{pl}(\boldsymbol{\beta})$ and the penalty,

$$\ell_{\text{pen}}(\boldsymbol{\beta}_\lambda) = \ell_{pl}(\boldsymbol{\beta}) - \frac{1}{2}\sum_j \lambda_j \int [f_j''(x_j)]^2 dx_j, \tag{5.8}$$

where $\ell_{pl}(\boldsymbol{\beta})$ is the partial log likelihood, and $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the smoothing parameters. The derivation of an integrated square second derivative penalty is described in Section 4.3.1. The j^{th} penalty term can be expressed as

$$\int [f_j''(x_j)]^2 dx_j = \boldsymbol{\beta}_{zj}^T \boldsymbol{Z}_j^T \boldsymbol{K}_j \boldsymbol{Z}_j \boldsymbol{\beta}_{zj},$$

where $\boldsymbol{Z}_j$ is a semi-orthogonal matrix of size $(n_k+1) \times (n_k-1)$, $\boldsymbol{K}_j$ is a square matrix of the j^{th} smoothing term of size $(n_k+1) \times (n_k+1)$, with (e, f) element being $12|x_{je}^* - x_{jf}^*|^3$, and the first row and column is equal to zero.

The model parameters can be obtained by maximizing the penalized partial log-likelihood, such that

$$\begin{aligned}
\ell_{pen}(\boldsymbol{\beta}_\lambda) &= \ell_{pl}(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_{zj}^T \boldsymbol{Z}_j^T \boldsymbol{K}_j \boldsymbol{Z}_j \boldsymbol{\beta}_{zj} \\
&= \ell_{pl}(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{\mathcal{K}} \boldsymbol{\beta} \\
&= \sum_{i=1}^{n} \delta_i [\boldsymbol{X}_i \boldsymbol{\beta}] - \sum_{i=1}^{n} \delta_i \log\left(\sum_{j \in R(t_i)} \exp[\boldsymbol{X}_j \boldsymbol{\beta}]\right) - \frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{\mathcal{K}} \boldsymbol{\beta},
\end{aligned} \tag{5.9}$$

where $\boldsymbol{\mathcal{K}}$ is the block diagonal matrix defined as $\boldsymbol{\mathcal{K}} = \operatorname{diag}(\mathbf{0}, \lambda_1 \boldsymbol{Z}_1^T \boldsymbol{K}_1 \boldsymbol{Z}_1, \ldots, \lambda_p \boldsymbol{Z}_p^T \boldsymbol{K}_p \boldsymbol{Z}_p)$. The parameters for linear effects, the parametric components remain unpenalized in the model, as only the parameters corresponding to smooth terms are penalized. The first and second derivative of the penalized partial log-likelihood with respect to $\boldsymbol{\beta}$ are given by

$$\begin{aligned}
U_{pen}(\boldsymbol{\beta}_\lambda) &= \frac{\partial \ell_{pen}(\boldsymbol{\beta}_\lambda)}{\partial \boldsymbol{\beta}_\lambda} = \frac{\partial \ell_{pl}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - \boldsymbol{\mathcal{K}}\boldsymbol{\beta} = U_{pl}(\boldsymbol{\beta}) - \boldsymbol{\mathcal{K}}\boldsymbol{\beta}, \\
I_{pen}(\boldsymbol{\beta}_\lambda) &= -\Big[\frac{\partial^2 \ell_{pen}(\boldsymbol{\beta}_\lambda)}{\partial \boldsymbol{\beta}_\lambda^2}\Big] = -\Big[\frac{\partial^2 \ell_{pl}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\Big] + \boldsymbol{\mathcal{K}} = I_{pl}(\boldsymbol{\beta}) + \boldsymbol{\mathcal{K}}.
\end{aligned}$$

For given values of smoothing parameters $\lambda_1, \lambda_2, \ldots, \lambda_p$, the penalized estimate for $\boldsymbol{\beta}_\lambda$ can be obtained by using Newton-Raphson algorithm, so an estimate of $\boldsymbol{\beta}_\lambda$ at the $(s+1)^{\text{th}}$ iteration procedure, $\hat{\boldsymbol{\beta}}_\lambda^{(s+1)}$ is calculated as

$$\hat{\boldsymbol{\beta}_\lambda}^{s+1} = \hat{\boldsymbol{\beta}_\lambda}^{s} + I_{pen}^{-1}(\hat{\boldsymbol{\beta}_\lambda}^{s})U_{pen}(\hat{\boldsymbol{\beta}_\lambda}^{s}), \tag{5.10}$$

where $U_{pen}(\hat{\boldsymbol{\beta}}_\lambda^s)$ is the penalized score vector and $I_{pen}^{-1}(\hat{\boldsymbol{\beta}}_\lambda^s)$ is the inverse of the penalized information matrix, both evaluated at $\hat{\boldsymbol{\beta}}_\lambda^s$. When the iterative procedure has converged, the inverse of the penalized information matrix $I_{pen}^{-1}(\hat{\boldsymbol{\beta}}_\lambda)$ can be used as an approximate variance-covariance matrix. Van Houwelingen and Verweij (1994) proposed the penalized standard Cox PH model, they used square root of the diagonal elements of $I_{pen}^{-1}(\hat{\boldsymbol{\beta}}_\lambda)$ as pesudo standard errors. Gray (1992) showed that $\hat{\boldsymbol{\beta}}_\lambda$ is asymptotically normal with mean 0 and variance-covariance matrix $\boldsymbol{V} = I_{pen}^{-1}(\hat{\boldsymbol{\beta}}_\lambda)I_{pl}(\hat{\boldsymbol{\beta}})I_{pen}^{-1}(\hat{\boldsymbol{\beta}}_\lambda)$. We wrote R code to carry out the Newton-Raphson algorithm in case of penalized additive Cox PH model using radial basis function as a smoothing term as an extension to our previous code for the standard Cox PH model. We did not use mgcv package because they represented the smoothing term using cubic regression spline, and thin plate regression spline.

The inference of the spline effect parameters is not important, because each spline term contains several spline parameters associated with the number of knots, and none of these spline effect parameters are statistically significant from zero in most cases. However, the estimation of spline effect parameters allows for visualizing the spline fit by plotting, which is described in Section 5.3.

5.3 The Smoothing Log Hazard Ratio

An excellent explanation of the hazard ratio in the standard Cox PH model and its interpretation can be found in Kay (2004) and Spruance et al. (2004). The standard Cox PH model only predicts relative risks between pairs of subjects, for the continuous covariate this means we can predict relative risk with respect to a reference value. The adjusted hazard ratio for a subject with continuous covariate value x_j compared to a subject with covariate value $x_{j,ref}$ can be expressed as

$$HR_j(x_j, x_{j,ref}) = \exp\{(x_j - x_{j,ref})\beta\},$$

where $x_{j,ref}$ is a particular value of the continuous covariate considered as the reference value, this reference value can be zero or the mean of the covariate. If the mean of continuous covariates is taken as a reference value, this effectively centers the covariates on their mean. The result of this is that the baseline hazard is evaluated at the mean of the continuous covariate, so basically it just redefines $h_0(t)$ in terms of the mean covariates, rather than zero. Plotting the logarithm of the hazard ratio (HR_j) against x_j yields a straight line.

In contrast, the explanation of construction of the log hazard ratio for continuous covariate in the additive Cox model, and its interpretation, can be found in Strasak et al. (2009), Cadarso-Surez et al. (2010) and Meira-Machado et al. (2013). The adjusted hazard ratio curve for a continuous covariate value x_j in the additive Cox PH model (5.1) can be expressed as

$$HR_j(x_j, x_{j,ref}) = \exp(f_j(x_j) - f_j(x_{j,ref})).$$

If we take $f_j(x_{j,ref})$ to be zero value, then we can replace $f_j(x_j)$ by the corresponding smoothing function estimate $\hat{\boldsymbol{f}}_j = \boldsymbol{X}_j\boldsymbol{Z}_j\hat{\boldsymbol{\beta}}_{zj}$, where $\boldsymbol{X}_j\boldsymbol{Z}_j$ is j^{th} smoothing matrix for j^{th} the covariate. The variance of the log hazard ratio estimate can be expressed in terms of the variance of the estimated parameter $\hat{\boldsymbol{\beta}}_{zj}$ and smoothing matrix of the smoothing function estimate $\hat{f}_j(x_j)$ as

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{f}}_{\boldsymbol{j}}) = \mathrm{Var}(\boldsymbol{X}_j\boldsymbol{Z}_j\hat{\boldsymbol{\beta}}_{zj}) &= \boldsymbol{X}_j\boldsymbol{Z}_j\mathrm{Var}(\hat{\boldsymbol{\beta}}_{zj})(\boldsymbol{X}_j\boldsymbol{Z}_j)^T \\ &= \boldsymbol{X}_j\boldsymbol{Z}_j\boldsymbol{I}_{\text{pen}}^{\boldsymbol{-1}}(\hat{\boldsymbol{\beta}}_{zj})(\boldsymbol{X}_j\boldsymbol{Z}_j)^T.\end{aligned}$$

Van Houwelingen and Verweij (1994) proposed the penalized standard Cox model, and they used the square root of the diagonal elements of $\boldsymbol{I}_{\text{pen}}^{-\mathbf{1}}$ as a pseudo-standard errors. We could compute the variance of the log hazard ratio using Gray's formula for variance covariance matrix, which makes the pointwise confidence band very narrow in the mean of the covariate. Relying on asymptotic normality of the estimated parameters $\hat{\boldsymbol{\beta}_{zj}}$, a 95% pointwise confidence band around the log hazard ratio can be calculated as

$$\hat{\boldsymbol{f}}_{ji} \pm 1.96\sqrt{\operatorname{Var}(\hat{\boldsymbol{f}}_{\boldsymbol{j}})_{ii}}. \tag{5.11}$$

5.4 Simulating Survival Time from Standard Parametric Distributions

We conducted simulation studies to investigate the performance of the additive Cox PH model using radial basis functions to construct the smoothing terms. The inverse methods to generate survival times for simulating the Cox proportional hazards model, for a variety of standard parametric distributions for the baseline hazard, are presented by Bender et al. (2005). Briefly, we describe the method for generating survival times to simulate the standard Cox PH models. The hazard function of the Cox PH model is defined in equation (3.1) as

$$h(t|\boldsymbol{X}) = h_0(t)\exp(\boldsymbol{X}\boldsymbol{\beta}), \tag{5.12}$$

where $h_0(t)$ is the baseline hazard function identified by some parametric distribution, $\boldsymbol{X}$ is a vector of one covariate, and $\boldsymbol{\beta}$ is the corresponding unknown fixed effect parameter. By assuming that the baseline hazard follows a parametric distribution, such as an Exponential or Weibull distribution, the survival time can be generated based on the inverse of the cumulative baseline hazard. The cumulative hazard $H(t|\boldsymbol{X})$, survival

$S(t|\boldsymbol{X})$, and cumulative distribution, $F(t|\boldsymbol{X})$, are defined as

$$\begin{aligned} H(t|\boldsymbol{X}) &= H_0(t)\exp(\boldsymbol{X}\beta), \quad \text{where} \quad H_0(t) = \int_0^t h_0(u)du, \\ S(t|\boldsymbol{x}) &= \exp\big\{-H(t|\boldsymbol{X})\big\}, \\ F(t|\boldsymbol{x}) &= 1 - \exp\big\{-H(t|\boldsymbol{X})\big\} \\ &= 1 - \exp\big\{-H_0(t)\exp(\boldsymbol{X}\beta)\big\}. \end{aligned}$$

If we let Y to be the random variable with distribution function F, then $U = F(Y)$ follows a uniform distribution, and if $U \sim U[0,1]$ then $1 - U \sim U[0,1]$. Let T be the simulated survival time of the Cox PH model. Then Bender et al. (2005) showed that

$$F(T|\boldsymbol{X}) = 1 - \exp\big\{H_0(t)\exp(\boldsymbol{X}\beta)\big\} = U, \quad \text{where} \quad U \sim U[0,1]$$

or equivalently

$$U = \exp\big\{-H_0(T)\exp(\boldsymbol{X}\beta)\big\} \sim U[0,1]. \tag{5.13}$$

If $h_0(T) > 0$, then equation (5.13) can be solved for the survival time T as long as $H_0(t)$ can be inverted. In this case,

$$T = H_0^{-1}\Big\{-\frac{\log(U)}{\exp(\boldsymbol{X}\beta)}\Big\}. \tag{5.14}$$

To simulate survival times of the Cox model in the case with a constant baseline hazard λ_{EXP}, equation (5.14) becomes

$$T = -\frac{\log(U)}{\lambda_{EXP} \times \exp(\boldsymbol{X}\beta)}. \tag{5.15}$$

Also to simulate survival times of a Cox model with the baseline hazard of a Weibull distribution with scale parameter λ_{Wei} and shape parameter ν, equation (5.14) can be written as

$$T = \Big(\frac{-\log(U)}{\lambda_{Wei} \times \exp(\boldsymbol{X}\beta)}\Big)^{\frac{1}{\nu}}. \tag{5.16}$$

5.4.1 Generating Survival Times to Simulate Cox PH models

We created the following algorithm to simulate the survival time from the standard Cox PH model:

1. Set the number of observations n.

2. Set $\{x_i : i = 1, \dots, n\}$.

3. Set $\breve{\beta}$.

4. Let n_{sim} be the number of simulations.

5. Generate $T_{ik} = H_0^{-1}\Big\{-\frac{\log(U)}{\exp(\boldsymbol{x}_i\breve{\boldsymbol{\beta}})}\Big\}$, where U is a random variable uniformly distributed on interval $[0, 1]$, and T_{ik} is the generated failure time for the i^{th} individual and for the k^{th} simulation $k = 1, \dots, n_{\text{sim}}$.

6. Generate censoring times C_{ik} randomly drawn from an exponential distribution with rate $\frac{1}{2}\min\limits_i(\exp(x_i\breve{\beta}))$, where C_{ik} is the censoring time for the i^{th} individual for the k^{th} simulation for $k = 1, \dots, n_{\text{sim}}$.

7. Evaluate observed value $t_{ik} = \min(T_{ik}, C_{ik})$, where t_{ik} is the observed survival time for the i^{th} individual for the k^{th} simulation.

8. Evaluate censoring indicator $\delta_{ik} = I(T_{ik} \leq C_{ik})$, where δ_{ik} is the censoring indicator of the i^{th} individual for the k^{th} simulation. $\delta_{ik} = 1$ if the event was observed $(T_{ik} \leq C_{ik})$, and zero if $(T_{ik} > C_{ik})$

5.5 Simulation Study

5.5.1 Generating Survival Times to Simulate Additive Cox PH Models

In this section we develop a general method for simulating data from the additive Cox model, we used a technique described in Bender et al. (2005) in more general form by replacing the linear predictor in the hazard function of standard Cox PH model by the known smoothing function, such as $f(x)$, so the Cox PH hazard function becomes

$$h(t|x_i) = h_0(t)\exp(f(x_i)).$$

If $f(x_i) = x_i\beta$, this is the standard Cox PH model. The general algorithm to simulate survival times from the additive Cox PH model is an extension to Bender et al. (2005) can be summarized as follows:

1. Set the number of observations n.

2. Set $\{x_i : i = 1, \dots, n\}$.

3. Define the known smoothing function $f(x_i)$.

4. Let n_{sim} be the number of simulations.

5. Generate $T_{ik} = H_0^{-1}\Big\{-\frac{\log(U)}{\exp(f(x_i))}\Big\}$, and $U \sim U[0,1]$, where T_{ik} is the generation failure time for the i^{th} observation and for the k^{th} simulation $k = 1, \dots, n_{\text{sim}}$.

6. Generate censoring times C_{ik} randomly drawn from an exponential distribution with rate $\frac{1}{2}\min_i(\exp(f(x_i)))$, where C_{ik} is the censoring time for the i^{th} individuals for the k^{th} simulation.

7. Evaluate observed value $t_{ik} = \min(T_{ik}, C_{ik})$, where t_{ik} is the observed survival time for the i^{th} individuals for the k^{th} simulation.

8. Evaluate censoring indicator $\delta_{ik} = I(T_{ik} \leq C_{ik})$, where δ_{ik} is the censoring indicator of the i^{th} individual for the k^{th} simulation. $\delta_{ik} = 1$ if the event was observed $(T_{ik} \leq C_{ik})$, and zero if $(T_{ik} > C_{ik})$

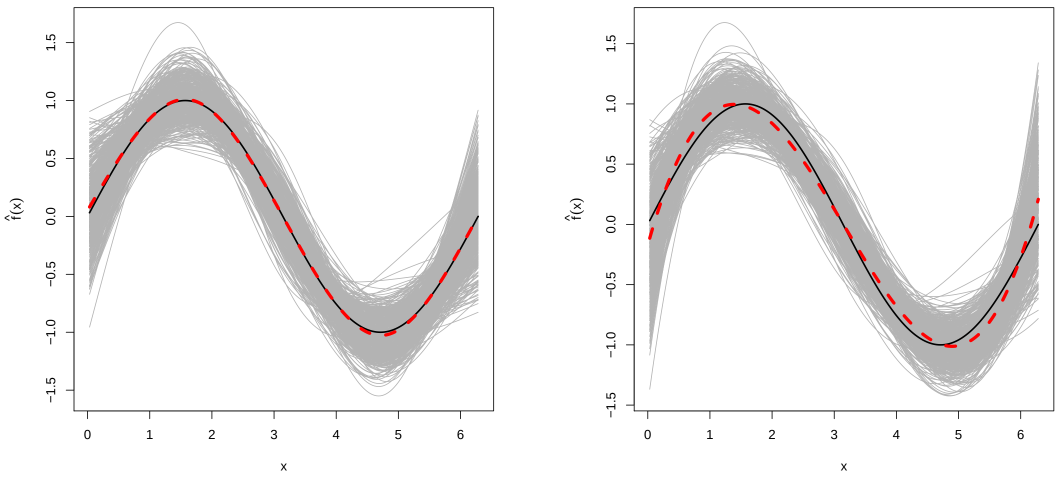
Once we have data $(t_{ik}, \delta_{ik}, x_i)_{i=1}^{n}$ for $k = 1, \dots, n_{\text{sim}}$, we fit the unpenalized additive Cox model to obtain the model parameters, and the estimate of the log hazard ratios are calculated separately for each simulation. This algorithm will be used to simulate the survival time to assess the performance of the proposed method.

5.5.2 Simulating Additive Cox Model for One Smooth Term

We ran 500 simulations with sample size $n = 200$, where the covariate was defined as $x_i = \frac{i}{n} \times 2\pi, \;\; i = 1, \dots, n$, and the true curve was the known smoothing function, which is defined by $f(x_i) = \sin(x_i)$. We used the algorithm described in Section 5.5.1 to generate survival data to fit the unpenalized additive Cox model for one smooth term using the radial basis function and the cubic B-spline. We generated survival times from the additive Cox model with constant baseline hazard $\lambda_{EXP} = 1$, which were calculated as $T_{ik} = -\frac{\log(U)}{\exp(f(x_i))}$, where $U \sim U[0, 1]$. The censoring times C_{ik} are randomly drawn from an exponential distribution with rate $\frac{1}{2}\min(\exp(f(x_i)))$. This generated data is such that approximately 17.63% of the observations were censored.

In order to compare the additive Cox PH model using radial basis function with additive Cox PH model using cubic B-spline, we fit the unpenalized additive Cox model with five equally spaced knots to obtain the model parameters using the radial basis function and cubic B-spline as the smoothing term separately, and then the estimation of the log hazard ratio for each simulation were calculated separately. The estimated log hazard ratios for each simulation were plotted versus the covariate x_i, and the average of these 500 estimated log hazard ratio was also plotted versus x_i.

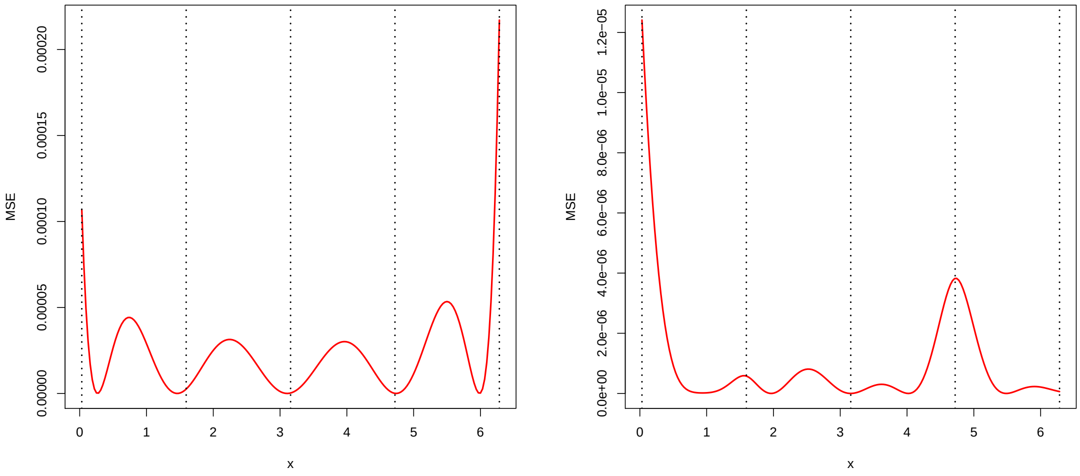
Figure 5.1 (a) shows the plot of the estimated log hazard ratio $\hat{f}(x)$ using radial basis function versus x. The gray lines are the estimated log hazard ratio $\hat{f}(x)$ for the 500 simulations, the black line indicates the true curve $f(x) = \sin(x)$, and the red dashed line is the average of $\hat{f}(x)$ for the 500 simulations, while Figure 5.1 (b) shows the plot of the estimated log hazard ratio $\hat{f}(x)$ using cubic B-spline versus x, which shows poor estimate of the log hazard ratio at the boundary knots. As a result, the average of the 500 estimated log hazard ratios was equal to the true curve, but the radial basis function has a better ending behavior, which means the estimated log hazard ratio is numerical stability at the knots. The mean square error was computed for both and the estimated log hazard ratio using radial basis function has a small mean square error compare to the estimated log hazard ratio using cubic B-spline as illustrated in Figure 5.2



(a) Radial basis function

(b) Cubic B-spline

Figure 5.1: The plot of the estimated log hazard ratio $\hat{f}(x)$ versus x. The gray lines are the estimated log hazard ratio $\hat{f}(x)$ for the 500 simulations, the black line indicates the true curve $f(x) = \sin(x)$, and the red dashed line is the average of $\hat{f}(x)$ for the 500 simulations.



(a) Cubic B-spline

(b) Radial basis function

Figure 5.2: The plot of the mean square error versus x, the dotted vertical lines indicates the locations of the knots.

5.5.3 Simulating Additive Cox Model for Two Smoothing Terms

This section presents the simulation example of fitting the additive Cox model that includes two smoothing terms. We ran 500 simulations with sample size $n = 200$, let $x_{1i} = \frac{i}{n} \times 2\pi$ for $i = 1, \ldots, n$ and $f_1(x_{1i}) = \sin(x_{1i})$. Let x_{2i} be the normal random variable with mean 0 and variance equal to 1, and $f_2(x_{2i}) = (x_{2i}^2 + x_{2i}^3)/3$. The algorithm in Section 5.5.1 was used to generate survival data to fit the unpenalized additive Cox PH model for two smoothing terms. The survival time of the additive Cox PH model with constant baseline hazard $\lambda_{EXP} = 1$ is given by $T_{ik} = -\frac{\log(U)}{\exp(f_1(x_{1i})+f_2(x_{2i}))}$, where $U \sim U[0, 1]$. The censoring times C_{ik} are randomly drawn from an exponential distribution with rate equal to $\min \exp\Big(f_1(x_{1i}) + f_2(x_{2i})\Big)$. This generated data where approximately 4.84% of the observations were censored. The unpenalized additive Cox PH model with five equally spaced knots for each smoothing term was fitted. The estimated log hazard ratio for the first and second smoothing terms, for each simulation, were computed. The results indicate that the mean of the 500 simulations for the two additive predictors is equal to the true curve. Our simulation results suggest the proposed model works well for more than one smoothing term.

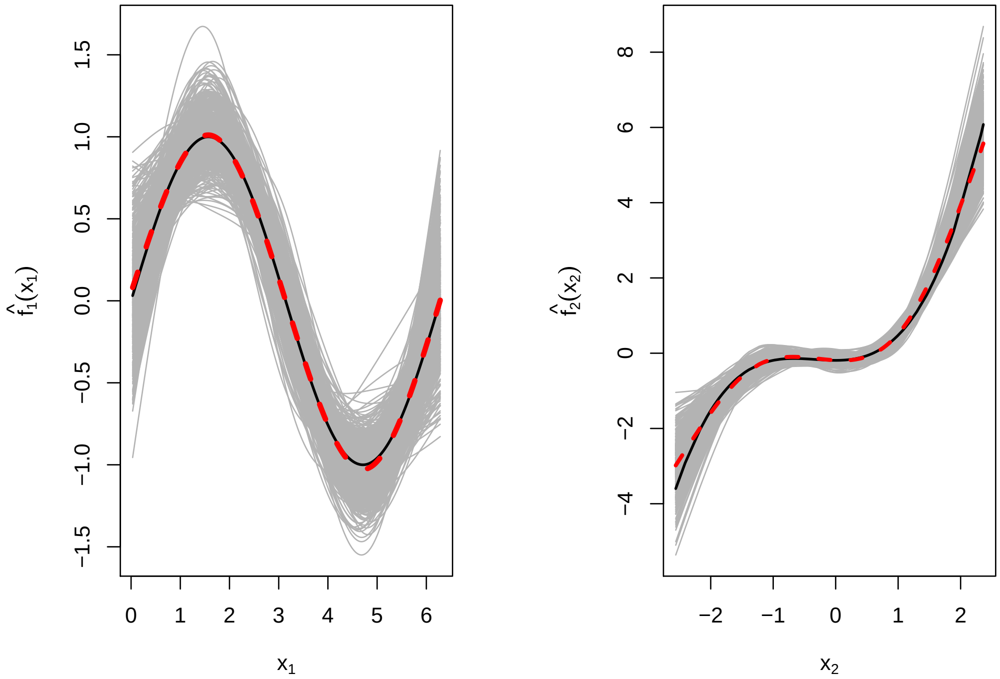


Figure 5.3: The left panel is the plot of the estimated log hazard ratio for the first smooth function $\hat{f}_1(x_1)$ evaluated at the observed x_1. The right panel is the plot of the estimate log hazard ratio of the second smooth function of $\hat{f}_2(x_2)$, evaluated at the observed x_2. The black solid line is the true curve and the red dashed lines indicate the average of 500 simulations.

5.5.4 Simulating Additive Cox Model with One Smoothing Term and One Categorical Variable

The additive Cox PH model that includes one categorical variable and one smooth term is

$$h(t_i|x_{1i}, x_{2i}) = \lambda \exp\Big(f(x_{1i}) + \theta I(x_{2i} = 1)\Big).$$

We ran 100 simulations with a sample size $n = 200$, let $x_{1i} = \frac{i}{n} \times 2\pi$ for $i = 1, \dots, n$, and $f(x_{1i}) = \sin(x_{1i})$, and let the categorical variable x_{2i} be a random sample of size n from a Bernoulli random variable which were equal to 1 with probability $p = 0.5$, and 0 with probability $1 - p$. The algorithm in Section 5.5.1 is used to generate survival times from the additive Cox model. The true categorical parameter $\breve{\theta}$ is equal to 1, so the survival time of additive Cox model with constant baseline hazard is given by $T_{ik} = -\frac{\log(U)}{\exp(f(x_{1i})+\breve{\theta}I(x_{2i}=1))}$, where $U \sim U[0, 1]$. The censoring times C_{ik} were randomly drawn from an exponential distribution with rate $\frac{1}{2}\min\Big(\exp\big(f(x_{1i})+\breve{\theta}I(x_{2i} = 1)\big)\Big)$. This generated data had approximately 9.60% of the observations were censored. The unpenalized additive Cox model with five equally spaced knots for the smoothing term was fitted. The estimated log hazard ratio for the first term, for each simulation, were computed and plotted versus x_{1i}. The results indicate that the mean of the 100 simulations for the additive predictors is equal to the true curve. Also, the histogram of the estimated categorical parameter $\hat{\theta}$ shows the mean of the histogram is equal to the true parameter $\breve{\theta}$, which suggests that the proposed model works well.

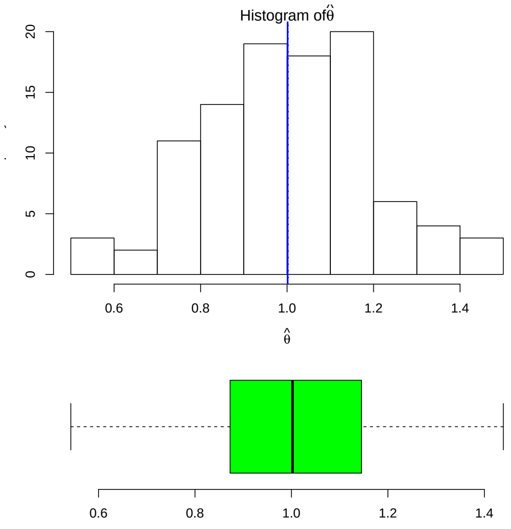


Figure 5.5: A histogram of the estimated parameter for the categorical term.

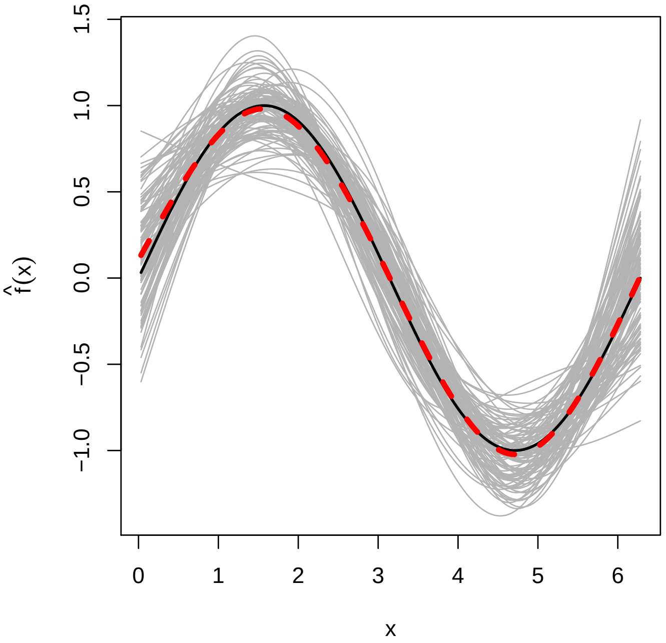


Figure 5.4: The plot of the estimated log hazard ratio of the smooth function of $\hat{f}_1(x_1)$ evaluated at the observed x_{1i}. The black solid line is the true curve $f(x_{1i}) = \sin(x_{1i})$, whereas the red dotted line indicate the mean of the 500 simulations.

5.6 Test for Covariates Effects

5.6.1 Test for the Spline Effect Parameters

In the additive Cox PH model there are fixed effect parameters and a spline effect parameter, we are interested in testing a hypothesis about subsets of $\boldsymbol{\beta}$ parameters, in particular the spline effect parameters. The hypothesis is then $H_0 : \boldsymbol{\beta}_z = \mathbf{0}$, where $\boldsymbol{\beta} = [\boldsymbol{\zeta}, \boldsymbol{\beta}_z^T]^T$, where $\boldsymbol{\zeta}$ is a vector of the fixed effect parameters and $\boldsymbol{\beta}_z$ is the vector of $(n_k - 1)$ unconstrained spline effect parameters. For fixed effect parameters, the null hypothesis is that $\boldsymbol{\zeta} = \mathbf{0}$, which can be tested by using the Wald statistic as described in Section 3.3.2. For the spline effect parameters, there are two hypotheses regarding the smooth function $f(x)$. Firstly, that the covariate has no effect, which can be expressed as $[\alpha_0, \alpha^T] = \boldsymbol{\beta}_c = \boldsymbol{\beta}_z = \mathbf{0}$, which means $f(x) = 0$ and secondly, that the covariate has a linear effect, so $[\alpha^T] = \mathbf{0}$, which means $f(x) = \alpha_0 x$. Gray (1992, 1994) recommended a penalized versions of the score, Wald, and likelihood ratio tests.

First consider the null hypothesis $H_0 : \boldsymbol{\beta}_z = \mathbf{0}$, where the unconstrained parameter vector $\boldsymbol{\beta}_z$ is equal to the zero. Let $\hat{\boldsymbol{\beta}}_\lambda = [\hat{\boldsymbol{\zeta}}, \hat{\boldsymbol{\beta}}_z^T]^T$ be the values of the parameters that maximize the penalized partial log-likelihood, and let $\hat{\boldsymbol{\zeta}}_0$ be the maximum unpenalized partial log likelihood estimator for $\boldsymbol{\zeta}$ when $\boldsymbol{\beta}_z = \mathbf{0}$. Denote the penalized partial score vector by

$$\begin{aligned} U_{pen}(\boldsymbol{\zeta}, \boldsymbol{\beta}_z) &= \left(\Big(\frac{\partial \ell_{pen}(\boldsymbol{\zeta}, \boldsymbol{\beta}_z)}{\partial \boldsymbol{\zeta}}\Big)^T, \Big(\frac{\partial \ell_{pen}(\boldsymbol{\zeta}, \boldsymbol{\beta}_z)}{\partial \boldsymbol{\beta}_z}\Big)^T \right)^T \\ &= \left(U^T_{pen,\boldsymbol{\zeta}}(\boldsymbol{\zeta}, \boldsymbol{\beta}_z), U^T_{pen,\boldsymbol{\beta}_z}(\boldsymbol{\zeta}, \boldsymbol{\beta}_z) \right)^T . \end{aligned}$$

Let $\boldsymbol{I}_{pl}$ be the Fisher information matrix from the unpenalized partial log-likelihood with subscript denoting the sub-matrices,

$$I_{pl} = \begin{bmatrix} I_{\boldsymbol{\zeta}\boldsymbol{\zeta}} & I_{\boldsymbol{\zeta}\boldsymbol{\beta}_z} \\ I_{\boldsymbol{\beta}_z\boldsymbol{\zeta}} & I_{\boldsymbol{\beta}_z\boldsymbol{\beta}_z} \end{bmatrix},$$

where $I_{\boldsymbol{\zeta}\boldsymbol{\zeta}} = -\Big(\frac{\partial^2 \ell_{pl}(\boldsymbol{\zeta},\boldsymbol{\beta}_z)}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}^T}\Big)$, $I_{\boldsymbol{\zeta}\boldsymbol{\beta}_z} = -\Big(\frac{\partial^2 \ell_{pl}(\boldsymbol{\zeta},\boldsymbol{\beta}_z)}{\partial\boldsymbol{\zeta}\partial\boldsymbol{\beta}_z^T}\Big)$, $I_{\boldsymbol{\beta}_z\boldsymbol{\zeta}} = -\Big(\frac{\partial^2 \ell_{pl}(\boldsymbol{\zeta},\boldsymbol{\beta}_z)}{\partial\boldsymbol{\zeta}^T\partial\boldsymbol{\beta}_z}\Big)$, and $I_{\boldsymbol{\beta}_z\boldsymbol{\beta}_z} = -\Big(\frac{\partial^2 \ell_{pl}(\boldsymbol{\zeta},\boldsymbol{\beta}_z)}{\partial\boldsymbol{\beta}_z\partial\boldsymbol{\beta}_z^T}\Big)$. Let the sub-vector $U_{pen,\boldsymbol{\beta}_z}(\hat{\boldsymbol{\zeta}}_0, \mathbf{0})$ denote the first derivative of the penalized log-likelihood evaluated at $\boldsymbol{\beta}_z = \mathbf{0}$, which is given by $U_{pen,\boldsymbol{\beta}_z}(\hat{\boldsymbol{\zeta}}_0, \mathbf{0}) = \frac{\partial \ell_{pl}(\hat{\boldsymbol{\zeta}}_0,\mathbf{0})}{\partial\boldsymbol{\beta}_z}$. The negative of the $\boldsymbol{\beta}_z\boldsymbol{\beta}_z$ portion of the second derivative of the penalized log-likelihood with respect to $\boldsymbol{\beta}_z$ is given by $\boldsymbol{I}_{\boldsymbol{\beta_z}\boldsymbol{\beta_z}} + \lambda\boldsymbol{\mathcal{K}}$. A penalized score statistic can be expressed as

$$T_{sc} = U^T_{pen,\boldsymbol{\beta}_z}(\hat{\boldsymbol{\zeta}}_0, \mathbf{0})(\boldsymbol{I}_{\boldsymbol{\beta_z}\boldsymbol{\beta_z}|\boldsymbol{\zeta}} + \lambda\boldsymbol{\mathcal{K}})^{-1} U_{pen,\boldsymbol{\beta}_z}(\hat{\boldsymbol{\zeta}}_0, \mathbf{0}), \tag{5.17}$$

where $\boldsymbol{I}_{\boldsymbol{\beta_z}\boldsymbol{\beta_z}|\boldsymbol{\zeta}} = \boldsymbol{I}_{\boldsymbol{\beta_z}\boldsymbol{\beta_z}} - \boldsymbol{I}_{\boldsymbol{\beta}_z\boldsymbol{\zeta}}\boldsymbol{I}^{-1}_{\boldsymbol{\zeta}\boldsymbol{\zeta}}\boldsymbol{I}_{\boldsymbol{\zeta}\boldsymbol{\beta}_z}$. The penalized likelihood ratio statistic can be defined by

$$T_r = 2[\ell_{pen}(\hat{\boldsymbol{\zeta}}, \hat{\boldsymbol{\beta}}_z) - \ell_{pen}(\hat{\boldsymbol{\zeta}}_0, \mathbf{0})]. \tag{5.18}$$

This penalized likelihood ratio statistic is similar to the deviance statistic that are discussed by Hastie and Tibshirani (1990a,b). The Wald-type test of this null hypothesis, $H_0 : \boldsymbol{\beta}_z = \mathbf{0}$, is based on the maximum penalized partial log-likelihood estimation of $\boldsymbol{\beta}$. The Wald-type statistic can be expressed as

$$T_w = \hat{\boldsymbol{\beta}}_z^T (\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\boldsymbol{\zeta}} + \lambda \boldsymbol{\mathcal{K}}) \hat{\boldsymbol{\beta}}_z. \tag{5.19}$$

Using the fact that $\hat{\boldsymbol{\beta}}_c = \boldsymbol{Z}\hat{\boldsymbol{\beta}}_z$, we can compute the variance of $\hat{\boldsymbol{\beta}}_c$ as

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}_c) &= \mathrm{Var}(\boldsymbol{Z}\hat{\boldsymbol{\beta}}_z) \\
&= \boldsymbol{Z}\mathrm{Var}(\hat{\boldsymbol{\beta}}_z)\boldsymbol{Z}^T \\
&= \boldsymbol{Z}(\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\boldsymbol{\zeta}} + \lambda \boldsymbol{\mathcal{K}})^{-1}\boldsymbol{Z}^T.
\end{aligned}$$

The Wald-type statistic for $\hat{\boldsymbol{\beta}}_c$ can be expressed as

$$T_c = (\boldsymbol{Z}\hat{\boldsymbol{\beta}}_z)^T \Big(\boldsymbol{Z}(\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\boldsymbol{\zeta}} + \lambda \boldsymbol{\mathcal{K}})^{-1}\boldsymbol{Z}^T\Big)^{-1} (\boldsymbol{Z}\hat{\boldsymbol{\beta}}_z), \tag{5.20}$$

The simulation result in Section 5.6.2 shows that $\hat{\boldsymbol{\beta}}_c$ is asymptotically normal with mean 0 and variance covariance matrix $(\boldsymbol{Z}\boldsymbol{I}_{pen}^{-1}(\hat{\boldsymbol{\beta}}_{z,\lambda})\boldsymbol{Z}^T)^{-1}$. These tests are rejected for large values of the statistic.

The test for the second hypothesis, which is for the effect being linear, $H_0 : f(x) = \alpha_0 x$, (i.e. $\boldsymbol{\alpha}^T = \mathbf{0}$), can be done in exactly the same way, but including α_0 with $\boldsymbol{\zeta}$ instead of $\boldsymbol{\beta}_z$.

The effective degrees of freedom replace the number of parameters in the model, since the penalty tends to reduce the number of parameters as λ gets larger. The effective degrees of freedom of the test is given by

$$df = \text{trace}[\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\zeta}(\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\zeta} + \lambda\boldsymbol{\mathcal{K}})^{-1}]. \tag{5.21}$$

This formula for the degrees of freedom corresponds more closely to Definition 3 in Buja et al. (1989), than to the formula in Hastie and Tibshirani (1990a). The effective degrees of freedom in the penalized standard Cox model proposed by Van Houwelingen and Verweij (1994) are identical to those in equation (5.21). The degrees of freedom for the Wald-type statistic regarding $\hat{\boldsymbol{\beta}}_c$ can be shown to be

$$\begin{aligned} df &= \text{trace}[(\boldsymbol{Z}\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\zeta}\boldsymbol{Z}^T)\big(\boldsymbol{Z}(\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\zeta} + \lambda\boldsymbol{\mathcal{K}})^{-1}\boldsymbol{Z}^T)], \\ &= \text{trace}[\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\zeta}(\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\zeta} + \lambda\boldsymbol{\mathcal{K}})^{-1}] \end{aligned}$$

This shows that the effective degrees of freedom for the Wald-type statistics regarding $\hat{\boldsymbol{\beta}}_c$ or $\hat{\boldsymbol{\beta}}_z$ are identical. The degrees of freedom are the trace of the matrix $[\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\zeta}(\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\zeta} + \lambda\boldsymbol{\mathcal{K}})^{-1}]$, and can be computed for both the penalized and the unpenalized problem.

The above definition of the effective degrees of freedom is motivated by the degrees of freedom for generalized additive models (GAMs). The degrees of freedom in GAMs can be defined as

$$\begin{aligned} df = \text{trace}(\boldsymbol{H}) &= \text{trace}\Big[\boldsymbol{X}(\boldsymbol{X^T W X} + \lambda\boldsymbol{\mathcal{K}})^{-1}\boldsymbol{X^T W}\Big], \\ &= \text{trace}\Big[\underbrace{\boldsymbol{X^T W X}}_{\boldsymbol{I_{pl}}}\underbrace{(\boldsymbol{X^T W X} + \lambda\boldsymbol{\mathcal{K}})^{-1}}_{\boldsymbol{I}_{pen}}\Big], \end{aligned}$$

where the hat matrix $\boldsymbol{H}$, the design matrix $\boldsymbol{X}$ and the weight matrix $\boldsymbol{W}$ are defined in Chapter 4, and λ and $\boldsymbol{\mathcal{K}}$ are as before.

Under the null hypothesis $H_0 : \boldsymbol{\beta}_z = \boldsymbol{0}$ the statistics T_{sc}, T_r and T_w all have the same asymptotic distribution for a fixed number of knots, which is $\sum e_i Y_i^2$, where the $\{e_i\}$ are the eigenvalues of $\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\zeta}(\boldsymbol{I}_{\boldsymbol{\beta_z\beta_z}|\zeta} + \lambda\boldsymbol{\mathcal{K}})^{-1}$ and Y_i are IID standard normal random variables. For the un-penalized Cox model, the e_i are all either 0 or 1, so the test statistic have a chi-squared distribution with $\sum e_i$ degrees of freedom. However, in the penalized additive Cox model, $0 \leq e_i \leq 1$, and the test statistics have mean $\sum e_i$ and variance $2\sum e_i^2 < 2\sum e_i$, so using the chi square distribution as an approximation to the distribution of the test statistic with $\sum e_i$ degrees of freedom will make it conservative. The variance and the degrees of freedom for the additive Cox model are computed as outlined in Van Houwelingen and Verweij (1994). However, the asymptotic variance estimator $\boldsymbol{V}$ can be used, but $\boldsymbol{I}_{pen}^{-1}$ tends to be larger then $\boldsymbol{V}$, and $\boldsymbol{I}_{pen}^{-1}$ is more conservative for the test statistic and for computing the confidence band. Therneau et al. (2003) suggested using $\boldsymbol{I}_{pen}^{-1}$ in the significance test as it is a more reliable choice, and we used $\boldsymbol{I}_{pen}^{-1}$ in our simulation examples and for the related problem of penalized additive Cox PH model.

5.6.2 Simulation Under the Null Hypothesis of the Tests in the Penalized Additive Cox PH Model

This section presents the simulation examples under the null hypothesis that there is no covariate effect in the survival model. The aim of the simulation examples is to assess the performance of the Wald-type statistic, penalized likelihood ratio statistic and the penalized score statistic. The univariate additive Cox PH model is expressed as

$$h(t|x_i) = h_0(t)\exp\Big(f(x_i)\Big), \quad i = 1, \dots, n. \tag{5.22}$$

The sample size is $n = 200$, and the covariate x is generated from a normal distribution with mean 0 and variance 1, under the null hypothesis $f(x_i) = 0$ so $h(t|x_i) = h_0(t)$. The algorithm in Section 5.5.1 was used to generate the survival time of the additive Cox model with constant baseline hazard $\lambda_{EXP} = 1$, and the censoring times were randomly drawn from an exponential distribution with rate $\frac{1}{2}\min\Big(\exp(f(x_i)\Big)$, which gave about 33% censored observations.

The null hypothesis is that $H_0 : f(x_i) = 0$, which means $\boldsymbol{\beta}_c = \boldsymbol{\beta}_z = \mathbf{0}$. For 5 knots, the penalized additive Cox PH model was fitted for $n_{\text{sim}} = 300$ simulations, and then the penalized additive Cox PH model test for no effect of the covariate was carried out as described in Section 5.6.1 without the fixed effect parameters, as no such terms were included in the model.

Firstly, the results of applying the Wald-type statistic to test the hypothesis $H_0 : \boldsymbol{\beta}_c = \boldsymbol{\beta}_z = \mathbf{0}$, for the value of the smoothing parameters $\lambda = 0.001$ are shown in Figures 5.6 and 5.7, while Figures 5.13 and 5.14 show the results of applying the Wald-type statistic to test the hypothesis $H_0 : \boldsymbol{\beta}_c = \boldsymbol{\beta}_z = \mathbf{0}$, for $\lambda = 0.01$.

The values of the Wald-type test statistics regarding the unconstrained spline parameter $\boldsymbol{\beta}_z$, and the constrained spline parameter $\boldsymbol{\beta}_c$ are identical, as well as the degrees of freedom and the p-values. This results can be seen in Figure 5.8 for $\lambda = 0.001$, and in Figures 5.15 for $\lambda = 0.01$.

