ABSTRACT

Title of dissertation: Regularized Variable Selection in Proportional Hazards Model Using Area under Receiver Operating Characteristic Curve Criterion

Wen-Chyi Wang
Doctor of Philosophy, 2009

Dissertation directed by: Professor Grace L. Yang
Department of Mathematics

The goal of this thesis is to develop a statistical procedure for selecting pertinent predictors among a number of covariates to accurately predict the survival time of a patient. There are available many variable selection procedures in the literature. This thesis is focused on a more recently developed "regularized variable selection procedure". This procedure, based on a penalized likelihood, can simultaneously address the problem of variable selection and variable estimation which previous procedures lack. Specifically, this thesis studies regularized variable selection procedure in the proportional hazards model for censored survival data.

Implementation of the procedure requires judicious determination of the amount of penalty, a regularization parameter λ, on the likelihood and the development of computational intensive algorithms. In this thesis, a new criterion of determining λ using the notion of "the area under the receiver operating characteristic curve (AUC)" is proposed. The conventional generalized cross-validation criterion (GCV) is based on the likelihood and its second derivative. Unlike GCV, the AUC criterion

is based on the performance of disease classification in terms of patients' survival times. Simulations show that performance of the AUC and the GCV criteria are similar. But the AUC criterion gives a better interpretation of the survival data.

We also establish the consistency and asymptotic normality of the regularized estimators of parameters in the partial likelihood of proportional hazards model. Some oracle properties of the regularized estimators are discussed under certain sparsity conditions. An algorithm for selecting λ and computing regularized estimates, $\hat{\boldsymbol{\beta}}$, is developed. The developed procedure is then illustrated with an application to the survival data of patients who have cancers in head and neck. The results show that the proposed method is comparable with the conventional one.

Regularized Variable Selection in Proportional Hazards Model Using Area under Receiver Operating Characteristic Curve Criterion

by

Wen-Chyi Wang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:

Professor Grace L. Yang, Chair/Advisor
Professor Paul J. Smith
Professor Eric V. Slud
Professor Ming T. Tan
Professor Yunfeng Zhang

DEDICATION

To my parents

ACKNOWLEDGMENTS

I would like to express my deepest appreciation and respect to my advisor, Dr. Grace Yang, for her guidance, advice, thoughtful consideration and persistent help. She carefully read and improved all of my writing, provided me opportunities to enrich my knowledge, and even spent numerous weekends discussing with me when she was on sabbatical. She is the talented and successful female role model for me. Without her support, this dissertation would not have been possible.

I am also grateful to Dr. Paul Smith, Dr. Eric Slud, Dr. Ming Tan, Dr. Yunfang Zhang and the late Dr. Ryszard Syski for their valuable time and insightful comments as members of my preliminary and final oral committees. Dr. Paul Smith is generously available to all students. I very much appreciate his encouragement and assistance to improve my research and to strengthen my application materials for fellowships. I would also like to express my gratitude to Dr. Ming Tan who enhanced my research experience, gave me judicious advice and provided me with the head and neck cancer data from Greenebaum Cancer Center at the University of Maryland at Baltimore to complete this research.

Specially thanks to Dr. Hongbin Fang for his perceptive suggestions and creative ideas. His inspiration on my study was of vital importance. I also very much appreciate the needed graduate assistantships and fellowships respectively supported by Dr. Ming Tan and Dr. Hongbin Fang, Department of Mathematics and Graduate

School at the University of Maryland.

I would like to thank my lovely friends, Mei-Yin Chou, Chun-Fang Chiang, Jeng-Daw Yu and Mindy Chuang who comforted and accompanied me through the darkest days at Maryland, Joyce Wang who encouraged and cheered up me in the last but the most important stage of writing, and Hwa-Lung Yu who made me laugh and forget my worries when I was down and frustrated.

Finally, the endless love and patience of my father and my mother are the greatest support for me in this long journey. They never complained about their only child staying thousands of miles away from them but had full confidence that I would succeed.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

In recent years there has been growing interest in the study of regularization for simultaneously carrying out variable selection and coefficient estimation in linear or non-linear regression models. The case under study in this thesis is that among a large member of variables (regressors) available to us, we wish to select a relatively small subset of significant variables to construct a model for analysis. The approach of many traditional variable selection techniques, such as forward selection, backward elimination and subset selection, is to select an "estimated model" of significant variables from a number of candidate models. To be concrete, take the multiple linear regression model

$$\mathbf{Y} = X\boldsymbol{\beta} + \varepsilon$$

as an example, assume that there are k unknown regression coefficients denoted by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$. Unlike these traditional selection methods, the regularized method will simultaneously perform variable selection by setting some estimated coefficients (in $\boldsymbol{\beta}$) zero and estimate other coefficients using shrinkage method in the sense of ridge regression. Those variables X with coefficients estimated to be zero are considered as insignificant variables.

The regularized estimates $\hat{\boldsymbol{\beta}}$ are the constrained minimizers of the sum of squared residuals:

$$\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 \text{ subject to } P(\boldsymbol{\beta}) \leq c, \tag{1.1}$$

where c is a specified constant. $P(\boldsymbol{\beta})$ is called a penalty function of $\boldsymbol{\beta}$. When

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{k} |\beta_j|, \tag{1.2}$$

one obtains the well-known least absolute shrinkage selection operator (Lasso) penalty (Tibshirani, 1996). The Lasso penalization approach is also called basis pursuit in signal processing (Chen, Donoho and Saunders, 2001).

For investigations and computations, it is convenient to express (1.1) in terms of the Lagrange multiplier λ. The optimizer from (1.1) is equivalent to

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + \lambda P(\boldsymbol{\beta}) \right\}. \tag{1.3}$$

The parameter λ is determined by c of (1.1) and vice versa, and is usually called a tuning parameter or a regularization parameter. The choice of an appropriate tuning parameter is related to how much prediction accuracy we pursue. There are several criteria used in the literature for selecting λ, for example, Akaike's information criterion (AIC), Bayesian information criterion (BIC), C_p criterion (Efron, Hastie, Johnstone and Tibshirani, 2004) and the general cross-validation (GCV) criterion (Tibshirani, 1996). In this thesis, we focus on the GCV and a new criterion called the AUC criterion.

Regularized methods in linear models have been modified and extended in recent years. These and their theoretical investigations can be found, e.g. in Fan

and Li (2001), Frank and Friedman (1993), Shen and Ye (2002) and Zou (2006). In particular, when the design matrix X is orthonormal, a closed form of regularized estimator $\hat{\boldsymbol{\beta}}$ can be written as a thresholding function of the ordinary least squares (OSL) estimator, in the terminology of wavelet theory (Donoho and Johnstone,1994). Knight and Fu (2000) considered the asymptotic behavior of the L_p estimator. Fan and Li (2001) discussed the oracle properties of the regularized estimator. Yuan and Lin (2005) connected the Lasso estimator with a particular hierarchical Bayesian framework.