For both values of the smoothing parameters, the estimation of the unconstrained spline effect parameter $\boldsymbol{\beta}_z$ and the constrained spline effect parameter $\boldsymbol{\beta}_c$ are asymptotically normal with mean zero. This results are summarized in Figures 5.9, 5.10, 5.11, and 5.12, for $\lambda = 0.001$, and Figures 5.16, 5.18 5.17, and 5.19 for $\lambda = 0.01$.

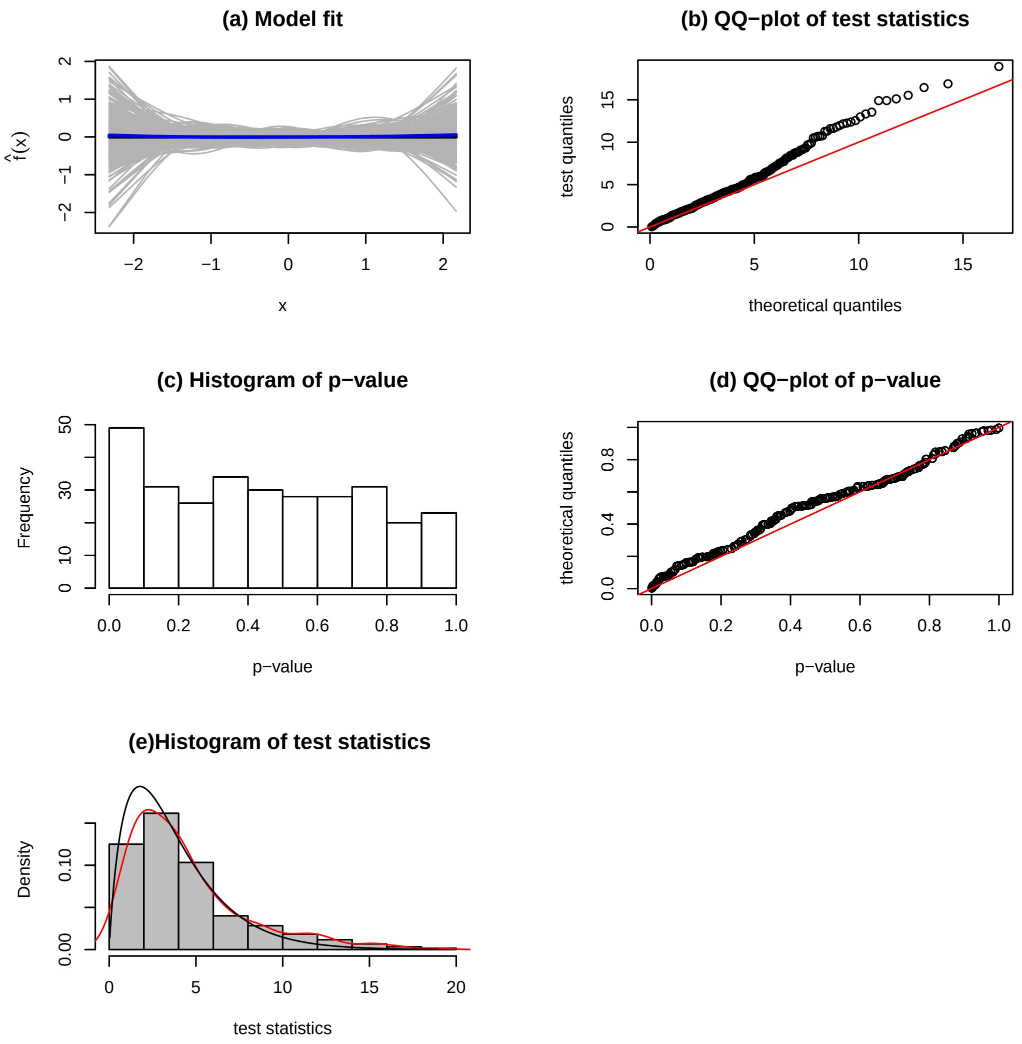


Figure 5.6: (a) The model fit for $\lambda = 0.001$. The solid black line is the mean of the estimated smoothing term for 300 simulations. (b) Chi-Square Q-Q plot of test statistics for $\hat{\boldsymbol{\beta}}_z$ with 3.704 degrees of freedom. (c) Histogram of p-values. (d) Uniform Q-Q plot of the p-values. (e) Histogram of the test statistic, the black solid line is the theoretical density, and the red solid line is the kernel density estimate of the test statistics. The rejection rate is 32 out of 300 simulations.

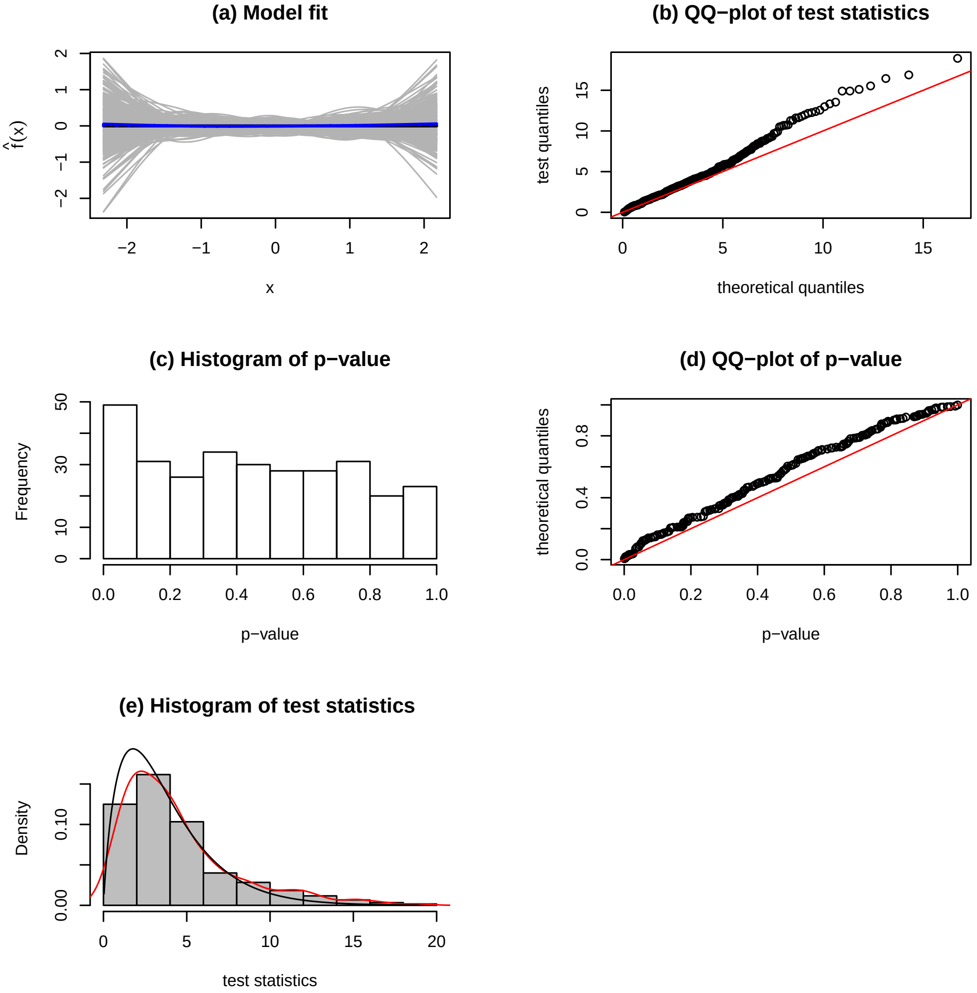


Figure 5.7: (a) The model fit for $\lambda = 0.001$. The solid black line is the mean of the estimated smoothing term for 300 simulations. (b) Chi-Square Q-Q plot of test statistics regarding $\hat{\boldsymbol{\beta}}_c$ with 3.704 degrees of freedom. (c) Histogram of p-values. (d) Uniform Q-Q plot of the p-values. (e) Histogram of the test statistic, the black solid line is the theoretical density, and the red solid line is the kernel density estimate of the test statistics. The rejection rate is 32 out of 300 simulations.

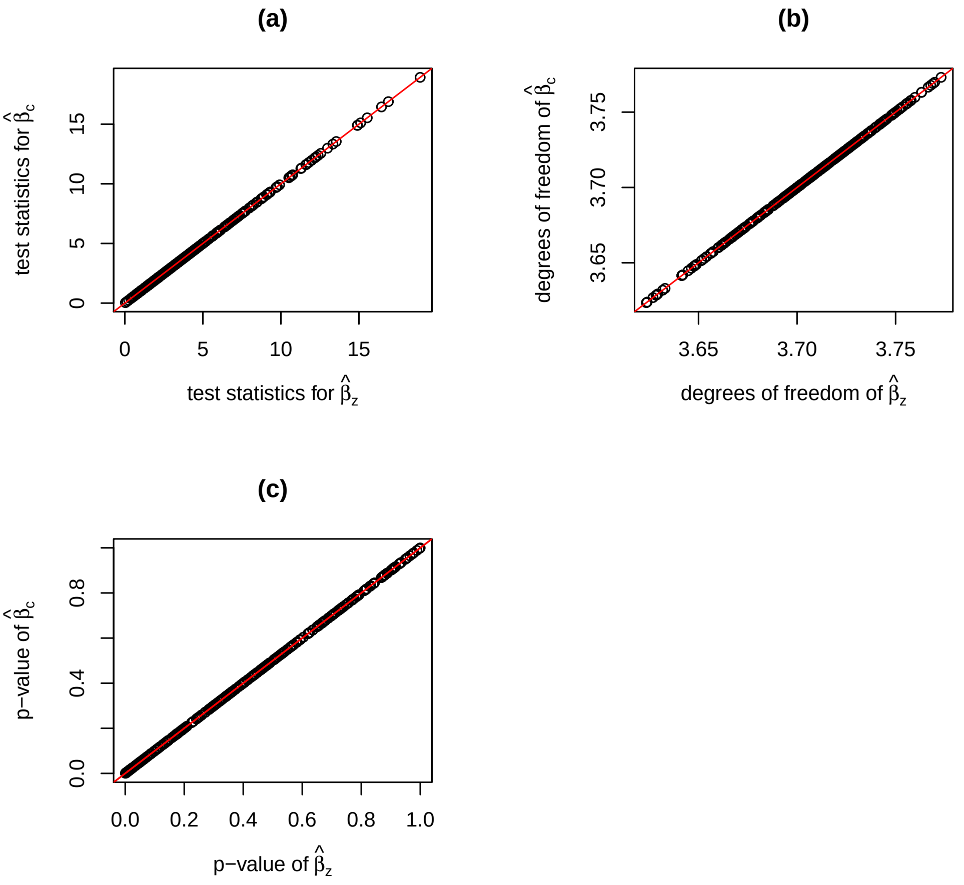


Figure 5.8: (a) Plot of the values of the Wald-type statistics for $\hat{\boldsymbol{\beta}}_z$ versus the values of the Wald-type statistic for $\hat{\boldsymbol{\beta}}_c$, when $\lambda = 0.001$. (b) Plot of the degrees of freedom for $\hat{\boldsymbol{\beta}}_z$ versus the degrees of freedom for $\hat{\boldsymbol{\beta}}_c$ for $\lambda = 0.001$, the range of the effective degrees of freedom is between 3.6235 and 3.7730. (c) Plot of the p-values of the test statistics for $\boldsymbol{\beta}_z$ versus the p-values of the test statistic for $\hat{\boldsymbol{\beta}}_c$

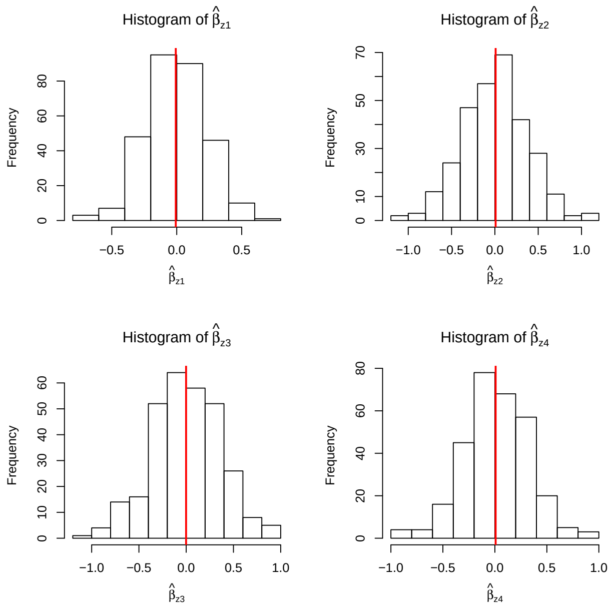


Figure 5.9: Histogram of the unconstrained estimated spline parameter $\hat{\boldsymbol{\beta}}_{zi}$ for $\lambda = 0.001$.

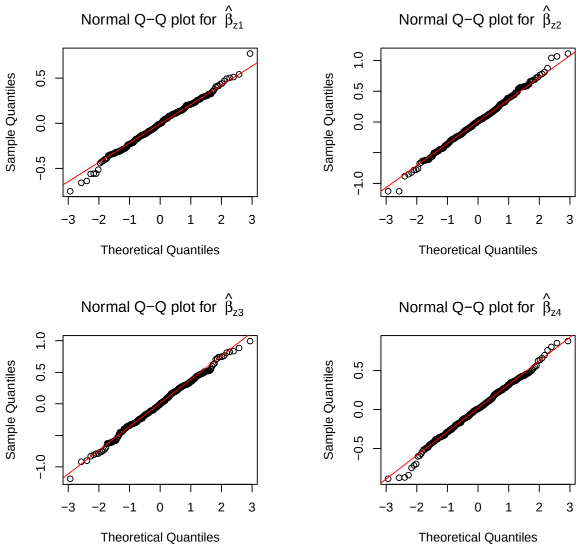


Figure 5.10: The Normal quantile-quantile plot of the unconstrained estimated spline parameter $\hat{\boldsymbol{\beta}}_{zi}$ for $\lambda = 0.001$.

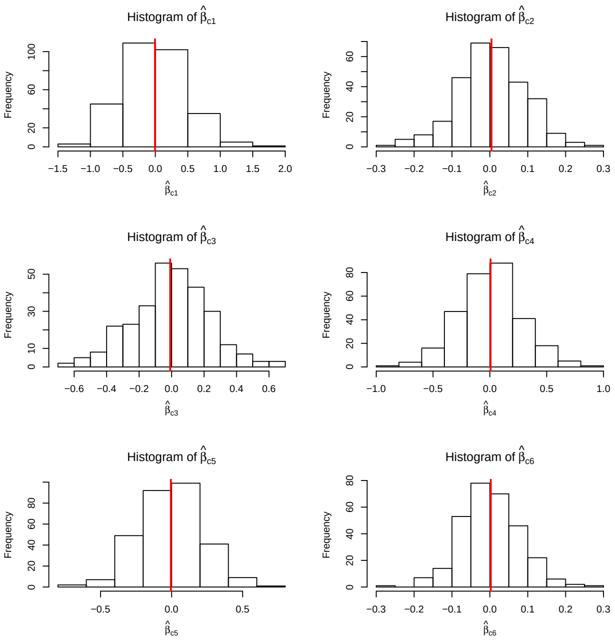


Figure 5.11: Histogram of the constrained estimated parameter $\hat{\boldsymbol{\beta}_{ci}}$ for $\lambda = 0.001$.

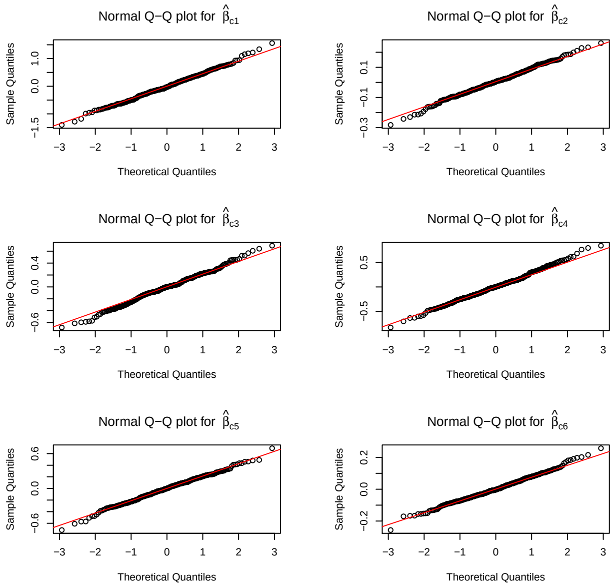


Figure 5.12: The Normal quantile-quantile plot of the constrained estimated parameter $\hat{\boldsymbol{\beta}}_{ci}$ for $\lambda = 0.001$.

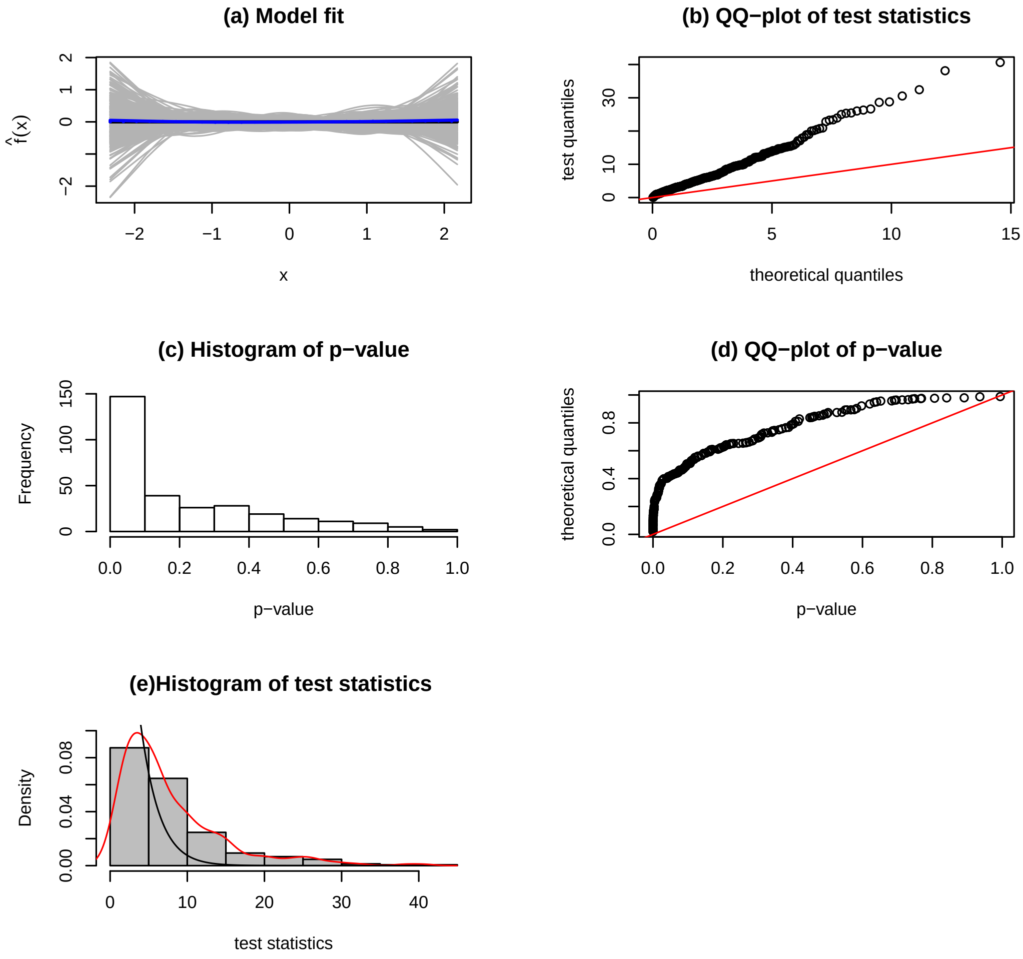


Figure 5.13: (a) The model fit for $\lambda = 0.01$. The solid black line is the mean of the estimated smoothing term for 300 simulations. (b) Chi-Square Q-Q plot of test statistics regarding $\hat{\boldsymbol{\beta}}_z$ with 2.723 degrees of freedom. (c) Histogram of the p-values. (d) Uniform Q-Q plot of the p-values. (e) Histogram of the test statistics, the black solid line is the theoretical density, and the red solid line is the kernel density estimate of the test statistic. The rejection rate is 112 out of 300 simulations.

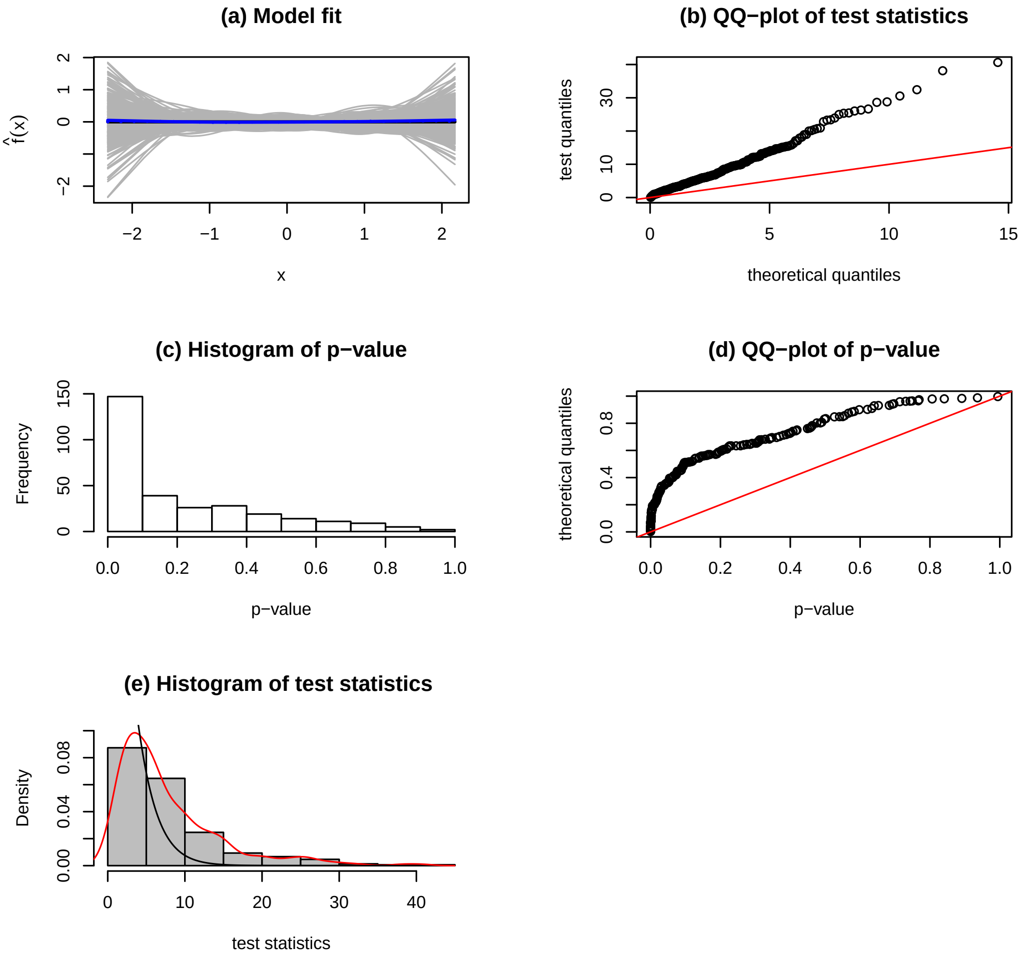


Figure 5.14: (a) The model fit for $\lambda = 0.01$. The solid black line is the mean of the estimated smoothing term for 300 simulations. (b) Chi-Square Q-Q plot of test statistics regarding $\hat{\boldsymbol{\beta}}_c$ with 2.723 degrees of freedom. (c) Histogram of the p-values. (d) Uniform Q-Q plot of the p-values. (e) Histogram of the test statistics. The black solid line is the theoretical density, and the red solid line is the kernel density estimate of the test statistic,the rejection rate is 112 out of 300 simulations.

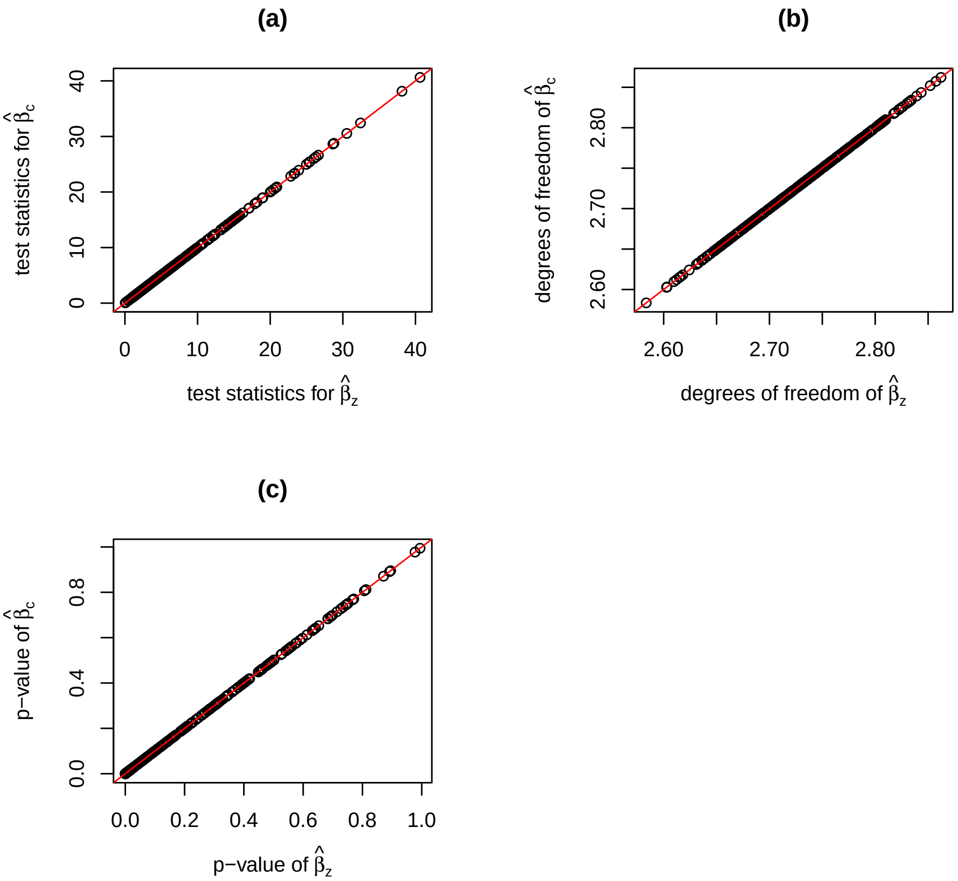


Figure 5.15: (a) Plot of the values of the Wald-type statistic for $\hat{\boldsymbol{\beta}}_z$ versus the values of the Wald-type statistic for $\hat{\boldsymbol{\beta}}_c$ for $\lambda = 0.01$. (b) Plot of the degrees of freedom for $\hat{\boldsymbol{\beta}}_z$ versus the degrees of freedom for $\hat{\boldsymbol{\beta}}_c$, the range of the effective degrees of freedom is between 2.583 and 2.862 for $\lambda = 0.01$. (c) the p-value of the test statistics for $\boldsymbol{\beta}_z$ versus the p-value of the test statistics for $\hat{\boldsymbol{\beta}}_c$.

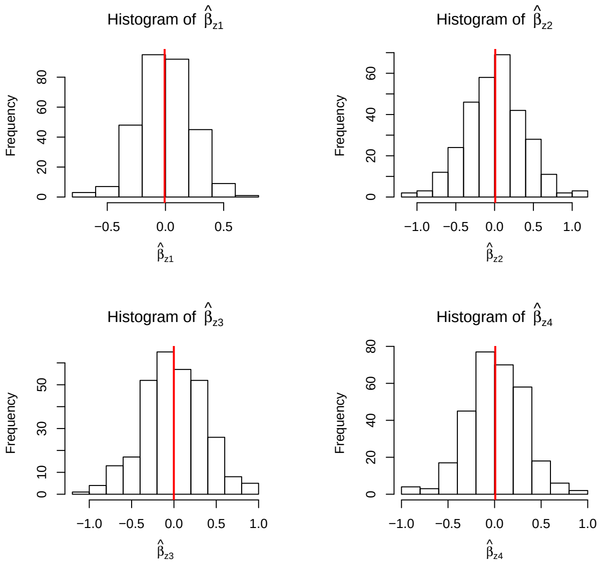


Figure 5.16: Histogram of the unconstrained estimated parameters $\hat{\boldsymbol{\beta}}_{zi}$ for $\lambda = 0.01$.

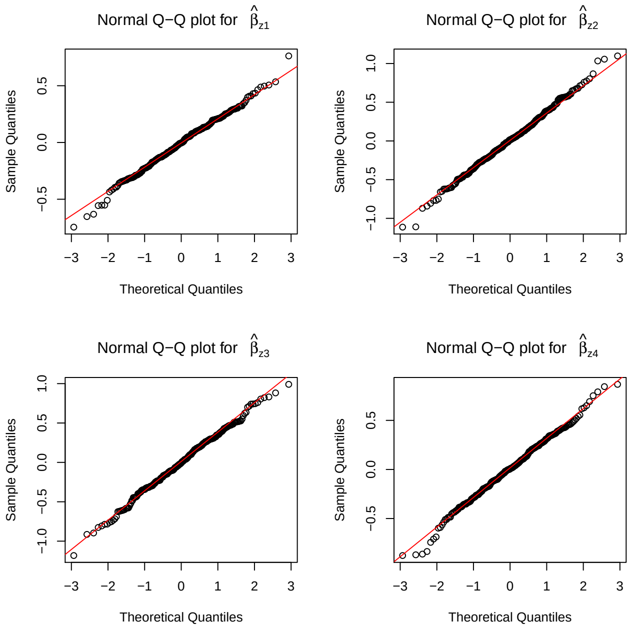


Figure 5.17: The Normal quantile-quantile plot of the unconstrained estimated parameters $\hat{\boldsymbol{\beta}}_{zi}$ for $\lambda = 0.01$.

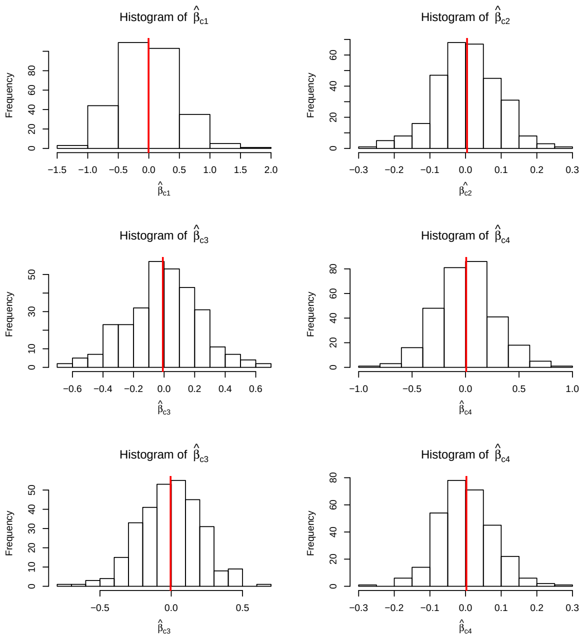


Figure 5.18: Histogram of the constrained estimated parameters $\hat{\boldsymbol{\beta}}_{ci}$ for $\lambda = 0.01$.

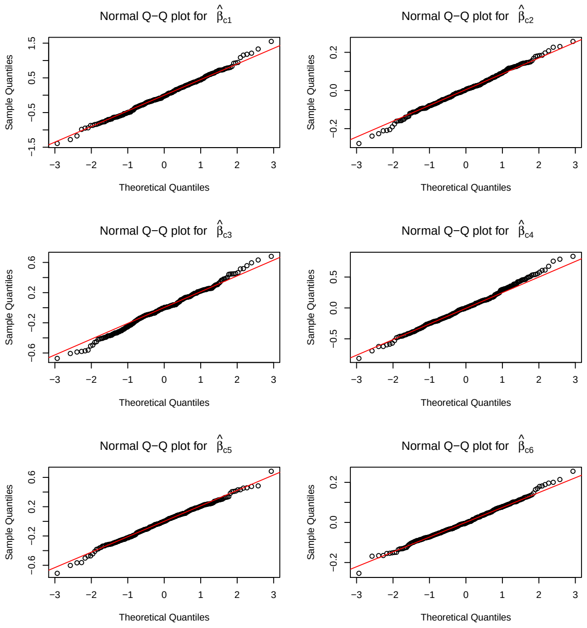


Figure 5.19: The Normal quantile-quantile plots of the constrained estimated parameters $\hat{\boldsymbol{\beta}}_{ci}$ for $\lambda = 0.01$.

Secondly, the results of applying the penalized likelihood ratio statistic to test the hypothesis $H_0 : \boldsymbol{\beta}_z = \mathbf{0}$, for the value of the smoothing parameters $\lambda = 0.1$ are shown in Figures 5.20, while Figures 5.21 for $\lambda = 0.2$.

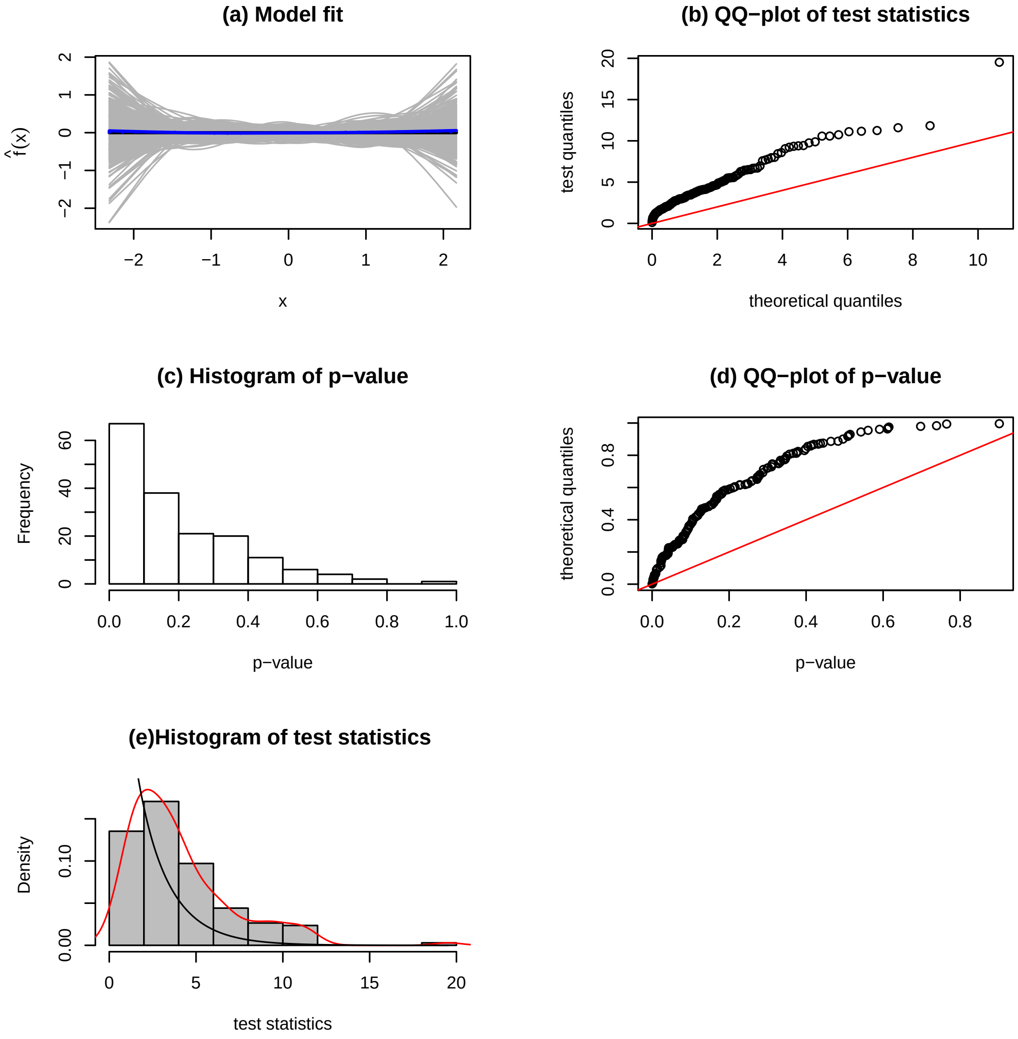


Figure 5.20: (a) The model fit for $\lambda = 0.1$. The solid black line is the mean of the estimated smoothing term for 300 simulations. (b) Chi-Square Q-Q plot of test statistics regarding $\hat{\boldsymbol{\beta}}_z$ with 1.611 degrees of freedom. (c) Histogram of the p-values. (d) Uniform Q-Q plot of the p-values. (e) Histogram of the test statistics, the black solid line is the theoretical density, and the red solid line is the kernel density estimate of the test statistic. The rejection rate is 110 out of 300 simulations.

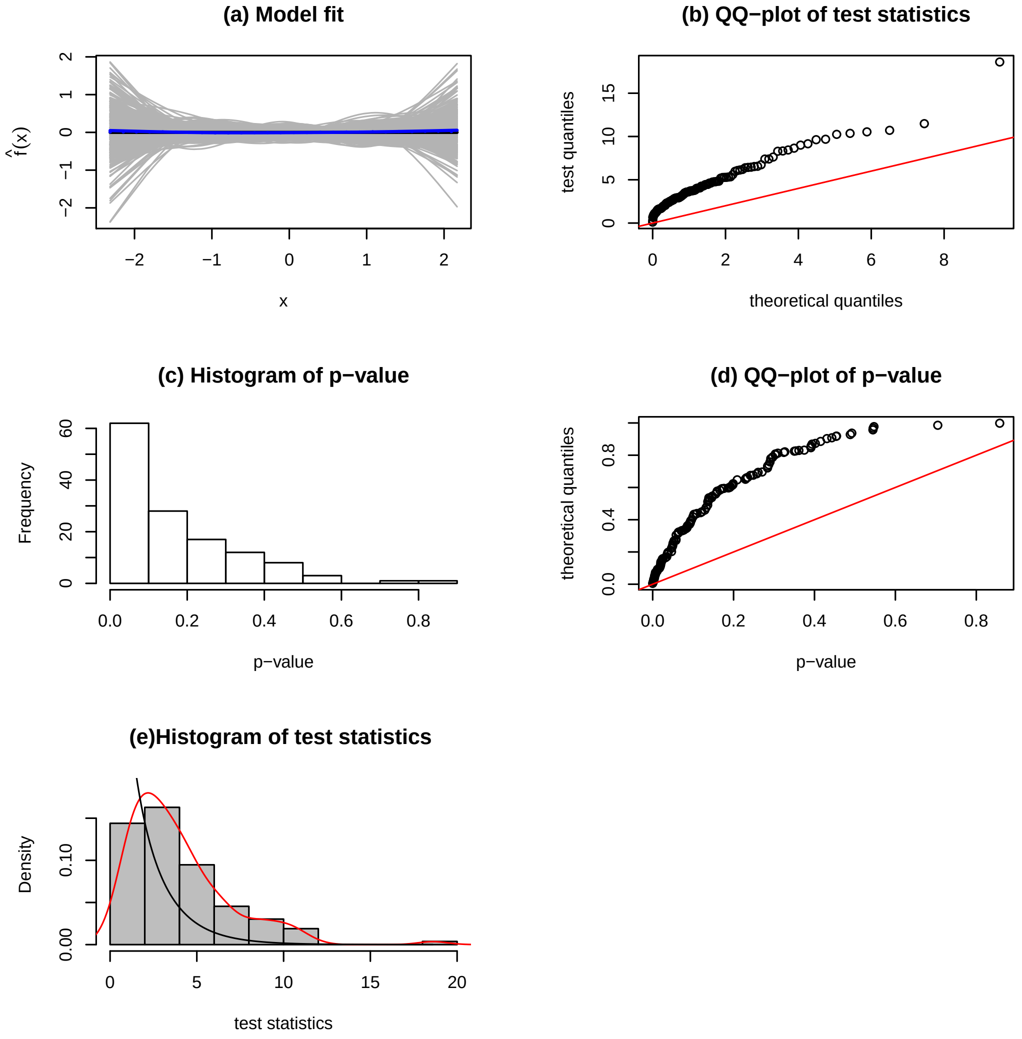


Figure 5.21: (a) The model fit for $\lambda = 0.2$. The solid black line is the mean of the estimated smoothing term for 300 simulations. (b) Chi-Square Q-Q plot of test statistics regarding $\hat{\boldsymbol{\beta}}_z$ with 1.378 degrees of freedom. (c) Histogram of the p-values. (d) Uniform Q-Q plot of the p-values. (e) Histogram of the test statistics, the black solid line is the theoretical density, and the red solid line is the kernel density estimate of the test statistic. The rejection rate is 110 out of 300 simulations.

For both values of the smoothing parameters, the estimation of the unconstrained spline effect parameter $\boldsymbol{\beta}_z$ are asymptotically normal with mean zero. This results are summarized in Figures 5.22, and 5.23 for $\lambda = 0.1$, and Figures 5.24, 5.25 for $\lambda = 0.2$.

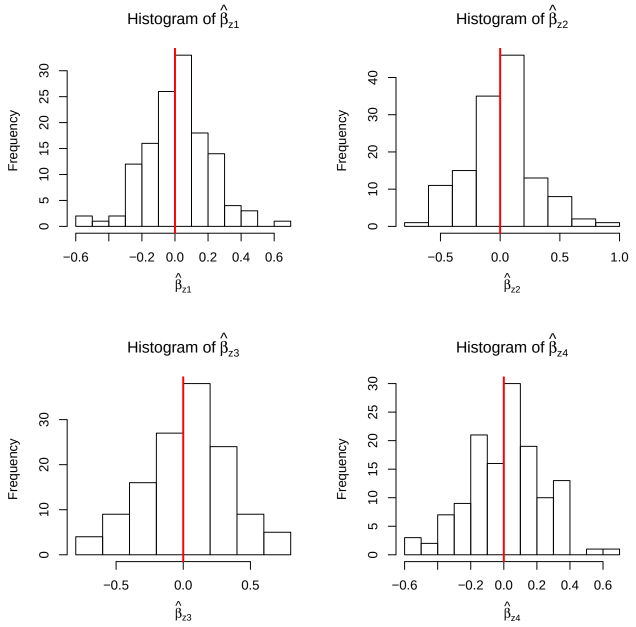


Figure 5.22: Histogram of the unconstrained estimated parameters $\hat{\beta_{zi}}$ for $\lambda = 0.1$.

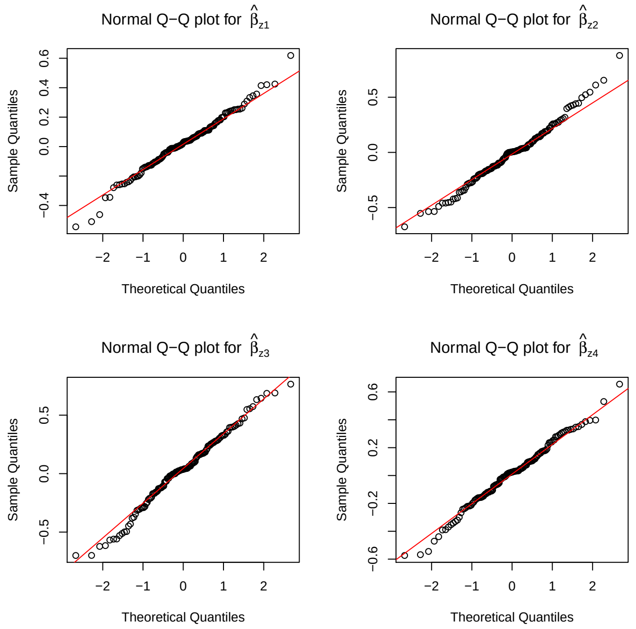


Figure 5.23: The Normal quantile-quantile plot of the unconstrained estimated parameters $\hat{\beta_{zi}}$ for $\lambda = 0.1$.

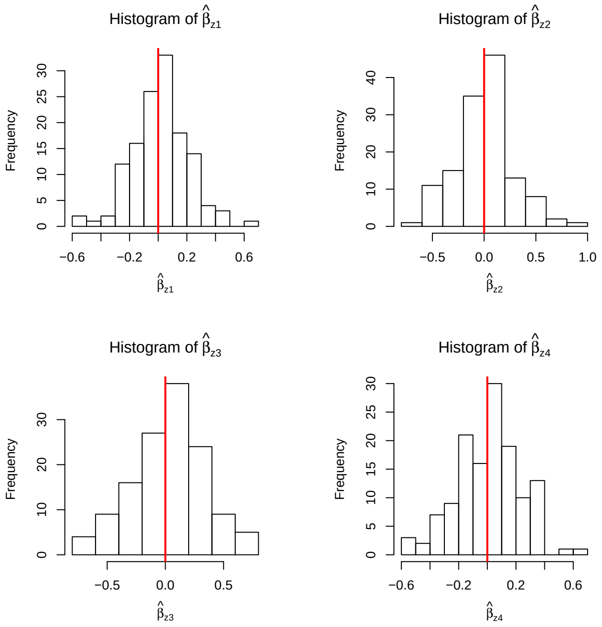


Figure 5.24: Histogram of the unconstrained estimated parameters $\hat{\beta}_{zi}$ for $\lambda = 0.2$.

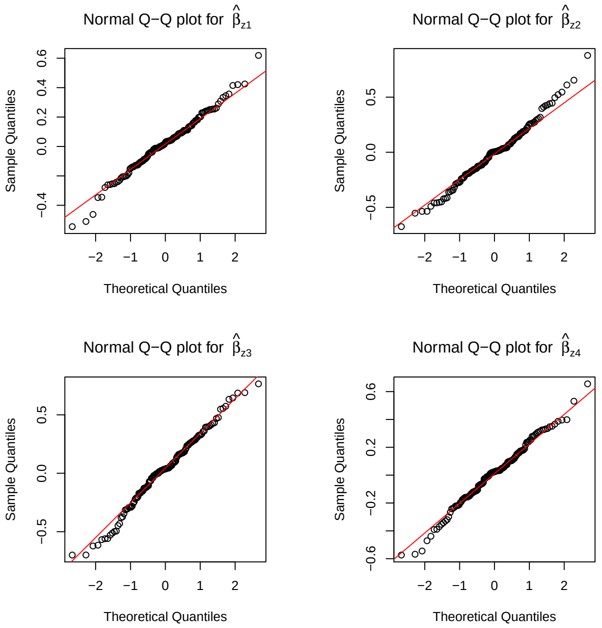


Figure 5.25: The Normal quantile-quantile plot of the unconstrained estimated parameters $\hat{\beta}_{zi}$ for $\lambda = 0.2$.

The main problem with the penalized version of the Wald statistics and likelihood ratio statistics is that the penalty term can induce bias in the estimated parameters $\hat{\boldsymbol{\beta}}_z$ and $\hat{\boldsymbol{\beta}}_c$, for the case $\mathcal{K}\boldsymbol{\beta}_{zT} \neq \mathbf{0}$ when the null hypothesis is true, where $\boldsymbol{\beta}_{zT}$ is the true value. In addition, if we test the hypothesis of linearity, under the null hypothesis $\mathcal{K}\boldsymbol{\beta}_z = \mathbf{0}$, so this clarify that this term itself does not contribute to the bias. Then careful attention must be given to choose the value of the smoothing parameter that gives the reasonable results of the test. However, the estimate of $\boldsymbol{\beta}_z$ depends on the value of the smoothing parameters. If the result shows some evidence of the bias in the estimation, Gray (1992) suggested to use smaller values of the smoothing parameters to reduce the possible bias in the estimation.

Thirdly, the penalized score test of the null hypothesis $H_0 : \boldsymbol{\beta}_z = \mathbf{0}$ was performed for the smoothing parameters values $\lambda = 0, 0.01, 0.1$ and 0.5. In the penalized score test, we do not have to estimate the parameters $\boldsymbol{\beta}_z$ to perform the test, so the penalized score tests can be computed using only the first iteration in the Newton-Raphson algorithm. The results are summarized in Figures 5.26, 5.27, 5.28, and 5.29 for the values of the smoothing parameters $\lambda = 0, 0.01, 0.1$ and 0.5 respectively, which confirm that the test statistics have chi-square distributions with degrees of freedom depending on the value of λ. Therefore, the result shows no bias. This confirm the bias estimate that we obtained in the Wald-type statistics due to estimation of $\boldsymbol{\beta}_z$ depending on the change value of λ. Because of this we use the penalized score test statistics for the variable selection later on in Chapter 6. We point out the histogram of p-values in Figures 5.27(b), 5.28(b), and 5.29(b) are deficiencies as impact of changing λ.

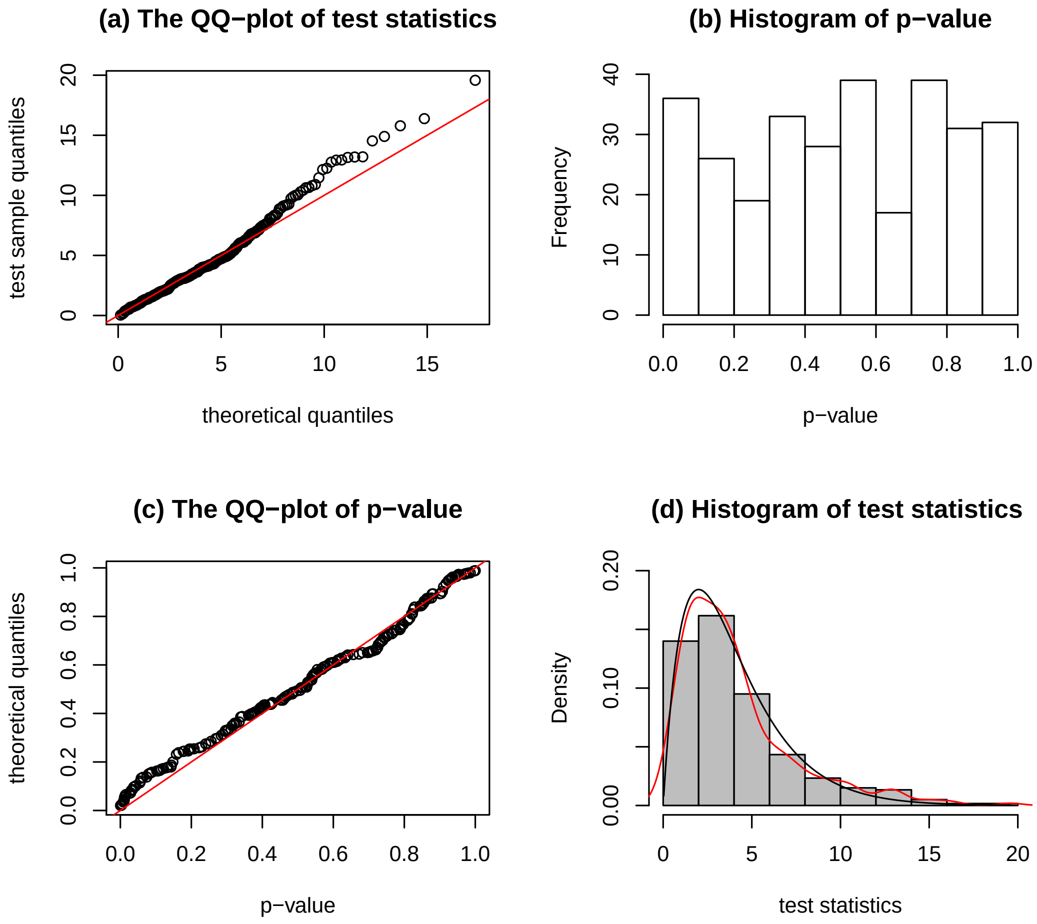


Figure 5.26: (a) Chi-Square Q-Q plot of tests statistics for $\lambda = 0$. (b) Histogram of p-values. (c) Uniform Q-Q plot of the p-values. (d) Histogram of the test statistic, the black solid line is the theoretical density, and the red solid line is kernel density estimate of the test statistic. The degrees of freedom is 4, the rejection rate is 24 out of 300 simulations.

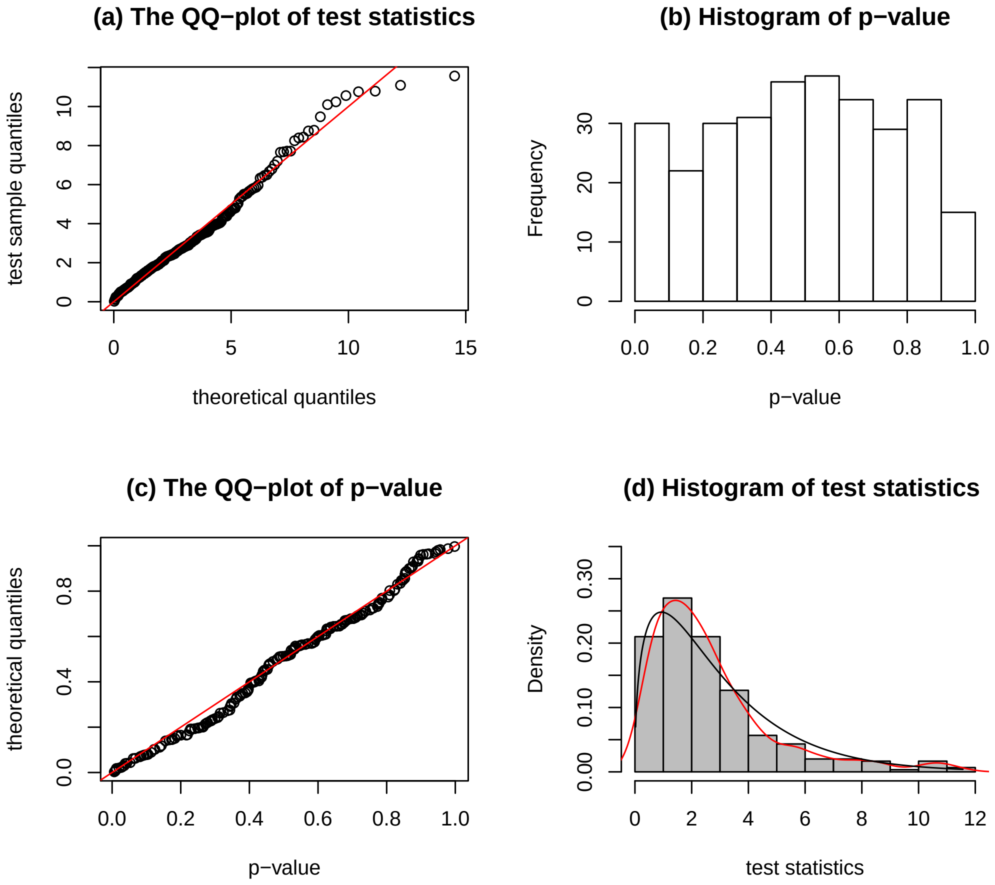


Figure 5.27: (a) Chi-Square Q-Q plot of test statistics for $\lambda = 0.01$. (b) Histogram of p-values. (c) Uniform Q-Q plot of the p-values. (d) Histogram of the test statistic, the black solid line is the theoretical density, and the red solid line is kernel density estimate of the test statistic. The range of the effective degrees of freedom is between 2.4797 and 2.9329. The rejection rate is 17 out of 300 simulations.

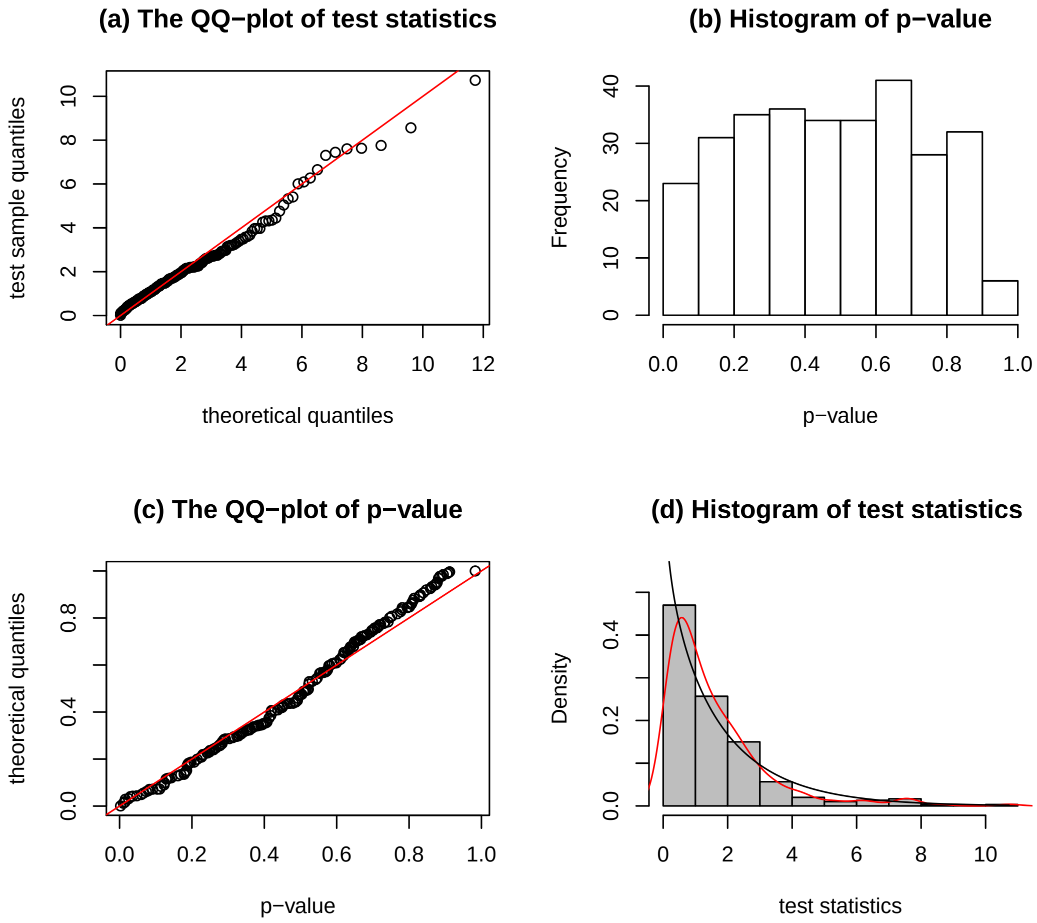


Figure 5.28: (a) Chi-Square Q-Q plot of test statistics for $\lambda = 0.1$. (b) Histogram of p-values. (c) Uniform Q-Q plot of the p-values. (d) Histogram of the test statistic, the black solid line is the theoretical density, and the red solid line is kernel density estimate of the test statistic. The range of the effective degrees of freedom is between 1.4634 and 1.7298. The rejection rate is 13 out of 300 simulations.

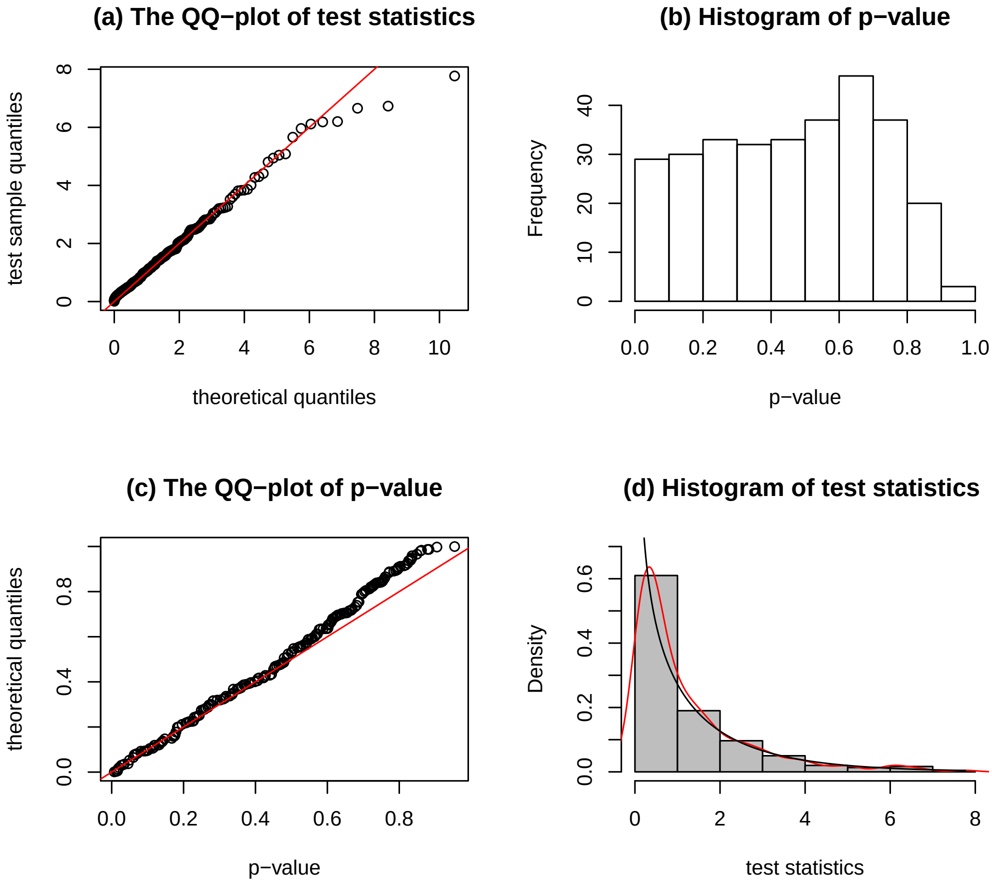


Figure 5.29: (a) Chi-Square Q-Q plot of test statistics for $\lambda = 0.5$. (b) Histogram of p-values. (c) Uniform Q-Q plot of the p-values. (d) Histogram of the test statistic, the black solid line is the theoretical density, and the red solid line is kernel density estimate of the test statistic. The range of the effective degrees of freedom is between 1.1335 and 1.2322. The rejection rate is 15 out of 300 simulations.

5.7 Estimating the Smoothing Parameters and the Number of Knots

In the penalized additive Cox PH model it is very important to determined the appropriate value of the smoothing parameters, which balance the trade-off between the goodness of fit and smoothness of the model parameters. To select the appropriate values of the smoothing parameters we need to identify a range of possible values of the smoothing parameters, and then select a criterion for assessing the model corresponding to each value of the smoothing parameter. There are various methods for choosing the amount of smoothing in a penalized additive Cox model, including Cross-validation (CV), generalized cross-validation (GCV), Akaike information criterion (AIC), or Bayesian information criterion (BIC). Wood et al. (2016)

O'Sullivan (1988) presented an approach for selecting the optimal smoothing parameter for smoothing spline estimators by minimizing the GCV under the assumption that the baseline cumulative function, $H_0(t) = \int_0^t h_0(u)du$, is known, and replacing $H_0(t)$ by the non-parametric estimator, the Breslow estimator, in the case of unknown baseline cumulative functions.

The degrees of freedom are determined by the smoothing parameter λ, so instead of specifying the value of the smoothing parameter, we can specify the number of the degrees of freedom and then solve for the value of the smoothing parameter that gives the specified degrees of freedom. This method has been used in Gray (1992, 1994), Cadarso-Surez et al. (2010), Meira-Machado et al. (2013), Therneau et al. (2003), and Wang et al. (2017a). They used the AIC criteria which is defined as $\mathrm{AIC} = -2\ell_{pl}(\hat{\boldsymbol{\beta}}_\lambda) + 2df$, where df is the effective degrees of freedom as presented earlier in section 5.6.1. However, AIC can be under-penalize, which leads to models with an excessively large number of degrees of freedom. Alternatively, Hurvich et al. (1998) proposed the corrected AIC, which adjusts for this over fitting by replacing the degrees of freedom by $n(df + 1)/(n - (df + 2))$, where n is the total number of events in the

Cox model. The corrected Akaike Information Criterion (AICc) is given by

$$\text{AICc} = -2\ell_{pl}(\hat{\boldsymbol{\beta}}_\lambda) + 2\frac{n(df+1)}{n-(df+2)}. \tag{5.23}$$

Minimization of the AIC and AICc is simple in the univariate setting and increasingly complicated as the number of variable increases.

Bayesian information criterion (BIC) is calculated as

$$\text{BIC} = -2\ell_{pl}(\hat{\boldsymbol{\beta}}_\lambda) + \log(n) \times df,$$

where n is the number of observations. Considering there are censored observations in survival data, Volinsky and Raftery (2000) corrected BIC by using the number of uncensored observations in place of the number of observation n. BIC for high-dimensional data tends to select overly sparse models, which means that BIC will select a model with few variables. Huang and Harrington (2002) proposed a resampling method such as bootstrap selection to choose the smoothing parameter. Wood et al. (2016) proposed the smoothing parameter estimation methods based on maximized Laplace approximate marginal likelihood. This methods can be used for generalized additive models for nonexponential family, Cox proportional hazards models and multivariate additive models.

The approach of leave-one-out cross validated partial log-likelihood (CVPL) for ridge estimators in the standard Cox PH model for high-dimensional data was invented by Van Houwelingen et al. (2006), which is based on the unpenalized partial log-likelihood. This approach is used in Tsujitani et al. (2012), Simon et al. (2011), and Bøvelstad et al. (2007). Generally speaking, in a k-fold cross validation, the data is randomly divided into k folds, so the size of these folds are as similar as possible. In turn, each fold is left out as the test set, while the remaining $k-1$ folds are used as the training set, to estimate $\boldsymbol{\beta}_\lambda$ using equation (5.10). The cross validation score for the fold is the negative partial log likelihood using equation (5.7). The overall cross validation score is then the sum of the scores across the k folds. The optimal value of

λ is the one that maximizes CVPL. Maximizing CVPL is obtained using grid values to represent λ. In mathematical form, the k-fold cross validation partial log-likelihood can be expressed as,

$$\text{CVPL}(\lambda) = \sum_{i=1}^{k} \left(\ell_{pl}^{i}(\hat{\boldsymbol{\beta}}_{\lambda}^{(-i)}) \right), \tag{5.24}$$

where $\hat{\boldsymbol{\beta}}_{\lambda}^{(-i)}$ is the penalized estimate for $\boldsymbol{\beta}$ for a given value of λ, with the i^{th} folds taken out as the test set and the remaining $k-1$ fold kept as the training set. ℓ_{pl}^{i} is the partial log likelihood for the i^{th} fold. In this thesis we used the 5-fold cross-validation partial log-likelihood to select the optimal smoothing parameters.

The shape of the log hazard ratio depends heavily on the number of knots, Nan et al. (2005) estimate the optimal number of knots, using a modification of the O'Sullivan (1988) GCV method, so that the optimal number of knots has the smallest GCV value. In this chapter, we obtain the optimal number of knots by computing the CVPL value for a range of the number of knots, and we chose the number of knots that maximizing the cross-validation partial log-likelihood.

All the three model selection methods we presented in this section aim to choose an optimal smoothing parameter that minimizes or maximizes the corresponding criterion. Both AIC and BIC are minimized with respect to df. AIC tends to choose relatively more complicated models than BIC. Therefore, the more complicated model is generally associated with smaller smoothing parameter values. Cross validation generally performs better because it is a data driven method where all data are used to find the optimal value of the smoothing parameters, which is used in this thesis to find the optimal values of the smoothing parameters. In large data sets, estimating the smoothing parameter requires considerable computational time, especially when CVPL methods are used for choosing the optimal smoothing parameters separately in the penalized additive Cox model.

5.8 Model Diagnostics

5.8.1 Estimation of $h_0(t)$ and $S(t|x)$

The most important part of the additive Cox model is predicting the survivor function $\hat{S}(t|x)$ for the model that contains the clinical characteristics and CNA profile (as smooth term). The estimation of $h_0(t)$ and $S(t|x)$ can be done in a similar manner as for the standard Cox PH model. Once we obtain the parameter estimates $\hat{\boldsymbol{\beta}}$, the estimate of the baseline hazard is given by

$$\hat{h}_0(t_i) = \frac{\delta_i}{\sum_{j \in R(t_i)} \exp(X_j \hat{\boldsymbol{\beta}})}. \tag{5.25}$$

The cumulative hazard function $H_0(t)$, and the baseline survivor function $S_0(t)$ can be estimated by

$$\hat{H}_0(t_i) = \sum_{i: t_i \le t_j} \frac{\delta_i}{\sum_{j \in R(t_i)} \exp(X_j \hat{\boldsymbol{\beta}})}, \tag{5.26}$$

$$\hat{S}_0(t) = \exp\{-\hat{H}_0(t)\}. \tag{5.27}$$

5.8.2 Cox-Snell Residuals

Plotting the Cox-Snell residuals provides a way of checking whether the additive Cox PH model is suitable for the data. Cox-Snell residuals, as discussed in Chapter 3, can be calculated for the additive Cox model as

$$r_i = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i). \tag{5.28}$$

where $\hat{H}_i(t_i)$ is the estimated cumulative hazard for individual at their failure (or censoring) time, and $\hat{S}_i(t_i|\boldsymbol{x}_i)$ is the estimated survivor function for the i^{th} individual at t_i.

The Martingales residuals are defined for the i-th individual as:

$$r_{Mi} = \delta_i - r_{Ci} = \delta_i - \hat{H}_i(t_i). \tag{5.29}$$

5.9 Real Data Analysis

For the time being the CNA covariates are ignored, and attention is focused on the clinical characteristics as an example of the proposed method. In order to reduce the value of the smoothing matrix, we divided age by 100 and we used 5 equally spaced knots to construct the smoothing matrix. Table 5.1 illustrates all possible 32 unpenalized additive Cox models. Each model contains one smooth term and one or more categorical variables. The values of -2 partial log likelihood, degrees of freedom, and AIC for each unpenalized additive Cox model are shown in Table 5.1. The best model is the model that contains Age as a smoothing function, Stage N, and Stage T.

#	parameters in the model	$-2\log\hat{L}_{pl}$	df	AIC
1	Intercept			478.567
2	f(Age/100)	471.586	4	479.586
3	Gender	477.982	1	479.982
4	Stage T	474.816	2	478.509
5	Stage N	475.810	2	479.810
6	Grade	477.087	2	481.087
7	f(Age/100)+Gender	471.045	5	481.045
8	f(Age/100)+Stage T	463.020	6	475.020
9	f(Age/100)+Stage N	465.387	6	477.387
10	f(Age/100)+Grade	470.092	6	482.092
11	Gender+Stage T	472.312	3	478.312
12	Gender+Stage N	475.530	3	481.530
13	Gender+Grade	476.384	3	482.384
14	Stage T+ Stage N	471.840	4	479.840
15	Stage T+Grade	473.098	4	481.098
16	Stage N+Grade	474.321	4	482.321
17	f(Age/100)+Gender+Stage T	461.147	7	475.1471
18	f(Age/100)+Gender+Stage N	465.302	7	479.302
19	f(Age/100)+Gender+Grade	469.456	7	483.456
20	f(Age/100)+Stage T+Stage N	455.017	8	471.017
21	f(Age/100)+Stage T+Grade	461.793	8	477.793
22	f(Age/100)+Stage N+Grade	463.611	8	479.611
23	Gender+Stage T+Stage N	470.289	5	480.289
24	Gender+Stage T+Grade	470.982	5	480.982
25	Stage T+Stage N+Grade	470.360	6	482.360
26	Gender+Stage N+Grade	473.942	5	483.942
27	Gender+Stage T+Stage N+Grade	468.871	6	482.871
28	f(Age/100)+Gender+Stage T+Stage N	454.069	9	472.069
29	f(Age/100)+Gender+Stage N+Grade	463.454	9	481.454
30	f(Age/100)+Stage T+Stage N+Grade	453.961	10	473.961
31	f(Age/100)+Gender+StageN+Grade	463.454	9	481.454
32	f(Age/100)+Gender+StageT+StageN+Grade	452.976	11	474.976

Table 5.1: The values of -2 log partial likelihood and the number of parameters, AIC for each fitted additive Cox model.

Cross validated partial log likelihood for the best model was computed for the number of knots 5, 6, 7 and 8 to obtain the optimal value of the smoothing parameter λ, and then to select the optimal number of knots that has the maximum cross-validation partial log-likelihood value. As shown in Figure 5.30, the penalized additive Cox model with 5 knots has the maximum cross-validation partial log-likelihood $\text{CVPL} = -125.015$, and the optimal smoothing parameter is equal to $\lambda_{opt} = 0.11$.

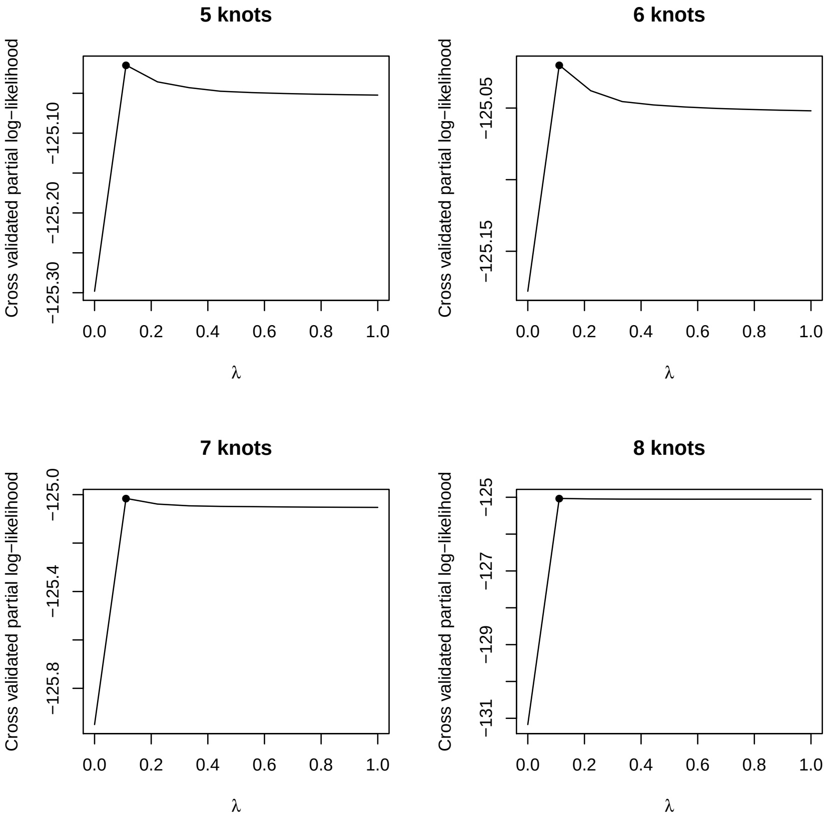


Figure 5.30: Plots of the Cross-validated partial log-likelihood versus λ for the penalized additive Cox model with 5, 6, 7, and 8 knots.

The estimates of the fixed effect and spline effect and their inferences, for the best model under conditions with or without penalization can be seen in Tables 5.2

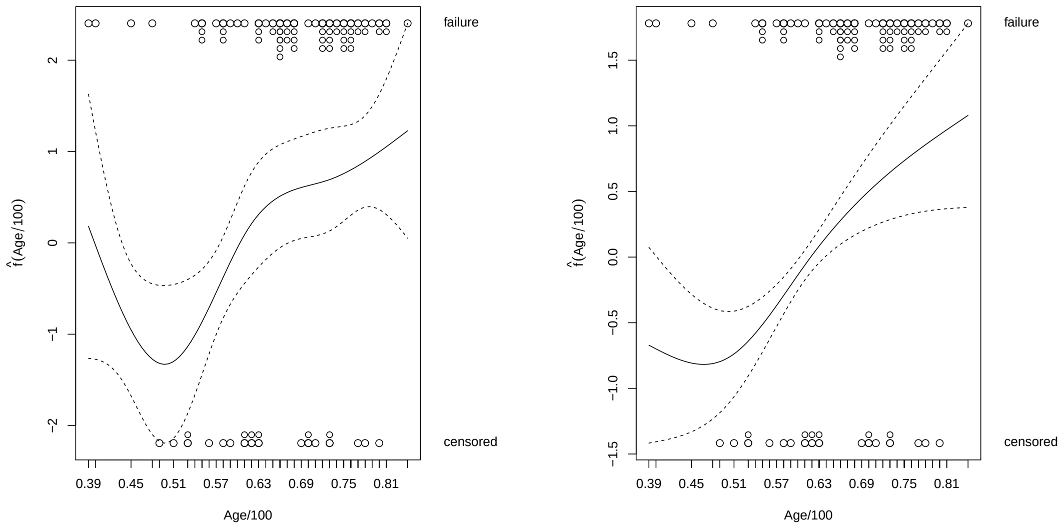
Predictor	Estimate	Exp	Standard error	$z$value	p-value
without smoothing term					
Age/100	5.311	202.585	1.558	3.409	0.001
Stage T2	0.150	1.162	0.302	0.499	0.618
Stage T3	1.800	6.052	0.576	3.123	0.002
Stage N1	0.345	1.411	0.284	1.212	0.225
Stage N2	1.336	3.804	0.478	2.797	0.005
with smoothing term					
$\lambda = 0.000$					
$f(Age/100)$					0.003
Stage T2	0.180	1.197	0.316	0.568	0.570
Stage T3	2.318	10.156	0.652	3.556	0.000
Stage N1	0.365	1.440	0.298	1.224	0.221
Stage N2	1.581	4.858	0.508	3.113	0.002
with smoothing term					
$\lambda_{opt} = 0.111$					
$f(Age/100)$					0.000
Stage T2	0.153	1.165	0.302	0.505	0.613
Stage T3	1.901	6.690	0.576	3.297	0.001
Stage N1	0.359	1.432	0.285	1.259	0.208
Stage N2	1.407	4.085	0.478	2.945	0.003

Table 5.2: Estimated values of the parameters on fitting additive Cox PH model.

For comparing for the estimate of the fixed effect both with and without a smoothing term of age, the result indicates that age, Stage-T and Stage-N are statistically significant with p-values< 0.05. To understand the effects of individual predictors for age, we look at the hazard rate for age $e^{0.01 \times 5.311} = 1.05$, which shows that the hazard ratio of death increases by about 5% as age increases by one year. The estimated coefficient of Stage-N2 is positive, which indicates that the wider spread of the cancer cell to the nearby lymph nodes increases the hazard relative to the baseline hazard Stage-N0. Similarly, the estimates of Stage-T3 indicate large tumor size increases the hazard relative to the baseline hazard Stage-T1.

Figure 5.31 shows the estimated log hazard ratio $\hat{f}(Age/100)$ versus Age/100 when $\lambda = 0$ (left panel) and $\lambda_{opt} = 0.111$ (right panel), and the small lines along the horizontal axis are the "rug", showing the values of the covariate of Age for each patient. The deep shape in $\hat{f}(Age/100)$ is between 48 to 55 years because the patient at 45

years had a short survival time (only 54 days) where the minimum survival time is 34. For both panels, the older patients had an increased hazard of death.



(a) $\lambda = 0$

(b) $\lambda_{opt} = 0.111$

Figure 5.31: (Left panel): the estimated smooth function $\hat{f}(Age/100)$ versus Age when $\lambda = 0$. (Right panel): the estimated smooth function $\hat{f}(Age/100)$ versus Age when $\lambda_{opt} = 0.111$. The points indicate the number of observations on each individual that were either censored or a failure. The dashed lines are the 95% point-wise confidence band.

As part of the model diagnostic, the plot of the cumulative hazard of the Cox-Snell residual from the best fitting model with $\lambda_{opt} = 0.111$ is shown in Figure 5.32. The figure shows the cumulative hazard line is very close to the identity line, which indicates that the additive Cox PH model is a reasonably good fit. The cumulative hazard steps near the top right corner are jagged, as a result of rare deaths near to the upper end of the distribution of the survival times. The martingale residuals were examined. Figure 5.33 display the log hazard ratio for age obtained by transforming LOWESS smoothing martingale residual for the penalized best model, which suggests that the age is modelled correctly.

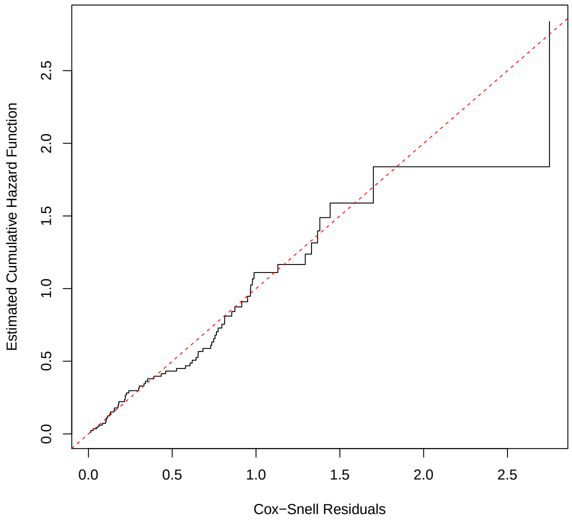


Figure 5.32: Cumulative hazard of Cox-Snell residual from the penalized additive Cox model fit. The red line is the identity line.

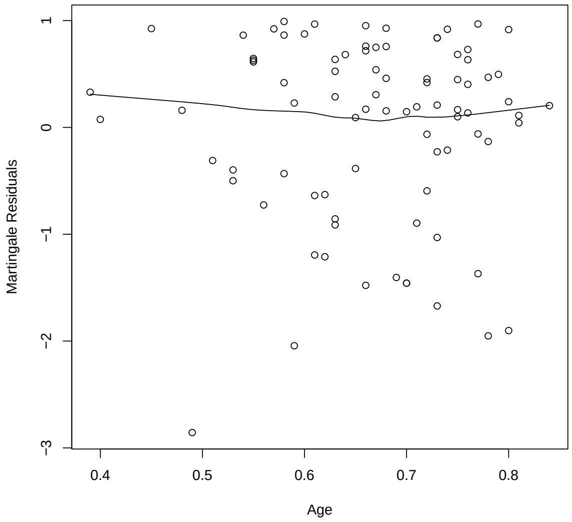


Figure 5.33: Martingale residuals versus Age, the solid black line is the smoothed curve using LOWESS method.

5.10 Conclusion

In this chapter, we proposed an extension of the standard Cox PH model by including smoothing terms. We demonstrated the use of the radial basis function as a smoothing term in the additive Cox PH model with a fixed number of knots. The radial basis function satisfies the natural cubic spline condition (linear tail constraints), so these constraints reduce the dimension of the spline effect parameter by 1 degree of freedom, this means if we have n_k equally spaced knots, the estimate of the unconstrained spline effect parameter is $n_k - 1$ degrees of freedom in the unpenalized problem, and less than that for the penalized problem. The estimate of the model parameters can be performed with or without penalization in the model, so the estimated parameters are obtained in a similar manner as for the standard Cox PH model, by maximizing the penalized partial log-likelihood using the Newton-Raphson algorithm. The asymptotic theory for the penalized additive Cox model with radial basis estimates is relatively straightforward.

Identification of the nonlinear effect of the continuous covariate enables us to estimate more accurately a patient's prognosis, whether the patient have lower or higher hazard, and thus to better determine lung cancer survival time. To better understand the nonlinear effect of the continuous covariate, we look at the effects of the spline parameters of a given continuous variable on survival time. The results of this effect can be expressed in terms of log hazard ratio curves, taking a specific covariate value as reference. Confidence bands for these log hazard ratio curves can also be calculated. The simulation examples show that the proposed method performed very well.

The model enables us to evaluate the significance of the fixed effect parameters, as well as a test for nonlinearity of the spline effect parameters. This chapter examined a penalized method for testing two hypothesis of spline effect. The first is the test for nonlinearity, which means testing the hypothesis of no covariate effects in the penalized additive Cox model. This test is based on the spline parameters associated with spline term all being equal to zero. The second test is the test for linearity, which means a test for zero slope in the spline parameter in the radial basis function. We demonstrate the penalized versions of the Wald test, likelihood ratio statistic, and score test. The Wald-type statistics can be used to test for no covariate effect for both constrained and unconstrained spline parameters. The degrees of freedom are the trace of the appropriate matrix, which can be computed for both penalized and unpenalized problem, and approximations to the distribution of these test statistic are used. The simulation examples are presented under the null hypothesis of no effect of the spline parameters to examine the performance of the hypothesis testing procedure.

In order to find the optimal value of the smoothing parameter, we used a grid search to select the optimal smoothing parameter value based on maximizing the cross-validated partial log-likelihood criterion. This is computationally demanding in the case of finding separate smoothing parameter value for each smoothing term. The optimal number of knots can be done by computing CVPL for a range of the number of knots, and we then select the number of knots that maximizes CVPL.

The use of the spline method with NSCLC survival data allows nonlinear covariate effects to be detected and tested easily. The clinical variables are considered as fixed effects in addition to age as spline effect. This approach proved to be useful in a real data example on lung cancer survival times. This work opens the possibility of future work for high-dimensional data. Some reports of the work on high-dimensional data analysis exist in the literature. However, most of them are focused on the standard Cox PH model. There is limited reporting on high-dimensional survival analysis in the literature for the penalized additive Cox model. Variable selection in high-dimensional penalized additive Cox model is introduced in the next chapter.

Chapter 6

Variable Selection for Penalized Additive Cox PH Model

6.1 Introduction

Extracting information from large number of covariates measured on patients for the purpose of prediction in survival model is an important aim in medical studies. Most of the standard statistical methods in survival analysis describe the relationship between the covariates and outcome assuming that the number of covariates p is less than the number of observations n. In the standard setting, $n > p$, the parameters in standard Cox PH model can be estimated by maximizing the partial log-likelihood. However, combining both the clinical characteristics as fixed effect predictors and the CNA genomic-window profiles as smoothing terms into a single prediction model will make the model unestimable, because estimating the model parameters by maximizing the partial log-likelihood is no longer possible.

The main challenge lies not only including high-dimensional CNA genomic-windows in the survival model, but it is also including the high-dimensional CNA genomic-windows as smoothing terms in the survival models. Each smoothing term can be expressed as matrix where the number of rows is the number of observations, and number of columns is the number of spline parameters, that associated with the number of knots. The size of the smoothing matrix for all the CNAs data is $n \times q(n_k - 1)$, where $n = 85$ is the number of patients, $q = 13253$ the number of CNAs genomic-windows, and n_k is the number of knots in each smoothing term. For 5 knots for each smoothing term, the size of the smoothing matrix for all CNA data is 85×53012, which is computational demanding. Tackling the high-dimensional smoothing problem by some form of variable selection is very important in our case.

Variable selection is a huge area of statistics, there are several way to do that. Firstly using filtering method, which based on select subsets of CNA variables as a pre-processing step. Secondly by using Wrapper methods and thirdly by using embedded methods. In this thesis we will use the filtering method because it is convenient and given a computational burden that we have and it is manageable. Wrapper methods consider the selection of a set of features as a search problem, a strategy is needed to explore the feature space where different combinations are prepared, evaluated and compared to other combinations. The problem with this approach is that feature space is vast and looking at every possible combination would be a computationally expensive. Two main approaches are mostly used in the wrapper methods which are backward elimination and forward selection, both of these apprpaches is difficult in the additive Cox PH model due to the huge smoothing matrix for all the CNA genomic-windows. Embedded methods learn which CNA genomic-windows features best contribute to the accuracy of the model. The most common type of embedded feature selection methods are regularization methods, which called penalization methods that introduce additional constraints into the optimization of a predictive algorithm (such as a regression algorithm) to shrink the model coefficients toward zero resulting fewer coefficients are not equal to zero. Embedded methods is also difficult to apply in the additive Cox PH model, but it can be done with the standard Cox PH model using ridge or lasso penalty. Wrapper approaches evaluation take into account the information dependency between the evaluated features, whereas, Filter approaches assume no dependency between the evaluated features. In this thesis we will use the filtering method because it is convenient and given a computational burden that we have and it is manageable. Filtering variable selection can enhance both easy interpretability and improved prediction accuracy.

Different strategies have been proposed for modifying the standard Cox PH model to deal with the high-dimensional setting. Some of these strategies are based on feature selection which can be either discrete or shrinkage selection. Discrete feature selection is the same as filtering method which will be discussed in this chapter, while the shrinkage feature selection is discussed in Chapter 7. Discrete feature selection dealing with high-dimension in the covariate space aims to determine which of the CNA genomic-windows have the strongest effects on the survival time. This can be done by a univariate score test for each of the CNA genomic-windows. Identifying significant CNA genomic-windows becomes more complicated due to the need to test thousands of hypotheses simultaneously, and this becomes harder if we use the penalized version of the test statistic. Additionally, a multiple testing correction is needed to properly adjustment for the number of tests performed.

Forward stepwise selection can be used to decide which of the subset of the CNA genomic-windows to include in a multivariate model sequentially. This forward stepwise selection is easy to implement, but it is not necessary to find the best model, instead it leads to find the locally optimal model rather than the best model (Klein, 2013)

This chapter is organized as follows. In Section 6.2, we provide two different univariate selection methods, the first method based on the standard Cox PH model in Section 6.2.1, and the second method based on penalized additive Cox PH model in Section 6.2.2. In Section 6.3 we deal with the dependence structure of significant CNA genomic-windows. The results of using forward stepwise selection are presented in Section 6.4.

6.2 Selecting Variables by Testing Individual Covariates

In this section we will consider the standard Cox PH model as well as the penalized additive Cox PH model in variable selection. The reason why we consider the standard Cox PH model is for comparison later on, whether some variables are selected by the penalized additive Cox PH model are more informative to the one that is selected by the standard Cox PH model.

6.2.1 Univariate Selection Based on Standard Cox PH Model

The univariate variable selection method proposed in Bøvelstad et al. (2007), which is based on testing the linear effect of each CNA genomics-windows values by itself on survival model. The univariate Cox model can be written as

$$h(t|\text{CNA}_j) = h_0(t)\exp(\beta_j \text{CNA}_j), \quad j = 1, \dots, 13253,$$

where β_j is the coefficient in a univariate Cox PH model where the j^{th} CNA variable is the only covariate included in the model. The null hypothesis is that there is no effect of each CNA. This test can be done by using the score test (Klein and Moeschberger (1997), Chapter 8.2), the test statistic is $T_c = U_{pl}(0)^T[I_{pl}^{-1}(0)]U_{pl}(0)$, where U_{pl}, and I_{pl} are the score vector and the information matrix of the partial log-likelihood. The score test has chi-square distribution with one degree of freedom, which enables us to compute the corresponding p value. Bøvelstad et al. (2007) used the score test to perform the univariate variable selection because this test does not require the estimate of the model parameter, which considerably reduces the computational time compared to the Wald and likelihood ratio tests.

Univariate variable selection was applied to the CNA genomic-windows. The number of CNA genomic-windows is equal to 13253, so this is the number of null hypothesis that we would like to test. The p-values were calculated for all 13253 hypothesis, then multiple testing correction used to determine how many of these p-value are rejected using both Bonferroni (1936) and Holm (1979) correction at the 5%. As a result, we rejected 612 hypothesis for both Bonferroni and Holm.

However, Cox PH model assumes a linearity of the covariates, so the univariate variable selection only identifies linear effect of the CNA genomic-windows, while we are interested in the non-linearity form of of the covariate. This motivated us to generalize the univariate selection of Bøvelstad et al. (2007) as described in the following section.

6.2.2 Univariate Selection Based on Penalized Additive Cox PH Model

We generalized the univariate variable selection method of Bøvelstad et al. (2007) by using penalized additive Cox PH model instead of the standard Cox PH model in the univariate variable selection method. The penalized univariate variable selection method is based on testing the effect of each CNA genomic-window variable by itself in the penalized additive Cox PH model. The univariate penalized additive Cox model can be expressed as

$$h(t|\text{CNA}_j) = h_0(t)\exp\left(f(\text{CNA}_j)\right), \quad j = 1, \ldots, 13253, \tag{6.1}$$

where $f(\text{CNA}_j)$ is the smoothing term for the j^{th} CNA genomic-window that is the only covariate included in the penalized additive Cox PH model. The null hypothesis of no covariate effect in the model is $H_{0j} : f(\text{CNA}_j) = \mathbf{0}$. The penalized univariate variable selection method can be done by testing the effect of each CNA genomic-window by itself in a univariate penalized additive Cox model with 5 equally spaced knots and the optimal value of the smoothing parameter choose separately for each CNA genomic-window in the CNA data.

As mentioned earlier, in Section 5.6.1, for this type of null hypothesis there are commonly three tests which can be used: penalized score test, Wald-type test and penalized likelihood ratio test. We used the penalized score test statistic as described in Section 5.6.1, which has a chi-square distribution under H_{0j}, with the degree of freedom depending on the value of the optimal smoothing parameter. The penalized score test does not require the estimate of the penalized spline effect, which reduces the computational time compared to Wald-type test and penalized likelihood ratio test. However, we still need to estimate the spline effect coefficients in order to compute five-fold Cross-Validated Partial Log-likelihood (CVPL) which is performed across a grid of values of smoothing parameter λ .