The regularization method has been applied to the proportional hazards model in which the regularized estimator $\hat{\boldsymbol{\beta}}$ is obtained by maximizing the log partial likelihood $\ell(\boldsymbol{\beta}, t)$ subject to a constraint. In terms of the Lagrange multiplier λ, the maximizer is given by

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \left\{ \ell(\boldsymbol{\beta}, t) - \lambda P(\boldsymbol{\beta}) \right\}. \tag{1.4}$$

Compared with the linear models, there are fewer articles discussing the regularization in the proportional hazards model (Tibshirani, 1997; Fan and Li, 2002; Gui and Li, 2005). This motivated us to focus on the development of regularized methods in proportional hazards model with a goal of giving accurate prediction of a patient's survival time.

1.2 Organization of the Dissertation

In Chapter 2, we discuss the regularization in the linear regression model and define the regularized estimator $\hat{\boldsymbol{\beta}}$. We review several estimators $\hat{\boldsymbol{\beta}}$ studied in the

literature that were obtained by different types of regularization including the L_p, the HARD and the SCAD. The L_p regularization includes the Lasso, the adaptive Lasso and the ridge regression estimator as special cases. An explicit form is available for each $\hat{\boldsymbol{\beta}}$ a graphical comparison of these estimates are made. These graphs show that the tuning parameter λ affects the size of the shrinkage. To emphasize the role of λ, we denote the estimate $\hat{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{\beta}}_\lambda$. A popular criterion to determine the value of λ is the GCV. We review the GCV and show how it is used to determine $\hat{\boldsymbol{\beta}}$ in the linear regression model.

In Chapter 3, we set out to develop a good method of predicting survival times, where the survival time is modeled by the proportional hazards model with parameter $\boldsymbol{\beta}$. We study the regularization in the proportional hazards model in which the estimator $\hat{\boldsymbol{\beta}}_\lambda$ is the maximizer of the regularized log partial likelihood function (see (1.4)). Under certain conditions, the consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}_\lambda$ with a fixed tuning parameter λ are proved and the oracle properties of $\hat{\boldsymbol{\beta}}_{\lambda_n}$ with n-dependent nonrandom λ are discussed. The tuning parameter λ affects the shrinkage of $\hat{\boldsymbol{\beta}}_\lambda$ in the proportional hazards model as well, but the GCV used in the linear model may not be an appropriate criterion for selecting λ.

In Chapter 4, a new criterion for selecting λ is proposed. This criterion, based on the receiver operating characteristic (ROC) curve, takes into consideration the maximum diagnostic performance of the model. The diagnostic performance is measured by AUC which stands for the area under the ROC curve. In predicting a patient's survival time, we generalize AUC to make it dependent on the survival time u. A time-dependent AUC(u) is used to develop a method of selecting the

tuning parameter λ. The regularized estimator $\hat{\boldsymbol{\beta}}_\lambda$ is selected by the maximum of the estimated AUC(u) value. Both parametric and nonparametric estimations are discussed.

In Chapter 5, we develop an algorithm for computing the regularized estimator $\hat{\boldsymbol{\beta}}_\lambda$ in the proportional hazards model. The algorithm allows the use of either the AUC or the GCV criteria to select the tuning parameter λ. We compare numerically the performance of the Lasso estimators $\hat{\boldsymbol{\beta}}_\lambda$ in three different scenarios using simulated data. Comparison is also made of different selecting methods. All programming codes for this chapter are written in the R language and are given in Appendix B. These codes include calculating the first two derivatives of the log partial likelihood function for a given $\boldsymbol{\beta}$, estimating of the Lasso estimator $\hat{\boldsymbol{\beta}}_\lambda$, calculating the GCV value with a given tuning parameter λ and the AUC(u) value with a given λ and a time u, and generating censored survival data.

Chapter 6 studies a real data set of survival times of patients who have squamous cell carcinoma. Some observations are right censored. We give a description of the data including how they are collected, what medical indexes in the data mean and a summary of patients' information. We then carry out the data analysis using the method and the algorithm developed in this thesis. The Lasso estimator of $\boldsymbol{\beta}$ is computed using both the AUC and the GCV criteria to select significant predicting variables.

Conclusions are given in Chapter 7.

Chapter 2

Regularization in Linear Regression

2.1 Definitions

Consider the multiple linear regression

$$\mathbf{y} = \beta_1 x_1 + \ldots + \beta_k x_k + \mathbf{e}, \tag{2.1}$$

where $\mathbf{y}$ is an observable random variable, x_i are known nonrandom regressors, β_i are unknown parameters for $i = 1, \ldots, k$, and $\mathbf{e}$ is a random error with mean 0 and finite variance σ^2.

Suppose we have a random sample of n independent observations $(\mathbf{y}_i, x_{i1}, x_{i2}, \ldots, x_{ik})$ from model (2.1). That is,

$$\mathbf{y}_i = \sum_{j=1}^{k} x_{ij}\beta_j + \mathbf{e}_i \;\; i = 1, \ldots, n.$$

In vector form, we denote the random sample and the regression model by

$$\mathbf{Y} = X\boldsymbol{\beta} + \varepsilon$$

where

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{pmatrix},$$

are respectively, an n-dimensional random vector, an $n \times k$ design matrix, a vector of unknown parameters, and a random error vector. It is assumed that $X^T X$ is a $k \times k$ non-singular matrix, and that ε has mean $\mathbf{0}$ and covariance $\sigma^2 I$, where $\mathbf{0}$ is an $n \times 1$ zero vector and I is an $n \times n$ identity matrix.

Let $P(\boldsymbol{\beta})$ denote a continuous function of $\boldsymbol{\beta}$ which serves as a penalty (or regularization) function in the estimation of $\boldsymbol{\beta}$. The function $P(\boldsymbol{\beta})$ is differentiable and takes positive values at all nonzero points. At the origin, $\boldsymbol{\beta} = \mathbf{0}$, $P(\boldsymbol{\beta})$ is zero and may be non-differentiable.

The regularized estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is defined as the constrained minimizer of the sum of squared residuals (RSS):

$$\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 \text{ subject to } P(\boldsymbol{\beta}) \leq c, \tag{2.2}$$

where $\|\cdot\|$ is the L_2-norm, and c is a specified nonnegative constant. The constraint restricts the value of the estimator to the set $\{\boldsymbol{\beta} : P(\boldsymbol{\beta}) \leq c\}$.