The p-values of the penalized score test were calculated for all 13253 hypotheses. Subsequently, multiple testing corrections were used to determine how many of these p-values can be rejected using either Bonferroni or Holm correction at the 5% significant level. As a result, we rejected 1056 out of 13253 hypotheses using either a Bonferroni and a Holm's multiple corrections.

This penalized univariate variable selection identified more significant CNA genomic-windows than the standard Cox PH model. This 1056 significant CNAs genomics-windows are from different block correlated data. The 1056 significant CNAs genomics-windows in each chromosome are presented in Appendix A Section A.1. There are 164 significant CNA genomic-windows are in common for both univariate variable selection and penalized univariate variable selection, which are presented in Appendix A, Section A.2. The relevant genes associated with Non-Small Cell Lung Cancer (NSCLC) from the 1056 significant CNA genomic-windows are summarized in Table 6.1. The gene which is in common for univariate variable selection and penalized univariate variable selection is written in red.

Gene	Chromosome	Reference
E2F2	1	Feliciano et al. (2017)
GALNT13	2	Nogimori et al. (2016)
SSFA2	2	Okayama et al. (2016)
PSMD2	3	Shi et al. (2017)
BEND4	4	Kettunen et al. (2017)
FEZF1	7	He et al. (2017)
NDUFAS	7	Xie et al. (2017)
CALU	7	Kundu et al. (2016)
TUSC3	8	Peng et al. (2017b)
ATAD2	8	Couto et al. (2017)
CPSF1	8	Kiehl et al. (2017)
RP11	10	Tang et al. (2017)
LMO3	12	Liu et al. (2017b)
PTPRE	10	Codreanu et al. (2017)
RSF1	11	Zhang et al. (2017)
BLID	11	Wang et al. (2015)
LRIG3	12	Roskoski (2017)
ATP8A2	13	Yan et al. (2017)
ERCC5	13	Perez-Ramirez et al. (2017)
YY1	14	Huang et al. (2017)
KIF23	15	Vikberg et al. (2017)

Table 6.1: Genes related to Non-Small Cell Lung Cancer. The gene which is in common for both univariate variable selection and penalized univariate variable selection is written in red.

Tables 6.2, and 6.3 show some genes from the windows which we found to be significant which are related to other types of cancer. The genes which are in common for both univariate variable selection and penalized univariate variable selection are written in red.

Gene	Chromosome	Related Cancer	Reference
B3GNT5	3	breast	Uehiro et al. (2016)
ATP11B	3	ovarian	Moreno-Smith et al. (2013)
KLHL6	3	leukemia	Sutton et al. (2015)
ABCCS	3	breast	Hofman et al. (2016)
SENP2	3	breast	Nait Achour et al. (2013)
TRA2B	3	breast	Liu et al. (2017a)
RPL39L	3	breast	Dave et al. (2014)
TP63	3	head and neck	Gleber-Netto et al. (2018)
GTF2H2	5	breast and ovarian	Walker C et al. (2017)
SMN1	5	tong	Upadhyay et al. (2017)
SERF1A	5	breast	Mustacchi et al. (2013)
STAG3L3	7	thyroid	Heß et al. (2011)
GTPBP10	7	prostate	Jin et al. (2016)
SYPL1	7	liver	Chen et al. (2017a)
ING3	7	prostate	Nabbi et al. (2017)
CPED1	7	breast	Peng et al. (2017a)
SULF1	8	bladder	Lee et al. (2017a)
ANXA13	8	collector	Jiang et al. (2017)
MTSS1	8	bladder	Du et al. (2017)
TRMT12	8	breast	Rodriguez et al. (2007)
SQLE	8	prostate	Stopsack H et al. (2017)
BCCIP	10	esophageal	Chen et al. (2017c)
MARCH8	11	gastric	Wang et al. (2017b)
PAK1	11	colorectal	Yuan He et al. (2017)
MMP7	11	gastric	Sandoval-Borquez et al. (2017)
TAGLN	11	esophageal	Chen et al. (2017b)
SIK3	11	breast	Amara et al. (2017)
RNF26	11	cervical	Zhang et al. (2005)
RERGL	12	colorectal	Liu and Zhang (2017)
MGAT4C	12	prostate	Demichelis et al. (2012)
CEP83	12	colorectal	Zhang et al. (2016)
MTERF2	12	cervical	Prasad et al. (2017)
ZIC5	13	colorectal	Satow et al. (2017)
EFNB2	13	breast	Schultz et al. (2017)

Table 6.2: Genes related to other types of the cancer. The gene which are in common for both univariate variable selection and penalized univariate variable selection are written in red.

Gene	Chromosome	Related Cancer	Reference
ARGLU1	13	breast	Zhang et al. (2011)
VRK1	14	liver	Lee et al. (2017b)
SETD3	14	liver	Cheng et al. (2017)
ANP32	15	hepatocellular	Ohno et al. (2017)
LRRC49	15	breast	De Souza Santos et al. (2008)
IQGAP	15	hepatocellular	Zoheir et al. (2016)
ABCC12	16	breast	Jalkh et al. (2017)
KIF2B	17	hepatocellular	Qu et al. (2016)
KDM4B	19	ovarian	Wilson et al. (2016)
SAFB2	19	breast	Hong et al. (2015)

Table 6.3: Genes related to other types of the cancer. The genes which are in common for univariate variable selection and penalized univariate variable selection are written in red.

Tables 6.4, 6.5, 6.6, and 6.7 show some genes that we found to be significant in our penalized univariate seelction, but there are no prior studies describing a relationship between these genes and any type of cancer. The genes which are in common for both univariate variable selection and penalized univariate variable selection are written in red.

<table>
<tr><td colspan="6">Chromosome 1</td></tr>
<tr><td>TEX46</td><td>ULZP1</td><td>HNRAPR</td><td>ZNF436</td><td>ASAP</td><td>ID3</td></tr>
<tr><td>RPL11</td><td>PITHD</td><td>CNR2</td><td>MYOM3</td><td>CRHL3</td><td>IL22RA1</td></tr>
<tr><td>NCMAP</td><td>ESRRG</td><td>ESPPG</td><td>GPATCH2</td><td>SPAT17</td><td>DUSP10</td></tr>
<tr><td>MDS4</td><td>PLDS</td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="6">Chromosome 2</td></tr>
<tr><td>ORCH4</td><td>MBD5</td><td>EPC2</td><td>KIFSC</td><td>LYPD6B</td><td>MMADHC</td></tr>
<tr><td>CSRNP3</td><td>SCN1A</td><td>SCN7A</td><td>SCN9A</td><td>XIRP2</td><td>UBE2E3</td></tr>
<tr><td>NEUROD1</td><td>CERKL</td><td>PPPIRIC</td><td>NFUROD1</td><td>DNAJC10</td><td>FRZB</td></tr>
<tr><td>NUP35</td><td>DUSP19</td><td>NEU4</td><td>PTR5</td><td>LYPP6</td><td>ZTGA4</td></tr>
<tr><td>NCKAP1</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="6">Chromosome 3</td></tr>
<tr><td>CNTN6</td><td>CNTN4</td><td>IL5RA</td><td>TRNT1</td><td>CRBN</td><td>SUMF1</td></tr>
<tr><td>GRM7</td><td>SSUH2</td><td>CAV3</td><td>UBE2E2</td><td>ROB02-206</td><td>ROB02-207</td></tr>
</table>

Table 6.4: Genes which are not related to any type of cancer. The genes which are in common for univariate variable selection and penalized univariate variable selection are written in red.

Chromosome 3					
ROBO1	DCVND1	KLHL24	LAMP3	MCCC1	PARL
ECCE	HTR3C	YEATS2	AP2M1	CAMK2N2	THPO
ABCF3	CHRD	VWA5B2	HTR3P	EPHB3	MAGEF1
EHHADE	MAP3KB	TMEM41A	ETV5	DGKG	CRYGS
ADIPOO	EIF4A4	ST6GAL1	DNAJB11	HRG	TBCCDI
LLP	ST6GAL1	DGKG	MAP3K13	FETUB	KNG1
SST	EDM1	HTR3E	MAP6D	ALG31	UPSB
RFC4	FGF12	BCL6			
Chromosome 4					
SLC10A6	DSSPP	SPP1	IBSP	PYURF	HERC3
FAM13A	DMP1	PKD2	HSD17B13	HSD17B11	NUDT9
HERC6	GPRIN3	SNCA	MMRN1	CCSSER	SMARCAD1
ATOH1	GRID2	EGF	GARI	LRIT3	RRH
FAM160A1	PRSS48	SH3D19	RPS3A		
Chromosome 5					
TEM161B	MEF2C				
Chromosome 7					
TYW1B	POM121	TR1M74	MAG12	ADAM22	SRI
STEAP4	TEX47	CDK17	STEAP1	CLDN12	STEAP2
PUS7	SRPK2	ATXN7L1	NAMPT	CCDC7	RRKAR
COG	PNPLA	GPR22	DOCK4	TMEM168	DLD
IFRD1	SLC26A4	LAMB1	LRRN3	BMT2	FOXP2
LSMEM1	CBLL1	THAP5	ZNF227	GPR85	NRCAM
MET	ASZ1	TES	CAV2	MDF1C	TEFC
LSM8	ANKRD7	TSPAN12	WNT16	FAM36	ZNF800
PTRRZ1	IQUB	SPAM	AASS	LMOD2	HYAL4
CADPS2	SLC13A	CRM8	AAF5	WASL	POT1
GCC1	SND1	P4X4	LEP	TNP03	HLPDA
CCDC136	FLNC	KCP	1MPDH1	METTL28	LEP
Chromosome 8					
ARFGEF	PREX2	PRDM14	NCOA2	NDUFAF6	PLEKHF2
HAS1	FAM91A1	ZHX2	ZHX1	TATDN	TMEM6
TBC1D31	FAM91A1	ZNF572	NSMC	TONSL	COMMDS
CYHRIGPT	ZNF34	ZNF7	ZNF16	ARHAP	
Chromosome 10					
ADMTS14	TBATA	PCBD1	UNC5B	SLC29A3	CDH23
VSTR	PSAR	CHST3	MICUI	ASCC1	P4A41
DNAJB12	TEX36	EDRF1	MMP21	UROS	CLRN3
TPRE	MGMT	EBF3	TCERGIL	RRC27	NKX6-2
JAKMMIP3	BNIP3	DPYSL4	ALOK5	QIT3	

Table 6.5: Genes which are not related to any type of cancer. The genes which are in common for univariate variable selection and penalized univariate variable selection are written in red.

Chromosome 11					
COA4	WASHC2C	FAM25E	PAAF1	DNAJB13	MRPL48
AGAP4	INPPSA	ADGRA1	PAB64	PLEKHB1	C2CD3
UCP3	C2CD	PPME1	LIPT2	P4HA3	PGM2L1
KCNE3	POLD3	RNF169	KLH35	TPBGL	OR2ATA
SLCO2B1	LRRC32	KCTD21-A51	AAMDC	INTS4	HYOU1
NDUFC2	ALG8	USP35	AQP11	GDPD4	TSKU
THAP12	NARS2	DEUP1	MTNRIB	SMCO4	CEP29S
TAFID	SMCO4	VSTMS	HEPHL1	TMEM123	DYNC2H1
CARD16	CARD17	CASP5	CASP1	CASP4	ZPR1
PC5K7	RNF214	CEP164	BACE1	FXYD6	FXYD2
JMML	MPZLZ	CD3E	ARCN1	CD3D	CD3G
ATP3L	UBE4A	KMT24	IFT46	BCL91	CCDC84
TRAPPC4					
Chromosome 12					
PIKK3C2G	TSFM	CY27B1	EEF1AKMT3	P4K2C	ARHGEF25
O59	TSPAN31	CTDSPS	B4GALNTI	CDK4	ATP23
SLC16A7	TPH2	TRHDE	TBC1D15	UBE2D	MRPL42
SOCS2	CRADD	PLXNC1			
Chromosome 13					
PARP4	ATP12A	RNF17	CENPJ	NUP58	AMER2
MTMR6	B3GLCT	RXFP2	SLITRK5	TM9SF2	CLYBL
PCCA	TMTC4	BIVH	SLC10A2	GGACT	ITGBL1
FGF14	KDELC1	TPP2	METTL21C	TEX30	LIG4
CDC16	RASA3	GAS6	ATP42	GRK1	TEDP1
Chromosome 14					
AKAP6	NPAS3	SLC35F4	ACTR10	LINC00216	TIMM9
K1AA0586	DACT1	ACTN1	NEK9	TGFB3	SMOC1
SUSD4	YLPM1	GALNT16	PROX2	ZFP36L1	PPF3
DCAF3	ZFYVE1	ADAM20	MED6	CIDC	TTC9
COQ6	PCNX1	ENTPD5	PTGR2	JDP2	VRTN
BDKRB2	GSKIP	PAPOLA	BDKRB1	ATG2B	TCF12
BCL11B	SETD3	CCNK	SETD3	HHIPL1	CCDC85C
EML1	EVL	DEGS2	BEGAIN	DLK1	MEG3
WDR25	MNS1	ZNF280D			
Chromosome 15					
MNS1	ZNF280D	TCF12	TCF12	ALDH1A	CGNL1
GCOM1	MYZAP	POLR2M	RNF111	SLTM	BN1P2
LDHAL6B	GTF2A2	RORA	GCNT	CCNB2	MINDY2
LIPC	ANXA2	C2CD48	PH1B	SNX1	TRIP4
SLC51B	MTFMT	SLC24A1	CLPXHACD3	ZNF609	CHD2

Table 6.6: Genes which are not related to any type of cancer. The genes which are in common for univariate variable selection and penalized univariate variable selection are written in red.

Chromosome 15					
IGDCC	TRIP4	CA12	CA12	USP3	MY01E
SLTM	GLCE	SPESP1	NOX5	PAQPS	ARRCD4
UACA	LARP6	HEXA	PKM	MY09A	NE01
HCN4	NPTN	ANPEP	AP352	ARPIN	NGRN
CIBI	FE2	HDDC3	IDH2	MAN2A2	UNC45A
CRTC3	VPS33B	RRC1	SLCO3A	RGMA	MCTP2
NR2F2	SPATA8	CHD2	ST8S1A	C15orf32	
Chromosome 16					
RBFOX1	PHKB	NOD2	CYLD	AKTIP	RBL2
SALL1	CHD9	ADCY7	BRD7	AMFR	GOT2
LONP2	USB1	GNAOL	CETP	PLLP	CPNE2
IRX5	CDH8	TOX3	SNX20	PRSSS5	PKDIL
VAC14	HYDIN	ZNF19	CALB2	ZNF13	TAT
CHST4	AP1GI	IST1	ATXNIL	CHST4	CMTR
Chromosome 17					
SPAG9	CA10	NME1	NME2	MBTD1	AKAP1
UTP18	KIF2B	TOMIL1	HLF	MMD	NOG
COX11	STXBP4	PCTP	TR1M26	COIL	DGKG
Chromosome 19					
LONP1	CATSPERD	UHRF1	PLN3	ARRDC5	GDF1
T1CAM1	ZNRF4	PRR22	DUS3L	RPL36	FUTC
LONP	NRTN	FUT3	VMAC	KHSRP	INSR
MBD3L2	GPR108	CAMSAP3	CLEC4C	ELAVL1	SUGP2
EVI5L	HNRNPM	STXBP2	CTXN1	STX10	TRIR
HOOK	JUNB	PFX1	SAMD1	TECR	GIPC1
PNK1	RLN3	MRI1	OR7A17	SL1A6	WIZ
NACC1	LYL1	TRMT1	FARSA	SYDE	HOMER3
UPF1	CERS1	COPE			
Chromosome 20					
TCF15	SCRT2	SLC52A3	ANGPT4	RSPO4	RAE1
PSMF1	BMP7	RBM38	ZBP1	PMEPA1	SPO11
CTCEL	PCK1				

Table 6.7: Genes which are not related to any type of cancer. The genes which are in common for univariate variable selection and penalized univariate variable selection are written in red.

6.3 Dependencies of Significant CNA genomic-windows

Generally, the CNA genomic-window is often correlated between neighboring genomic-windows, and some of our significant neighboring windows of CNA are highly correlated. Figure 6.1 shows the correlation map for all the 1056 significant CNA genomic-windows, each value on the x and y axes correspond to the significant CNA genomic-widows and the color of the figure at each index $(i, j),\ i, j = 1, \dots, 1056$, represents the correlation value as seen in the legend.

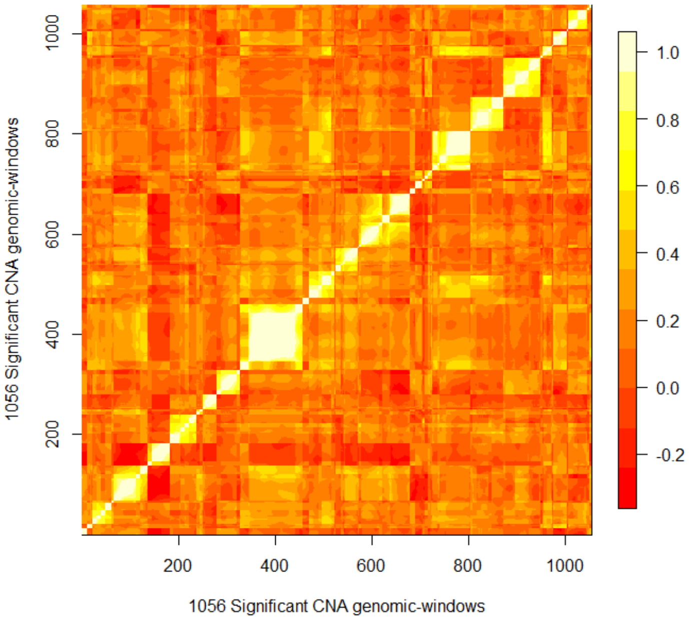


Figure 6.1: Correlation heat map all 1056 significant CNA genomic-windows.

To deal with this correlation structure, we need to choose one variable from each highly-correlated block as representative of that block, and then we include the not correlated significant CNA genomic-windows in the multivariate penalized additive Cox PH model. Because the shapes of the estimated log hazard ratios for the correlated significant CNA genomic-windows are similar, we need to choose the uncorrelated significant CNA genomic-windows to visualize different shapes of the estimated log hazard ratios. To choose one variable from each highly-correlated block, we must select a threshold value of correlation, in order to define the blocks. The threshold value is selected as follows.

- We pick one reference variable from the middle of the highly-correlated block of Chr 7, as we can see in black box in the Correlation heat map all 1056 significant CNA genomic-windows in Figure 6.2.

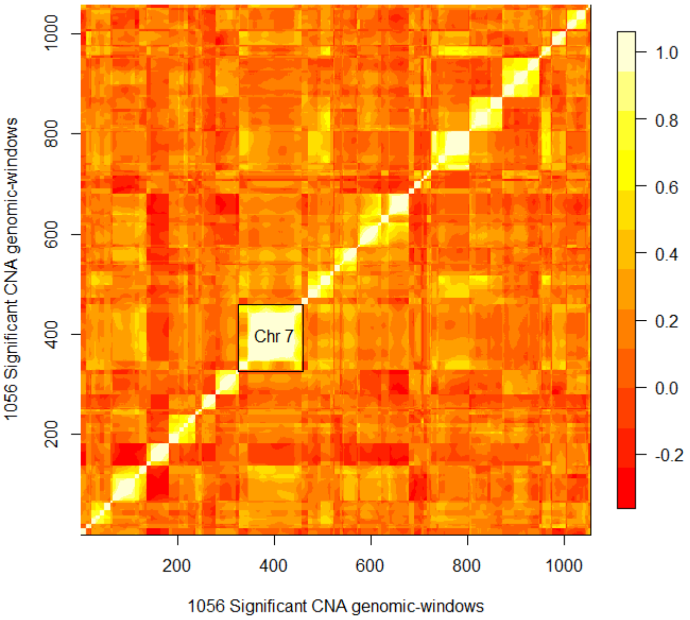


Figure 6.2: Correlation heat map all 1056 significant CNA genomic-windows, the black box represents the highly-correlated block of Chromosome 7.

then we fit the model with an optimal value of the smoothing parameter, and 5 equally spaced knots. The model is

$$h(t, \text{CNA}_{6456}) = h_0(t) \exp\Big(f(\text{CNA}_{6456})\Big), \tag{6.2}$$

we have four standard error of the parameters estimate, in order to compute one value of the standard error of the parameter estimates, we used the geometric mean because it has an advantage over the arithmetic mean in that it is not affected much by fluctuations (i.e is not affected by extreme values). The geometric mean for the standard error of the parameter estimates of model (6.2) is calculated as

$$s_1 = \Big(\mathrm{se}(\beta_{z1}) \times \mathrm{se}(\beta_{z2}) \times \mathrm{se}(\beta_{z3}) \times \mathrm{se}(\beta_{z4})\Big)^{1/4}.$$

- We fit the models that contains $f(\mathrm{CNA}_{6456})$, and only one other significant CNA genomic-window j

$$h(t, \mathrm{CNA}_{6456}, \mathrm{CNA}_j) = h_0(t) \exp\big(f_1(\mathrm{CNA}_{6456}) + f_2(\mathrm{CNA}_j)\big), \qquad (6.3)$$

  for each j in our list of 1056 significant CNA genomic-windows except CNA_{6456}, with optimal values of the smoothing parameters and 5 equally spaced knots. The geometric mean for the standard error of the parameter estimates of the first term in the model $f(\mathrm{CNA}_{6456})$ is calculated as

$$\mathrm{s}_{1|j} = \Big(\mathrm{se}(\beta_{z1|j}) \times \mathrm{se}(\beta_{z2|j}) \times \mathrm{se}(\beta_{z3|j}) \times \mathrm{se}(\beta_{z4|j})\Big)^{1/4},$$

  where $\mathrm{s}_{1|j}$ is the geometric mean of the first variable in model (6.3) where variable j is also present.

- Calculate the correlation between $r_j = \mathrm{cor}(\mathrm{CNA}_{6456}, \mathrm{CNA}_j)$

- Plot the correlation r_j versus $\log(\mathrm{s}_{1|j}/s_1)$, and then fit a line using loess with smoothing parameter 0.1 (Figure 6.3).

The standard error $\mathrm{s}_{1|j}/s_1$ starts to increase at about 0.6 correlation $r = 0.6$. There is an infusion of standard error from the first variable in model as result of including the second variable, which means that above 0.6 correlation, the other variable starts to influence the standard error. Figure 6.3 shows the plot of the $\log(\mathrm{s}_{1|j}/s_1)$ versus the correlation. The points at correlation values 1 are from the neighboring significant CNA genomic-windows from Chr7, the red solid line is the fitted loess line. The standard error of the estimated parameters starts to increase at about correlation $r = 0.6$.

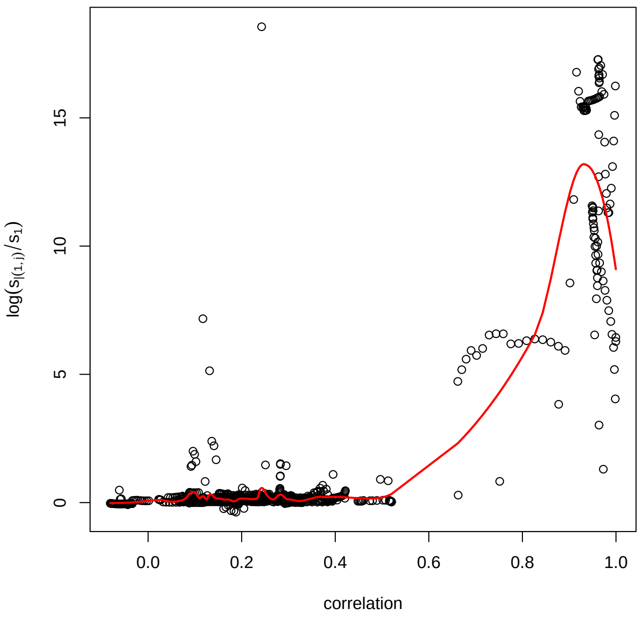


Figure 6.3: The plot of $\log(se/s)$ versus the correlation r. The red solid line is the fitted loess line.

We define a block of significant CNA genomic-windows to be a group of consecutive windows in our list of significant CNA genomic-windows where each window has a correlation of at least 0.6, then we select the first CNA genomic-window from each block to represent that block. As a result we have a list of 41 significant CNA genomic windows, which are not correlated with each other. Table 6.8 illustrates the 41 significant uncorrelated CNA genomic-windows with the corresponding chromosomes, and p-value.

#	chr	window	p-value	#	chr	window	p-value
1	Chr 1	108	0.027921	21	Chr8	7264	0.049593
2	Chr 1	949	0.041385	22	Chr10	8125	0.033214
3	Chr2	1815	0.032956	23	Chr10	8443	0.029323
4	Chr2	1979	0.046475	24	Chr11	8804	0.047742
5	Chr2	2284	0.043557	25	Chr11	8928	0.046319
6	Chr3	2291	0.013653	26	Chr11	9024	0.033174
7	Chr3	2676	0.042087	27	Chr11	9084	0.035453
8	Chr3	3094	0.046057	28	Chr11	9143	0.023786
9	Chr3	3186	0.046428	29	Chr12	9316	0.029923
10	Chr4	3474	0.048623	30	Chr12	9508	0.046966
11	Chr4	3797	0.034302	31	Chr12	9579	0.033303
12	Chr4	4001	0.043624	32	Chr12	9649	0.043065
13	Chr5	4528	0.034418	33	Chr12	9686	0.047636
14	Chr5	4614	0.033966	34	Chr13	9752	0.049974
15	Chr6	5137	0.040975	35	Chr13	10231	0.041216
16	Chr6	5481	0.046758	36	Chr14	10278	0.041364
17	Chr7	6253	0.038643	37	Chr15	10418	0.048141
18	Chr7	6410	0.046685	38	Chr16	10964	0.047407
19	Chr8	6528	0.048067	39	Chr17	11217	0.037782
20	Chr8	6992	0.049689	40	Chr18	11581	0.029107
				41	Chr19	12356	0.045674

Table 6.8: 41 significant CNA genomic windows

6.4 Forward Stepwise Selection

To improve on the penalized univariate variable selection, forward stepwise selection can be used to decide which of the 41 significant CNA genomic-windows to include in a multivariate model sequentially. We start with the model that contains the significant clinical data, Stage-T, stage-N, and age with only one of the significant CNA genomic windows as smoothing terms. The model is

$$h(t) = h_0(t)\exp\Big(StageT + StageN + f(Age) + f_1(\text{CNA}_j)\Big), \tag{6.4}$$

for each j is on our list of 41 significant CNA genomic-window in Table 6.8. We have two smoothing parameters to estimate one each for age and the significant CNA variable. Let $\lambda = 0.1 \times \exp(\log(100)/n_\lambda)^i$ for $i = 0, \ldots, n_\lambda$ be a vector of values of the smoothing parameter, and $n_\lambda = 20$ is the length of the smoothing parameter. In practice, the smoothing parameter is often chosen by a grid search of the parameter space, because we need to find global maxima in grid. However, gradient method are very useful for convex optimization problems, which leads to select the smoothing parameter more accuracy. The smoothing parameters λ_1, λ_2, for the first and second smoothing terms in the model (6.4) are obtained separately using five-fold Cross-Validated partial log-likelihood. In order to estimate two smoothing parameters λ_1 and λ_2 separately, we used two-dimensional grid search for all possible pairs of the two smoothing parameters λ_1 and λ_2. The values of the smoothing parameters $\lambda_{1,opt}$ and $\lambda_{2,opt}$ that maximizing CVPL are selected.

Subsequently, the penalized additive Cox PH model that contains Stage-T, stage-N, age and one significant CNA as smoothing terms with the optimal values of the smoothing parameters are fitted with 5 equally spaced knots. Testing the hypothesis of no effect of the smoothing terms is carried out using a penalized score test as described in Section 5.6.1 for all covariates. The results of including only one significant window of the CNA in the penalized additive Cox PH model that contains Stage-T, Stage-N and age are presented in Table 6.9.

According to Table 6.9, model 31 has the smallest p-value, CNA$_{9579}$ is the first significant CNA genomic-window covariate included in the model.

#	chr	window	p-value	#	chr	window	p-value
1	chr1	108	0.001986	21	chr8	7264	0.001056
2	chr1	949	0.000382	22	chr10	8125	0.001503
3	chr2	1815	0.003874	23	chr10	8443	0.002914
4	chr2	1979	0.001871	24	chr11	8804	0.002765
5	chr2	2284	0.003728	25	chr11	8928	0.000223
6	chr3	2291	0.000128	26	chr11	9024	0.002302
7	chr3	2676	0.001571	27	chr11	9084	0.004603
8	chr3	3094	0.000109	28	chr11	9143	0.000325
9	chr3	3186	0.001683	29	chr12	9316	0.000463
10	chr4	3474	0.001529	30	chr12	9508	0.001101
11	chr4	3797	0.000432	31	chr12	9579	9.3×10^{-05}
12	chr4	4001	0.001435	32	chr12	9649	0.000425
13	chr5	4528	0.001183	33	chr12	9686	0.002184
14	chr5	4614	0.003752	34	chr12	9752	0.001297
15	chr6	5137	0.000806	35	chr13	10231	0.001498
16	chr6	5481	0.003395	36	chr13	10278	0.000477
17	chr7	6253	0.001174	37	chr14	10418	0.000517
18	chr7	6410	0.001543	38	chr15	10964	0.000961
19	chr8	6528	0.000548	39	chr16	11217	0.000426
20	chr8	6992	0.001091	40	chr17	11581	0.000539
				41	chr19	12356	0.002509

Table 6.9: The result of p-value for testing each of the selected CNA in the multivariate penalized Cox model.

To include the second covariate in the model, we fit the model with three smoothing terms, the model is

$$h_i(t) = h_0(t) \exp\Big(StageT + StageN + f(Age) + f_1(\mathrm{CNA}_{9579}) + f_2(\mathrm{CNA}_j))\Big).$$

for each j is on the list of 41 significant CNA genomic-window in Table 6.8 except CNA_{9579}, and the optimal values of the smoothing parameters λ_1, λ_2, and λ_3 for each smoothing terms are obtained separately. This is can be done by three-dimensional grid searching for all possible pairs of the three smoothing parameters. The smoothing parameters $\lambda_{1,opt}$, $\lambda_{2,opt}$, and $\lambda_{3,opt}$ that maximizing the CVPL are selected. However, $\lambda_{1,opt}$ from the three-dimensional searching does not change much compared to the $\lambda_{1,opt}$ in the two-dimensional searching, this mean that we can fixed $\lambda_1 = \lambda_{1,opt}$ according to the previous two-dimensional searching, and only searching for λ_2, and λ_3 that maximizing the CVPL. This reduces the computational time compared to three-dimensional searching.

Once we have the optimal values of the smoothing terms $\lambda_{1,opt}, \lambda_{2,opt}, \lambda_{3,opt}$, we can calculate the penalized score test to select which of the remaining significant CNA covariate can be added to the most improve model. The result of including the second significant CNA covariate is presented in Table 6.10. We continue this process until our model is not significant in terms of p-value.

#	chr	window	p-value	#	chr	window	p-value
1	chr1	108	0.000161	21	chr8	7264	4.1×10^{-05}
2	chr1	949	$9.6e \times 10^{-05}$	22	chr10	8125	0.000137
3	chr2	1815	0.000621	23	chr10	8443	0.000316
4	chr2	1979	0.00018	24	chr11	8804	0.000448
5	chr2	2284	0.000138	25	chr11	8928	3×10^{-05}
6	chr3	2291	1.5×10^{-05}	26	chr11	9024	0.000242
7	chr3	2676	6.1×10^{-05}	27	chr11	9084	0.000188
8	chr3	3094	1.1×10^{-05}	28	chr11	9143	3.3×10^{-05}
9	chr3	3186	0.001143	29	chr12	9316	6.5×10^{-05}
10	chr4	3474	0.000335	30	chr12	9508	6.7×10^{-05}
11	chr4	3797	3×10^{-05}	32	chr12	9649	0.000378
12	chr4	4001	0.000135	33	chr12	9686	2×10^{-06}
13	chr5	4528	0.000108	34	chr12	9752	0.000251
14	chr5	4614	0.000206	35	chr13	10231	0.00011
15	chr6	5137	1×10^{-05}	36	chr13	10278	4.7×10^{-05}
16	chr6	5481	0.000111	37	chr14	10418	8×10^{-06}
17	chr7	6253	7.7×10^{-05}	38	chr15	10964	0.00019
18	chr7	6410	0.000251	39	chr16	11217	3.3×10^{-05}
19	chr8	6528	1.5×10^{-05}	40	chr17	11581	3.5×10^{-05}
20	chr8	6992	3.6×10^{-05}	41	chr19	12356	0.000425

Table 6.10: The result of p-value for testing each of the significant CNA genomic-windows in the multivariate penalized Cox model.

The challenge here is how to estimate the smoothing parameters for each smoothing terms separately in the model. We have fixed the smoothing parameter of the smoothing terms that been in the model for several iterations, and we only search for the last two smoothing parameters that are included in the model. Although this process is time consuming, it is less time consuming the optimizing all 42 smoothing parameters simultaneously. However, our approach is not ideal since we do not search the entire possible parameter space. The 42-dimensional search were to computationally demanding to be feasible in practice. Our approach is a pragmatic middle ground between the impractical full search and the overly restricted assumption that the smoothing parameters are the same for each smoothing terms.

This forward stepwise selection ended with 40 significant CNA genomic-windows with p-value$= 0.046$. The final model can be written as

$$\begin{aligned} h(t) = h_0(t) \exp \Big(& StageT + StageN + f(Age) + f_1(\text{CNA}_{9579}) + f_2(\text{CNA}_{9686}) \\ & + f_3(\text{CNA}_{8928}) + f_4(\text{CNA}_{2291}) + f_5(\text{CNA}_{10278}) + f_6(\text{CNA}_{949}) + f_7(\text{CNA}_{9143}) \\ & + f_8(\text{CNA}_{10964}) + f_9(\text{CNA}_{8125}) + f_{10}(\text{CNA}_{11217}) + f_{11}(\text{CNA}_{10418}) + f_{12}(\text{CNA}_{5137}) \\ & + f_{13}(\text{CNA}_{3474}) + f_{14}(\text{CNA}_{9752}) + f_{15}(\text{CNA}_{6528}) + f_{16}(\text{CNA}_{3797}) + f_{17}(\text{CNA}_{4001}) \\ & + f_{18}(\text{CNA}_{3094}) + f_{19}(\text{CNA}_{10231}) + f_{20}(\text{CNA}_{4614}) + f_{21}(\text{CNA}_{9508}) + f_{22}(\text{CNA}_{12356}) \\ & + f_{23}(\text{CNA}_{3186}) + f_{24}(\text{CNA}_{9024}) + f_{25}(\text{CNA}_{108}) + f_{26}(\text{CNA}_{6992}) + f_{27}(\text{CNA}_{2676}) \\ & + f_{28}(\text{CNA}_{9316}) + f_{29}(\text{CNA}_{8804}) + f_{30}(\text{CNA}_{9084}) + f_{31}(\text{CNA}_{9649}) + f_{32}(\text{CNA}_{1815}) \\ & + f_{33}(\text{CNA}_{5481}) + f_{34}(\text{CNA}_{2284}) + f_{35}(\text{CNA}_{8443}) + f_{36}(\text{CNA}_{6410}) + f_{37}(\text{CNA}_{4528}) \\ & + f_{38}(\text{CNA}_{1979}) + f_{39}(\text{CNA}_{11581}) + f_{40}(\text{CNA}_{6253}) \Big) \end{aligned} \tag{6.5}$$

The optimal smoothing parameter of age is equal to 0.100 and 1.928 degrees of freedom, while the optimal values of the smoothing parameters for all the 40 significant CNA are summarized in Table 6.11.

Using the optimal values of the smoothing parameters for each smoothing term in the model (6.5), the estimate of the fixed effect parameter and their inferences can be obtain, which are summarized in Table 7.8 in Section 7.6 for comparison with the shrinkage variable selection approach later on.

#	Variable	λ_{opt}	#	Variable	λ_{opt}
1	9579	0.100	21	9508	0.630
2	9686	0.398	22	12356	0.630
3	8928	0.158	23	3186	0.630
4	2291	0.100	24	9024	0.630
5	10278	0.251	25	108	0.630
6	949	2.154	26	6992	0.630
7	9143	1.584	27	2676	0.630
8	10964	0.630	28	9316	0.215
9	8125	0.630	29	8804	0.630
10	11217	0.630	30	9084	0.630
11	10418	0.630	31	9649	0.630
12	5137	3.981	32	1815	0.630
13	3474	0.630	33	5481	0.630
14	9752	1.584	34	2284	0.316
15	6528	0.100	35	8443	0.215
16	3797	1.584	36	6410	0.630
17	4001	0.630	37	4528	0.630
18	3094	0.630	38	1979	0.630
19	10231	0.630	39	11581	0.630
20	4614	0.630	40	6253	0.630

Table 6.11: The optimal values of smoothing parameters λ for each of the significant windows of the CNA in the multivariate penalized Cox PH model according to the significant CNA genomic-windows ordering in the multivariate penalized Cox PH model.

Figure 6.4 shows the plot of the log hazard ratio of age versus age. The solid black line is the estimated log hazard ratio for age, the dashed lines are the 95% point-wise confidence band. The points indicate the number of observations on each individual that were either censored or a failure. Figures 6.5, 6.6, 6.7, and 6.8 show equivalent plots of the estimated log hazard ratio for each of the significant CNA genomic-windows, x-axis represents the observed CNA genomic-window, and y-axis represents the estimated log hazard ratio of the significant CNA genamic-window. The number of the CNA window is shown in the legend.

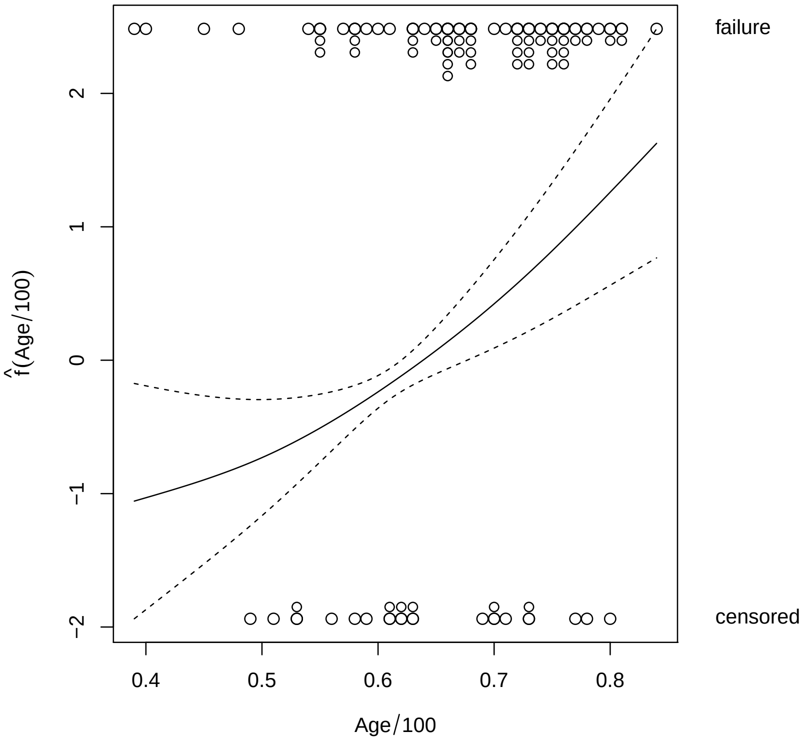


Figure 6.4: The plot of the estimated log hazard ratio for $\hat{f}(age)$. The solid black line is the estimated log hazard ratio, the points indicate the number of observations on each individual that were either censored or a failure. The dashed lines are the 95% point-wise confidence band.

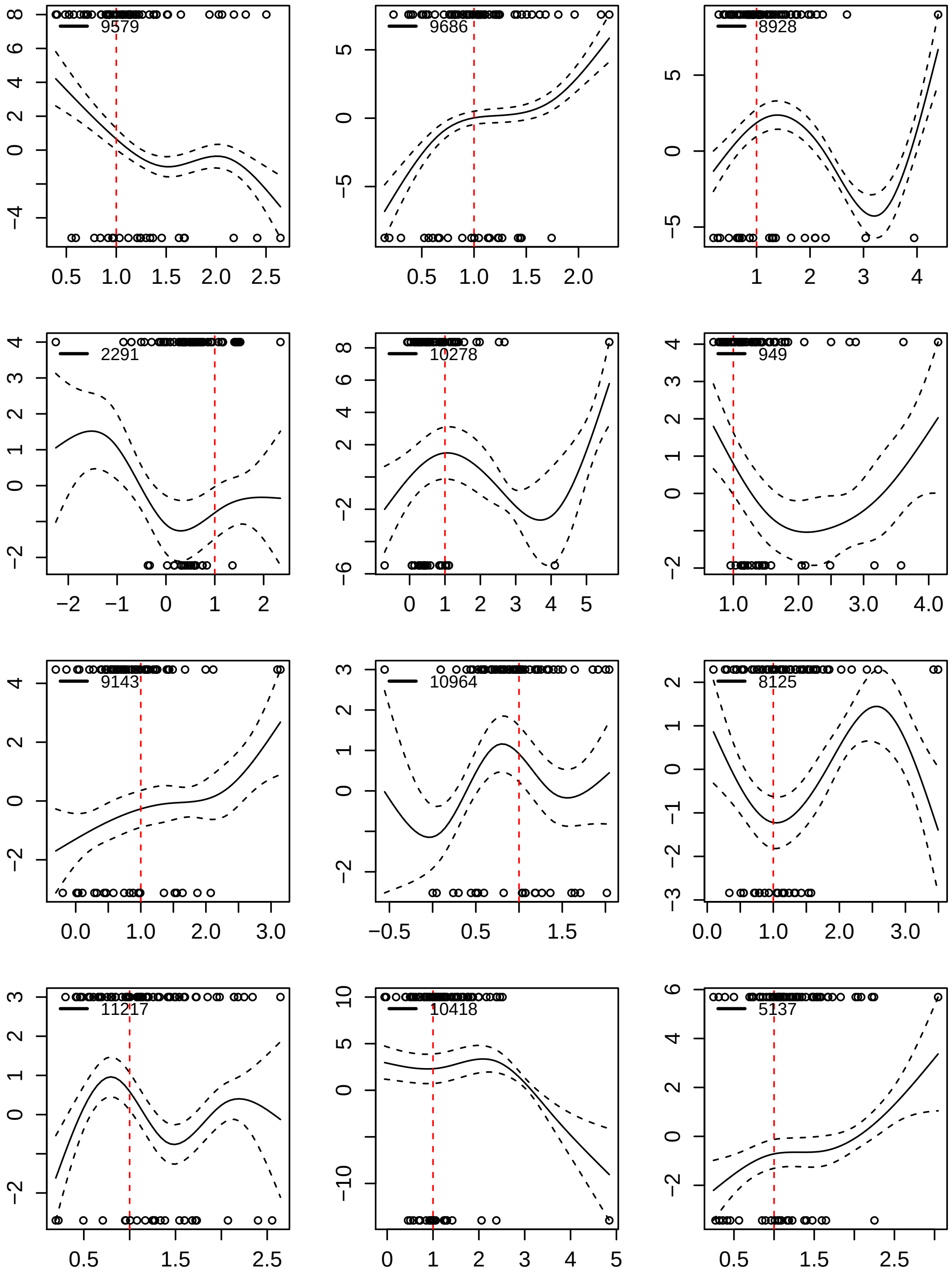


Figure 6.5: Plots of the estimated log hazard ratio for each of the significant window of the CNA. The solid black line is the estimated log hazard ratio, the dashed lines are the 95% point-wise confidence band. The points indicate the number of observations on each individual that were either censored or a failure.

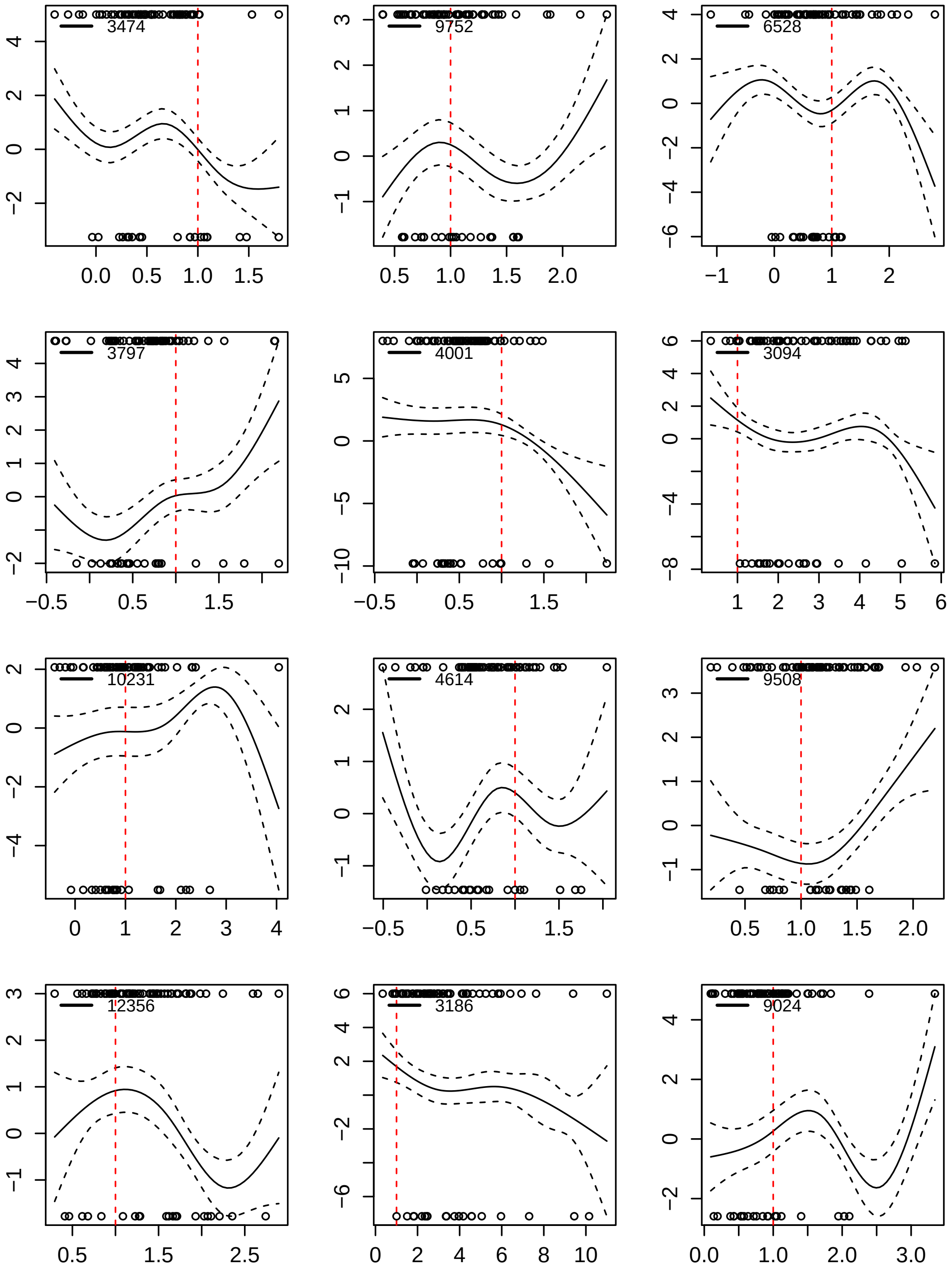


Figure 6.6: Plots of the estimated log hazard ratio for each of the significant window of the CNA. The solid black line is the estimated log hazard ratio, the dashed lines are the 95% point-wise confidence band. The points indicate the number of observations on each individual that were either censored or a failure.

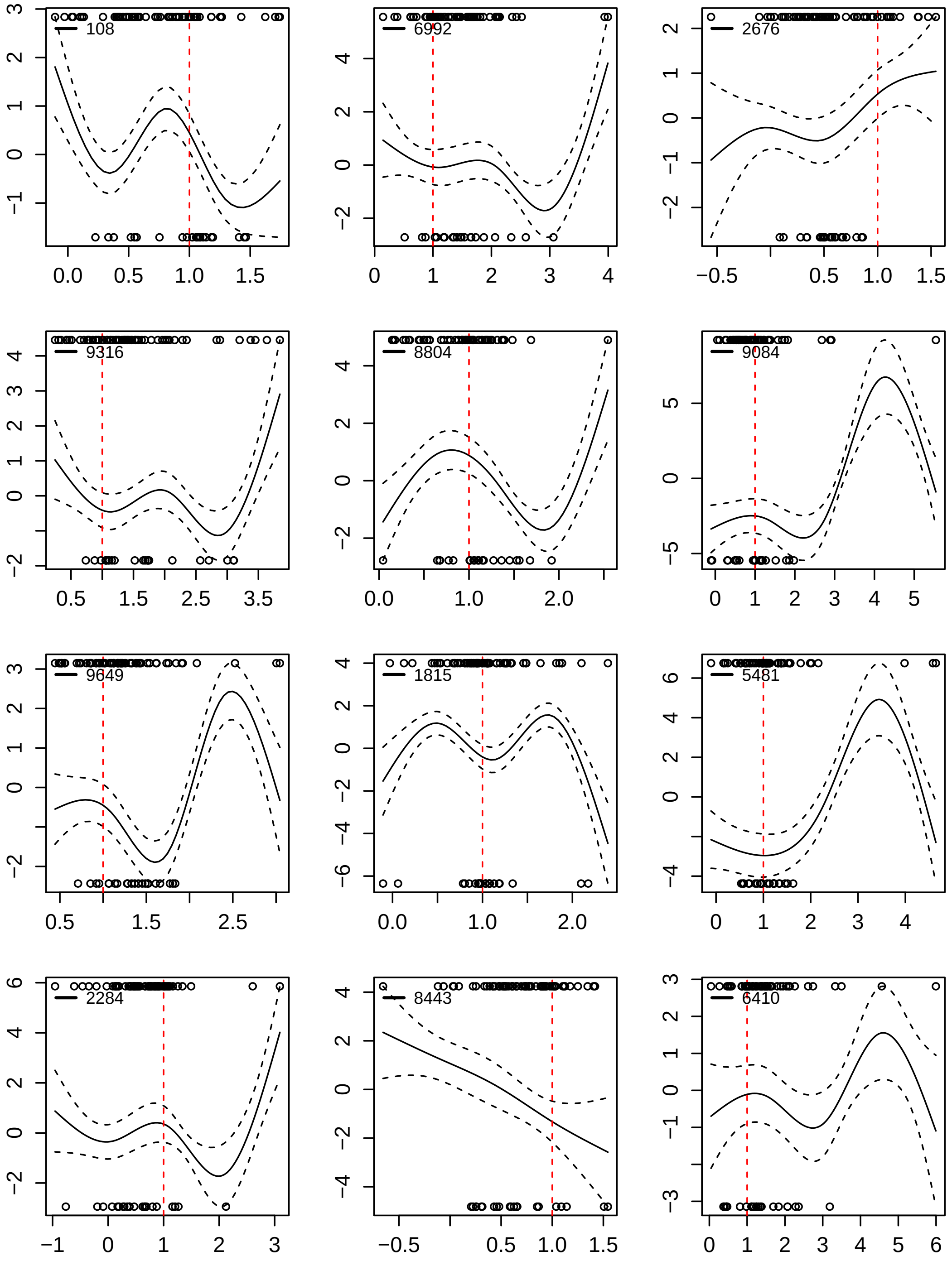


Figure 6.7: Plots of the estimated log hazard ratio for each of the significant window of the CNA. The solid black line is the estimated log hazard ratio, the dashed lines are the 95% point-wise confidence band. The points indicate the number of observations on each individual that were either censored or a failure.

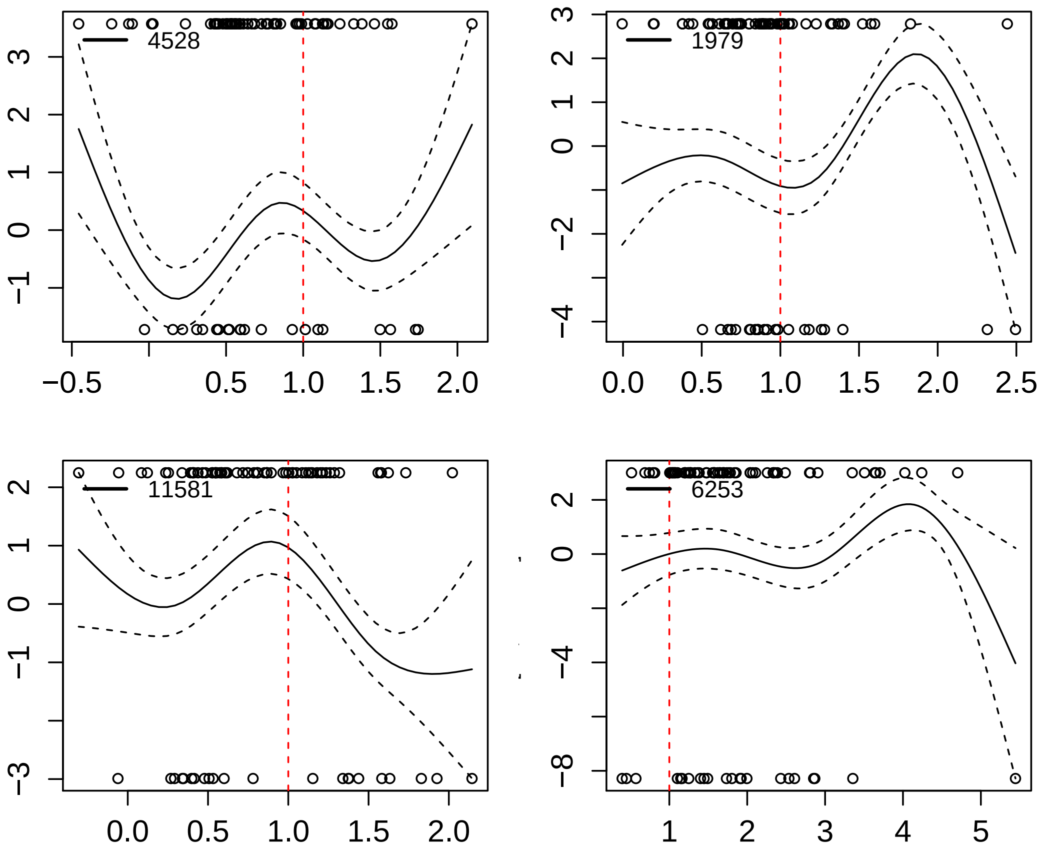


Figure 6.8: Plots of the estimated log hazard ratio for each of the significant window of the CNA. The solid black line is the estimated log hazard ratio, the dashed lines are the 95% point-wise confidence band. The points indicate the number of observations on each individual that were either censored or a failure.

Some of the plots of the estimated log hazard ratios for the significant CNA window showed a similar pattern, especially at the normal ploidy of one. However, these significant CNA genomic-windows that have similar patterns are not correlated, which means there are some windows in the genome have the similar effects on the survival model.

Eight plots of the estimated log hazard ratios for the significant CNA genomic-windows indicate the lower risk at ploidy one, which are $\hat{f}_4(\text{CNA}_{2292})$, $\hat{f}_9(\text{CNA}_{8125})$ $\hat{f}_{15}(\text{CNA}_{6528})$, $\hat{f}_{21}(\text{CNA}_{9508})$, $\hat{f}_{28}(\text{CNA}_{9316})$, $\hat{f}_{32}(\text{CNA}_{1815})$, $\hat{f}_{33}(\text{CNA}_{5481})$, and $\hat{f}_{38}(\text{CNA}_{1979})$. Twelve CNA genomic-windows indicate higher risk at ploidy one, which are shown in the plots of the estimated log hazard ratios of the following significant CNA windows, $\hat{f}_3(\text{CNA}_{8982})$, $\hat{f}_5(\text{CNA}_{10278})$,$\hat{f}_8(\text{CNA}_{10964})$, $\hat{f}_{10}(\text{CNA}_{11217})$, $\hat{f}_{14}(\text{CNA}_{9752})$, $\hat{f}_{20}(\text{CNA}_{4614})$, $\hat{f}_{22}(\text{CNA}_{12356})$, $\hat{f}_{25}(\text{CNA}_{108})$, $\hat{f}_{29}(\text{CNA}_{8804})$, $\hat{f}_{34}(\text{CNA}_{2284})$, $\hat{f}_{37}(\text{CNA}_{4528})$, and $\hat{f}_{39}(\text{CNA}_{11581})$. Approximately linear pattern of the estimated log hazard ratio can be seen in the estimated log hazard ratio for age, $\hat{f}_{12}(\text{CNA}_{5137})$, $\hat{f}_{23}(\text{CNA}_{3186})$. For testing for linearity of these, we perform the penalized score test, as described in Section 5.6.1, we reject the null hypothesis, so non of these significant CNA genomic windows have a linear log hazard ratio estimate, the p-value is 0.023.

The cumulative hazard of the Cox-Snell residuals from the model fitting (6.5) is plotted as part of the model diagnostics, which can be seen in Figure 6.9.

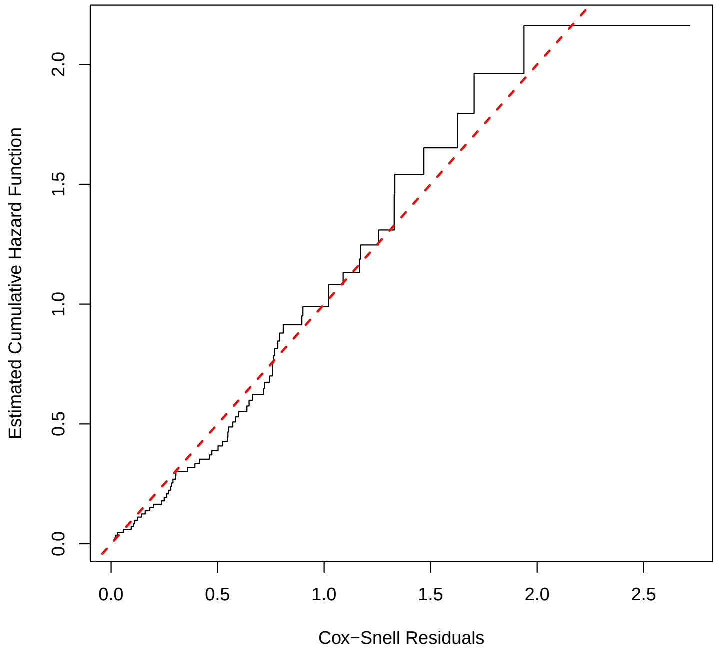


Figure 6.9: The solid black line is the cumulative hazard of Cox-Snell residual from the fitted the penalized additive Cox PH model (6.5), based on fitting the model with 40 significant windows from CNA and age as smoothing terms and the significant fixed effect from the clinical characteristics, comparing to the identity dashed red line.

To ensure that all the 40 significant CNA genomic-windows are important in the model, we try to add two more variable to evaluate the selecting of the forward selection process. Firstly we add one variable to the model equation (6.5), this variable is one of the non-significant CNA genomic-windows. The resulting plot of the estimated log hazard ratio for the non-significant CNA window is shown in Figure 6.10 (p-value$=0.310$). Secondly, we also add one variable to the model (6.5), which is not related to the survival time, let $y \sim N(0,1)$, we include this variable to the model in equation (6.5). This term is not significant p-value$=0.402$; the plot of the estimated log hazard ratio for y is shown in Figure 6.11.

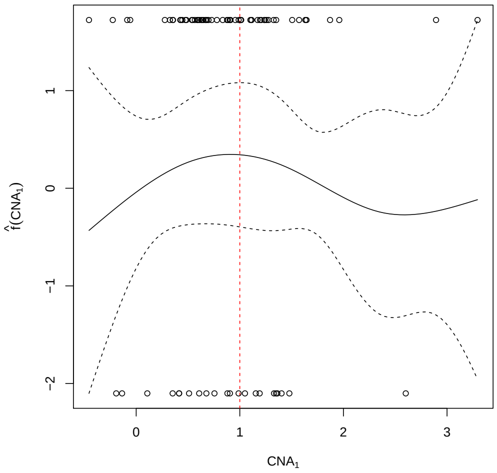


Figure 6.10: Plot of the estimated log hazard ratio for the non-significant window of the CNA.

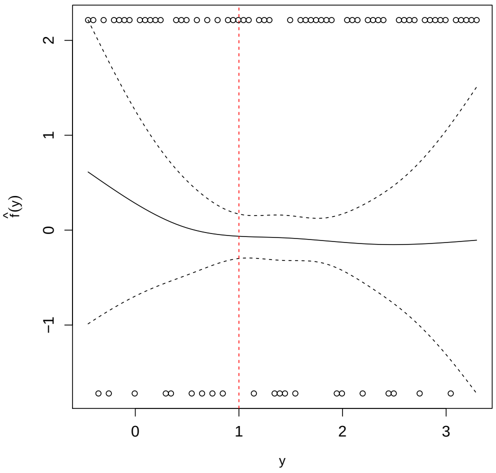


Figure 6.11: Plot of the estimated log hazard ratio for y.

6.4.1 Cluster Analysis

After estimating the log hazard ratios for each of the continuous variables in the model (6.5), cluster analysis can then be used to identify groups of significant windows of the CNA that have similar log hazard ratio shapes at normal ploidy 1. For example, the shape of the log hazard ratio of $\hat{f}_{32}(\text{CNA}_{1815})$ in Figure 6.7 is similar to the log hazard ratio $\hat{f}_{15}(\text{CNA}_{6528})$ in Figure 6.6, but the location of the normal ploidy one is different, so we can not consider them as a group. However, the shapes of the log hazard ratios $\hat{f}_{30}(\text{CNA}_{9084})$, $\hat{f}_{31}(\text{CNA}_{9649})$, and $\hat{f}_{33}(\text{CNA}_{5481})$ in Figure 6.7 are similar at normal ploidy one, although they are from different chromosomes (Chr11, Chr12, and Chr6 respectively), and from different regions, and different scales of the CNA genomic-windows.

Before using clustering analysis on our final multivariate additive Cox PH model that contain Stage-T, Stage-N, age and 40 significant windows of the CNA, we would like to fit the model with correlated significant windows of the CNA. The first 11 neighboring significant windows of the CNA was fit using 5 equally spaced knots, with optimal values of the smoothing parameters. The model is

$$h(t) = h_0(t) \exp\Big(StageN + StageT + \sum_{j=108}^{118} f_j(\mathrm{CNA}_j)\Big).$$

The first 11 neighboring significant CNA genomic-windows are from Chromosome 1, and they have the same scales of the CNA genomic-windows. The plots of the estimated log hazard ratios for the 11 neighboring significant CNA windows are presented in Figure 6.12. This can clearly be considered as one cluster.

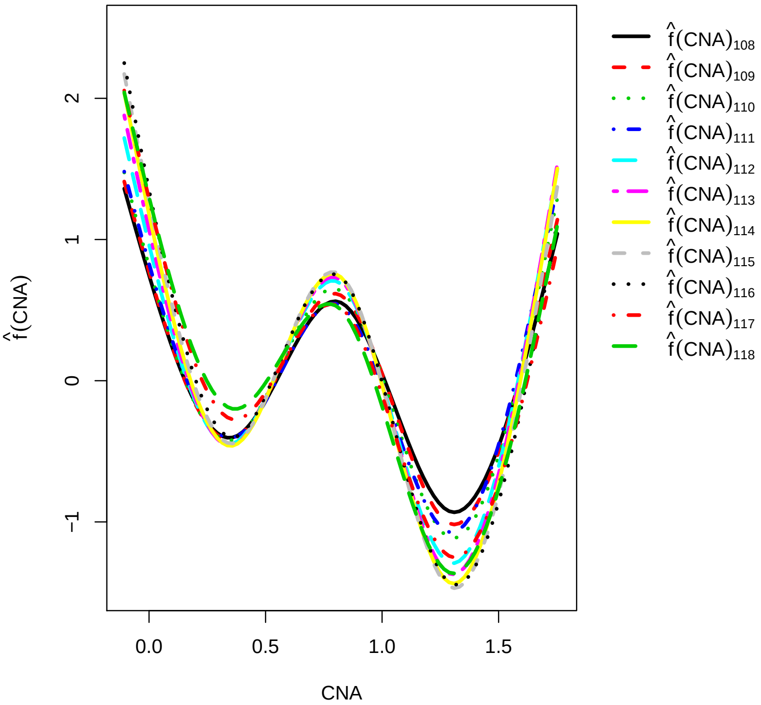


Figure 6.12: The plots of the estimated log hazard ratio of the first 11 significant neighboring windows CNA.

Hierarchical clustering is the most common clustering method. In this method, significant CNAs genomic-windows whose estimated log hazard ratio patterns are similar across patients can be gathered into one cluster, taking into account the location of the normal ploidy one. Each of the significant CNA genomic-window have a different scale of the observe significant CNA window, in order to cluster a group of the significant CNA genomic-windows that have a similar shape of the log hazard ratio for the significant CNA windows at ploidy one, we create selection points, and then we evaluate the log hazard ratio for each significant CNA window using some of these selection points to ensure that the estimated log hazard ratio share the similar shape at normal ploidy one.

The minimum and maximum values of all the significant CNA genomic-windows in model (6.5) are -2.255 and 11.031 respectively, these minimum and maximum values are used to create the selection points by 0.05, so the length of this selection points is 266 observations. Then we evaluate each of the estimate of the log hazard ratio at some but not all of the selection points and NA otherwise, this can be recorded as a row in matrix, where the number of rows is the number of the significant CNA genomic-windows, which is 40, and number of columns is length of the selection points, which is 266. Three different cluster method as used, the first method is that we compute the Euclidean distance to perform the complete linkage.

The second method, we would consider a maximum weighted difference. For two smoothing terms $\hat{f}_1(x_1)$ and $\hat{f}_2(x_2)$, which are observed at some but not all of the selection points, we define the distance measure

$$d_{12} = \max_i \frac{|\hat{f}_1(x_{1i}) - \hat{f}_2(x_{2i})|}{\sqrt{\mathrm{Var}\Big(f_1(x_{1i})_{ii}\Big) + \mathrm{Var}(f_2(x_{2i})_{ii}\Big)}},$$

where the maximum is over all i such that both $f(x_{1i})$ and $f(x_{2i})$ are not NA. The third method is described in Bozkus (2017). Complete linkage clustering is used for all the different methods, which is used as explanatory tool that we use to suggest group (in this case, pairs) of CNA windows which seems to have similar effects. We are interested to see whether the result of the clustering is meaningful and interpretable. The objective is to find any small cluster that have very similar shape of the estimated log hazard ratio of the significant CNA genomic-windows at normal ploidy 1. As a result, some of the clusters are not necessarily show a consistent similar shape estimated log hazard ratio of the significant CNA genomic-windows at normal ratio 1, while the other clustering shows a consistent similar shape estimated log hazard ratio at normal ratio 1 between the significant CNA genomic-windows. For example Figure 6.13 shows they are not necessarily similar to each other although they are found to be in the same cluster. while Figure 6.14 shows a similar estimated log hazaerd ratio of the significant CNA genomic-windows at normal ploidy 1 and they clustered together.

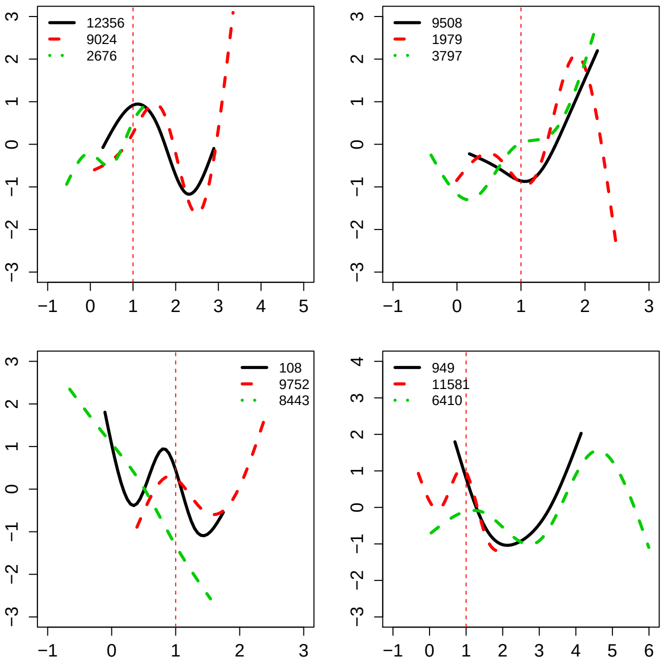


Figure 6.13: The clusters of the log hazard ratio of the significant CNA genomic-windows.

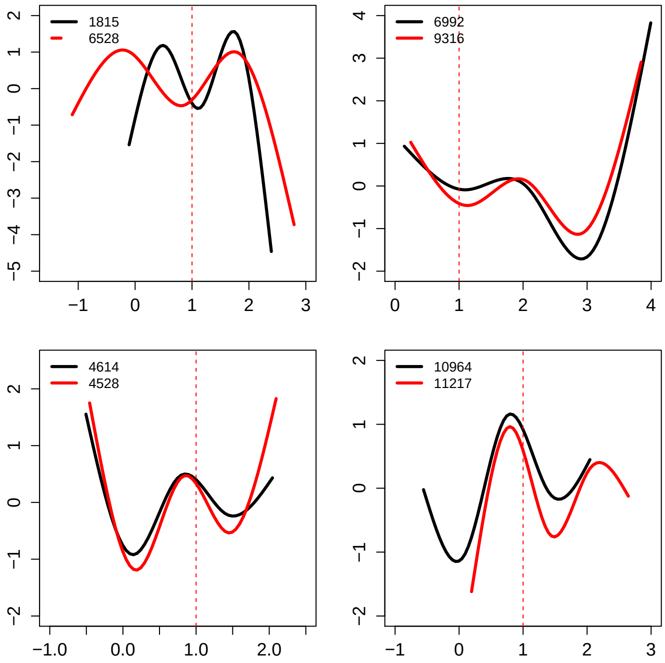


Figure 6.14: The clusters of the log hazard ratio of the significant CNA genomic-windows.

The common results from all three different types of clustering are presented in Figure 6.15. As a result, there are seven clusters, each cluster is a pairs of two estimated log hazard ratio that is have a very similar shapes of the estimated log hazard ratio at ploidy one. Table 6.12 shows the seven clusters with the corresponding chromosome for each significant CNA genomic-windows.

Cluster	# Window	Chr	Cluster	# Window	Chr
1	6992 9316	8 12	5	8804 11581	11 18
2	4614 4528	5 5	6	12356 10278	19 14
3	10964 11217	16 17	7	8125 9508	10 12
4	9024 2676	11 13			

Table 6.12: The seven clusters of the estimated log hazard ratio of the significant CNA genomic-windows.

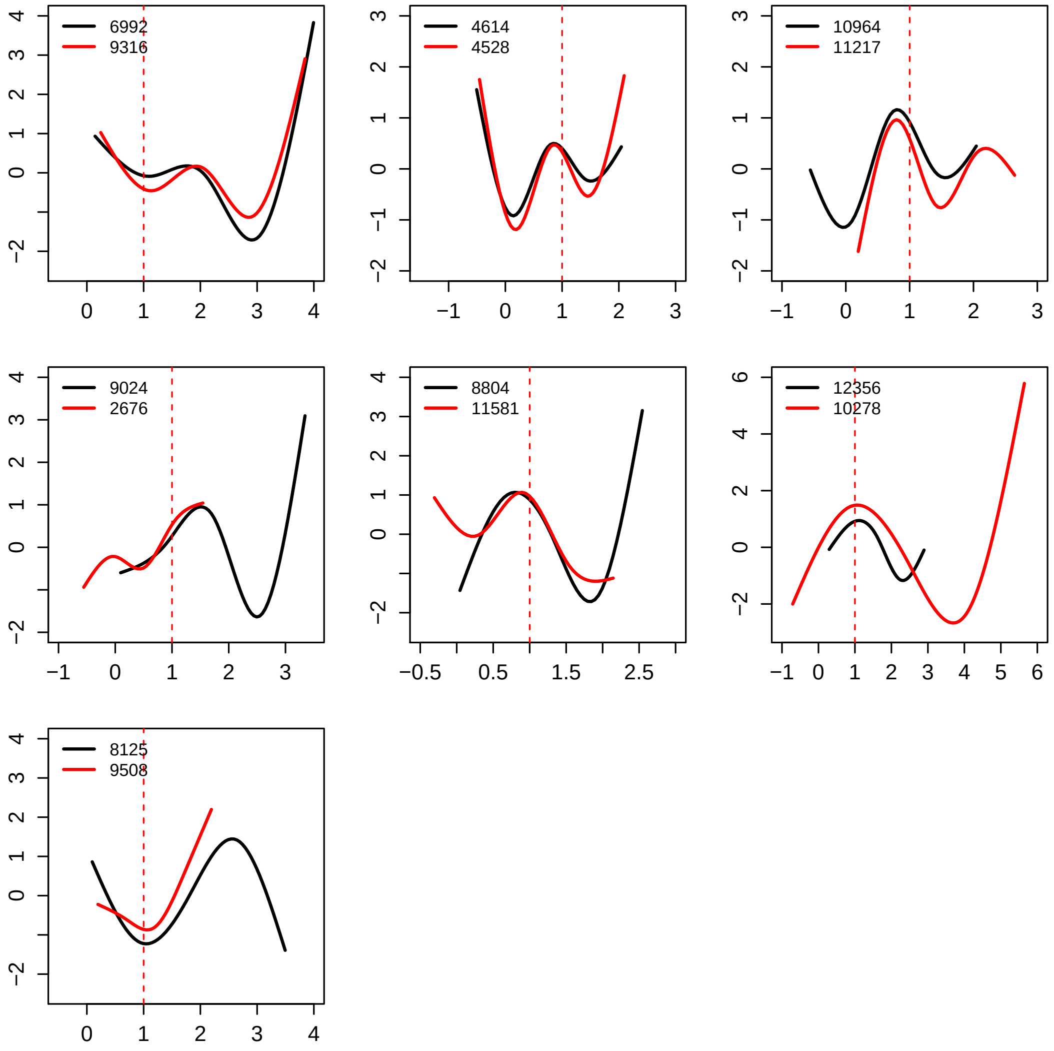


Figure 6.15: The seven clusters of the log hazard ratio of the significant CNA genomic-windows.

The estimated log hazard ratios for the significant CNA genomic-windows $\hat{f}_{26}(\mathrm{CNA}_{6992})$, $\hat{f}_{28}(\mathrm{CNA}_{9316})$, $\hat{f}_{9}(\mathrm{CNA}_{8125})$ and $\hat{f}_{21}(\mathrm{CNA}_{9508})$ indicate the lower risk at ploidy one, While $\hat{f}_{20}(\mathrm{CNA}_{4614})$, $\hat{f}_{37}(\mathrm{CNA}_{4528})$, $\hat{f}_{8}(\mathrm{CNA}_{10964})$, $\hat{f}_{10}(\mathrm{CNA}_{11217})$, $\hat{f}_{29}(\mathrm{CNA}_{8804})$, $\hat{f}_{39}(\mathrm{CNA}_{11581})$, $\hat{f}_{22}(\mathrm{CNA}_{12356})$, and $\hat{f}_{5}(\mathrm{CNA}_{10278})$ indicate the higher risk at ploidy one.

6.5 Conclusion

In this chapter, we consider the inclusion of genome-wide CNA profiles as a smoothing terms in the penalized additive Cox PH model, in addition to using the significant clinical characteristics as fixed predictors. The genome-wide CNA profile had 13,256 genomic-windows. Including them all in the model would run into computational difficulties. Therefore, we did a screening test using a generalized method of the univariate selection described in Bøvelstad et al. (2007) to determine which of the CNA profiles have the strongest effects on the survival time. This reduce the CNA profile to 1056 genomic-windows.

We have compared our penalized univariate variable selection method with Bøvelsted et al's univariate variable selection method. Our penalized univariate variable selection method identifies more significant CNA genomic-windows than their univariate variable selection method. The significant CNA genomic-windows identified some which overlap with genes associated with Non-Small Cell Lung Cancer (NSCLC). Table 6.1 shows that some of these genes are related to NSCLC as found in previous studies, also some of these genes have been found to be related to other types of cancer, in previous studies, and we found other genes with no prior studies displaying any link to any type of cancer.

We deal with the dependencies between the significant neighboring CNA genomic-windows by defining a block of correlated windows, after finding a value of correlation to use as a threshold. As a result we have a list of 41 significant CNA genomic-windows. The smoothing parameter λ, for each smoothing terms is obtained separately in the model using five-fold Cross-Validated partial log-likelihood. The results of the forward stepwise selection method enable us to assess the significance of the clinical characteristics as fixed predictors, and to identify the significant CNA genomic-windows, that exhibit higher or lower risk at normal ploidy one. Clustering techniques are used to express the similar log hazard ratio shapes of the significant CNA genomic-windows across patients.

Chapter 7

Shrinkage Penalty Variable Selection

7.1 Introduction

The method of variable selection by shrinkage is based on penalizing coefficients in the model, which leads to all of the spline coefficients for some variable being equal to zero. Earlier papers on shrinkage methods for variable selection and penalized coefficients in the standard Cox PH model have been written by Van Houwelingen and Verweij (1994); Tibshirani (1997); Fan and Li (2002); Segal (2006); Van Houwelingen et al. (2006); Zhang and Lu (2007). The performance of univariate variable selection method was compared with performance of shrinkage (using ridge and lasso regression), and of summary variable (using principal components regression (PCR), and supervised principal components regression) by Bøvelstad et al. (2007). They concluded that methods based on shrinkage, especially ridge regression, tend to perform better than univariate variable selection.

Benner et al. (2010) presented various regularization methods for fitting the standard Cox PH model in high-dimensional cases. These regularization methods were ridge, lasso, adaptive lasso, elastic net and SCAD; they recommend the use of lasso or elastic net in data applications. The main advantage of using the ridge penalty in the standard Cox PH model is to prevent degeneracy due to multi-collinearity of the covariates (Van Houwelingen and Putter, 2012), and using ridge penalty in the standard

Cox model, as it is excellent at handling correlated predictors, is suggested by Simon et al. (2011). In the case of the penalized additive Cox PH model, the lasso and ridge penalties are not the right choice to use as the smoothing terms contain several spline coefficents, so it is not practical to remove one spline coefficents out of the set of spline coefficients for a given variable. We need to create a penalty that can shrink all spline coefficients for one variable toward zero.

There is a lack of literature on methods of variable selection in the additive Cox PH model. However, in the GAM setting Marra and Wood (2011) presented two effective shrinkage methods and an extension of the nonnegative garrote estimator. The non-negative garrote component selection of Breiman (1995) is used as an approach for variable selection in Marra and Wood (2011). This method is based on estimating the model parameters by ordinary last square (OLS) and subsequently shrinking the parameters by non-negative factors. However, in order to obtain the OLS estimator, the number of covariates has to be smaller than the number of observations, therefore this method can not be used in high-dimensional data. In this chapter we present a shrinkage methods for variable selection in the penalized additive Cox PH model based on the shrinkage methods in Marra and Wood (2011).

This chapter is organized as follows. In Section 7.2, we present two shrinkage approaches, the double penalty approach in Section 7.2.1, and shrinkage approach in Section 7.2.2. A simulation study to assess the shrinkage approach is discussed in Section 7.3. Selecting the optimal value of the regularization parameters is presented in Section 7.4. The results and evaluation of the significant CNA genomic-windows using the shrinkage approach are found in Section 7.6. A comparison between the forward variable selection, which is discussed in Chapter 6, and shrinkage method is presented in Section 7.7.

7.2 Shrinkage Method

Discrete feature selection methods may not capture well the joint effect of the significant CNA genomic windows. To overcome this problem, various regularization methods can be used, which aim to maximize a penalized partial log likelihood with a penalty accounting for the model shrinkage. A penalized partial log-likelihood is expressed as

$$\ell_{\text{pen}}(\boldsymbol{\beta}_\lambda) = \ell_{pl}(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j}^{p}\lambda_j \int [f_j''(x_j)]^2 dx_j, \tag{7.1}$$

where $\ell_{pl}(\boldsymbol{\beta})$ is the partial log likelihood, and $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the smoothing parameters. The derivation of an integrated square second derivative penalty is described in Section 4.3.1. The j^{th} penalty term can be expressed as

$$\int [f_j''(x_j)]^2 dx_j = \boldsymbol{\beta}_{zj}^T \boldsymbol{Z}_j^T \boldsymbol{K}_j \boldsymbol{Z}_j \boldsymbol{\beta}_{zj},$$

where $\boldsymbol{Z}_j$ is a semi-orthogonal matrix of size $(n_k+1)\times(n_k-1)$, and n_k is the number of the knots for the smoothing term, $\boldsymbol{K}_j$ is a squared matrix of the j^{th} smoothing term of size $(n_k+1)\times(n_k+1)$, with (e,f) element being $12|x_{je}^* - x_{jf}^*|^3$, and the first row and column is equal to zero. A penalized partial log-likelihood is expressed as

$$\ell_{\text{pen}}(\boldsymbol{\beta}_\lambda) = \ell_{pl}(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j=1}^{p}\lambda_j \boldsymbol{\beta}_{zj}^T \boldsymbol{Z}_j^T \boldsymbol{K}_j \boldsymbol{Z}_j \boldsymbol{\beta}_{zj}, \tag{7.2}$$

As mentioned earlier in Chapter 5, the trade-off between the goodness of fit and smoothness of the estimated model parameters is driven by the penalty term, which has the form $\lambda_j \boldsymbol{\beta}_{zj}^T \boldsymbol{Z}_j^T \boldsymbol{K}_j \boldsymbol{Z}_j \boldsymbol{\beta}_{zj}$. However, even if λ_j goes to infinity, there is no guarantee that any j^{th} smoothing term is completely removed from the model. This removal of the j^{th} variable from the model corresponds to estimating the corresponding vector of the spline coefficients, $\boldsymbol{\beta}_{zj}$, as $\mathbf{0}$.

To shrink the whole smoothing term to zero, and perform automatic model selection for the continuous covariates based on the values of the smoothing parameter, we need to include a second penalty in the penalized partial log likelihood (7.2). This idea was proposed in Marra and Wood (2011) for the GAM setting. The general idea is that the penalty term can be decomposed into the sum of two components. The first component is associated with the function in the penalty null space, and the second component is associated with the function in the range space. As a result, the smoothing penalty shrinks functions in the range space to zero if the smoothing parameter is big enough without shrinkage the penalty null space, to shrink the whole smoothing term to zero we need to penalized the penalty null space. However, the parametric components in the penalized additive Cox PH model are not affected by this shrinkage selection process. As a result in this method the estimation of the model coefficients and variable selection can be carried out simultaneously.

7.2.1 Double Penalty Approach

The smoothing penalty matrix for the j^{th} smoothing term is $\boldsymbol{Z}_j^T \boldsymbol{K}_j \boldsymbol{Z}_j$, which is not a full rank matrix of size $(n_k - 1) \times (n_k - 1)$. The eigen decomposition of $\boldsymbol{Z}_j^T \boldsymbol{K}_j \boldsymbol{Z}_j$ can be written as,

$$\boldsymbol{Z_j^T K_j Z_j} = \boldsymbol{\Gamma_j \Lambda_j \Gamma_j^T}, \tag{7.3}$$

where $\boldsymbol{\Gamma}_j$ is $(n_k - 1) \times (n_k - 1)$ matrix of eigenvectors, and $\boldsymbol{\Lambda}_j$ is a diagonal matrix of eigenvalues, which contains one zero eigenvalue due to the radial basis penalty null space. The second penalty can be formed as follows

$$\boldsymbol{K}_j^* = \boldsymbol{\Gamma}_j^* \boldsymbol{\Gamma}_j^{*T}, \tag{7.4}$$

where $\boldsymbol{\Gamma}_j^*$ is a matrix where columns are the eigenvectors corresponding to the zero eigenvalues of $\boldsymbol{\Lambda}_j$, and $\boldsymbol{K}_j^*$ is the second penalty matrix of size $(n_k - 1) \times (n_k - 1)$ for the j^{th} smoothing term, where the diagonal of $\boldsymbol{K}_j^*$ is the square elements of Γ_j^{*T}

column that corresponding to the zero eigenvalues of $\mathbf{\Lambda}_j$.

The double penalized additive Cox PH model can be expressed as

$$\ell_{\text{dp}}(\boldsymbol{\beta}_{\boldsymbol{\lambda},\boldsymbol{\lambda}^*}) = \ell_{pl}(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_{zj}^T \boldsymbol{Z}_j^T \boldsymbol{K}_j \boldsymbol{Z}_j \boldsymbol{\beta}_{zj} - \frac{1}{2}\sum_{j=1}^{p} \lambda_j^* \boldsymbol{\beta}_{zj}^T \boldsymbol{K}_j^* \boldsymbol{\beta}_{zj}. \tag{7.5}$$

The second term in (7.5) penalizes only smoothing components in the range space, and hence can shrink these to zero, while the third term in (7.5) penalizes only smoothing components in the null space, and can shrink these to zero. Consequently, the second term would penalize (towards zero) smoothing components representing departure from straight line behavior, while the third term in (7.5) would penalize straight line components to zero.

There are two smoothing parameters in the proposed double penalty procedure. The first smoothing parameters λ is associated with the roughness penalty and the regularization parameters λ^* is associated with second penalty. The smoothing parameter $\lambda > 0$ balances smoothness of $f(x)$, and $\lambda^* > 0$ is a regularization parameters controlling the amount of shrinkage used in the variable selection. However, maximizing (7.5) is not practical as fitting each smoothing term requires to estimation of both λ and λ^*.

7.2.2 Shrinkage Approach

Marra and Wood (2011) proposed a shrinkage approach as an alternative to the double penalty method to avoid doubling the number of smoothing parameters that are associated with each smoothing term. The shrinkage approach replaces the smoothing penalty $\boldsymbol{Z}_j^T \boldsymbol{K}_j \boldsymbol{Z}_j$ by $\tilde{\boldsymbol{K}}_j$, where

$$\tilde{\boldsymbol{K}}_j = \mathbf{\Gamma}_j \tilde{\mathbf{\Lambda}}_j \mathbf{\Gamma}_j^T. \tag{7.6}$$

Here, $\tilde{\mathbf{\Lambda}}_j$ is similar to $\mathbf{\Lambda}_j$ but we replace the zero eigenvalue by ϵ, where ϵ is a small value in proportion to the smallest positive eigenvalue of $\boldsymbol{Z}_j^T \boldsymbol{K}_j \boldsymbol{Z}_j$. Hence,

$$\begin{aligned}
\tilde{\lambda}_j \boldsymbol{\beta}_{zj}^T \tilde{\boldsymbol{K}}_j \boldsymbol{\beta}_{zj} &= \tilde{\lambda}_j \boldsymbol{\beta}_{zj}^T \boldsymbol{\Gamma}_j \tilde{\Lambda}_j \boldsymbol{\Gamma}_j^T \boldsymbol{\beta}_{zj} \\
&= \tilde{\lambda}_j \boldsymbol{\beta}_{zj}^T \boldsymbol{\Gamma}_j \begin{bmatrix} l_1 & 0 & 0 & \dots & 0 \\ 0 & l_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \boldsymbol{\Gamma}_j^T \boldsymbol{\beta}_{zj} + \tilde{\lambda}_j \boldsymbol{\beta}_{zj}^T \boldsymbol{\Gamma}_j \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \epsilon \end{bmatrix} \boldsymbol{\Gamma}_j^T \boldsymbol{\beta}_{zj} \\
&= \tilde{\lambda}_j \boldsymbol{\beta}_{zj}^T \boldsymbol{\Gamma}_j \boldsymbol{\Lambda}_j \boldsymbol{\Gamma}_j^T \boldsymbol{\beta}_{zj} + \tilde{\lambda}_j \boldsymbol{\beta}_{zj}^T \epsilon \boldsymbol{\Gamma}_j^* \boldsymbol{\Gamma}_j^{*T} \boldsymbol{\beta}_{zj} \\
&= \tilde{\lambda}_j \boldsymbol{\beta}_{zj}^T \underbrace{\left(\boldsymbol{Z_j K_j Z_j^T} + \epsilon \boldsymbol{\Gamma}_j^* \boldsymbol{\Gamma}_j^{*T} \right)}_{\tilde{\boldsymbol{K}}_j} \boldsymbol{\beta}_{zj},
\end{aligned}$$

where $\boldsymbol{\Gamma}_j^*$ is a matrix where columns are the eigenvectors corresponding to the zero eigenvalues of $\boldsymbol{\Lambda}_j$. This is equivalent to fixing $\lambda_j^* = \epsilon \tilde{\lambda}_j$, so each smoothing term has one smoothing parameter that needs to be estimated. The penalized partial log-likelihood is expressed as

$$\begin{aligned}
\ell_{shrink}(\boldsymbol{\beta}_{\tilde{\lambda}}) &= \ell_{pl}(\boldsymbol{\beta}) - \frac{1}{2} \sum_{j=1}^{p} \tilde{\lambda}_j \boldsymbol{\beta}_{zj}^T \tilde{\boldsymbol{K}}_j \boldsymbol{\beta}_{zj} \\
&= \ell_{pl}(\boldsymbol{\beta}) - \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{\mathcal{K}}_{shrink} \boldsymbol{\beta} \\
&= \sum_{i=1}^{n} \delta_i [\boldsymbol{X}_i \boldsymbol{\beta}] - \sum_{i=1}^{n} \delta_i \log \left(\sum_{j \in R(t_i)} \exp[\boldsymbol{X}_j \boldsymbol{\beta}] \right) - \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{\mathcal{K}}_{shrink} \boldsymbol{\beta}.
\end{aligned}$$

where $\boldsymbol{\mathcal{K}}_{shrink}$ is the block diagonal matrix for the shrinkage penalty matrix for each smoothing term in the model, $\boldsymbol{\mathcal{K}}_{shrink} = \text{diag}(\boldsymbol{0}, \tilde{\lambda}_1 \tilde{\boldsymbol{K}}_1, \dots, \tilde{\lambda}_p \tilde{\boldsymbol{K}}_p)$. For given values of smoothing parameters $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_p$, the penalized estimate for $\boldsymbol{\beta}_{\tilde{\lambda}}$ can be obtained by using the Newton-Raphson algorithm.

7.3 Simulation Study

In Section 5.5.2 we presented a simulation study to assess the performance of the unpenalized additive Cox PH model with only one smoothing term. This simulation study is used here to assess the effect of our shrinkage approach.

7.3.1 Simulation Setting

As in Section 5.5.2, for sample size $n = 200$, we define $x_i = \frac{i}{n} \times 2\pi$ for $i = 1, \ldots, n$ as covariate vector, and the true smooth function is $f(x_i) = \sin(x_i)$. The algorithm in Section 5.5.1 is used to generated survival times from the additive Cox PH model with constant baseline hazard $\lambda_{EXP} = 1$. This generated data are such that approximately 17.63% of the observations were censored. The eigenvalues of the penalty matrix are 603.613, 73.304, 19.473, and 0. In the following scenarios, the simulated data are modeled with 5 equally spaced knots.