For investigations and computations, it is convenient to express (2.2) in terms of the Lagrange multiplier λ. Then minimizing (2.2) is equivalent to that of minimizing the so called penalized (or regularized) RSS with respect to $\boldsymbol{\beta}$,

$$\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + \lambda P(\boldsymbol{\beta}), \tag{2.3}$$

where given c, there exists a λ which can be solved as a function of c and vice versa. This will be made precise when we consider specific penalty functions.

Throughout this chapter, we will use the ordinary least squares (OLS) estimator $\tilde{\boldsymbol{\beta}}$ in establishing properties for the regularized estimator $\hat{\boldsymbol{\beta}}$. The OLS estimator

$\tilde{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}$ is obtained by minimizing $\|\mathbf{Y} - X\boldsymbol{\beta}\|^2$ without imposing any constraint.

The following lemma proves that λ must be nonnegative.

Lemma 2.1 *In view of equation (2.3), for a nonzero OLS estimator $\tilde{\boldsymbol{\beta}}$, we have (a) if the penalty $P(\tilde{\boldsymbol{\beta}}) \leq c$, then $\lambda = 0$; (b) if $P(\tilde{\boldsymbol{\beta}}) > c$, then $\lambda > 0$.*

Proof: (a) It is obvious that $P(\tilde{\boldsymbol{\beta}}) \leq c$ if and only if $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$. This implies that $\frac{\partial}{\partial \beta} \|\mathbf{Y} - X\boldsymbol{\beta}\|^2|_{\beta=\hat{\beta}} = \frac{\partial}{\partial \beta} \|\mathbf{Y} - X\boldsymbol{\beta}\|^2|_{\beta=\tilde{\beta}} = \mathbf{0}$. Since $\frac{\partial}{\partial \beta} \left(\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + \lambda P(\boldsymbol{\beta})\right)|_{\beta=\hat{\beta}} = \mathbf{0}$, we have either $\lambda = 0$ or $\frac{\partial}{\partial \beta} P(\boldsymbol{\beta})|_{\beta=\hat{\beta}} = \mathbf{0}$. By the definition of $P(\boldsymbol{\beta})$, $\mathbf{0}$ is the minimum. Therefore we have $\lambda = 0$ because $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$ is nonzero.

(b) If $\lambda < 0$, then $\lambda P(\tilde{\boldsymbol{\beta}}) < \lambda c \leq \lambda P(\hat{\boldsymbol{\beta}})$. Since $\|\mathbf{Y} - X\tilde{\boldsymbol{\beta}}\|^2 \leq \|\mathbf{Y} - X\boldsymbol{\beta}\|^2$ for all $\boldsymbol{\beta}$, we have $\|\mathbf{Y} - X\tilde{\boldsymbol{\beta}}\|^2 + \lambda P(\tilde{\boldsymbol{\beta}}) < \|\mathbf{Y} - X\hat{\boldsymbol{\beta}}\|^2 + \lambda P(\hat{\boldsymbol{\beta}})$ lead to a contradiction. □

By Lemma 2.1, the tuning parameter $\lambda \geq 0$. Therefore $\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + \lambda P(\boldsymbol{\beta})$ tends to infinity as $\|\boldsymbol{\beta}\| \to \infty$. This implies the minimizing solution $\hat{\boldsymbol{\beta}}$ of (2.3) exists. If $P(\boldsymbol{\beta})$ is a strictly convex function in $\boldsymbol{\beta}$, then $\hat{\boldsymbol{\beta}}$ is unique because $\|\mathbf{Y} - X\boldsymbol{\beta}\|^2$ is strictly convex as well. If $P(\boldsymbol{\beta})$ is not strictly convex, there is no guarantee that the unique global minimum exists. But under the special condition of orthonormal design matrix X, we shall, in the next section, present unique solutions of some regularized estimators for several (convex or non-convex) penalty functions.

The following are two examples of regularized estimators.

Example 2.1 *(Lasso). Let $P(\boldsymbol{\beta}) = \sum_{j=1}^{k} |\beta_j|$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$. The mini-*

mizer of (2.3),

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right\},$$

is called the Lasso estimator of $\boldsymbol{\beta}$ (Tibshirani, 1996).

Example 2.2 *(Ridge regression). Let $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|^2 = \sum_{j=1}^{k} \beta_j^2$. Let I be a $k \times k$ identity matrix. The minimizer of (2.3),*

$$\hat{\boldsymbol{\beta}} = (X^T X + \lambda I)^{-1} X^T \mathbf{Y},$$

is the well-studied ridge regression estimator (Frank and Friedman, 1993).

2.2 Penalty Functions

In this section we shall present various penalty functions $P(\boldsymbol{\beta})$ and the associated regularized estimators $\hat{\boldsymbol{\beta}}$ discussed in the literature. The literature on this subject is huge. Particularly relevant to our investigation are papers by Fan and Li (2001), Frank and Friedman (1993), Fu (1998), Knight and Fu (2000), Tibshirani (1996) and Zou (2006). In the following, we consider a more general form of the penalty $P_\lambda(\boldsymbol{\beta})$ than the product $\lambda P(\boldsymbol{\beta})$; that is, we discuss the problem of minimizing

$$\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + P_\lambda(\boldsymbol{\beta}) \tag{2.4}$$

instead of (2.3). This assumption accommodates some of the literature in which the penalized function is not the product of λ and $P(\boldsymbol{\beta})$ but a function of λ and $\boldsymbol{\beta}$, for example, the HARD penalty in Section 2.2.2 and the SCAD penalty (Fan and Li, 2001) in Section 2.2.3.

Write (2.4) in the form of

$$\|\mathbf{Y} - X\tilde{\boldsymbol{\beta}}\|^2 + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T X^T X(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + P_\lambda(\boldsymbol{\beta}), \tag{2.5}$$

where $\tilde{\boldsymbol{\beta}}$ is the OLS estimator. If k columns of the design matrix X are orthonormal; that is, $X^T X$ equals an identity matrix I, then minimizing (2.5) is equivalent to that of minimizing

$$(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + P_\lambda(\boldsymbol{\beta}). \tag{2.6}$$

Orthonormalization simplifies the calculation and make it easier to design a penalty function for obtaining a shrinkage estimator $\hat{\boldsymbol{\beta}}$. For instance, in Example 2.2, if $X^T X = I$, we obtain an explicit form for $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \frac{\tilde{\boldsymbol{\beta}}}{1 + \lambda}.$$

It is easily seen the amount of shrinkage of $\tilde{\boldsymbol{\beta}}$ in $\hat{\boldsymbol{\beta}}$ for $\lambda > 0$. We will show more shrinkage forms later in this chapter.