1. Scenario 1: Fitting the penalized additive Cox PH model where ϵ is fixed, and the smoothing parameter λ increases.

2. Scenario 2: Fitting the penalized additive Cox PH model where λ is fixed, and ϵ increases.

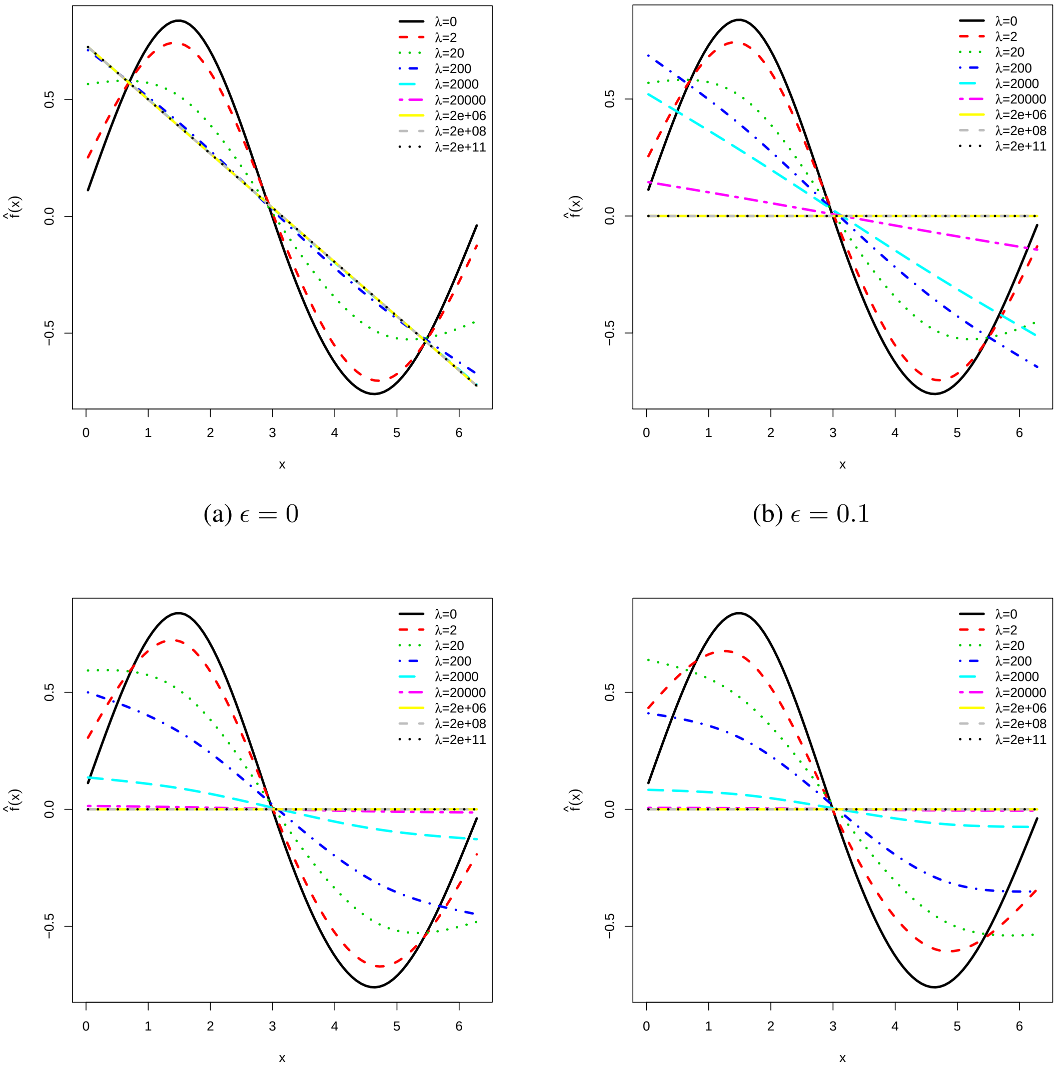
3. Scenario 3: Fitting the penalized additive Cox PH model where λ increases, and ϵ decreases to maintain a constant $\lambda^* = \lambda\epsilon$.

7.3.2 Simulation Results

To evaluate the impact of λ in the shrinkage approach, we fixed the value ϵ, then we fit the shrinkage penalized additive Cox PH model with 5 equally spaced knots. For zero values of ϵ, the un-shrunk penalty matrix is used, so as λ increases the estimated log hazard ratio shrinks from the penalized curve to a straight line only. This straight line is not shrunk to zero which can be seen in Figure 7.1 (a), where the solid black line is the unpenalized additive Cox PH model, and the dashed lines are the estimated log hazard ratio for different values of λ. The values of the smoothing parameters are shown in the legend. However, when $\epsilon > 0$, the estimate of the log hazard ratio shrinks to a straight line and then to zero as the value of the smoothing parameter λ increases. On the other hand, the large value of ϵ will lead to the faster shrinks toward zero which can be seen in Figure 7.1 (b), (c), and (d). Different values of ϵ help λ to shrink the log hazard curve to a straight line and then to zero, this is raising the question which value of ϵ do we choose? The optimal value of ϵ is discussed in Section 7.4.

To evaluate the impact of ϵ, we fixed the smoothing parameter λ and we change ϵ. For the small value of the smoothing parameter λ, the estimated log hazard ratio is not shrunk to a straight line, even if the value of ϵ increases, this can be seen in Figure 7.2 (a). The solid black line is the unpenalized additive Cox PH model, the dashed lines are the estimated log hazard ratio for different values of ϵ. The values of the ϵ are shown in the legend. However, for large values of the smoothing parameter λ and large value of ϵ, the estimated log hazard ratio is shrunk fast to a straight line and then to zero, which can be seen in Figure 7.2 (b), (c), and (d). We can see that, the value of the smoothing parameter plays an important role for the shrinkage approach.

In order to evaluate the impact of both λ and ϵ, we fitted the penalized additive Cox PH model with $\lambda\epsilon$ is fixed but varying λ and ϵ. Figure 7.3 shows the impact of changing λ and ϵ, in this figure the solid black line is the unpenalized additive Cox PH model, the dashed lines are the estimated log hazard ratio for different values of λ and ϵ. The values of the λ and ϵ are shown in the legend. This is confirms that large values of λ shrunk the estimated log hazard ratio to zero, even if ϵ is very small.



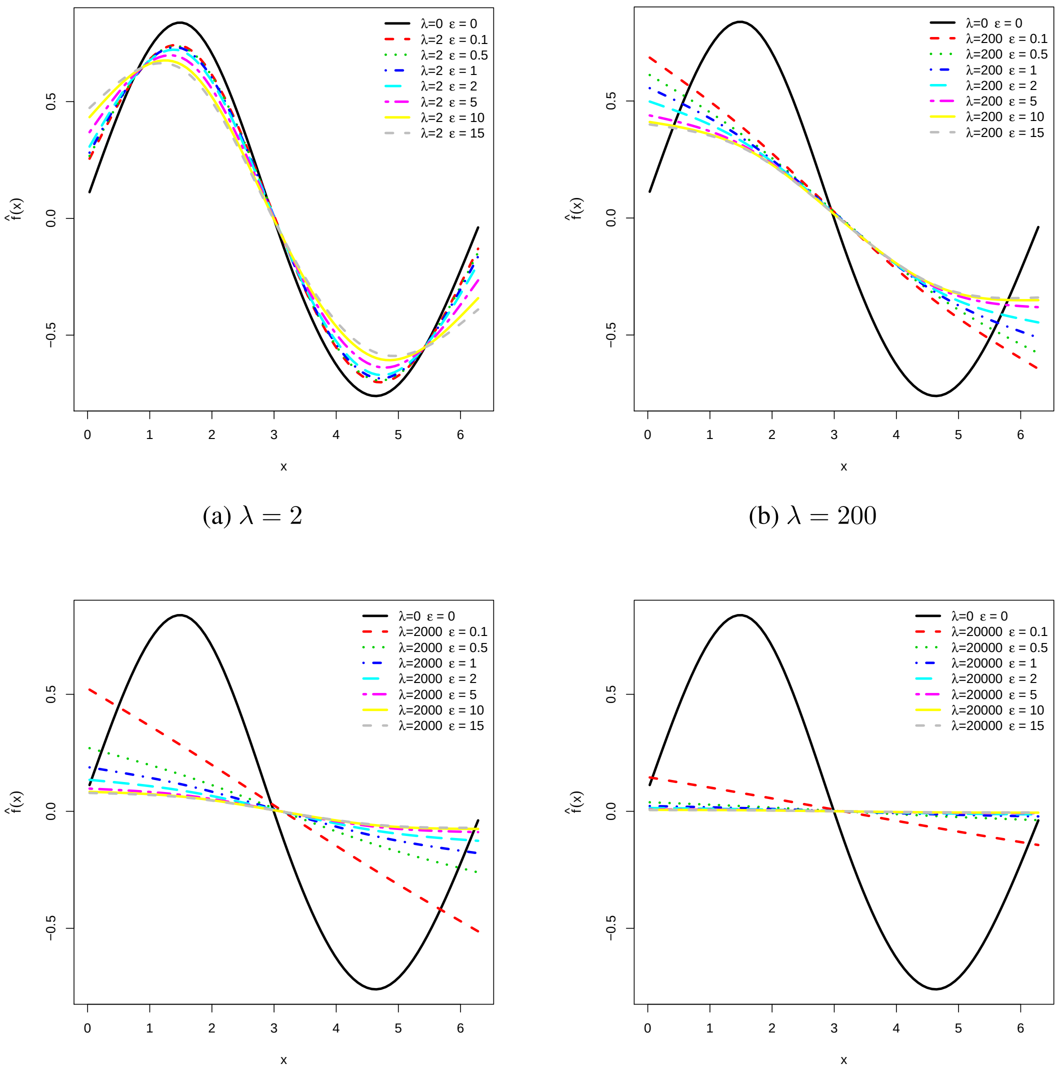
(a) $\epsilon = 0$

(b) $\epsilon = 0.1$

(c) $\epsilon = 1.91$

(d) $\epsilon = 10$

Figure 7.1: Smoothing function estimates obtained applying the shrinkage approach for fixing ϵ and increasing λ.



(a) $\lambda = 2$

(b) $\lambda = 200$

(c) $\lambda = 2000$

(d) $\lambda = 20000$

Figure 7.2: Smoothing function estimates obtained applying the shrinkage approach for fixing λ and increasing ϵ.

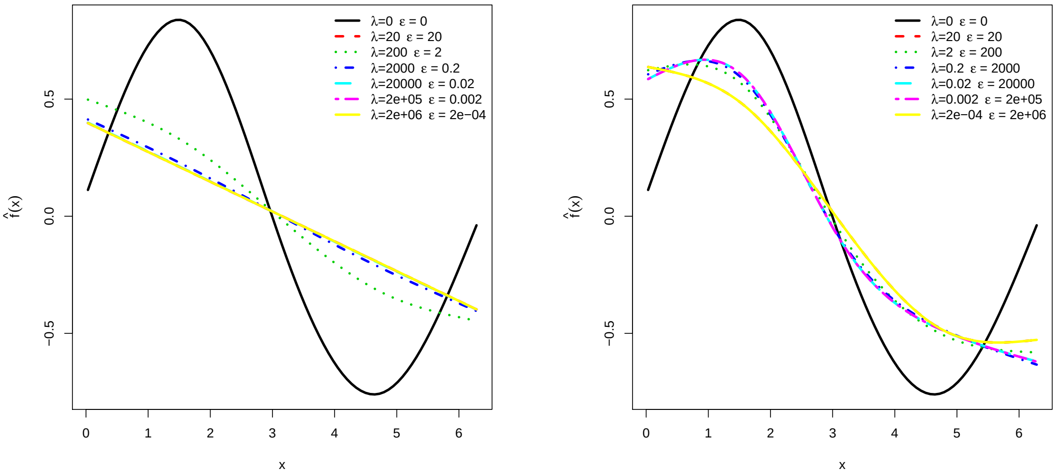


Figure 7.3: Smoothing function estimates obtained applying the shrinkage approach for increasing λ and decreasing ϵ and vise versus in the shrinkage approach.

7.4 Choosing the Optimal Value of ϵ

Marra and Wood (2011) suggest choosing ϵ to be a small relative to the smallest positive eigenvalue of the penalty matrix $\boldsymbol{Z}_j^T \boldsymbol{K}_j \boldsymbol{Z}_j$ to ensure that

$$\boldsymbol{\beta}_{zj}^T \boldsymbol{Z}_j^T \boldsymbol{K}_j \boldsymbol{Z}_j \boldsymbol{\beta}_{zj} \approx \boldsymbol{\beta}_{zj}^T \tilde{\boldsymbol{K}}_j \boldsymbol{\beta}_{zj}.$$

This means we want the value to be small enough that extra penalization only has effect when the penalty is already penalizing strongly toward the null space, but large enough that the value of the penalty does not plateau for some range of the smoothing parameter values before the null space penalization takes effect. To find the optimal value of ϵ, we conducted the following simulation study.

7.4.1 Simulation Study

We conducted a simulation study to find the optimal values of ϵ in the shrinkage approach. In Section 7.3 we evaluated the impact of λ, ϵ, or both λ and ϵ. In this section we used this simulation study again, aiming to find the optimal values of smoothing parameter λ and ϵ that maximize the CVPL (5.24). Figure 7.4 shows the five-fold CVPL values for different values of λ and ϵ. The optimal value of λ, ϵ and the corresponding CVPL's are $\lambda_{opt} = 2$, $\epsilon_{opt} = 0$, and CVPL$(\lambda_{opt}, \epsilon_{opt}) = -452.842$. The zero value of ϵ means there is no linear relationship between the estimated log hazard ratio and the covariate.

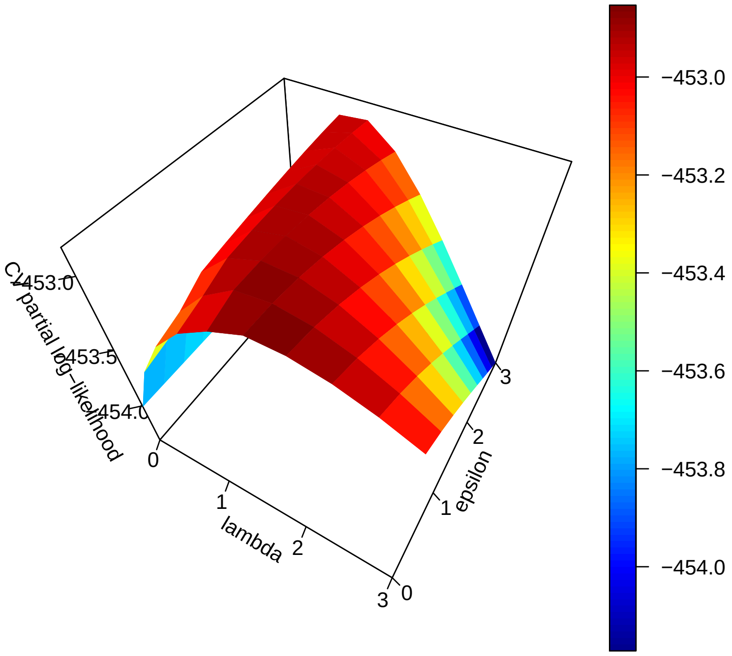
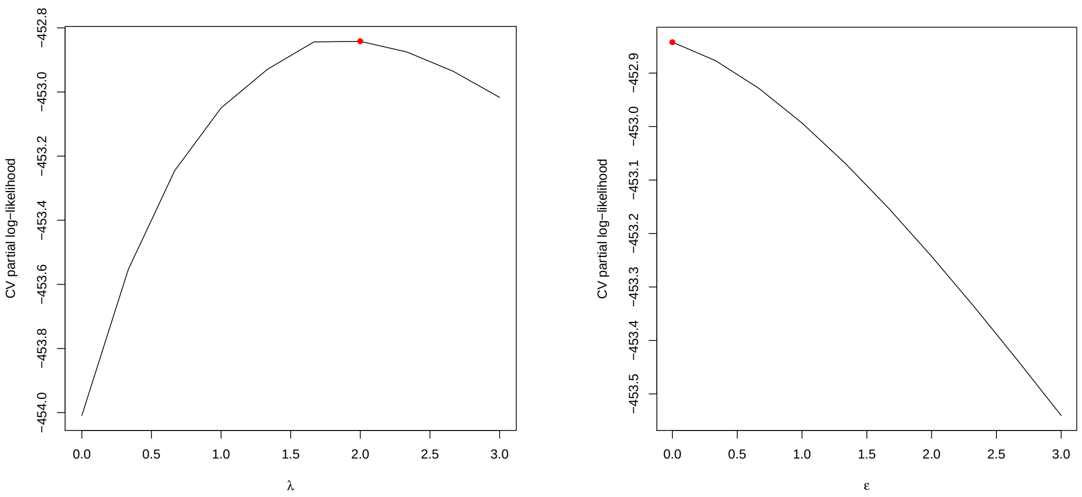


Figure 7.4: Five-fold CVPL for different values of λ and ϵ.

The left panel in Figure 7.5 shows the five-fold CVPL for different values of λ at the optimal value of $\epsilon_{opt} = 0$. The red points represent the optimal smoothing parameter λ_{opt}. The right panel in Figure 7.5 illustrates the five-fold CVPL for different values of ϵ at the optimal value of $\lambda_{opt} = 2$. The red points represent the optimal values of ϵ_{opt}.



(a) $\epsilon_{opt} = 0$

(b) $\lambda_{opt} = 2$

Figure 7.5: (a) Five-fold CVPL for different values of λ at the optimal value of $\epsilon_{opt} = 0$. (b) Five-fold CVPL for different values of ϵ at the optimal value of $\lambda_{opt} = 2$.

The plots of the different estimated log hazard ratio at the optimal value of the smoothing parameter $\lambda_{opt} = 2$, with different values of ϵ are presented in Figure 7.6. The solid black line is the unpenalized additive Cox PH model, the dashed lines are the estimated log hazard ratio for different values of ϵ. The values of ϵ are shown in the legend.

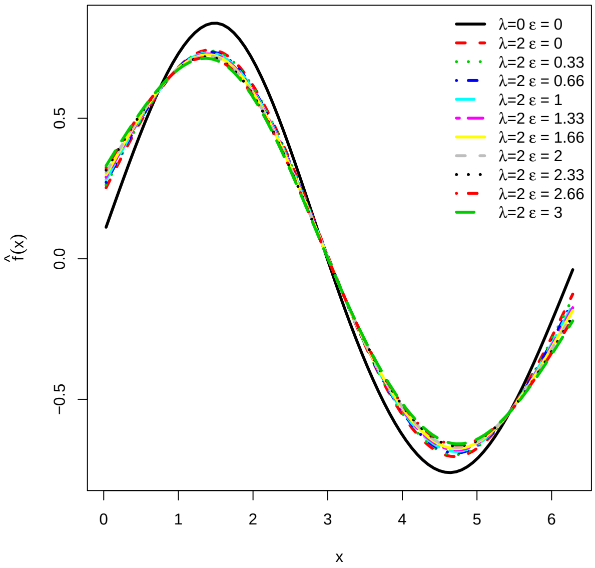


Figure 7.6: Plots of the estimated log hazard ratio at optimal values of $\lambda_{opt} = 2$ and different values of ϵ.

7.4.2 Simulation Study Using Real Data

In this section we used the significant CNA genomic-windows data in order to find the optimal value of ϵ. The results of obtaining the optimal value of ϵ for each univariate penalized additive Cox PH model that includes one significant CNA genomic-windows are summarized in Table 7.1. As result, the optimal values of ϵ is approximately equal to the 10% of the smallest non zero eigenvalues. The optimal values of λ and ϵ for univariate penalized additive Cox PH model for Age are zeros.

#	Chr	Variable	smallest non-zero eigenvalues	λ_{opt}	ϵ_{opt}
1	Chr 1	108	0.511	0.010	0.048
2	Chr 1	949	0.324	0.655	0.030
3	Chr2	1815	1.270	0.020	0.112
4	Chr2	1979	1.331	0.081	0.118
5	Chr2	2284	5.637	10.483	0.553
6	Chr3	2291	7.878	0.081	0.675
7	Chr3	2676	0.792	0.010	0.070
8	Chr3	3094	13.530	0.020	1.202
9	Chr3	3186	97.280	1.310	8.666
10	Chr4	3474	0.906	0.020	0.089
11	Chr4	3797	1.402	0.020	0.124
12	Chr4	4001	1.611	0.020	0.143
13	Chr5	4528	1.404	10.485	0.124
14	Chr5	4614	0.891	1.310	0.079
15	Chr6	5137	1.794	167.772	0.159
16	Chr6	5481	8.782	0.020	0.878
17	Chr7	6253	10.560	0.081	1.056
18	Chr7	6410	17.500	0.010	1.598
19	Chr8	6528	10.110	167.772	1.010
20	Chr8	6992	4.661	83.030	0.451
21	Chr8	7264	18.817	83.886	1.587
22	Chr10	8125	3.333	1.310	0.285
23	Chr10	8443	0.926	0.655	0.091
24	Chr11	8804	1.298	20.971	0.126
25	Chr11	8928	6.059	0.327	0.591
26	Chr11	9024	2.851	0.655	0.244
27	Chr11	9084	15.062	0.327	1.506
28	Chr11	9143	3.384	0.655	0.298
29	Chr12	9316	3.916	0.163	0.335
30	Chr12	9508	0.659	0.081	0.059
31	Chr12	9579	0.976	0.081	0.086
32	Chr12	9649	1.463	0.000	0.139
33	Chr12	9686	0.893	1.310	0.076
34	Chr13	9752	0.706	0.163	0.069
35	Chr13	10231	7.205	0.163	0.710
36	Chr14	10278	20.676	0.163	2.045
37	Chr15	10418	9.594	0.081	0.899
38	Chr16	10964	1.439	0.020	0.123
39	Chr17	11217	1.229	0.327	0.105
40	Chr18	11581	1.219	0.163	0.117
41	Chr19	12356	1.463	0.655	0.125

Table 7.1: The optimal values of smoothing parameters λ, and ϵ for each univariate penalized additive Cox PH model.

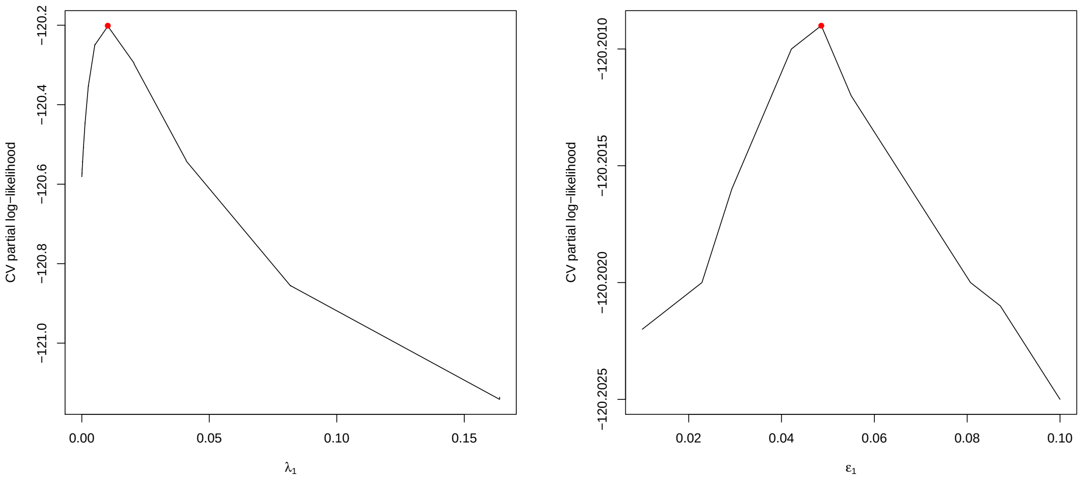
The result of only one univariate penalized additive Cox PH model will be presented as an illustrative example.

One Smoothing Term in the Model

The model is

$$h(t) = h_0(t) \exp\Big(f_1(\mathrm{CNA}_{108}))\Big). \tag{7.7}$$

The eigenvalues of the penalty matrix are 15.867, 1.927, 0.511, and 0, We choose a range values of ϵ_1 to be between zero and 20% of the smallest non-zero eigenvalue. Two-dimensional grid searching is carried out to find the optimal values of λ_1 and ϵ_1 that maximize CVPL, which are $\lambda_{1,opt} = 0.010$, $\epsilon_{1,opt} = 0.048$, and the corresponding CVPL is -120.200. We note that, in this case the optimal value of ϵ_1 is approximately 10% of the smallest non-zero eigenvalue of the penalty matrix ($0.048/0.511 = 0.09$). The left panel in Figure 7.7 shows the five-fold CVPL for different values of λ_1, the red point represents the optimal smoothing parameter $\lambda_{1,opt}$. The right panel in Figure 7.7 illustrates the five-fold CVPL for different values of ϵ_1 at the optimal value of $\lambda_{1,opt} = 0.0102$. The red point represents the optimal values of $\epsilon_{1,opt}$. The magnitude of change in CVPL due to changing ϵ_1 is very small, as it can be seen in Figure 7.7(b). The right panel of Figure 7.8 shows CVPL for different values of λ_1 and ϵ_1.



(a) $\epsilon_{1,opt} = 0.048$

(b) $\lambda_{1,opt} = 0.010$

Figure 7.7: (a) Five-fold CVPL for different values of λ_1 at the optimal value pf $\epsilon_{1,opt}$. (b) Five-fold CVPL for different values of ϵ_1 at the optimal value of $\lambda_{1,opt} = 0.0102$.

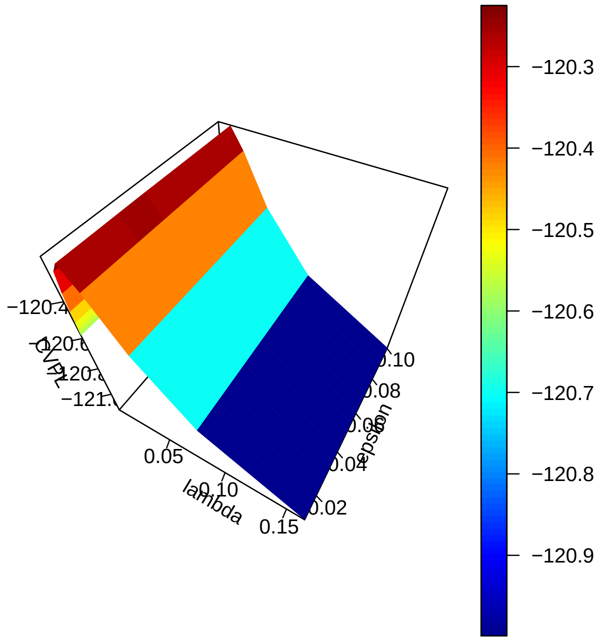


Figure 7.8: Five-fold CVPL for different values of λ and ϵ.

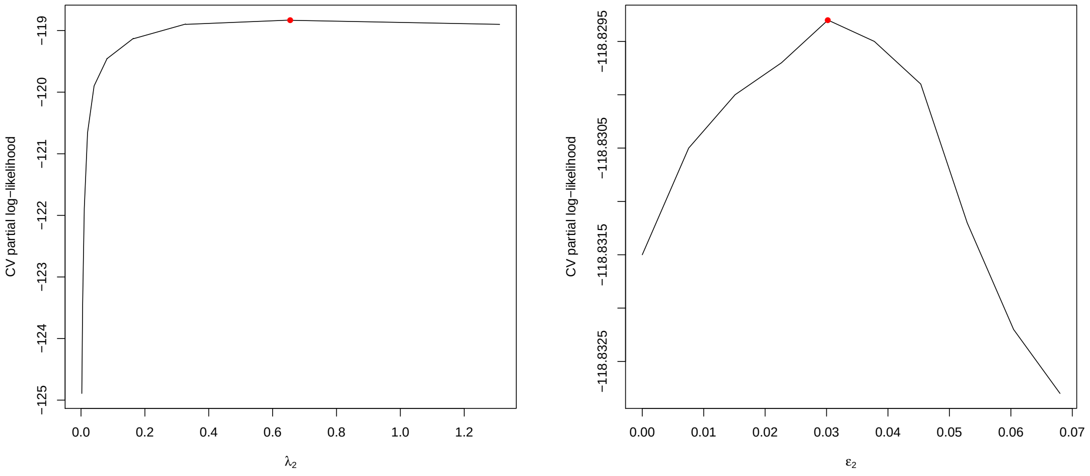
Two Smoothing Terms in the Model

The second smoothing term is added to the model 7.7, the additive Cox PH model with two smoothing term is

$$h(t) = h_0(t)\exp\Big(f_1(\text{CNA}_{108}) + f_2(\text{CNA}_{949}))\Big). \tag{7.8}$$

The smoothing parameter $\lambda_{1,opt}$, and $\epsilon_{1,opt}$ are kept fixed, and we only search for optimal values of λ_2, and ϵ_2 which maximize CVPL. The eigenvalues of the second penalty matrix are 24.561, 2.982, 0.324, and 0. The range of values of ϵ_2 is between zero and 20% of the smallest non-zero eigenvalue. The optimal values of λ_2, and ϵ_2 are $\lambda_{2,opt} = 0.655$, $\epsilon_{2,opt} = 0.030$, with CVPL -118.829. The optimal value of $\epsilon_{2,opt}$ is approximately 10% of the smallest non-zero eigenvalue of the penalty matrix $(0.03/0.3240 = 0.09)$.

The left panel in Figure 7.9 shows the five-fold CVPL for different values of λ_2 at the optimal value of $\epsilon_{2,opt}$, the red point represents the optimal smoothing parameter $\lambda_{2,opt}$. The right panel in Figure 7.7 illustrates the five-fold CVPL for different values of ϵ_2 at the optimal value of $\lambda_{2,opt} = 0.655$. The red point represents the optimal values of $\epsilon_{2,opt}$. The magnitude of change in CVPL due to changing ϵ_2 is very small, as it can be seen in Figure 7.9(b). The right panel of Figure 7.8 shows CVPL for different values of λ_2 and ϵ_2. When we switch the order for considering $f_1(\text{CNA}_{949})$ and $f_2(\text{CNA}_{108})$ in equation (7.8), the optimal values of ϵ_1 and ϵ_2 are approximately equal to 10% of the smallest non-zero eigenvalues of the penalties matrix. In practice it is difficult to obtain the optimal values of each smoothing parameters and ϵ separately in the model, we can fix each ϵ to be 10% of the smallest non-zero eigenvalue for each penalty matrix separately, and only obtaining the optimal values of the smoothing parameters separately in the model.



(a) $\epsilon_{2,opt}$=0.03

(b) $\lambda_{2,opt} = 0.655$

Figure 7.9: (a) Five-fold CVPL for different values of λ_2 at the optimal value of $\epsilon_{2,opt}$. (b) Five-fold CVPL for different values of ϵ_2 at the optimal value of $\lambda_{2,opt} = 0.655$.

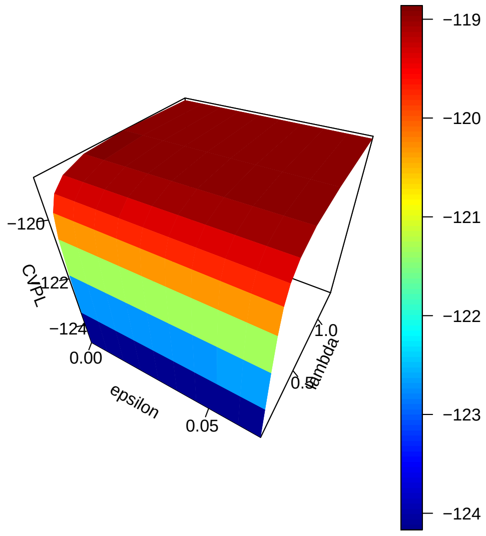


Figure 7.10: Five-fold CVPL for different values of λ_2 and ϵ_2.

7.5 Real Data Analysis

7.5.1 Fit the Shrinkage Penalized Additive Cox PH Model with One Smoothing Parameter

The basic idea of the shrinkage approach is that the variable selection can be carried out in one single model. As a result in this method the estimation of the model coefficients and variable selection can be carried out simultaneously. There are 41 significant CNA genomic-windows out of 1056, each of these 41 significant CNA genomic-windows represents a block. However, to apply the shrinkage approach to the significant CNA genomic-windows simultaneously, we need to assume that the smoothing parameters for each smoothing term will have the same value, and we fixed ϵ to be 10% of the smallest non-zero eigenvalue for each penalty matrix separately.

Let $\lambda = 10^{-5} \times 2^{(i-1)}$ for $i = 1, \ldots, n_\lambda$ be a vector of values of the smoothing parameter, and $n_\lambda = 30$ is the length of the smoothing parameter. Plot of the five-fold cross-validated partial log-likelihood as a function of the smoothing parameter λ is given in Figure 7.11. The optimal value of the smoothing parameter is $\lambda_{opt} = 83.88$, CVPL$(\lambda_{opt}) = -118.23$.

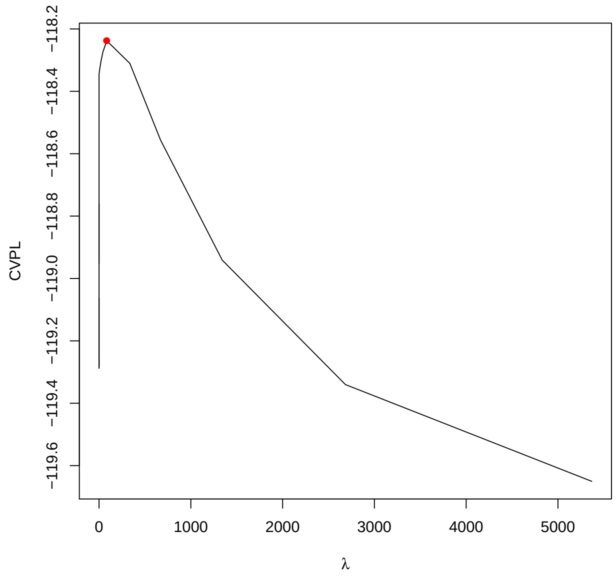


Figure 7.11: The cross-validated partial log likelihood for the shrinkage approach.

The penalized additive Cox PH model is fitted with 5 equally spaced knots for each smoothing term, and the optimal value of the smoothing parameter λ. This model includes Stage-T, Stage-N, age and 41 significant windows of CNA as smoothing terms. It is easy to optimize the CVPL for one smoothing parameter, but it gives a very smooth fit to all the significant CNA genomic-windows. As a result, the estimated log hazard ratio for all the smoothing terms are fitted as a linear terms. In particular we keep only 27 significant CNA genomic-windows, and 14 CNA windows are removed from the model, because the confidence band of these 14 CNA windows include $f(x) \equiv \mathbf{0}$, so we can just plot a horizontal line at zero.

Figure 7.12 shows the estimated log hazard ratio for age. The estimated log hazard ratios for all of the 41 significant windows of CNA are presented in Figures 7.13 - 7.16, x-axis represents the observed CNA window, and the y-axis represents the estimated log hazard ratio of CNA window. The solid black line is the estimated log hazard ratio, the points indicate the number of observations on each individual that were either censored or a failure. The dashed lines are the 95% point-wise confidence band. The CNA genomic-windows number is shown in the legend.

To overcome this over smoothed fit we will use forward variable selection, and estimate the smoothing parameters for each smoothing term separately, which is discussed in the following section.

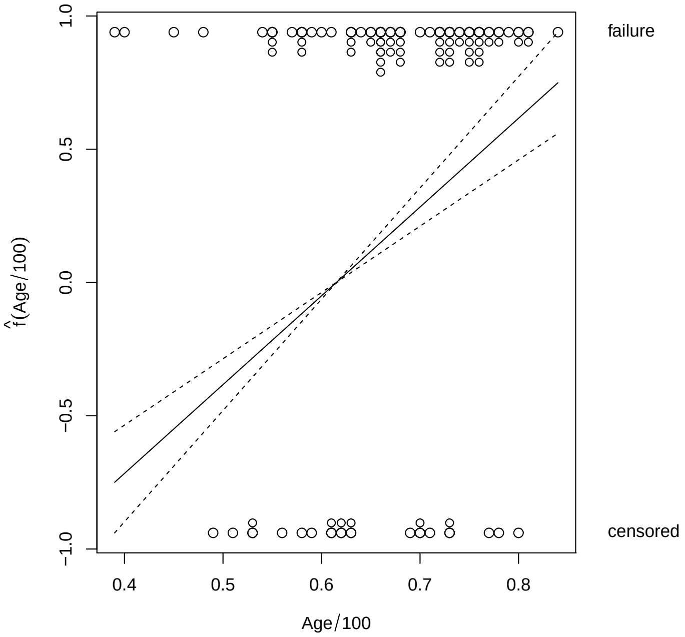


Figure 7.12: Plots of the estimate log hazard ratio for age. The sold black line is the estimated log hazard ratio, the points indicate the number of observations on each individual that were either censored or a failure. The dashed lines are the 95% pointwise confidence band.

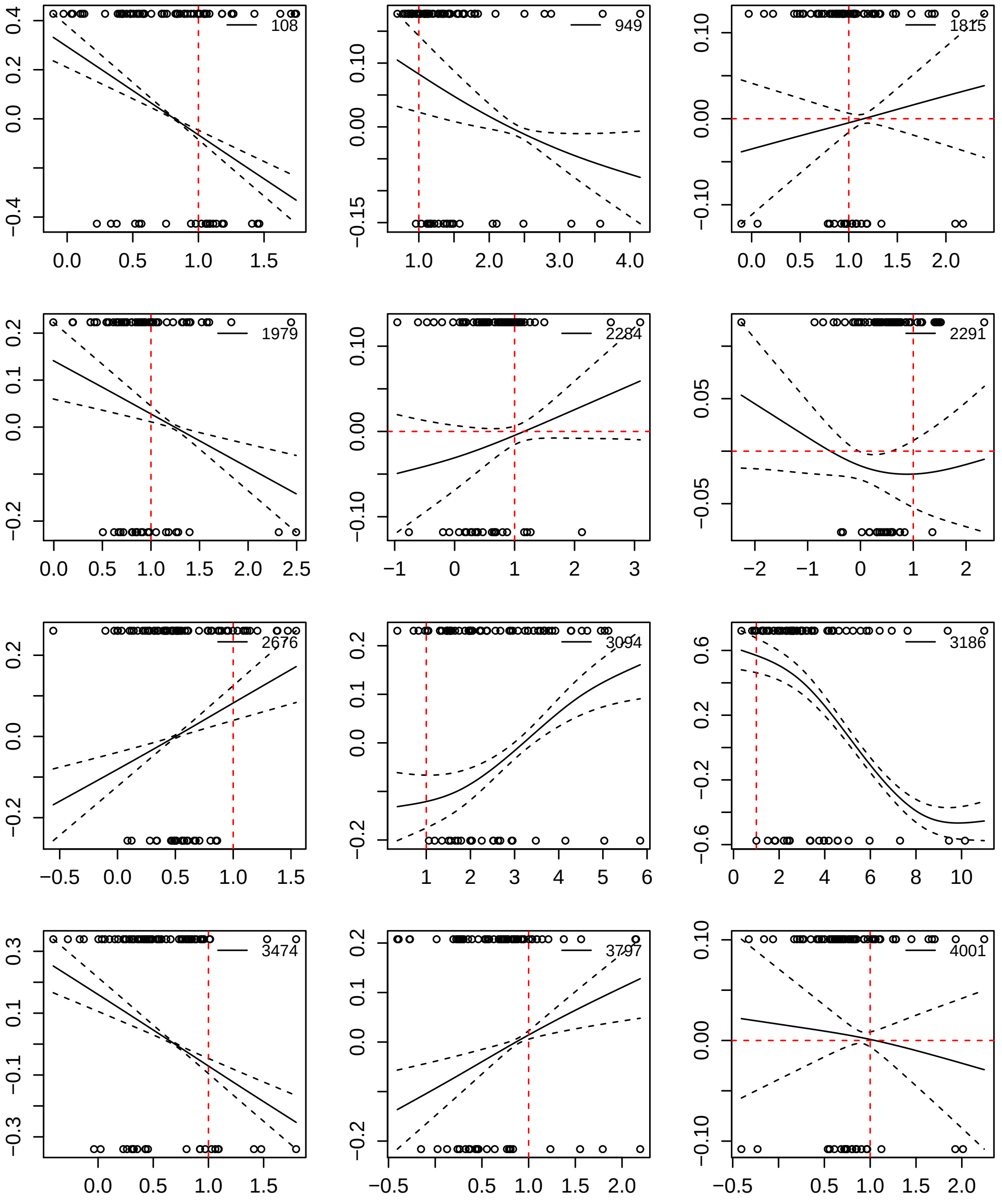


Figure 7.13: Plots of the estimate log hazard ratios. The sold black line is the estimated log hazard ratio, the points indicate the number of observations on each individual that were either censored or a failure. The dashed lines are the 95% point-wise confidence band. The CNA genomic-windows number is shown in the legend.

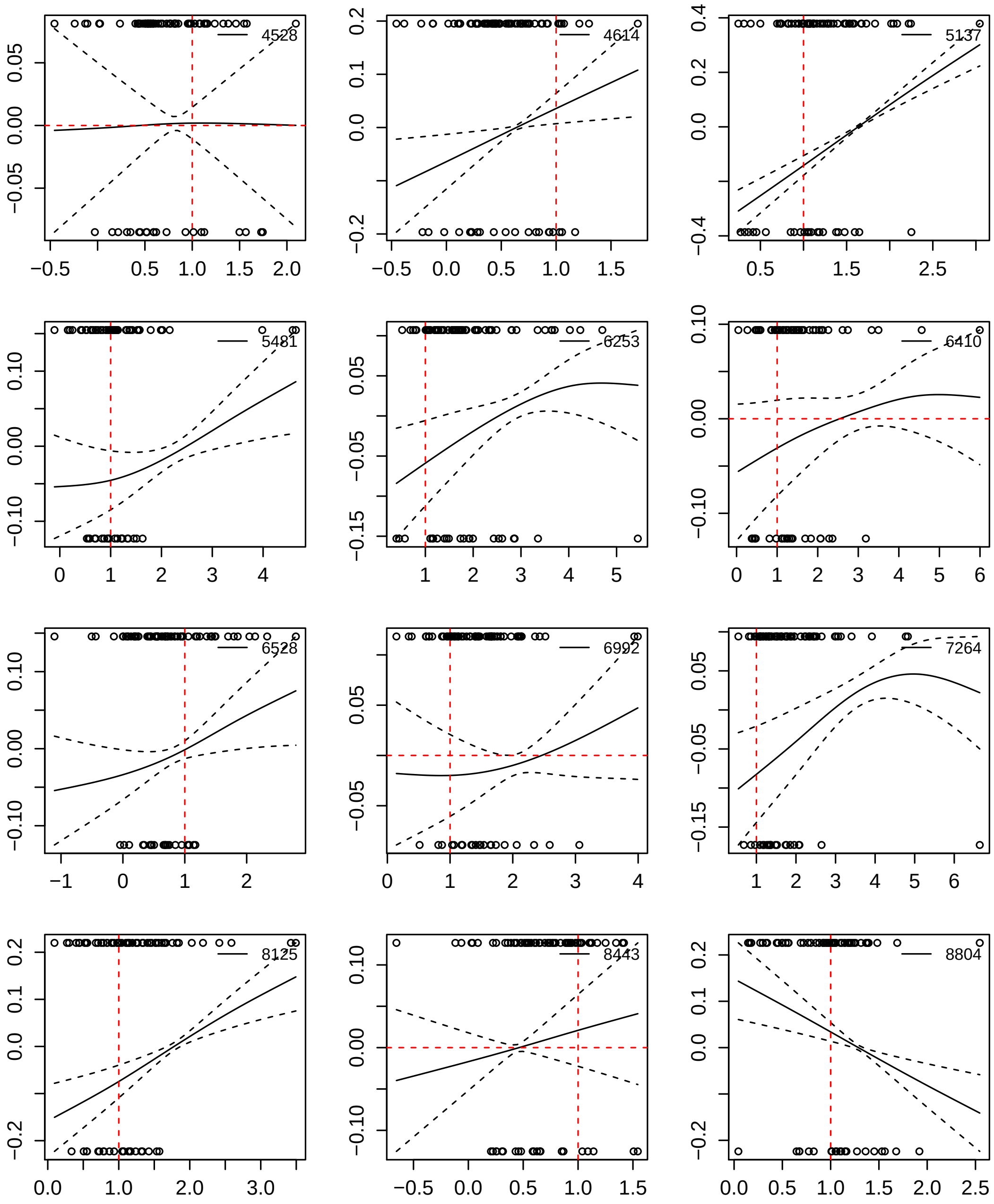


Figure 7.14: Plots of the estimate log hazard ratios. The sold black line is the estimated log hazard ratio, the points indicate the number of observations on each individual that were either censored or a failure. The dashed lines are the 95% point-wise confidence band. The CNA genomic-windows number is shown in the legend.

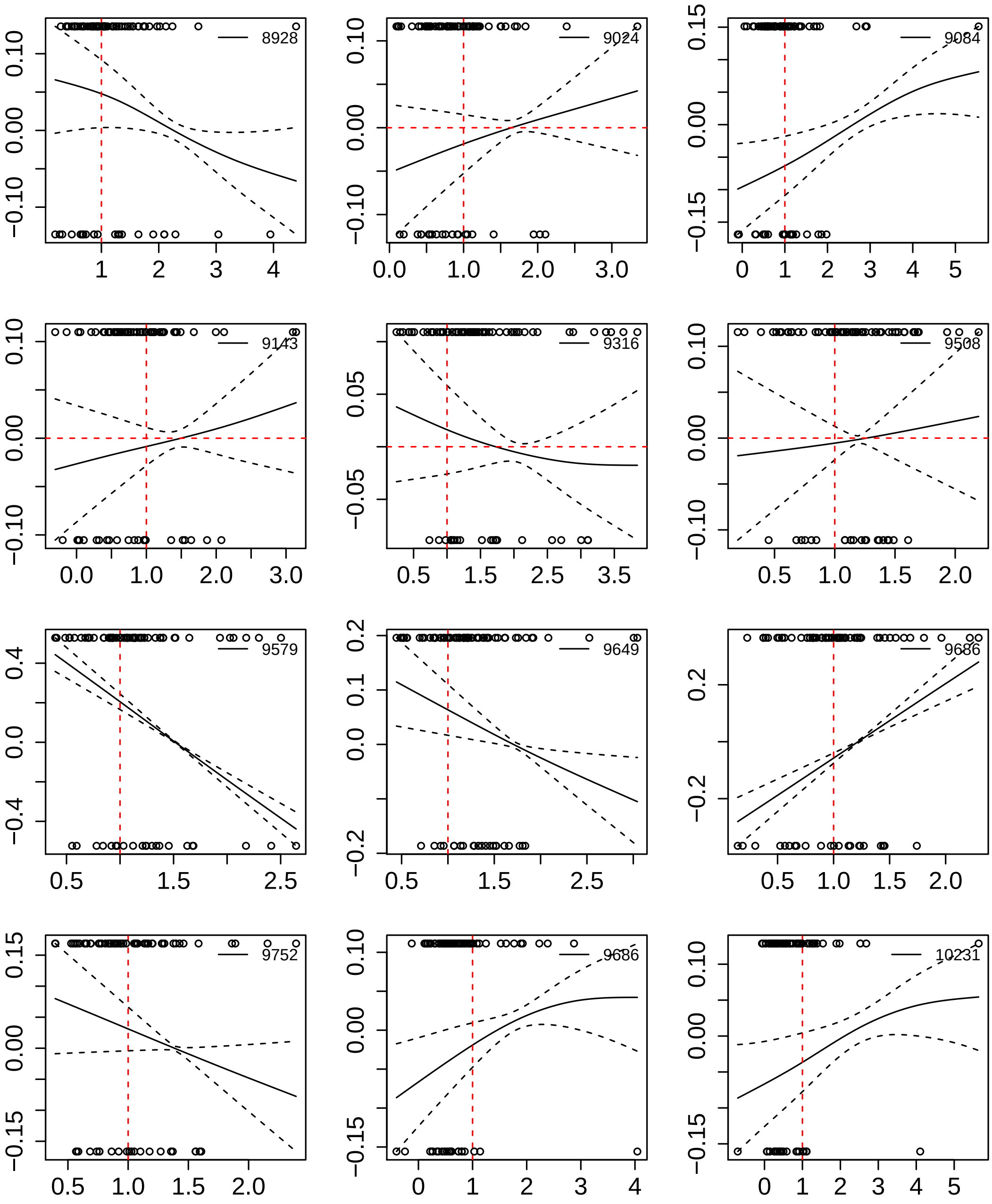


Figure 7.15: Plots of the estimate log hazard ratios. The sold black line is the estimated log hazard ratio, the points indicate the number of observations on each individual that were either censored or a failure. The dashed lines are the 95% point-wise confidence band. The CNA genomic-windows number is shown in the legend.