A popular choice in the literature is to let the penalty function be of an additive form

$$P_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{k} p_\lambda(\beta_j), \tag{2.7}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$. For each $\lambda \geq 0$, $p_\lambda(\beta_j)$ is continuous on the real line, differentiable at nonzero values of β_j, and the values of $p_\lambda(\beta_j)$ are positive for all nonzero β_j and zero otherwise. Under this additive assumption and orthonomal of X, (2.6) can be written as

$$\sum_{j=1}^{k} \left[(\beta_j - \tilde{\beta}_j)^2 + p_\lambda(\beta_j)\right]$$

which simplifies the computation to the extent that the minimization can be carried out component-wise. In other words, we only need to consider the problem of minimizing a single component

$$(\beta_j - \tilde{\beta}_j)^2 + p_\lambda(\beta_j). \tag{2.8}$$

The minimizer $\hat{\beta}_j$ of (2.8) is also called a thresholding function of $\tilde{\beta}_j$ since it takes value zero within some set of $\tilde{\beta}_j$ and has value less or equal to $\tilde{\beta}_j$ otherwise.

In the following Sections 2.2.1, 2.2.2 and 2.2.3, we will present several penalty functions studied in the literature. The corresponding regularized estimators will be given below in equations (2.10), (2.11), (2.12), (2.14) and (2.16).

2.2.1 The L_p Penalty

A widely used penalty function is the L_p-penalty with $p > 0$, also known as the Lasso-type penalty, given by

$$P_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{k} p_\lambda(\beta_j) = \sum_{j=1}^{k} \lambda|\beta_j|^p. \tag{2.9}$$

This includes the Lasso penalty (Tibshirani, 1996) when $p = 1$, and the ridge regression penalty (Frank and Friedman, 1993) when $p = 2$ (see Example 2.1 and Example 2.2). When $p \to 0$, the limiting case of the L_p penalty can be viewed as penalization by the number of nonzero parameters, yielding the AIC and the BIC criteria (Burnham and Anderson, 2002), since

$$\lim_{p \to 0} \sum_{j=1}^{k} |\beta_j|^p = \sum_{j=1}^{k} I(\beta_j \neq 0).$$

For the Lasso penalty, $p_\lambda(\beta_j) = \lambda|\beta_j|$, the minimizer of (2.8) is given by

$$\hat{\beta}_j = \text{sign}(\tilde{\beta}_j)\left(|\tilde{\beta}_j| - \frac{\lambda}{2}\right)_+ \tag{2.10}$$

where

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \quad \text{and} \quad (x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Figure 2.1(a) shows the Lasso estimators (2.10) with $\lambda = 0, 2$ and 4. If $\lambda = 0$, then $\hat{\beta}_j = \tilde{\beta}_j$ and no changes happen. When $\lambda > 0$, shrinkage occurs. It is seen that estimator $\hat{\beta}_j$ is zero for $\tilde{\beta}_j \in [-\lambda/2, \lambda/2]$, and the magnitude of the estimator is shrunk to $(|\tilde{\beta}_j| - \lambda/2)$ for $\tilde{\beta}_j$ outside the interval $[-\lambda/2, \lambda/2]$. The Lasso estimator $\hat{\beta}_j$ is a continuous function of $\tilde{\beta}_j$ and follows a either "shrink" or "kill" regulation. This is called a soft thresholding rule (Donoho and Johnstone, 1994) in the wavelet shrinkage literature.

For a ridge regression penalty, the corresponding estimator is given by

$$\hat{\beta}_j = \frac{\tilde{\beta}_j}{1 + \lambda}. \tag{2.11}$$

A graph of $\hat{\beta}_j$ with $\lambda = 0, 2$ and 4 is shown in Figure 2.1(b). The absolute value of the estimator $\hat{\beta}_j$ is a shrinkage of $|\tilde{\beta}_j|$. However, $\hat{\beta}_j$ is zero only if $\tilde{\beta}_j$ is, which may not be helpful for variable selection. The reason is that $\tilde{\beta}_j$ is rarely zero. Thus the ridge regression estimator will not be zero. Note that a zero value for the estimator $\hat{\beta}_j$ is the criterion in regularization method for eliminating β_j from the model. This property is laking in the ridge regression estimation.

Zou (2006) considered the case where the tuning parameter λ in (2.9) varies

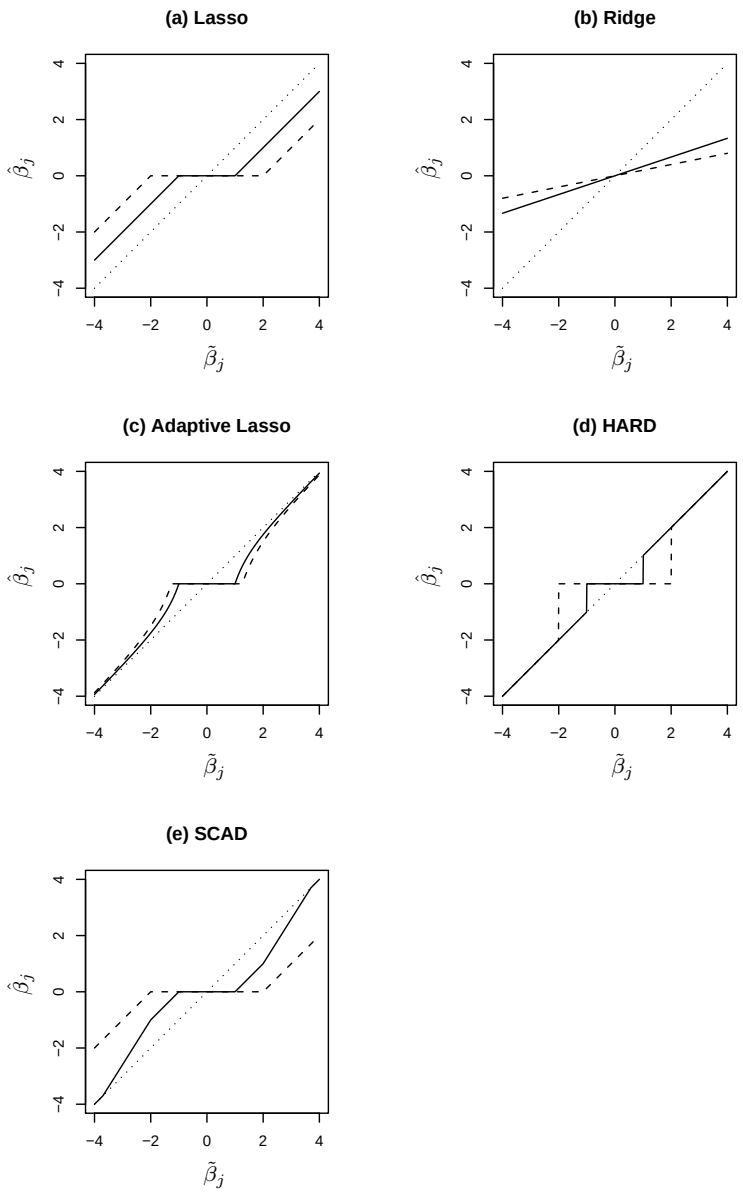

Figure 2.1: The dotted lines ($\lambda = 0$), solid lines ($\lambda = 2$) and dash lines ($\lambda = 4$) are plots of the minimizer $\hat{\beta}_j$ of (2.8) versus the OLS estimator $\tilde{\beta}_j$ for (a) the Lasso (b) the ridge regression (c) the adaptive Lasso with $\hat{w}_j = 1/|\tilde{\beta}_j|^2$ (d) the HARD and (e) the SCAD with $a = 1.85$.

with the sample size n to be denoted by λ_n. Suppose $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0k})^T$ is the true parameter in the linear regression model (2.1). Let $\mathcal{A} = \{j : \beta_{0j} \neq 0, j = 1, \ldots, k\}$, and assume that $|\mathcal{A}| = k_0 < k$. So that the true model depends only on a subset of regressors and $(k - k_0)$ components of $\boldsymbol{\beta}_0$ are zero. Let $\hat{\boldsymbol{\beta}}^{(n)} = (\hat{\beta}_1^{(n)}, \ldots, \hat{\beta}_k^{(n)})^T$ be the Lasso estimator of $\boldsymbol{\beta}$ which minimizes

$$\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + \sum_{j=1}^{k} \lambda_n |\beta_j|.$$

Let $\mathcal{A}_n = \{j : \hat{\beta}_j^{(n)} \neq 0, j = 1, \ldots, k\}$. By example, Zou showed that

$$\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) \neq 1.$$

In other words, the Lasso estimation procedure is not consistent in variable selection.

Zou, therefore, proposed an adaptive Lasso estimator $\hat{\boldsymbol{\beta}}^{*(n)}$ which is the minimizer of

$$\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + \sum_{j=1}^{k} \lambda_n \hat{w}_j |\beta_j|,$$

where the weight is a function of the OLS estimator: $\hat{w}_j = 1/|\tilde{\beta}_j|^\gamma$ and $\gamma > 0$. Let $\mathcal{A}_n^* = \{j : \hat{\beta}_j^{*(n)} \neq 0, j = 1, \ldots, k\}$. Suppose that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$. Then the adaptive Lasso satisfies the oracle properties:

1. $\lim_{n\to\infty} P(\mathcal{A}_n^* = \mathcal{A}) = 1$;
2. $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{*(n)} - \boldsymbol{\beta}_{0\mathcal{A}})$ converges to a normal distribution asymptotically, where $\boldsymbol{\beta}_{0\mathcal{A}}$ is the vector of those k_0 nonzero components in $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{*(n)}$ is an adaptive Lasso estimator of $\boldsymbol{\beta}_{0\mathcal{A}}$.

This means that asymptotically the procedure performs as well as if the true model were known in advance.

The minimizer of (2.8) with the adaptive Lasso penalty, $\sum_{j=1}^{k} \lambda_n \hat{w}_j |\beta_j|$, becomes

$$\hat{\beta}_j = \text{sign}(\tilde{\beta}_j) \left(|\tilde{\beta}_j| - \frac{\lambda_n}{2} \hat{w}_j \right)_+ . \tag{2.12}$$

See Figure 2.1(c) for (2.12) with $\lambda_n = 0, 2$ or 4 and $\hat{w}_j = 1/|\tilde{\beta}_j|^2$. We observe that the gaps between the dotted line and solid line and between the dotted line and dashed line are smaller in Figure 2.1(c) as compared with the corresponding ones in Figure 2.1(a). This indicates that the bias of the estimator $\hat{\beta}_j$ is smaller in (c) than in (a).

2.2.2 The HARD Penalty

Fan (1997) proposed the HARD (thresholding) penalty function

$$P_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{k} p_\lambda(\beta_j) = \sum_{j=1}^{k} \left[(\frac{\lambda}{2})^2 - (|\beta_j| - \frac{\lambda}{2})^2 \mathbf{1} \left(|\beta_j| < \frac{\lambda}{2} \right) \right], \tag{2.13}$$

where $\mathbf{1}(\cdot)$ is an indicator function of the set $\{|\beta_j| < \lambda/2\}$. Note that unlike the L_p-penalty (2.9), the HARD penalty is no longer a product of a constant λ and a function β_j. The name hard threshold is adapted from Donoho and Johnstone (1994) who obtained the minimizer of (2.8),

$$\hat{\beta}_j = \tilde{\beta}_j \mathbf{1} \left(|\tilde{\beta}_j| > \frac{\lambda}{2} \right). \tag{2.14}$$

Figure 2.1(d) shows the estimator $\hat{\beta}_j$ with $\lambda = 0, 2$ and 4. Donoho and Johnstone's rule of "keep" or "kill" is to keep the estimator $\hat{\beta}_j$ at the value of $\tilde{\beta}_j$ if $\tilde{\beta}_j \notin [-\lambda/2, \lambda/2]$, and to set (or kill) $\hat{\beta}_j$ to zero otherwise. The estimator $\hat{\beta}_j$ (2.14) in Figure 2.1 (d) looks better than the Lasso estimator (2.10) in Figure 2.1 (a) because there are no gaps between the dotted line and others when $\tilde{\beta}_j \notin [-\lambda/2, \lambda/2]$.

However, there is a drawback in (2.14). It is discontinuous at $|\lambda/2|$. Discontinuity induces instability in model selection in that a small change in the data can result in very different regressors being selected and hence reduces prediction accuracy.

2.2.3 The SCAD Penalty

In order to improve the discontinuity problem in HARD, Fan and Li (2001) proposed the Smoothly Clipped Absolute Deviation (SCAD) penalty function

$$P_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{k} p_\lambda(\beta_j) = \sum_{j=1}^{k} \int_0^{|\beta_j|} \left[\mathbf{1}\left(x \le \frac{\lambda}{2}\right) + \frac{2(a\lambda - x)_+}{(2a-1)\lambda} \mathbf{1}\left(x > \frac{\lambda}{2}\right) \right] dx \quad (2.15)$$

for some $a > 1$. The estimator $\hat{\beta}_j$ is continuous in $\tilde{\beta}_j$ and given by

$$\hat{\beta}_j = \begin{cases} \text{sign}(\tilde{\beta}_j)\left(|\tilde{\beta}_j| - \lambda/2\right)_+ & |\tilde{\beta}_j| \le \lambda \\ \left[(2a-1)\tilde{\beta}_j - a\lambda \text{sign}(\tilde{\beta}_j)\right] / \left[2(a-1)\right] & \lambda < |\tilde{\beta}_j| \le a\lambda \\ \tilde{\beta}_j & |\tilde{\beta}_j| > a\lambda. \end{cases} \quad (2.16)$$

See Figure 2.1(e) with $\lambda = 0, 2$ and 4 and $a = 1.85$. Beside λ, however, one more tuning parameter a needs to be chosen. Fan and Li (1999) recommended $a = 1.85$ based on a Bayesian argument.