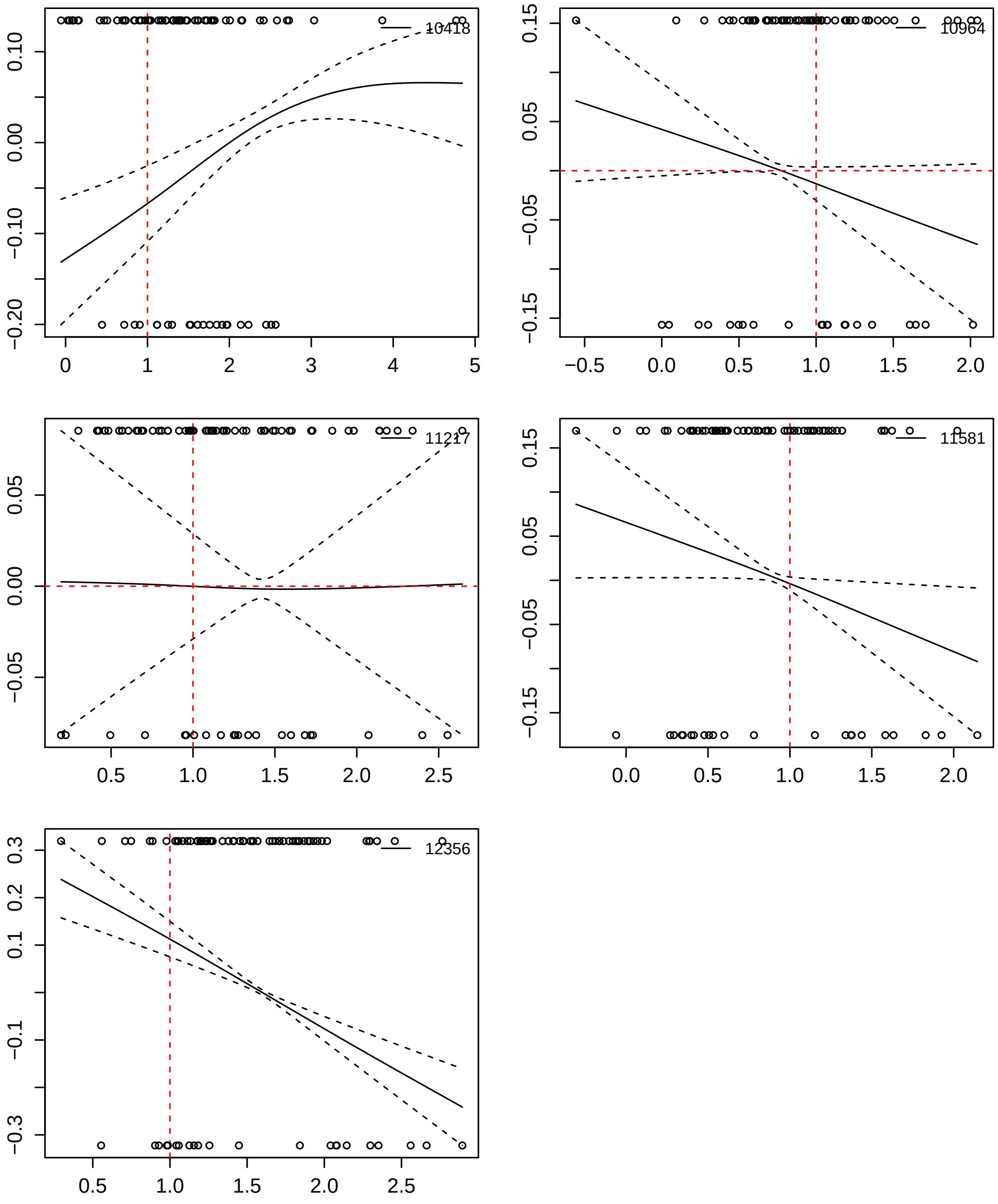


Figure 7.16: Plots of the estimate log hazard ratios. The sold black line is the estimated log hazard ratio, the points indicate the number of observations on each individual that were either censored or a failure. The dashed lines are the 95% point-wise confidence band. The CNA genomic-windows number is shown in the legend.

7.6 Forward Stepwise Selection Using the Shrinkage Approach

The aim of this section is to apply the shrinkage approach to the 41 significant CNA genomic-windows, obtaining the smoothing parameters λ separately for each smoothing term. The main question here is which variable should we include first in the model; we include the most significant CNA genomic-windows based on the smallest p-value from the penalized univariate variable selection, as described in Section 6.2.2. The order of the significant CNA genomic-windows based on their p-values are presented in Table 7.2.

#	chr	window	p-value	#	chr	window	p-value
1	Chr3	2291	0.013653	21	Chr3	2676	0.042087
2	Chr11	9143	0.023786	22	Chr12	9649	0.043065
3	Chr 1	108	0.027921	23	Chr2	2284	0.043557
4	Chr18	11581	0.029107	24	Chr4	4001	0.043624
5	Chr10	8443	0.029323	25	Chr19	12356	0.045674
6	Chr12	9316	0.029923	26	Chr3	3094	0.046057
7	Chr2	1815	0.032956	27	Chr11	8928	0.046315
8	Chr11	9024	0.033174	28	Chr3	3186	0.046428
9	Chr10	8125	0.033214	29	Chr2	1979	0.046475
10	Chr12	9579	0.033303	30	Chr7	6410	0.046685
11	Chr5	4614	0.033966	31	Chr6	5481	0.046758
12	Chr4	3797	0.034302	32	Chr12	9508	0.046966
13	Chr5	4528	0.034418	33	Chr16	10964	0.047407
14	Chr11	9084	0.035453	34	Chr12	9686	0.047636
15	Chr17	11217	0.037782	35	Chr11	8804	0.047742
16	Chr7	6253	0.038643	36	Chr8	6528	0.048067
17	Chr6	5137	0.040975	37	Chr15	10418	0.048141
18	Chr13	10231	0.041216	38	Chr4	3474	0.048623
19	Chr14	10278	0.041364	39	Chr8	7264	0.049593
20	Chr 1	949	0.041385	40	Chr8	6992	0.049689
				41	Chr13	9752	0.049974

Table 7.2: The ordering of the significant CNA genomic-windows.

We fixed ϵ for each smoothing term to be 10% for the smallest non-zero eigenvalue of the penalty matrix. For the j^{th} smoothing term, we fixed ϵ_j to be 10% of the smallest non-zero eigenvalue of the j^{th} penalty matrix. Let $\lambda = 10^{-5} \times 2^{(i-1)}$ for $i = 1, \dots, n_\lambda$ be a vector of values of the smoothing parameter, and $n_\lambda = 30$ is the length of the smoothing parameter, to obtain the optimal values of the smoothing parameters in the model, we used two-dimensional grid searching. For the j^{th} smoothing term term include in the model we chose $(\lambda_{j-1}, \lambda_j)$ by two dimensional searching, with re-optimize the previous λ. The optimal value $(\lambda_{j-1,opt}, \lambda_{j,opt})$ that maximizing CVPL are selected. Basically, we used the same technique of estimating the smoothing parameters as we did in Chapter 6. The final model is

$$\begin{aligned} h(t) = & h_0(t) \exp \Big(StageT + StageN + f(Age) + f_1(\text{CNA}_{2291}) + f_2(\text{CNA}_{9143}) \\ & + f_3(\text{CNA}_{108}) + f_4(\text{CNA}_{11581}) + f_5(\text{CNA}_{9316}) + f_6(\text{CNA}_{4614}) \\ & + f_7(\text{CNA}_{3797}) + f_8(\text{CNA}_{4528}) + f_9(\text{CNA}_{9084}) + f_{10}(\text{CNA}_{10231}) \\ & + f_{11}(\text{CNA}_{10278}) + f_{12}(\text{CNA}_{949}) + f_{13}(\text{CNA}_{2676}) + f_{14}(\text{CNA}_{9649}) \\ & + f_{15}(\text{CNA}_{2284}) + f_{16}(\text{CNA}_{12356}) + f_{17}(\text{CNA}_{6528}) + f_{18}(\text{CNA}_{10418}) \Big). \end{aligned} \tag{7.9}$$

The p-value for testing of the null hypothesis there is no covariate effect of the smoothing terms is equal to 0.0165. The optimal values of the smoothing parameter of age $\lambda = 0.163$, and $\epsilon = 0.0007$. The optimal values of the smoothing parameters, and ϵ for the each of the significant CNA genomic-window in model (7.9) can be summarized in Table 7.3.

#	window	λ_{opt}	ϵ	#	window	λ_{opt}	ϵ
1	2291	0.010	0.7878	10	10231	0.010	0.7205
2	9143	0.163	0.3384	11	10278	0.163	2.0676
3	108	0.010	0.0511	12	949	0.010	0.3247
4	11581	0.004	0.1219	13	2676	0.163	0.0792
5	9316	0.327	0.3916	14	9649	0.327	0.1463
6	4614	0.010	0.0891	15	2284	0.655	0.5849
7	3797	0.005	0.1507	16	12356	0.163	0.1463
8	4528	0.327	0.1404	17	6528	0.010	0.4869
9	9084	0.163	1.5062	18	10418	0.163	0.9594

Table 7.3: The optimal values of smoothing parameters λ for each of the selected significant windows of the CNA in the multivariate shrinkage penalized Cox model.

Using the optimal values of the smoothing parameters for each smoothing term in the model, the estimate of the fixed effect parameter and their inferences can be summarized in Table 7.8 in Section 7.7 for comparison with the penalized univariate selection method using forward variable selection. The effective degrees of freedom for each smoothing significant CNA genomic-windows are summarized in Table 7.4.

#	window	edf	#	window	edf
1	2291	2.020	10	10231	2.003
2	9143	3.657	11	10278	1.035
3	108	3.069	12	949	1.010
4	11581	3.216	13	2676	3.940
5	9316	3.956	14	9649	1.213
6	4614	1.859	15	2284	3.049
7	3797	3.414	16	12356	1.160
8	4528	1.495	17	6528	3.135
9	9084	2.468	18	10418	1.178

Table 7.4: The effective degrees of freedom for each smoothing significant CNA genomic-windows.

The plot of the estimated log hazard for $\hat{f}(Age)$ is presented in Figure 7.17, the solid black line is the estimated log hazard ratio, the points indicate the number of observations on each individual that were either censored or a failure. The dashed lines are the 95% point-wise confidence band. Figures 7.18 and 7.19 show the plots of the estimated log hazard ratio for each of the 18 significant window of the CNA, x-axis represents the observed CNA window, and the y-axis represents the estimated log hazard ratio of CNA window. The solid black line is the estimated log hazard ratio, the points indicate the number of observations on each individual that were either censored or a failure. The dashed lines are the 95% point-wise confidence band. The CNA genomic-windows number is shown in the legend.

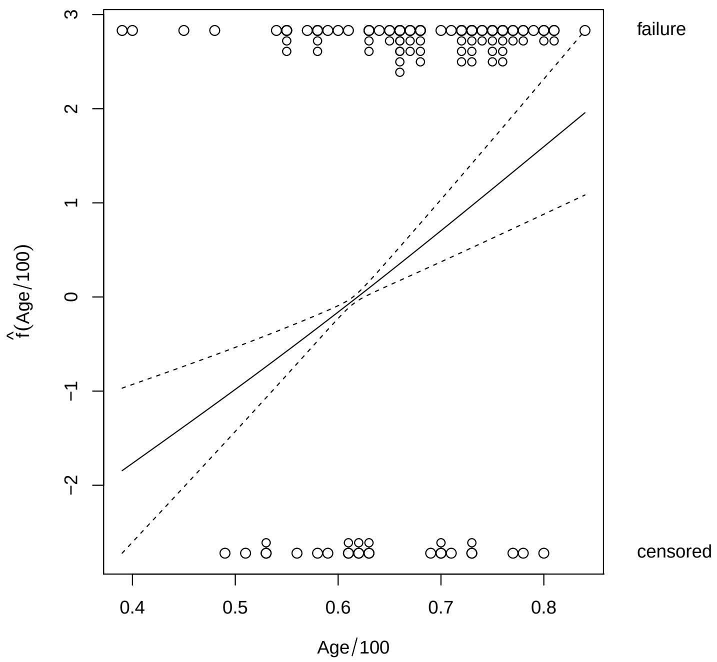


Figure 7.17: The estimated log hazard ratios for the $\hat{f}(Age)$ versus Age.

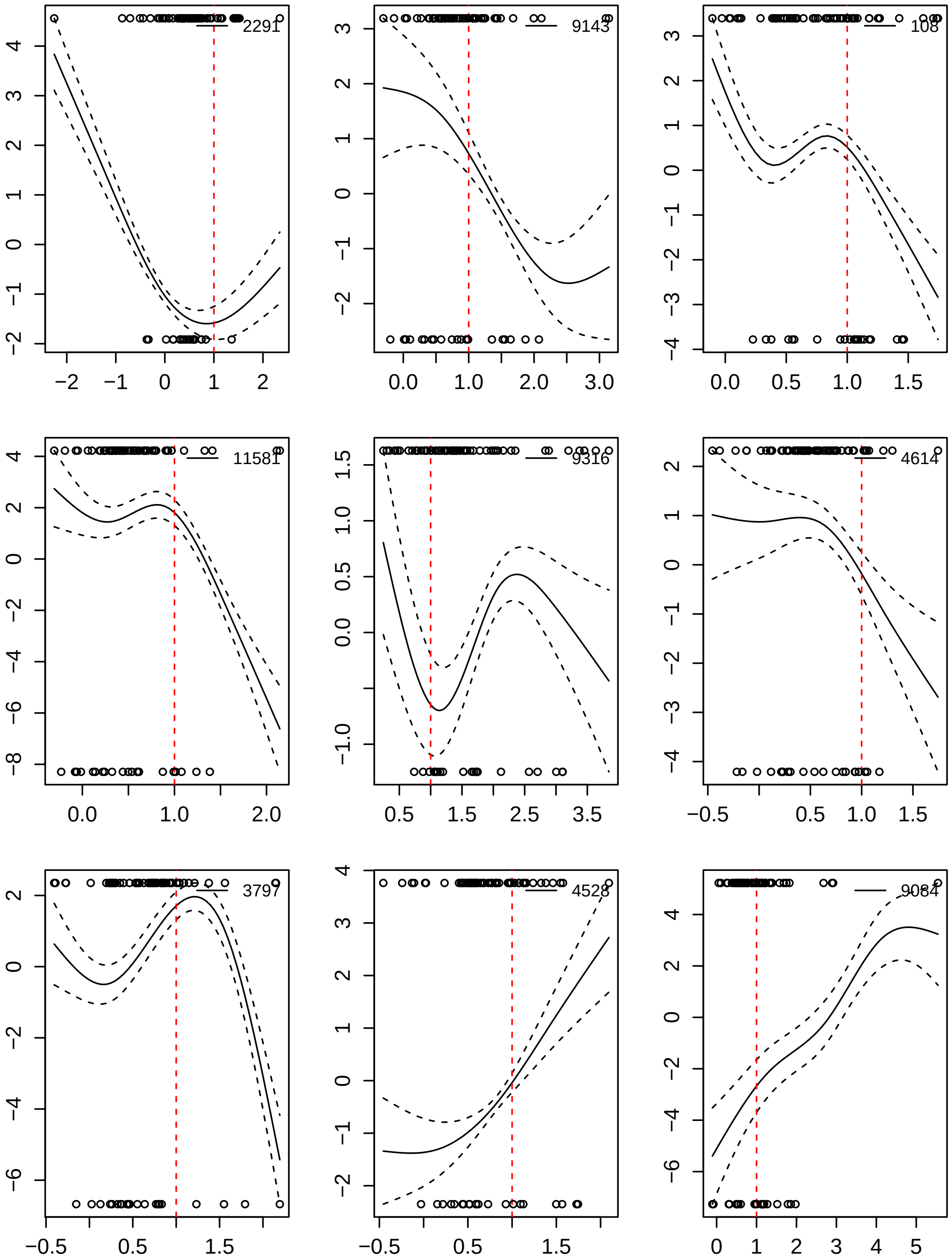


Figure 7.18: Plots of the estimated log hazard ratio for each of the significant window of the CNA. The sold black line is the estimated log hazard ratio, the dashed lines are the 95% point-wise confidence band. The points indicate the number of observations on each individual that were either censored or a failure. The CNA genomic-windows number is shown in the legend.

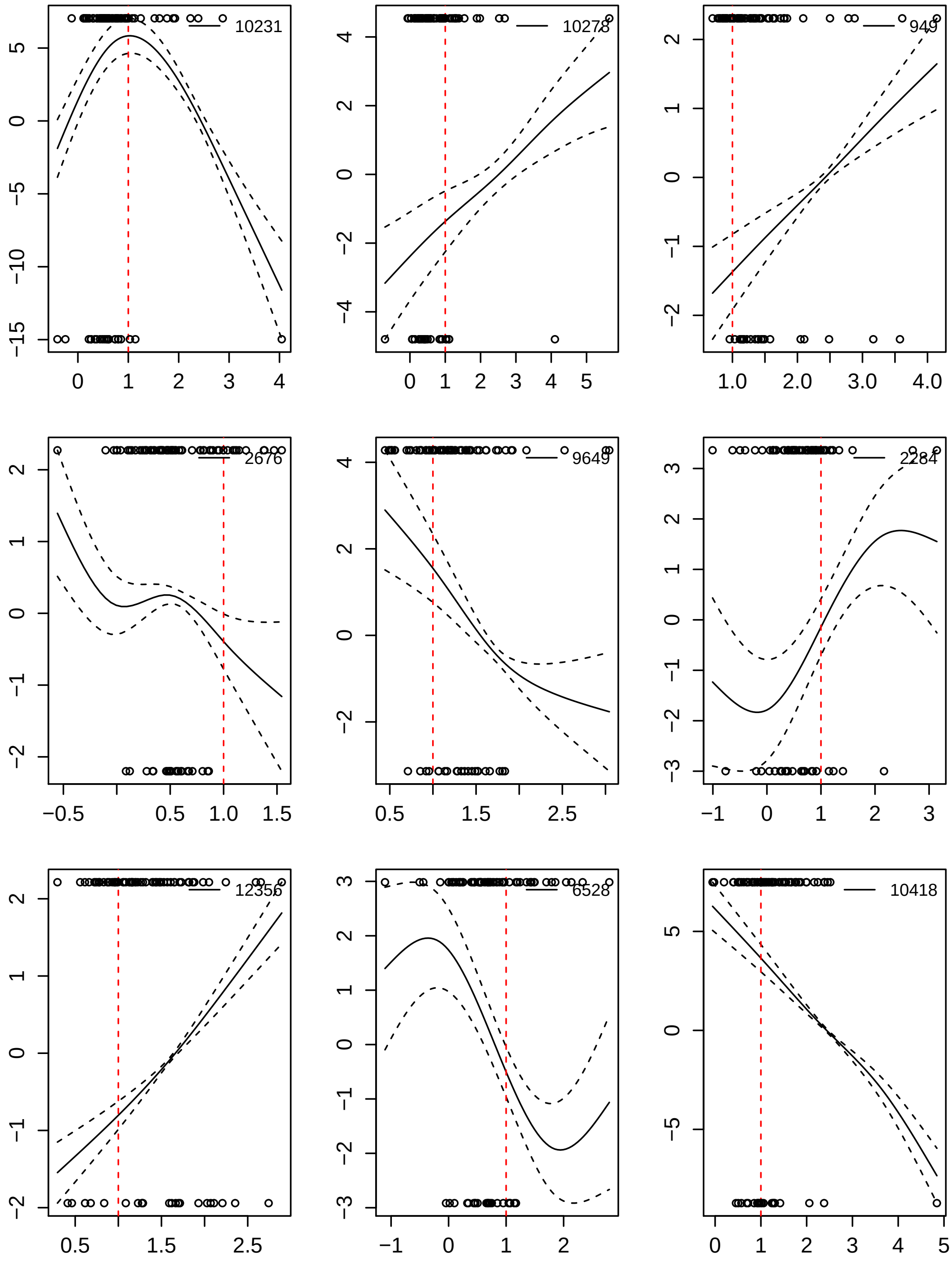


Figure 7.19: Plots of the estimated log hazard ratio for each of the significant window of the CNA. The sold black line is the estimated log hazard ratio, the dashed lines are the 95% point-wise confidence band. The points indicate the number of observations on each individual that were either censored or a failure. The CNA genomic-windows number is shown in the legend.

The smoothing hazard ratio of Age, CNA_{10278}, CNA_{949}, CNA_{9649}, CNA_{12356}, and CNA_{10418} exhibit a linear pattern, hence these terms can be replace in the model by linear terms. Two significant windows of CNA identify the lower risk at normal ploidy 1, which are CNA_{2291}, and CNA_{9316}, while CNA_{108}, CNA_{11581}, CNA_{3797}, and CNA_{10231} significant windows of CNA identify a higher risk at normal ploidy 1.

The cumulative hazard of the Cox-Snell residuals from the model fitting based on Stage-T, Stage-N, age, and 18 significant windows of the CNA and as smoothing terms in model (7.9) is plotted as part of the model diagnostics, which can be seen in Figure 7.20.

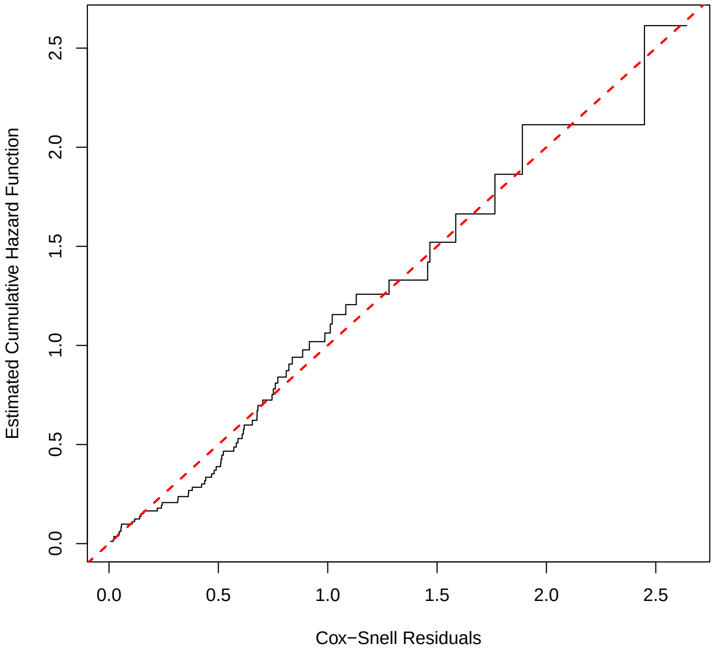


Figure 7.20: Cumulative hazard of Cox-Snell residual, the solid line, from the shrink additive Cox PH model fit, comparing to the identity line the red dashed line .

Some of the plots of the estimated log hazard ratios for the significant CNA window showed a similar pattern, especially at the normal ploidy of one. However, these significant CNA window that have similar patterns are not correlated, which means there are some windows in the gene are have similar effects on the survival model.

Hierarchical clustering is the most common clustering method. In this method, significant CNAs genomic-windows whose estimated log hazard ratio patterns are similar across patients can be gathered into one cluster, taking into account the location of the normal ploidy one. Each of the significant CNA genomic-window have a different scale of the observe significant CNA window, in order to cluster a group of the significant CNA genomic-windows that have a similar shape of the log hazard ratio for the significant CNA windows at ploidy one, we create selection points, and then we evaluate the log hazard ratio for each significant CNA window using some of these selection points to ensure that the estimated log hazard ratio share the similar shape at normal ploidy one.

The minimum and maximum values of all the significant CNA genomic-windows in model (6.5) are -2.255 and 11.031 respectively, these minimum and maximum values are used to create the selection points by 0.05, so the length of this selection points is 266 observations. Then we evaluate each of the estimate of the log hazard ratio at some but not all of the selection points and NA otherwise, this can be recorded as a row in matrix, where the number of rows is the number of the significant CNA genomic-windows, which is 18, and number of columns is length of the selection points, which is 266. Three different cluster method as used, the first method is that we compute the Euclidean distance to perform the complete linkage. The second method, we would consider a maximum weighted difference. For two smoothing terms $\hat{f}_1(x_1)$ and $\hat{f}_2(x_2)$, which are observed at some but not all of the selection points, we define the distance measure

$$d_{12} = \max_i \frac{|\hat{f}_1(x_{1i}) - \hat{f}_2(x_{2i})|}{\sqrt{\mathrm{Var}\Big(f_1(x_{1i})_{ii}\Big) + \mathrm{Var}(f_2(x_{2i})_{ii}\Big)}},$$

where the maximum is over all i such that both $f(x_{1i})$ and $f(x_{2i})$ are not NA. The third method is described in Bozkus (2017). Complete linkage clustering is used for all the different methods, which is used as explanatory tool that we use to suggest group (in this case, pairs) of CNA windows which seems to have similar effects. We are interested to see whether the result of the clustering is meaningful and interpretable.

The objective is to find any small cluster that have very similar shape of the estimated log hazard ratio of the significant CNA genomic-windows at normal ploidy 1. As a result, some of the clusters do not necessarily show a consistent similar shape estimated log hazard ratio of the significant CNA genomic-windows at normal ratio 1, while other clustering shows a consistent similar shape estimated log hazard ratio at normal ratio 1 between the significant CNA genomic-windows. For example the top panel of Figure 7.21 shows the estimated log hazard ratios are not necessarily similar to each other although they are found to be in the same cluster. while down panel of Figure 7.21 shows a similar estimated log hazard ratio of the significant CNA genomic-windows at normal ploidy 1 and they clustered together.

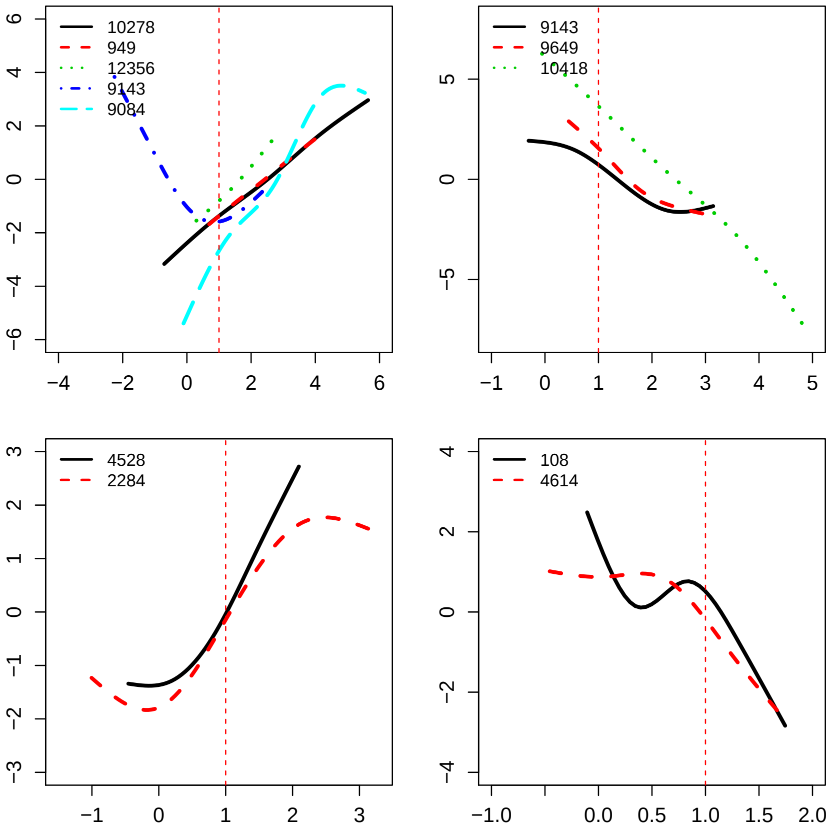


Figure 7.21: Two clusters of the estimated log hazard ratio of the significant CNA genomic-windows.

The common results from all different type of clustering are presented in Figure

7.22. Table 7.5 shows the six clusters with the corresponding chromosome for each significant CNA genomic-windows.

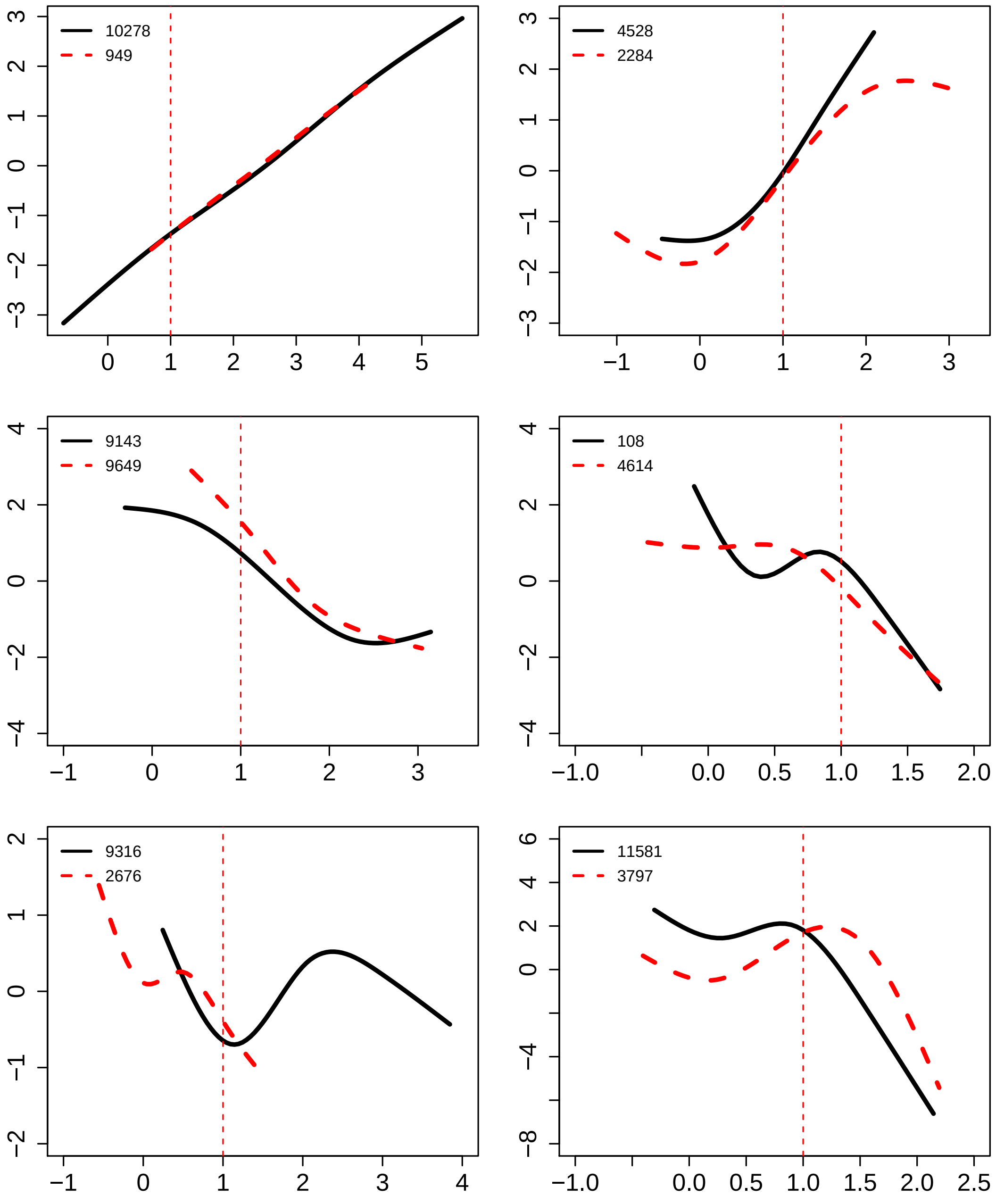


Figure 7.22: The six clusters of the estimated log hazard ratio of the significant CNA genomic-windows.

Cluster	Window	Chromosome
1	10278	14
	949	1
2	4528	5
	2284	2
3	9143	11
	9649	12
4	108	1
	4614	5
5	9316	12
	2676	3
6	11581	18
	3797	4

Table 7.5: The six clusters of the estimated log hazard ratio of the significant CNA genomic-windows.

7.7 A Comparison between the Discrete Feature Selection and Shrinkage Feature Selection

The main objective of this section is to investigate and compare the use of the discrete feature selection and shrinkage feature selection in the penalized additive Cox PH model, where the clinical data are considered as fixed effect predictors, and the 41 significant CNA genomic-windows as a smoothing terms. In the case of fitting the standard Cox PH model to the clinical data, Stage-T, Stage-N and age were significant (p-value < 0.05); these variables were included in both variable selection methods. In both discrete feature selection and shrinkage feature selection, no variable selection have been applied to the clinical covariates. The variable selection methods are applied to the high-dimensional CNA genomic-windows. A List of the 18 common significant CNA genomic-windows by the order on the genome are presented in Table 7.6. There are 23 significant CNA genomic-windows in the forward stepwise selection, which are eliminated using the shrinkage approach.

For both methods, five-fold CVPL was used to obtain optimal values of the smoothing parameters. Table 7.7 presents the optimal smoothing parameter values of each of

#	Chr	#window	#	Chr	# window
1	1	108	10	11	9084
2	1	949	11	11	9143
3	2	2284	12	12	9316
4	3	2291	13	12	9649
5	3	2676	14	13	10231
6	4	3797	15	14	10278
7	5	4528	16	15	10418
8	5	4614	17	18	11581
9	8	6528	18	19	12356

Table 7.6: List of the common significant CNA genomic-windows for both forward variable selection and shrinkage selection.

the common significant CNA genomic-windows.

#	Chr	window	forward	shrinkage	
			λ_{opt}	λ_{opt}	ϵ
1	1	108	0.630	0.010	0.051
2	1	949	2.154	0.010	0.324
3	2	2284	0.341	0.655	0.584
4	3	2291	0.100	0.010	0.787
5	3	2676	0.630	0.163	0.079
6	4	3797	1.584	0.005	0.150
7	5	4528	0.630	0.327	0.140
8	5	4614	0.630	0.010	0.089
9	8	6528	0.100	0.010	0.486
10	11	9084	0.630	0.163	1.506
11	11	9143	1.584	0.163	0.338
12	12	9316	0.215	0.327	0.391
13	12	9649	0.630	0.327	0.146
14	13	10231	0.630	0.010	0.726
15	14	10278	0.251	0.163	2.067
16	15	10418	0.630	0.163	0.959
17	18	11581	0.630	0.004	0.121
18	19	12356	0.630	0.163	0.146

Table 7.7: List of the common significant CNA genomic-windows for both forward variable selection and shrinkage selection.

Using the optimal values of the smoothing parameters for each smoothing term in the two different approaches, the fixed effect parameters for both forward variable selection and shrinkage selection were estimated. Estimates of the fixed effect parameters and their inferences using both methods can be seen in Table 7.8. This table shows that Stage-T3 and Stage-N2 are statistically significant (p-value< 0.05). The estimate of Stage-T3 is positive, which indicate that the larger tumour size increases the hazard relative to the baseline Stage-T1. Likewise, the positive estimates of Stage-N2 suggest that a wider spread of cancer to the lymph nodes increases the hazard significantly relevant to the baseline Stage-N0. The estimated fixed effect parameters from fitting the two approaches, both forward variable selection and shrinkage selection, are similar in some aspect, and the confidence interval of the fixed effect parameters in both approaches overlap. The p-values for the estimated fixed effect parameters are similar, except Stage-N1 in the forward stepwise selection, which appears to to be significant. However, additional investigations are needed on order to confirm our findings and draw a final conclusions.

Predictor	Estimate	Exp	Standard error	Wald test	p-value
without smoothing term					
Age	5.311	202.585	1.558	3.409	0.001
Stage T2	0.150	1.162	0.302	0.499	0.618
Stage T3	1.800	6.052	0.576	3.123	0.002
Stage N1	0.345	1.411	0.284	1.212	0.225
Stage N2	1.336	3.804	0.478	2.797	0.005
with forward selection					
StageT2	0.465	1.593	0.318	1.465	0.143
StageT3	2.207	9.087	0.591	3.736	0.000
StageN1	0.722	2.059	0.323	2.233	0.026
StageN2	1.887	6.602	0.498	3.790	0.000
with shrinkage selection					
StageT2	0.115	1.122	0.304	0.378	0.706
StageT3	1.958	7.082	0.576	3.400	0.001
StageN1	0.144	1.154	0.288	0.499	0.618
StageN2	1.442	4.228	0.481	2.996	0.003

Table 7.8: Estimated values of the parameters on fitting additive Cox PH model for both forward variable selection and shrinkage selection.

The models are not nested except in special cases, this excludes using the test statistics for both forward variable selection and shrinkage selection of no effect of the age and significant CNA genomic-windows, the p-values are calculated for both forward variable selection and shrinkage selection which are 0.0165, and 0.046 for shrinkage, forward stepwise selection respectively. The likelihood function for both approaches is different, this implies that likelihood based model selection criteria such as AIC can not be used in this case.

To compare the estimated log hazard ratio of the two variable selection methods, we plot the estimated log hazard ratio for the common significant CNA genomic-windows for both methods for each significant CNA genomic-windows, which can be seen in Figures 7.23 and 7.24, the red dashed line represent the estimated log hazard ratio for each of the significant window of the CNA using forward stepwise selection, and the black solid line represent the estimated log hazard ratio for each of the significant window of the CNA using shrinkage approach. The significant CNA genomic-windows number are shown in the legend.

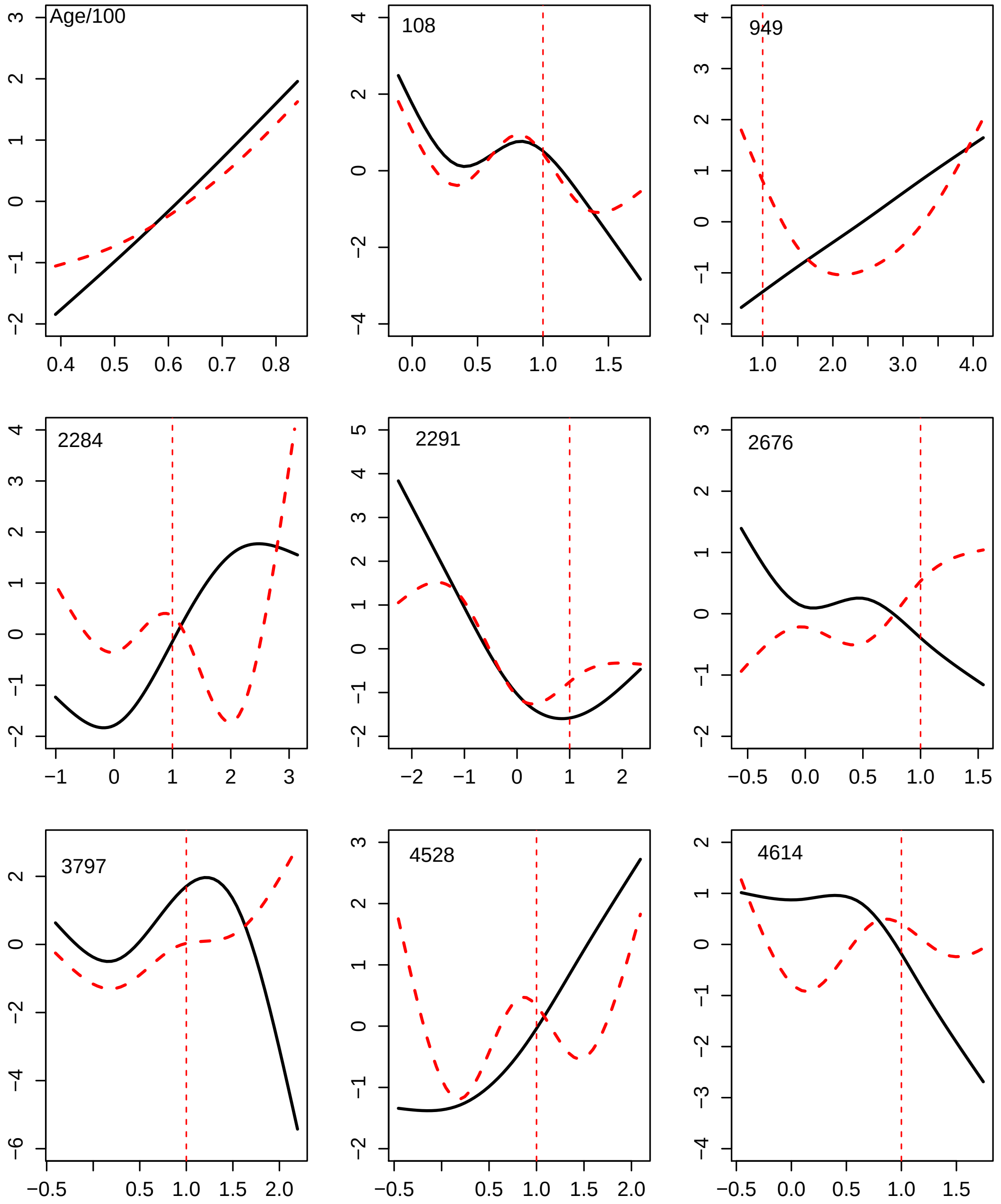


Figure 7.23: The plot of the estimated log hazard ratio for each of the significant window of the CNA using forward stepwise selection (red dashed line) and shrinkage approach (black solid line).The CNA genomic-windows number is shown in the legend.

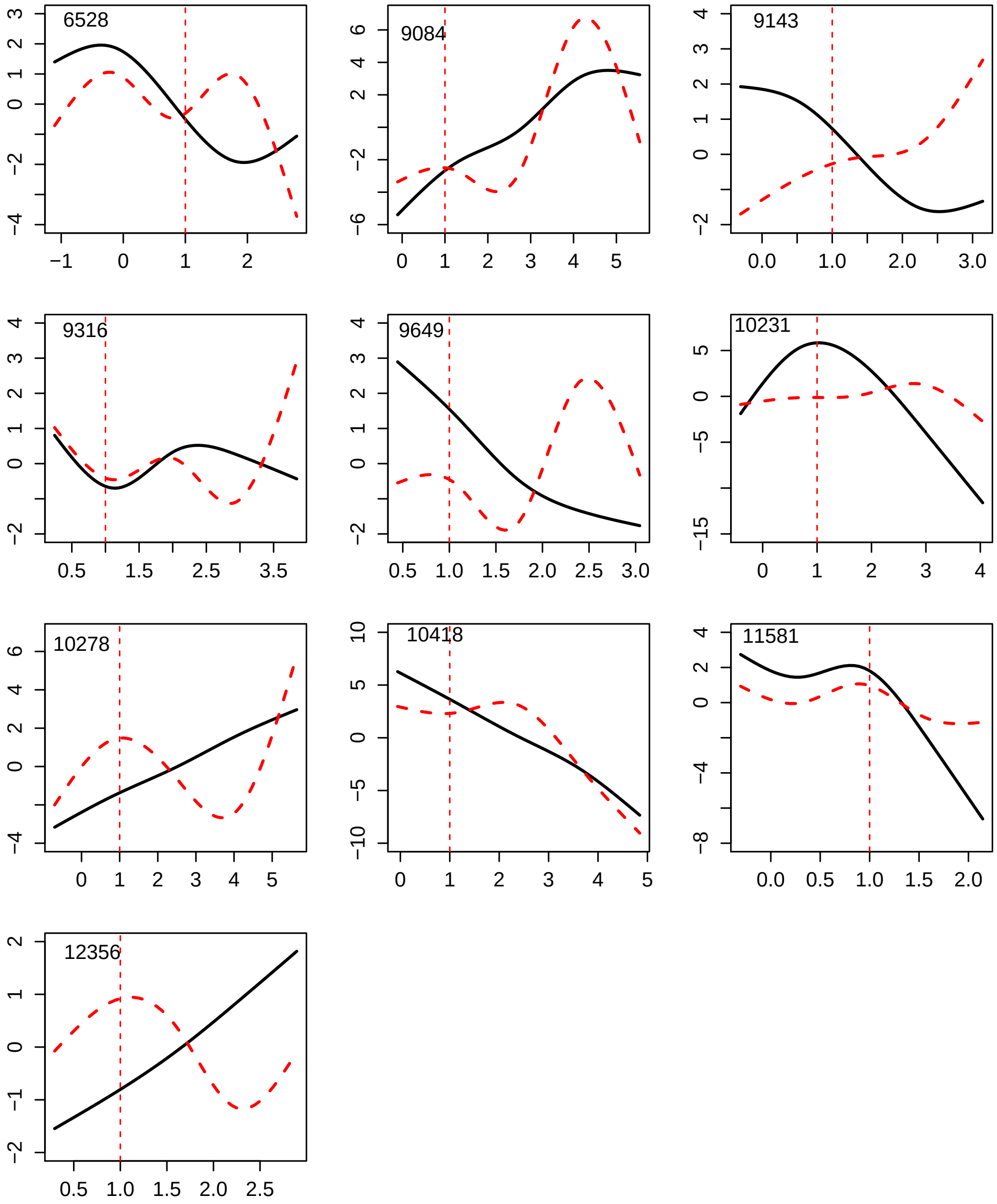


Figure 7.24: The plot of the estimated log hazard ratio for each of the significant window of the CNA using forward stepwise selection (red dashed line) and shrinkage approach (black solid line). The CNA genomic-windows number is shown in the legend.