A good penalty function should have properties of *unbiasedness*, *sparsity solution* and *continuity*. *Unbiasedness* will ensure no penalization for large coefficients thus avoiding unnecessary modeling bias. *Sparsity solution* refers to estimating insignificant regression coefficients by zero. Thus it reduces model complexity. *Continuity* provides stability in model prediction. That is, small change in the data will not result in a drastic change of variable selection. From Figure 2.1, we see that Lasso satisfies *sparsity* and *continuity* but not *unbiasedness*, ridge regression satisfies

only *continuity*, and HARD satisfies *unbiasedness* and *sparsity* but not *continuity*. SCAD and adaptive Lasso penalties satisfy all three properties. But they require the determination of two tuning parameters, λ and a, while other penalties require the determination of only one parameter λ.

Moreover, we notice that different values of λ result in different sizes of shrinkage in $\boldsymbol{\beta}$. However, shrinking too many variables may reduce prediction accuracy of the model. Model selection, therefore, is a necessary process of attaining regularized estimators.

In the next section, we shall turn our attention to the determination of the tuning parameter λ in (2.4). A conventional way to select the tuning parameter λ in the linear regression model is the generalized cross-validation (GCV) criterion.

2.3 Generalized Cross-Validation Criterion for Determining λ

The idea behind the cross-validation is to break up the data into several groups and use one group of the data to predict the rest of the data, and then to find the tuning parameter λ which gives the smallest prediction error. When the original data is partitioned into n groups, we call the cross-validation the n-fold cross-validation. The GCV is a modified form of n-fold cross-validation. Let us first discuss n-fold cross-validation.

2.3.1 n-Fold Cross-Validation

Consider n independent random variables $\mathbf{y}_1, \ldots, \mathbf{y}_n$. Suppose we set $\mathbf{y}_i$ aside for some arbitrarily fixed i, and $\mathbf{y}_i$ will be used for validation. We shall use the remaining $(n-1)$ observations as training data. Let the training data be denoted by $\mathbf{Y}^{-i} = (\mathbf{y}_1, \ldots, \mathbf{y}_{i-1}, \mathbf{y}_{i+1}, \ldots, \mathbf{y}_n)^T$. Let $\hat{\mathbf{y}}_\lambda^{-i}$ be the predictor of $\mathbf{y}_i$, computed from a procedure $\mathcal{M}_\lambda$ based on the data $\mathbf{Y}^{-i}$ and depending on the parameter λ. For example, in the linear regression model (2.1), we have

$$\hat{\mathbf{y}}_\lambda^{-i} = \mathbf{x}_i \hat{\boldsymbol{\beta}}_\lambda^{-i},$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$ and $\hat{\boldsymbol{\beta}}_\lambda^{-i} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{j \neq i} (\mathbf{y}_j - \mathbf{x}_j \boldsymbol{\beta})^2 + P_\lambda(\boldsymbol{\beta}) \right\}$ obtained by minimizing (2.4).

For a given λ, we repeat the procedure $\mathcal{M}_\lambda$ n times until each observation in $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$ is used once for validation. The ordinary cross-validation (OCV) function is defined as the average of squared discrepancies between $\mathbf{y}_i$ and its estimator $\hat{\mathbf{y}}_\lambda^{-i}$:

$$\mathrm{OCV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_\lambda^{-i})^2. \tag{2.17}$$

The minimizer of (2.17) with respect to λ is the desired value of $\hat{\lambda}$ which gives the smallest average prediction error.

To compute OCV for each λ, we need to repeat the procedure $\mathcal{M}_\lambda$ n times. This is usually computationally intensive. A generalized cross-validation criterion is introduced to ease the computation.

2.3.2 Generalized Cross-Validation

The idea of generalized cross-validation is to find a more easily computed variable to substitute for $\hat{\mathbf{y}}_\lambda^{-i}$ in (2.17). This was proposed by Craven and Wahba (1979). The following description of GCV is adapted from Wang (2004) which is more suitable for our purpose.

Let $\hat{\mathbf{Y}} = (\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_n)^T$ be the vector estimated from the procedure $\mathcal{M}_\lambda$ based on the complete data $\mathbf{Y}$. Assume that there exists an $n \times n$ matrix $A_\lambda = (a_{ij})_{n\times n}$ depending on λ such that $\hat{\mathbf{Y}}$ can be represented as a linear function of $\mathbf{Y}$,

$$\hat{\mathbf{Y}} = A_\lambda \mathbf{Y}. \tag{2.18}$$

For example, in the ridge regression (Example 2.2), given a λ,

$$\hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}}_\lambda = X(X^TX + \lambda I)^{-1}X^T\mathbf{Y}.$$

We choose $A_\lambda = X(X^TX + \lambda I)^{-1}X^T$. For the same λ, let

$$\mathring{\mathbf{Y}}^{-i} = (\mathbf{y}_1, \ldots, \mathbf{y}_{i-1}, \hat{\mathbf{y}}_\lambda^{-i}, \mathbf{y}_{i+1}, \ldots, \mathbf{y}_n)^T$$

be our data with $\mathbf{y}_i$ replaced by $\hat{\mathbf{y}}_\lambda^{-i}$. Using the data $\mathring{\mathbf{Y}}^{-i}$, we compute another estimator of $\mathbf{y}_i$, $\mathring{\mathbf{y}}_\lambda^{-i}$, which is obtained the same way as that of $\hat{\mathbf{y}}_\lambda^{-i}$ from the procedure $\mathcal{M}_\lambda$. We shall prove in Lemma 2.2 that for $\hat{\mathbf{y}}_\lambda^{-i}$ computed with regularized RSS (2.4),

$$\hat{\mathbf{y}}_\lambda^{-i} = \mathring{\mathbf{y}}_\lambda^{-i} \text{ for } i = 1, \ldots, n. \tag{2.19}$$

Since Wang (2004) did not provide a proof of (2.19), we will give it in the following.

Lemma 2.2 *Let* $\mathbf{Y}^{-i}$, $\mathring{\mathbf{Y}}^{-i}$ *and* $\mathbf{x}_i$ *be defined as above, and let* $\hat{\boldsymbol{\beta}}_\lambda^{-i}$ *and* $\mathring{\boldsymbol{\beta}}_\lambda^{-i}$ *be obtained from minimizing the regularized RSS (2.4) using respectively the data* $\mathbf{Y}^{-i}$ *and* $\mathring{\mathbf{Y}}^{-i}$. *Let* $\hat{\mathbf{y}}_\lambda^{-i} = \mathbf{x}_i\hat{\boldsymbol{\beta}}_\lambda^{-i}$ *and* $\mathring{\mathbf{y}}_\lambda^{-i} = \mathbf{x}_i\mathring{\boldsymbol{\beta}}_\lambda^{-i}$ *be the estimators of* $\mathbf{y}_i$. *Then* $\hat{\mathbf{y}}_\lambda^{-i} = \mathring{\mathbf{y}}_\lambda^{-i}$ *for* $i = 1, \ldots, n$.

Proof: Note that $\hat{\boldsymbol{\beta}}_\lambda^{-i}$ is the minimizer of $\left\{\sum_{j\neq i}(\mathbf{y}_j - \mathbf{x}_j\boldsymbol{\beta})^2 + P_\lambda(\boldsymbol{\beta})\right\}$ and $\mathring{\boldsymbol{\beta}}_\lambda^{-i}$ is the minimizer of $h(\boldsymbol{\beta}) = \left\{\sum_{j\neq i}(\mathbf{y}_j - \mathbf{x}_j\boldsymbol{\beta})^2 + (\hat{\mathbf{y}}_\lambda^{-i} - \mathbf{x}_i\boldsymbol{\beta})^2 + P_\lambda(\boldsymbol{\beta})\right\}$. Substituting $\hat{\boldsymbol{\beta}}_\lambda^{-i}$ for $\boldsymbol{\beta}$ in h gives

$$\begin{aligned} h(\hat{\boldsymbol{\beta}}_\lambda^{-i}) &= \sum_{j\neq i}(\mathbf{y}_j - \mathbf{x}_j\hat{\boldsymbol{\beta}}_\lambda^{-i})^2 + (\hat{\mathbf{y}}_\lambda^{-i} - \mathbf{x}_i\hat{\boldsymbol{\beta}}_\lambda^{-i})^2 + P_\lambda(\hat{\boldsymbol{\beta}}_\lambda^{-i}) \\ &= \sum_{j\neq i}(\mathbf{y}_j - \mathbf{x}_j\hat{\boldsymbol{\beta}}_\lambda^{-i})^2 + P_\lambda(\hat{\boldsymbol{\beta}}_\lambda^{-i}) \\ &\leq \sum_{j\neq i}(\mathbf{y}_j - \mathbf{x}_j\boldsymbol{\beta})^2 + P_\lambda(\boldsymbol{\beta}) \text{ for all } \boldsymbol{\beta} \\ &\leq h(\boldsymbol{\beta}) \text{ for all } \boldsymbol{\beta}. \end{aligned}$$

Hence $\hat{\boldsymbol{\beta}}_\lambda^{-i} = \mathring{\boldsymbol{\beta}}_\lambda^{-i}$ which entails $\hat{\mathbf{y}}_\lambda^{-i} = \mathring{\mathbf{y}}_\lambda^{-i}$. □

Remark: Although our proof of Lemma 2.2 follows the approach of Craven and Wahba (1979), our method is different from theirs. They consider the spline smoothing model

$$\mathbf{y} = f(x) + \mathbf{e}, \quad x \in [0, 1], \tag{2.20}$$

where $\mathbf{e}$ is a random error with mean 0 and finite variance σ^2, and f is a function in the Sobolev space $W_2^m[0, 1]$ for a given finite m. A function f in $W_2^m[0, 1]$ if the $(m-1)$-th derivative of f is absolutely continuous in the interval $[0, 1]$ and its m-th derivative, $f^{(m)}$, is finite in $L^2[0, 1]$. Consider a random sample of n independent observations $(\mathbf{y}_i, x_i)$ from (2.20). The smoothing spline estimator $\hat{f}$ of f is the

minimizer in $W_2^m[0,1]$ of

$$\frac{1}{n}\sum_{j=1}^{n}(\mathbf{y}_j - f(x_j))^2 + \lambda\int_0^1 (f^{(m)}(x))^2 dx.$$

Craven and Wahba (1979) considered two estimators of f, $\hat{f}_\lambda^{-i}$ and $\mathring{f}_\lambda^{-i}$. $\hat{f}_\lambda^{-i}$ is obtained by minimizing

$$\frac{1}{n}\sum_{j\neq i}(\mathbf{y}_j - f(x_j))^2 + \lambda\int_0^1 (f^{(m)}(x))^2 dx, \tag{2.21}$$

while $\mathring{f}_\lambda^{-i}$ is obtained by minimizing

$$\frac{1}{n}\left[(\hat{f}_\lambda^{-i}(x_i) - f(x_i))^2 + \sum_{j\neq i}(\mathbf{y}_j - f(x_j))^2\right] + \lambda\int_0^1 (f^{(m)}(x))^2 dx. \tag{2.22}$$

In their Lemma 3.1, it is shown that $\hat{f}_\lambda^{-i}(x) = \mathring{f}_\lambda^{-i}(x)$ for all $x \in [0,1]$.

However, since $\hat{f}_\lambda^{-i}$ uses $(n-1)$ observations, (2.21) should be modified as

$$\frac{1}{n-1}\sum_{j\neq i}(\mathbf{y}_j - f(x_j))^2 + \lambda\int_0^1 (f^{(m)}(x))^2 dx. \tag{2.23}$$

If so, (2.22) and (2.23) fail to imply $\hat{f}_\lambda^{-i} = \mathring{f}_\lambda^{-i}$ as claimed in their Lemma 3.1. This is corrected in our Lemma 2.2 for the regularized RSS. □

By the assumptions of (2.18) and definitions of $\hat{\mathbf{y}}_\lambda^{-i}$, we have

$$\begin{aligned}\hat{\mathbf{y}}_i &= \sum_{j=1}^{n} a_{ij}\mathbf{y}_j \\ \hat{\mathbf{y}}_\lambda^{-i} &= \mathring{\mathbf{y}}_\lambda^{-i} = \sum_{j\neq i} a_{ij}\mathbf{y}_j + a_{ii}\hat{\mathbf{y}}_\lambda^{-i}\end{aligned}$$

for all i. Then

$$\mathbf{y}_i - \hat{\mathbf{y}}_i = (1 - a_{ii})(\mathbf{y}_i - \hat{\mathbf{y}}_\lambda^{-i}). \tag{2.24}$$

Assume that $1 - a_{ii} \neq 0$. By (2.24), the OCV function (2.17) can be written as

$$\mathrm{OCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbf{y}_i - \hat{\mathbf{y}}_i}{1 - a_{ii}}\right)^2. \tag{2.25}$$

Now replace a_{ii} by the average of the trace of A_λ (2.18), $\sum_{i=1}^n a_{ii}/n = tr(A_\lambda)/n$. Note that both $\hat{\mathbf{y}}_i$ and A_λ depend on λ. The generalized cross-validation function is a modified OCV function (2.25) defined by

$$\begin{aligned} \text{GCV}(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbf{y}_i - \hat{\mathbf{y}}_i}{1 - tr(A_\lambda)/n}\right)^2 \\ &= \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2}{n[1 - tr(A_\lambda)/n]^2}. \end{aligned} \tag{2.26}$$

In GCV, it is not necessary to compute $\hat{\mathbf{y}}_\lambda^{-i}$. We only need to compute the trace of A_λ (see (2.18)) and compute $\hat{\mathbf{Y}}$ once with the complete data $\mathbf{Y}$ for each given λ. Then an optimal value of λ is obtained by minimizing GCV(λ) over λ.