For both methods, we have the same conclusion about the estimated log hazard ratios for three significant CNA genomic-windows, the significant CNA genomic windows CNA_{108}, and CNA_{11581} display a higher risk at ploidy one, while the log hazard ratio estimate for CNA_{9316} shows a lower hazard at ploidy one. The shapes of the estimated log hazard ratio for most cases are different. This may be because the set of variables in the two methods are different, and there may be correlation with combinations of other variables. These different log hazard ratios for individual significant CNAs genomic-windows need to be investigated further in order to achieve a better biological understanding of the process.

Method based on shrinkage approach of significant CNA genomic-windows have fewer variable compare to a forward stepwise selection, but the ”forward stepwise leads to locally optimal model rather than the best model” Klein (2013) page 96.

7.8 Conclusion

In this chapter, two shrinkage approach are introduced, which are based on the shrinkage approach explained in Marra and Wood (2011). The smoothing components can be selected based on their smoothing parameter values. This shrinkage approach can be carried out in one single model. However, obtaining one optimal smoothing parameter in the model by assuming that the each smoothing term have the same smoothing parameter value leads to an over-smoothed model. Alternatively, a forward stepwise method can be used, which allow us to obtain the optimal smoothing parameter value for each smoothing component separately.

In addition to the methods that are described in this chapter, clustering techniques are used to identify possible similar patterns of the significant genomic windows across patients. Comparison between the desecrate feature selection (penalized univariate selection) and shrinkage feature selection is presented in terms of the estimated fixed effect parameters, and comparing the shapes of the log hazard ratio of these two different methods.

Generally, for both prediction and interpretation, the shrinkage approach has the better performance, because it is fit a method that includes fewer significant genomic windows of CNA in contrast to the forward stepwise selection of the high dimension of significant genomic windows of CNA.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

The main objective of our thesis was to establish and develop statistical methods that allows us to include the high-dimensional CNA genomic-windows in the Cox PH model as smooth functions. In the literature, there are existing approach to include the high-dimensional data in the standard Cox PH model, and no existing approach in the additive Cox PH model. Different strategies have been proposed for modifying the standard Cox PH model to deal with the high-dimensional setting. Some of these strategies are based on feature selection which can be either discrete or shrinkage. Discrete feature selection can be done by penalized univariate score test for each of the CNA genomic-windows. Variable selection by shrinkage is based on penalizing coefficients in the model, which leads to all of the spline coefficients for some variable being equal to zero. Identification of the nonlinear effect of the CNA genomic-windows enables us to estimate more accurately a patient's prognosis, and thus better determine NSCLC survival time. We addressed our objective by achieving the following points:

- In Chapter 2, we reviewed the method of estimating the CNA of our lung cancer dataset. The estimation of the CNA as ratio of tumour sample to the normal sample are obtained from each patients, which are recorded as rows in a matrix, with dimension 85×13253, where 85 is the number of patients and 13253 is the number of CNA genomic-windows.

- In Chapter 3, we presented the standard Cox PH model with some related concept, and we applied that to the clinical characteristics of the NSCLC data.

- Generalized Additive Models are presented in Chapter 4. We applied the logistic GAM to the clinical characteristics of the NSCLC data with a binary response variable.

- In Chapter 5, we presented a detailed discussion of new an extension to the standard Cox proportional hazard model. We demonstrated the use of radial basis function to represent the smoothing term in the additive Cox proportional hazard models. The clinical characteristics are considered as fixed predictors, and the genomic-wide CNA as smoothing terms in the model. The results of the smoothing effect can be expressed in terms of the log hazard ratio curve. Testing the hypothesis of no effect of the smoothing terms is carried out using the penalized version of the test statistic. Our proposed model used a grid search to select the optimal smoothing parameter value based on five-fold Cross-Validated partial log-likelihood (CVPL).

- In Chapter 6, we described the discrete variable selection method, to select subset of CNA genomic-windows that have a strong effect on the survival time. We generalized the univariate selection method explained in Bøvelstad et al. (2007), testing the CNA genomic-windows individually in the penalized additive Cox PH model using the penalized score test. We dealt with the dependence structure of the neighboring significant windows by selecting the correlation values between the significant CNA gemomic windows. We used the five-fold cross-validated partial log likelihood to perform the choice of the smoothing parameter. To improve the penalized univariate variable selection, we include the CNA genomic-windows in a multivariate penalized additive Cox PH model sequentially, analogous to forward stepwise selection. Finally, a clustering technique is used to identify the groups of the significant CNA genomic windows that have the similar log hazard ratio shapes around normal ploidy one.

- Another methods of the variable selection based on the shrinkage approach described in Marra and Wood (2011), are presented in Chapter 7. The smooth component can completely removed from the model based on its smoothing parameter value. Simulation studies to assess the shrinkage approach are included. The shrinkage approach is used for the variable selection of the significant CNA genomics windows. We compared both methods, forward stepwise selection and shrinkage selection.

8.2 Future Work

8.2.1 Selecting the Optimal Values of the Smoothing Parameters

The smoothing parameter plays an important role in the regularization procedure in the penalized additive Cox PH model. To choose the optimal values of the smoothing parameter, we need to specify a range of possible value for the smoothing parameters, and select the criterion in order to evaluate the model fit corresponding to each value. In this thesis, five-fold Cross-Validated Partial Log-likelihood (CVPL) procedure is performed for each value of the smoothing parameter, and the value of the smoothing parameter which maximizing the CVPL is selected. Optimizing all the smoothing parameter in the high-dimensional CNA genomic-windows is computational demanding, we only perform the two-dimensional grid searching to select the smoothing parameters that maximize CVPL. We fixed the smoothing parameter of the smoothing terms that been in the model for several iterations, and we only search for the last two smoothing parameters that are included in the model. In the following subsection, we present an example of four-dimensional searching.

Two Smoothing Terms in the Model Using 4 Dimensional Searching

The model is

$$h(t) = h_0(t) \exp \Big(f_1(\text{CNA}_{108}) + f_2(\text{CNA}_{949})) \Big). \tag{8.1}$$

We would like to obtain optimal values of λ_1, λ_2, ϵ_1, and ϵ_2 using a four dimensional grid searching. The eigenvalues of the first penalty matrix are 15.867, 1.927, 0.511, and 0, and the eigenvalues of the second penalty matrix are 24.561, 2.982, 0.324, and 0. Thus all possible values of ϵ_1 and ϵ_2 are between zero and 20% of the smallest non-zero eigenvalues that corresponding to the first and the second penalty respectively.

The optimal values are $\lambda_{1,opt} = 1.310$, $\epsilon_{1,opt} = 0.05$, $\lambda_{2,opt} = 0.655$, $\epsilon_{2,opt} = 0.034$, and CVPL -121.702. The optimal values of $\epsilon_{1,opt}$, and $\epsilon_{2,opt}$ are approximately 10% of the smallest eigenvalues of the second penalty matrix ($0.034/0.34 = 0.10$), ($0.05/0.51 = 0.09$).

Figure 8.1 represent the five fold CVPL values for different values of λ_1 and λ_2. The red point represent the optimal values for both λ_1 and λ_2. Figure 8.2 show CVPL for different values of ϵ_1 and ϵ_2. The red point represent the optimal values for both ϵ_1 and ϵ_2.

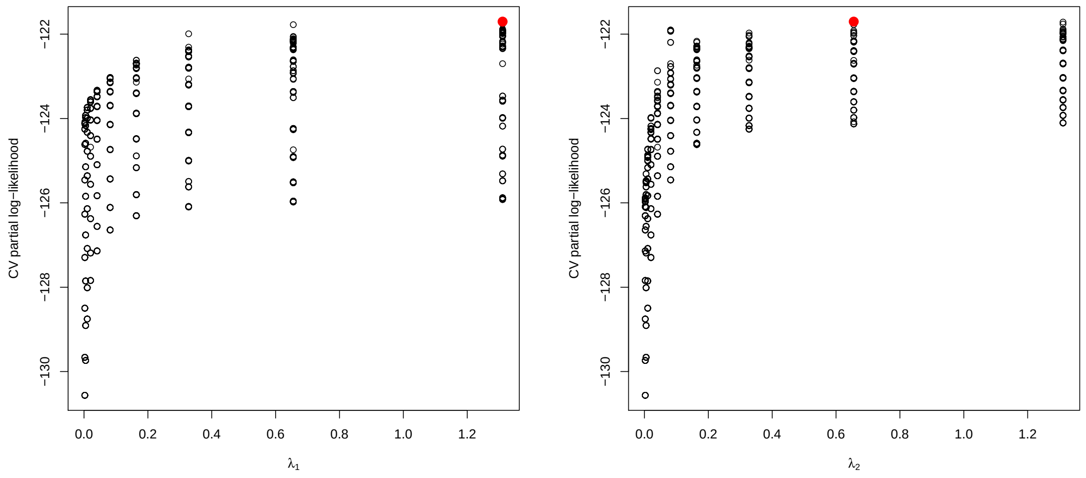


Figure 8.1: Five-fold CVPL for different values of λ_1 and λ_2.

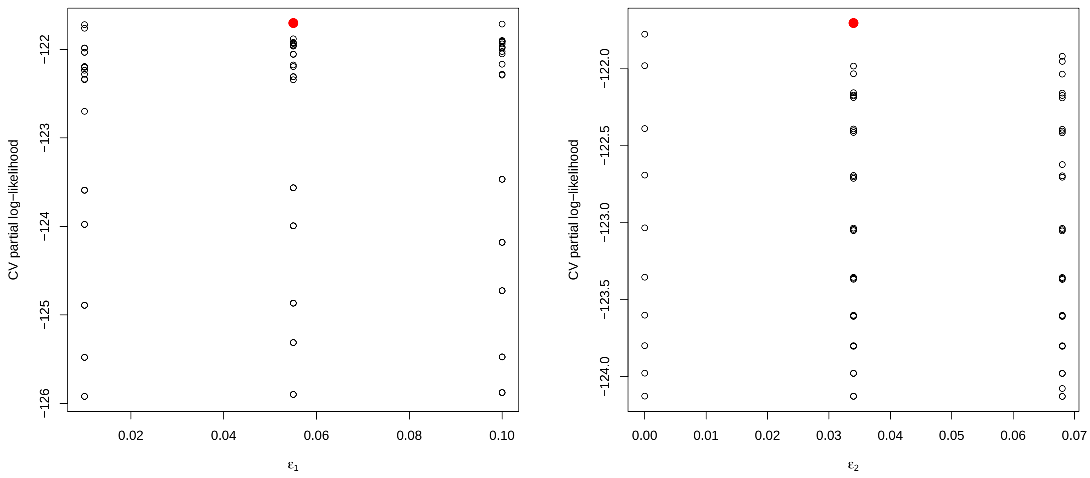


Figure 8.2: Five-fold CVPL for different values of ϵ_1 and ϵ_2.

However, the value of λ_1 is not maximizing CVPL. We increase the values of possible values of λ_1 and repeated the 4 dimensional grid searching, but again we did not find a optimal value of λ_1. Figure 8.3 represent the five fold CVPL values for different values of λ_1 and λ_2. The red point represent the optimal values for both λ_1 and λ_2. Figure 8.4 show CVPL for different values of ϵ_1 and ϵ_2. The red point represent the optimal values for both ϵ_1 and ϵ_2.

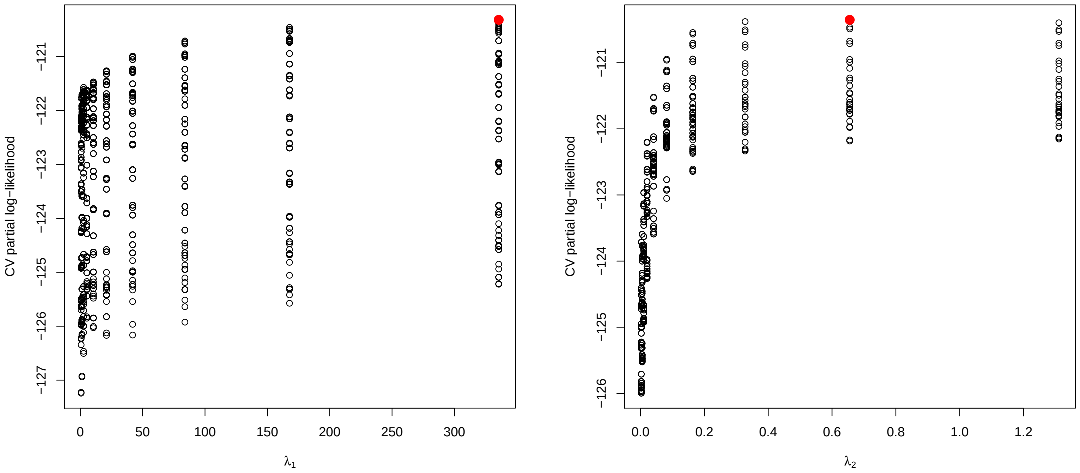


Figure 8.3: Five-fold CVPL for different values of λ_1 and λ_2.

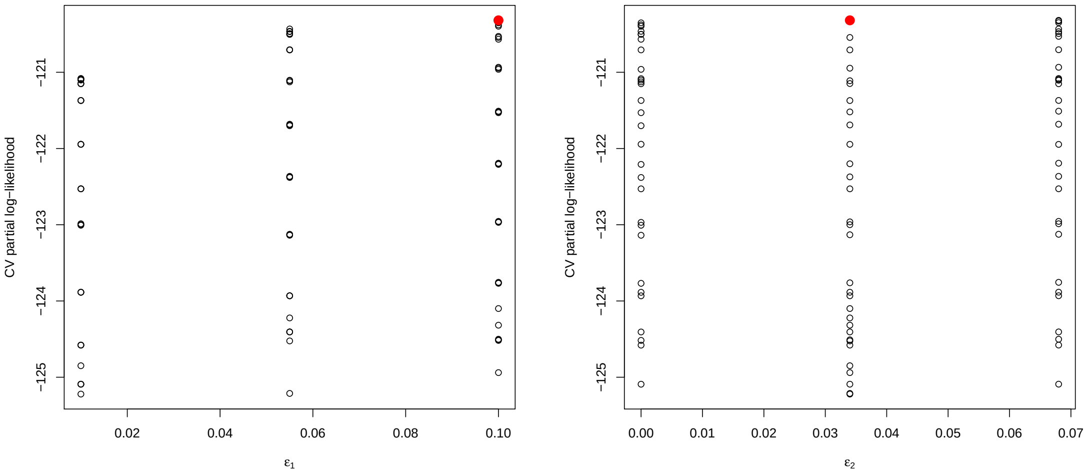


Figure 8.4: Five-fold CVPL for different values of ϵ_1 and ϵ_2.

We need to develop the effective algorithm that can obtain the optimal smoothing parameters for each smoothing term separately.

Bibliography

Amara, S., Majors, C., Roy, B., Hill, S., Rose, K., Myles, E., and Tiriveedhi, V. (2017). Critical role of SIK3 in mediating high salt and IL-17 synergy leading to breast cancer cell proliferation. *Plos One*, 12(6):1–21.

Belvedere, O., Berri, S., Chalkley, R., Conway, C., Barbone, F., Pisa, F., MacLennan, K., Daly, C., Alsop, M., Morgan, J., Menis, J., Tcherveniakov, P., Papagiannopoulos, K., Rabbitts, P., and Wood, H. (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(1):18–24.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723.

Benner, A., Zucknick, M., Hielscher, T., Ittrich, C., and Mansmann, U. (2010). High-dimensional Cox models: The choice of penalty as part of the model building process. *Biometrical Journal*, 52(1):50–69.

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J.-P., Janoueix-Lerosey, I., Delattre, O., and Barillot, E. (2011). Control-free calling of copy number alterations in deep-sequencing data using gc-content normalization. *Bioinformatics*, 27(2):268–269.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62.

Bøvelstad, H., Nygård, S., and Borgan, Ø. (2009). Survival prediction from clinico-genomic models - a comparative study. *Bioinformatics*, 10(1):413–413.

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

Bozkus, N. (2017). Lifting on clustering. *Biostatistics and Machine Learning in Omics Research, proceeding of the 34th LASR Workshop*, page 34.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30:89–99.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510.

Cadarso-Surez, C., Meira-Machado, L., Kneib, T., and Gude, F. (2010). Flexible hazard ratio curves for continuous predictors in multi-state models: an application to breast cancer data. *Statistical Modelling*, 10(3):291–314.

Chen, D.-H., Wu, Q.-W., Li, X.-D., Wang, S.-J., and Zhang, Z.-M. (2017a). SYPL1 overexpression predicts poor prognosis of hepatocellular carcinoma and associates with epithelial-mesenchymal transition. *Oncology Reports*, 38(3):1533–1542.

Chen, J.-Y., Xu, L., Fang, W.-M., Han, J.-Y., Wang, K., and Zhu, K.-S. (2017b). Identification of PA28 β as a potential novel biomarker in human esophageal squamous cell carcinoma. *Tumor Biology*, 39(10).

Chen, L., Ni, S., Li, M., Shen, C., Lin, Z., Ouyang, Y., Xia, F., Liang, L., Jiang, W., Ni, R., and Zhang, J. (2017c). High expression of BCCIP β can promote proliferation of esophageal squamous cell carcinoma. *Digestive Diseases and Sciences*, 62(2):387–395.

Chen, Y., Liu, T., Yu, C., Chiang, T., and Hwang, C. (2013). Effects of GC bias in next-generation-sequencing data on De Novo genome assembly. *Plos One*, 8(4):1–20.

Cheng, X., Hao, Y., Shu, W., Zhao, M., Zhao, C., Wu, Y., Peng, X., Yao, P., Xiao, D., Qing, G., Pan, Z., Yin, L., Hu, D., and Du, H. (2017). Cell cycle-dependent degradation of the methyltransferase SETD3 attenuates cell proliferation and liver tumorigenesis. *Journal of Biological Chemistry*, 292(22):9022–9033.

Codreanu, S., Hoeksema, M., Slebos, R., Zirnmerman, L., Rahman, S., Li, M., Chen, S., Chen, H., Eisenberg, R., Liebler, D., and Massion, P. (2017). Identification of proteomic features to distinguish benign pulmonary nodules from lung adenocarcinoma. *Journal of Proteome Research*, 16(9):3266–3276.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. CRC Press, Boca Raton, Fl, 2nd edition.

Couto, P., Bastos-Rodrigues, L., Schayek, H., Melo, F., Lisboa, R., Miranda, D., Vilhena, A., Bale, A., Friedman, E., and De Marco, L. (2017). Spectrum of germline mutations in smokers and non-smokers in brazilian non-small-cell lung cancer (NSCLC) patients. *Carcinogenesis*, 38(11):1112–1118.

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

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):248–275.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerischr Mathematik*, 31(4):377–403.

Dave, B., Granados-Principal, S., Zhu, R., Benz, S., Rabizadeh, S., Soon-Shiong, P., Yu, K.-D., Shao, Z., Li, X., Gilcrease, M., Lai, Z., Chen, Y., Huang, T. H. ., Shen, H., Liu, X., Ferrari, M., Zhan, M., Wong, S. T. C., Kumaraswami, M., Mittal, V., Chen, X., Gross, S. S., and Chang, J. C. (2014). Targeting RPL39 and MLF2 reduces tumor initiation and metastasis in breast cancer by inhibiting nitric oxide synthase signaling. *Proceedings of the National Academy of Sciences of the United States of America*, 111(24):8838–8843.

De Souza Santos, E., Bessa Garcia, S., M Netto, M., and Nagai, M. (2008). Silencing of LRRC49 and THAP10 genes by bidirectional promoter hypermethylation is a frequent event in breast cancer. *International Journal of Oncology*, 33(1):25–31.

Demichelis, F., Setlur, S. R., Banerjee, S., Chakravarty, D., Chen, J. Y. H., Chen, C. X., Huang, J., Beltran, H., Oldridge, D. A., Kitabayashi, N., Stenzel, B., Schaefer, G., Horninger, W., Bektic, J., Chinnaiyan, A. M., Goldenberg, S., Siddiqui, J., Regan, M. M., Kearney, M., Soong, T. D., Rickman, D. S., Elemento, O., Wei, J. T., Scherr, D. S., Sanda, M. A., Bartsch, G., Lee, C., Klocker, H., and Rubin, M. A. (2012). Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. *Proceedings of the National Academy of Sciences of the United States of America*, 109(17):6686–6691.

Du, P., Wang, S., Tang, X., AN, Chaoand Yang, Y., and Jiang, W. (2017). Reduced expression of metastasis suppressor-1 (MTSS1) accelerates progression of human bladder uroepithelium cell carcinoma. *Anticancer Research*, 37(8):4499–4505.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1):74–99.

Feliciano, A., Garcia-Mayea, Y., Jubierre, L., Mir, C., Hummel, M., Castellvi, J., Hernández-Losa, J., Paciucci, R., Sansano, I., Sun, Y., Cajal, R., Kondon, H., Soriano, A., Segura, M., Lyakhovich, A., and LLeonart, M. (2017). miR-99a reveals two novel oncogenic proteins E2F2 and EMR2 and represses stemness in lung cancer. *Cell Death and Disease*, 8(10):e3141.

Gleber-Netto, F., Zhao, M., Trivedi, S., Wang, J., Jasser, S., McDowell, C., Kadara, H., Zhang, J., Wang, J., N. William, W., Lee, J., Ly Nguyen, M., Pai, S., M. Walline, H., Shin, D., L. Ferris, R., Carey, T., Myers, J., and Pickering, C. (2018). Distinct pattern of TP53 mutations in human immunodeficiency virus-related head and neck squamous cell carcinoma. *Cancer*, 124(1):84–94.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526.

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(420):942–951.

Gray, R. J. (1994). Spline-based tests in survival analysis. *Biometrics*, 50(3):640–652.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall, New York, London.

Gusnanto, A., Wood, H. M., Pawitan, Y., Rabbitts, P., and 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(1):40–47.

Hastie, T. and Tibshirani, R. (1990a). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 46(4):1005–1016.

Hastie, T. J. and Tibshirani, R. (1990b). *Generalized Additive Models*. Chapman and Hall, London.

He, R., Zhang, F. H., and Shen, N. (2017). LncRNA FEZF1-AS1 enhances epithelial-mesenchymal transition (EMT) through suppressing E-cadherin and regulating WNT pathway in non-small cell lung cancer (NSCLC). *Biomedicine and Pharmacotherapy*, 95:331 – 338.

Heß, J., Thomas, G., Braselmann, H., Bauer, V., Bogdanova, T., Wienberg, J., Zitzelsberger, H., and Unger, K. (2011). Gain of chromosome band 7q11 in papillary thyroid carcinomas of young patients is associated with exposure to low-dose irradiation. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23):9595–9600.

Hofman, J., Kucera, R., Neumanova, Z., Klimes, J., Ceckova, M., and Staud, F. (2016). Placental passage of olomoucine II, but not purvalanol a, is affected by p-glycoprotein (ABCB1), breast cancer resistance protein (ABCG2) and multidrug resistance-associated proteins (ABCCs). *Xenobiotica*, 46(5):416–423.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Hong, E., Best, A., Gautrey, H., Chin, J., Razdan, A., Curk, T., Elliott, D. J., and Tyson-Capper, A. J. (2015). Unravelling the RNA-binding properties of SAFB proteins in breast cancer cells. *BioMed Research International*, 2015:1–9.

Huang, J., Gusnanto, A., O'Sullivan, K., Staaf, J., Borg, Å., and Pawitan, Y. (2007). Robust smooth segmentation approach for array CGH data analysis. *Bioinformatics*, 23(18):2463–2469.

Huang, J. and Harrington, D. (2002). Penalized partial likelihood regression for right-censored data with bootstrap selection of the penalty parameter. *Biometrics*, 58(4):781–791.

Huang, T., Wang, G., Yang, L., Peng, B., Wen, Y., Ding, G., and Wang, Z. (2017).

MiR-186 inhibits proliferation, migration, and invasion of non-small cell lung cancer cells by downregulating Yin Yang 1. *Cancer Biomarkers*, 21(1):221–228.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):271–293.

Jalkh, N., Chouery, E., Haidar, Z., Khater, C., Atallah, D., Ali, H., Marafie, M., Al-Mulla, M., Al-Mulla, F., and Megarbane, A. (2017). Next-generation sequencing in familial breast cancer patients from Lebanon. *BMC Medical Genomics*, 10(1):8.

Jiang, G., Wang, P., Wang, W., Li, W., Dai, L., and Chen, K. (2017). Annexin A13 promotes tumor cell invasion in vitro and is associated with metastasis in human colorectal cancer. *Oncotarget*, 8(13):21663–21673.

Jin, H., Jung, S., DebRoy, A., and Davuluri, R. (2016). Identification and validation of regulatory SNPs that modulate transcription factor chromatin binding and gene expression in prostate cancer. *Oncotarget*, 7(34):54616–54626.

Kay, R. (2004). An explanation of the hazard ratio. *Pharmaceutical Statistics*, 3(4):295–297.

Kettunen, E., Hernández-Vargas, H., Cros, M.-P., Durand, G., Calvez-Kelm, F., Stuopelytė, K., Jarmalaite, S., Salmenkivi, K., Anttila, S., Wolff, H., Herceg, Z., and Husgafvel-Pursiainen, K. (2017). Asbestos-associated genome-wide DNA methylation changes in lung cancer. *International Journal of Cancer*, 141(10):2014–2029.

Kiehl, S., Zimmermann, T., Savai, R., Pullamsetti, S., Seeger, W., Bartkuhn, M., and Dammann, R. (2017). Epigenetic silencing of downstream genes mediated by tandem orientation in lung cancer. *Scientific Reports*, 7(1):3896.

Klein, J. P. (2013). *Handbook of survival analysis*. CRC Press, Boca Raton, Florida.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival analysis: techniques for censored and truncated data*. Springer, New York.

Kleinbaum, D. G. and Klein, M. (2005). *Survival analysis: a self-learning text*. Springer, New York, NY, 2nd edition.

Kundu, S. T., Byers, L. A., Peng, D. H., Roybal, J. D., Diao, L., Wang, J., Tong, P., Creighton, C. J., and Gibbons, D. L. (2016). The miR-200 family and the miR-$183-96-182$ cluster target Foxf2 to inhibit invasion and metastasis in lung cancers. *Oncogene*, 35(2):173–186.

Lancaster, P. and Šalkauskas, K. (1986). *Curve and Surface Fitting: an Introduction*. Academic Press, London.

Lee, H., Yeh, B., Chan, T., Yang, K., Li, W., Huang, C., Ke, H., Li, C., Yeh, H., Liang, P., Shiue, Y., Wu, W., and Li, C. (2017a). Sulfatase-1 overexpression indicates poor prognosis in urothelial carcinoma of the urinary bladder and upper tract. *Oncotarget*, 8(29):47216–47229.

Lee, N., Kim, D.-K., Hyun Han, S., Ryu, H. G., Park, S. J., Kim, K.-T., and Yong Choi, K. (2017b). Comparative interactomes of VRK1 and VRK3 with their distinct roles in the cell cycle of liver cancer. *Molecules and Cells*, 40(9):621–631.

Liu, H. Y. and Zhang, C. J. (2017). Identification of differentially expressed genes and their upstream regulators in colorectal cancer. *Cancer Gene Therapy*, 24(6):244.

Liu, T., Sun, H., Zhu, D., Dong, X., Liu, F., Liang, X., Chen, C., Shao, B., Wang, M., Wang, Y., and Sun, B. (2017a). TRA2A promoted paclitaxel resistance and tumor progression in triple-negative breast cancers via regulating alternative splicing. *Molecular Cancer Therapeutics*, 16(7):1377–1388.

Liu, Z., Yanagisawa, K., Griesing, S., Iwai, M., Kano, K., Hotta, N., Kajino, T., Suzuki, M., and Takahashi, T. (2017b). TTF-1/NKX2-1 binds to DDB1 and confers replication stress resistance to lung adenocarcinomas. *Oncogene*, 36(26):3740–3748.

Mallows, C. L. (1973). Some comments on c_p. *Technometrics*, 15(4):661–675.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis*, 55(7):2372–2387.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, London, 2nd edition.

Meira-Machado, L., Cadarso-Suarez, C., Gude, F., and Araujo, A. (2013). smoothHR: An R package for pointwise nonparametric estimation of hazard ratio curves of continuous predictors. *Computational and Mathematical Methods in Medicine*, 2013:1–11.

Moreno-Smith, M., Halder, J., Meltzer, P., Gonda, T., Mangala, L., Rupaimoole, R., Lu, C., Nagaraja, A., Gharpure, K., Kang, Y., Rodriguez-Aguayo, C., Vivas-Mejia, P., Zand, B., Schmandt, R., Wang, H., Langley, R., Jennings, N., Ivan, C., Coffin, J., Armaiz, G., Bottsford-Miller, J., Kim, S., Halleck, M., Hendrix, M., Bornman, W., Bar-Eli, M., Lee, J., Siddik, Z., Lopez-Berestein, G., and Sood, A. (2013). ATP11B mediates platinum resistance in ovarian cancer. *Journal of Clinical Investigation*, 123(5):2119–2130.

Mustacchi, G., Sormani, M. P., Bruzzi, P., Gennari, A., Zanconati, F., Bonifacio, D., Monzoni, A., and Morandi, L. (2013). Identification and validation of a new set of five genes for prediction of risk in early breast cancer. *International Journal of Molecular Sciences*, 14:9686–702.

Nabbi, A., McClurg, U., Thalappilly, S., Almami, A., Mobahat, M., Bismar, T., Binda, O., and T. Riabowol, K. (2017). ING3 promotes prostate cancer growth by activating the androgen receptor. *BMC Medicine*, 15(1):103.

Nait Achour, T., Sentis, S., Teyssier, C., Philippat, A., Lucas, A., Corbo, L., Cavaillès, V., and Jalaguier, S. (2013). Transcriptional repression of estrogen receptor α signaling by SENP2 in breast cancer cells. *Molecular Endocrinology*, 28(2):183–196.

Nan, B., Lin, X., Lisabeth, L. D., and Harlow, S. D. (2005). A varying-coefficient Cox model for the effect of age at a marker event on age at menopause. *Biometrics*, 61(2):576–583.

Nogimori, K., Hori, T., Kawaguchi, K., Fukui, T., Mii S, Nakada, H., Matsumoto, Y., Yamauchi, Y., Takahashi, M., Furukawa, K., Tetsuya, O., Yokoi, K., Hasegawa, Y., and Furukawa, K. (2016). Increased expression levels of PpGalNAc-T13 in lung cancers: Significance in the prognostic diagnosis. *International Journal Oncology*, 49(4):1369–76.

Ohno, Y., Koizumi, M., Nakayama, H., Watanabe, T., Hirooka, M., Tokumoto, Y., Kuroda, T., Abe, M., Fukuda, S., Higashiyama, S., Kumagi, T., and Hiasa, Y. (2017). Downregulation of ANP32B exerts anti-apoptotic effects in hepatocellular carcinoma. *Plos One*, 12(5):e0177343.

Okayama, A., Kimura, Y., Miyagi, Y., Oshima, T., Oshita, F., Ito, H., Nakayama, H., Nagashima, T., Rino, Y., Masuda, M., Ryo, A., and Hirano, H. (2016). Relationship between phosphorylation of sperm-specific antigen and prognosis of lung adenocarcinoma. *Journal of Proteomics*, 139:60–66.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *Society for Industrial and Applied Mathematics. SIAM Journal on Scientific and Statistical Computing*, 9(3):531.

Peng, C., Lou, H.-L., Liu, F., Shen, J., Lin, X., Zeng, C.-P., Long, J.-R., Su, K.-J., Zhang, L., Greenbaum, J., Deng, W.-F., Li, Y.-M., and Deng, H.-W. (2017a). Enhanced identification of potential pleiotropic genetic variants for bone mineral density and breast cancer. *Calcified Tissue International*, 101(5):489–500.

Peng, Y., Cao, J., Yao, X., Wang, J., Zhong, M., Gan, P., and Li, J. (2017b). TUSC3 induces autophagy in human non-small cell lung cancer cells through wnt/β catenin signaling. *Oncotarget*, 8(32):52960–52974.

Perez-Ramirez, C., Cañadas Garre, M., Molina-Vila, M., Robles, A., José Faus-Dáder, M., and Ángel Calleja-HernáFndez, M. (2017). Contribution of genetic factors to platinum-based chemotherapy sensitivity and prognosis of non-small cell lung cancer. *Mutation Research-Reviews in Mutation Research*, 771:32–58.

Poirier, D. (1973). Piecewise regression using cubic splines. *Journal of the American Statistical Association*, 68(343):515–524.

Prasad, C., Prasad, S., Yadav, S., Pandey, L., Singh, S., Pradhan, S., and Narayan, G. (2017). Olaparib modulates DNA repair efficiency, sensitizes cervical cancer cells to cisplatin and exhibits anti-metastatic property. *Scientific Reports*, 7:1–5.

Qu, L.-S., Jin, F., Guo, Y.-M., Liu, T.-T., Xue, R.-Y., Huang, X.-W., Xu, M., Chen, T.-Y., Ni, Z.-P., and Shen, X.-Z. (2016). Nine susceptibility loci for hepatitis B virus-related hepatocellular carcinoma identified by a pilot two-stage genome-wide association study. *Oncology Letters*, 11(1):624–432.

Rodriguez, V., Chen, Y., Elkahloun, A., Dutra, A., Pak, E., and Chandrasekharappa, S. (2007). Chromosome 8 BAC array comparative genomic hybridization and expression analysis identify amplification and overexpression of TRMT12 in breast cancer. *Genes Chromosomes and Cancer*, 46(7):694–707.

Roskoski, R. (2017). ROS1 protein-tyrosine kinase inhibitors in the treatment of ROS1 fusion protein-driven non-small cell lung cancers. *Pharmacological Research*, 121:202–212.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Uinversity Press.

Sandoval-Borquez, A., Polakovicova, I., Carrasco-Veliz, N., Lobos-Gonzalez, L., Riquelme, I., Carrasco-Avino, G., Bizama, C., Norero, E., Owen, G., Roa, J., and

Corvalan, A. (2017). MicroRNA-335-5p is a potential suppressor of metastasis and invasion in gastric cancer. *Clinical Epigenetics*, 17(9):114.

Satow, R., Inagaki, S., Kato, C., Shimozawa, M., and Fukami, K. (2017). Identification of zinc finger protein of the cerebellum 5 as a survival factor of prostate and colorectal cancer cells. *Cancer Science*, 108(12):2405–2412.

Schemper, M. (1992). Cox analysis of survival data with non-proportional hazard functions. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 41(4):455–465.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241.

Schultz, D., Muluhngwi, P., Alizadeh-Rad, N., Green, M., Rouchka, E., Waigel, S., and Klinge, C. (2017). Genome-wide miRNA response to anacardic acid in breast cancer cells. *Plos One*, 12(9):1–29.

Segal, M. (2006). Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. *Biostatistics*, 7(2):268–285.

Shi, Y.-X., Yin, J.-Y., Shen, Y., Zhang, W., Zhou, H.-H., and Liu, Z.-Q. (2017). Genome-scale analysis identifies NEK2, DLGAP5 and ECT2 as promising diagnostic and prognostic biomarkers in human lung cancer. *Scientific Reports*, 7(1):8072.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Sleeper, L. A. and Harrington, D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, 85(412):941–949.

Spruance, S., Reid, J., Grace, M., and Samore, M. (2004). Hazard ratio in clinical trials. *Antimicrobial Agents and Chemotherapy*, 48(8):2787–2792.

Stopsack H, K., Gerke, T., Andrn, O., Andersson, S.-O., L Giovannucci, E., Mucci, L., and R Rider, J. (2017). Cholesterol uptake and regulation in high-grade and lethal prostate cancers. *Carcinogenesis*, 38(8):806–811.

Strasak, A. M., Lang, S., Kneib, T., Brant, L. J., Klenk, J., Hilbe, W., Oberaigner, W., Ruttmann, E., Kaltenbach, L., Concin, H., Diem, G., Pfeiffer, K. P., Ulmer, H., Grp, V. S., and Group, V. S. (2009). Use of penalized splines in extended Cox-type additive hazard regression to flexibly estimate the effect of time-varying serum uric acid on risk of cancer incidence: A prospective, population-based study in 78,850 men. *Annals of Epidemiology*, 19(1):15–24.

Sutton, L.-A., Ljungström, V., Mansouri, L., Young, E., Cortese, D., Navrkalova, V., Malcikova, J., Muggen, A., Trbusek, M., panayiotidis, P., Davi, F., Belessi, C., Langerak, A., Ghia, P., Pospisilova, S., Stamatopoulos, K., and Rosenquist, R. (2015). Targeted next-generation sequencing in chronic lymphocytic leukemia: A high-throughput yet tailored approach will facilitate implementation in a clinical setting. *Haematologica*, 100(3):370–376.

Tang, R.-X., Chen, W.-J., He, R.-Q., Zeng, J.-H., Liang, L., Li, S.-K., Ma, J., Luo, D.-Z., and Chen, G. (2017). Identification of a RNA-seq based prognostic signature with five lncRNAs for lung squamous cell carcinoma. *Oncotarget*, 8(31):50761–50773.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 12(1):156–175.

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

Tsujitani, M. and Tanaka, Y. (2013). Analysis of heart transplant survival data using generalized additive models. *Computational and Mathematical Methods in Medicine*, 2013:609857.

Tsujitani, M., Tanaka, Y., and Sakon, M. (2012). Survival data analysis with time-dependent covariates using generalized additive models. *Computational and Mathematical Methods in Medicine*, 2012(986176):1–9.

Uehiro, N., Sato, F., Pu, F., Tanaka, S., Kawashima, M., Kawaguchi, K., Sugimoto, M., Saji, S., and Toi, M. (2016). Circulating cell-free DNA-based epigenetic assay can detect early breast cancer. *Breast Cancer Research*, 18(1):129.

Upadhyay, P., Gardi, N., Desai, S., Chandrani, P., Joshi, A., Dharavath, B., Arora, P., Bal, M., Nair, S., and Dutt, A. (2017). Genomic characterization of tobacco/nut chewing HPV-negative early stage tongue tumors identify MMP10 as a candidate to predict metastases. *Oral Oncology*, 73:56–64.

Van Houwelingen, H. and Putter, H. (2012). *Dynamic prediction in clinical survival analysis*. CRC Press, Boca Raton.

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

Van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van't Veer, L. J., and Wessels, L. F. A. (2006). Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*, 25(18):3201–3216.

Vikberg, A.-L., Vooder, T., Lokk, K., Annilo, T., and Golovleva, I. (2017). Mutation analysis and copy number alterations of KIF23 in non-small-cell lung cancer exhibiting KIF23 over-expression. *Onco Targets and Therapy*, 10:4969–4979.

Volinsky, C. and Raftery, A. (2000). Bayesian information criterion for censored survival models. *Biometrics*, 56(1):256–262.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, Pa.

Wahba, G., Golub, G. H., and Heath, M. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Walker C, L., Marquart, L., F Pearson, J., Wiggins, G., O'Mara, T., T Parsons, M., Barrowdale, D., McGuffog, L., Dennis, J., BenÃtez, J., P Slavin, T., Radice, P., Frost, D., Godwin, A., Meindl, A., Katharina Schmutzler, R., Isaacs, C., N Peshkin, B., CaldÃs, T., and Spurdle, A. (2017). Evaluation of copy-number variants as modifiers of breast and ovarian cancer risk for brca1 pathogenic variant carriers. *European Journal of Human Genetics*, 25(4):432–438.

Wang, C., Liu, H., and Gao, S. (2017a). A penalized Cox proportinal hazard model with multiple time-varying exposures. *Annals of Applied Statistics*, 11(1):185–201.

Wang, H., Yan, C., Shi, X., Zheng, J., Deng, L., Yang, L., Yu, F., Shao, Y., and Lv, F. (2015). MicroRNA-575 targets BLID to promote growth and invasion of non-small cell lung cancer cells. *FEBS Letters*, 589(7):805–811.

Wang, Q., Chen, Q., Zhu, L., Chen, M., Xu, W., Panday, S., Wang, Z., Li, A., Re, O., Chen, R., Wang, S., Zhang, R., and Zhou, J.-w. (2017b). JWA regulates trail-induced apoptosis via MARCH8-mediated DR4 ubiquitination in cisplatin-resistant gastric cancer cells. *Oncogenesis*, 6(7):e353.

Wilson, C., Qiu, L., Hong, Y., Karnik, T., Tadros, G., Mau, B., Ma, T., Mu, Y., New, J., J Louie, R., Gunewardena, S., Godwin, A., W Tawfik, O., Chien, J., F Roby, K., and

J Krieg, A. (2016). The histone demethylase KDM4B regulates peritoneal seeding of ovarian cancer. *Oncogene*, 36(18):2565–2576.

Wood, S. N. (2006). *Generalized Additive Models: an Introduction with R*. Chapman and Hall CRC.

Wood, S. N. and Augustin, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 157(2):157–177.

Wood, S. N., Pya, N., and Sfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.

Xie, K., Zhu, M., Xiang, P., Chen, X., Kasimumali, A., Lu, R., Wang, Q., Mou, S., Ni, Z., Gu, L., and Pang, H. (2017). Protein kinase A/CREB signaling prevents adriamycin-induced podocyte apoptosis via upregulation of mitochondrial respiratory chain complexes. *Molecular and Cellular Biology*, 38(1):e00181–17.

Yan, H., Guan, Q., He, J., Lin, Y., Zhang, J., Li, H., Liu, H., Gu, Y., Guo, Z., and He, F. (2017). Individualized analysis reveals CpG sites with methylation aberrations in almost all lung adenocarcinoma tissues. *Journal of Translational Medicine*, 15(1):26.

Yuan He, P., Kien Yip, W., Lee Chai, B., Faisal Jabar, M., Dusa, N., Mohtarrudin, N., and Seow, H. (2017). Inhibition of cell migration and invasion by mir?29a?3p in a colorectal cancer cell line through suppression of CDC42BPA mRNA expression. *Oncology Reports*, 38(6):3554–3566.

Zhang, D., Jiang, P., Xu, Q., and Zhang, X. (2011). Arginine and glutamate-rich 1 (ARGLU1) interacts with mediator subunit 1 (MED1) and is required for estrogen receptor-mediated gene transcription and breast cancer cell growth. *The Journal of Biological Chemistry*, 286(20):17746–54.

Zhang, F., Gu, W., Hurles, M., and Lupski, J. (2009). Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, 10(1):451–481.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Zhang, M., Duan, T., Wang, L., Tang, J., Luo, R., Zhang, R., and Kang, T. (2016). Low expression of centrosomal protein 78 (CEP78) is associated with poor prognosis of colorectal cancer patients. *Chinese Journal of Cancer*, 35(10):509–517.

Zhang, X., Fu, L., Xue, D., Zhang, X., Hao, F., Xie, L., He, J., Gai, J., Liu, Y., Xu, H., Li, Q., and Wang, E. (2017). Overexpression of Rsf-1 correlates with poor survival and promotes invasion in non-small cell lung cancer. *Virchows Archiv*, 470:553–560.

Zhang, Z., S Gerhard, D., Nguyen, L., Li, J., Traugott, A., C Huettner, P., and Rader, J. (2005). Fine mapping and evaluation of candidate genes for cervical cancer on 11q23. *Genes Chromosomes and Cancer*, 43(1):95–103.

Zoheir, K., Abd-Rabou, A., Harisa, G., Kumar, A., Fayaz Ahmad, S., Ansari, M., and Abd-Allah, A. (2016). IQGAP1 gene silencing induces apoptosis and decreases the invasive capacity of human hepatocellular carcinoma cells. *Tumor Biology*, 37(10):13927–13939.

Appendices

Appendix A

Significant CNA genomics windows

A.1 Significant CNA Genomics Windows in the Penalized Univariate Selection

Chr	total number of windows	number of significant window	windows
Chr 1	1109	21	108-118
			948-949
			953-956
			973
			976
			999
			1075
			1079
Chr2	1181	45	1815-1824
			1840-1846
			1903-1913
			1979-1992
			2282-2287
Chr 3	974	115	2291-2336
			2341
			2403-2410
			2679-3102
			3186-3221

Chr	total number of windows	number of significant window	windows
Chr4	934	57	3474
			3679-3697
			3715-3716
			3718
			3794-3996
			4001-4006
Chr5	873	12	4528-4533
			4614-4620
Chr 6	839	77	5317
			5157
			5177
			5181
			5209
			5213-5236
			5481-5489
			5507-5538
			5798-5804
Chr 7	769	131	6253
			6280
			6284
			6327-6341
			6405
			6410-6505
			6513-6528
Chr8	708	66	6742-6754
			6992-7009
			7131-7137
			7200
			7264-7282
			7378-7386
Chr 9	549	0	
Chr 10	643	47	8125
			8258-8269
			8321
			8323
			8326
			8443-8445
			8526-8528
			8543-8566
			8568

Chr	total number	number of	windows
	of windows	significant window	
Chr 11	655	109	8804-8806
			888-8890
			8928-8965
			9024-9029
			9070-9084
			9119-9120
			9143-9182
Chr 12	653	46	9316-9325
			9508-9521
			9538
			9579-9598
			9649-9654
			9686-9691
			9732
			9752-9755
Chr 13	471	78	9002-9003
			9911-9919
			9945-9948
			10218-10231
			10278-10324
			10351-10352
Chr 14	437	69	10418-10422
			10517
			10546-10553
			10596-10619
			10639-10644
			10738-10762
Chr 15	394	77	10964-10967
			10991-11011
			11023-11050
			11125-11147
			11169-11170
Chr 16	381	25	11217-11220
			11257-11358
			11376
			11448-11454
			11473-11479
			11565-11568

Chr	total number of windows	number of significant window	windows
Chr 17	388	29	11581
			11806
			11808-11821
			11827-11839
Chr 18	374	4	12104-12107
Chr 19	280	39	12356-12374
			12391-12408
			1242-12425
Chr 20	300	9	12614-12617
			12882-12886
Chr21	174	0	
Chr22	169	0	
total	13253	1056	

A.2 The Common significant CNA Genomic-Windows for both Univariate Variable Selection and Penalized Univariate Variable Selection

Chr	total number of windows	number of significant window	windows
Chr 3	974	80	2300-2336
			2403-2410
			3190-3221
Chr 6	839	14	5175-5181
			5209
			5213
			5217-5227
			5229
Chr 10	643	12	8258-8269
Chr 11	655	30	9070-9076
			9077
			9079
			9154-9174
Chr 13	471	2	10351-10352
Chr 15	394	13	11134-11147
Chr 17	388	11	11813-11821
			11827-11829