The following two examples illustrate the use of GCV functions in selecting the tuning parameter λ in linear regressions.

Example 2.3 *(Lasso). The L_1 penalty $P(\boldsymbol{\beta}) = \sum_{j=1}^k |\beta_j|$ is not differentiable at the origin. To carry out the Newton-Raphson, Tibshirani (1996) argued that $P(\boldsymbol{\beta})$ can be approximated by $\sum_{j=1}^k \beta_j^2/|\tilde{\beta}_j| = \boldsymbol{\beta}'\mathbf{W}\boldsymbol{\beta}$, where $\tilde{\beta}_j$ is the OLS estimator of β_j and $\mathbf{W}$ is a diagonal matrix with entries $(1/|\tilde{\beta}_1|, \ldots, 1/|\tilde{\beta}_k|)$ if $\tilde{\beta}_j \neq 0$. This device has been widely used in the literature. Then $X\hat{\boldsymbol{\beta}}_\lambda \approx X(X^TX + \lambda\mathbf{W})^{-1}X^T\mathbf{Y} = A_\lambda^l\mathbf{Y}$ and*

$$GCV(\lambda) = \frac{\|\mathbf{Y} - X\hat{\boldsymbol{\beta}}_\lambda\|^2}{n[1 - tr(A_\lambda^l)/n]^2}.$$

Example 2.4 *(Ridge regression). The L_2 penalty is $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|^2$. Then we have $X\hat{\boldsymbol{\beta}}_\lambda = X(X^TX + \lambda I)^{-1}X^T\mathbf{Y} = A_\lambda^r\mathbf{Y}$ and*

$$GCV(\lambda) = \frac{\|\mathbf{Y} - X\hat{\boldsymbol{\beta}}_\lambda\|^2}{n[1 - tr(A_\lambda^r)/n]^2}.$$

In summary, the regularization in linear regression is to estimate coefficients $\boldsymbol{\beta}$ by minimizing the regularized RSS $\{\|\mathbf{Y} - X\boldsymbol{\beta}\|^2 + P_\lambda(\boldsymbol{\beta})\}$. For each given value of the tuning parameter λ, we can compute $\hat{\boldsymbol{\beta}}_\lambda$, the estimate of $\boldsymbol{\beta}$, and calculate its corresponding GCV(λ). Then the λ which yields the smallest value of GCV is our best choice.

In the following chapter, we shall extend the regularized method to the proportional hazards model for censored survival data.

Chapter 3

Regularization in Proportional Hazards Regression

3.1 Definition

Let T be a nonnegative random variable denoting the survival time of an individual in the study population. Let C be a nonnegative random variable independent of T. The observation of T is subject to right censoring by C in the sense that T is observable up to $\tilde{T}$ where $\tilde{T} = \min(T, C)$. Let $\delta = I[T \leq C]$ be the censoring indicator of the event $[T \leq C]$.

Let $\mathbf{Z} = (Z_1, \ldots, Z_k)^T$ denote the k-dimensional covariate of T. The hazard function of T conditioning on the covariate $\mathbf{Z} = \mathbf{z}$ is defined by

$$h(t|\mathbf{z}) = \lim_{x \to 0} \frac{1}{x} P(t \leq T < t + x | T \geq t, \mathbf{z}), \quad \text{for } t \geq 0.$$

If an arbitrary distribution function $H_0(t)$ possesses a density, its hazard function is given by

$$h_0(t) = -\frac{d}{dt} \log \left[1 - H_0(t)\right], \quad \text{for } t \geq 0.$$

The proportional hazards model, also known as the Cox regression model (Cox, 1975), is the product

$$h(t|\mathbf{Z} = \mathbf{z}) = h_0(t) \exp(\boldsymbol{\beta}^T \mathbf{z}), \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ is a k-dimensional column vector of unknown regression coefficients.

Assume that we have a sample of n independent and identically distributed (i.i.d.) random vectors $(\tilde{T}_i, \delta_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, from a given population. Estimation of regression coefficients, $\boldsymbol{\beta}$, can be performed by using the partial likelihood method. Let $R_i = \{j : \tilde{T}_j \geq \tilde{T}_i\}$ denote the risk set at time $\tilde{T}_i$. That is, R_i contains all of those individuals in the sample that are alive and not censored at time $\tilde{T}_i$. The partial likelihood function of $\boldsymbol{\beta}$ is defined by

$$\prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}^T \mathbf{Z}_i)}{\sum_{j \in R_i} \exp(\boldsymbol{\beta}^T \mathbf{Z}_j)} \right]^{\delta_i}. \tag{3.2}$$

It is well-known that the maximizer $\tilde{\boldsymbol{\beta}}$ of the partial likelihood (3.2) is an asymptotically normal and efficient estimator of $\boldsymbol{\beta}$. The estimation of $\boldsymbol{\beta}$ using (3.2) does not depend on the unknown nuisance hazard function $h_0(t)$.

A popular approach to study the properties of $\tilde{\boldsymbol{\beta}}$ is to formulate the problem in terms of counting processes. Consider two counting processes $\mathbf{N} = \{\mathbf{N}(t) : t \geq 0\}$ and $\mathbf{Y} = \{\mathbf{Y}(t) : t \geq 0\}$, where

$$\mathbf{N}(t) = I[T \leq t, T \leq C], \tag{3.3}$$

$$\mathbf{Y}(t) = I[T \geq t, C \geq t]. \tag{3.4}$$

These two processes monitor the survival and possible censoring time of an individual over time t. The sample, $\{(\tilde{T}_i, \delta_i, \mathbf{Z}_i), i = 1, \ldots, n\}$, gives rise to a family of counting processes $\{\mathbf{N}_i, \mathbf{Y}_i, i = 1, \ldots, n\}$, where $\mathbf{N}_i$ and $\mathbf{Y}_i$ are defined as $\mathbf{N}$ and $\mathbf{Y}$ with T and C replaced by T_i and C_i. Let $\Delta\mathbf{N}(t) = \mathbf{N}(t) - \mathbf{N}(t-)$ denote the jump of $\mathbf{N}$ at time t. For any fixed t, we consider all $(\tilde{T}_i, \delta_i, \mathbf{Z}_i)$ that are observed by time t. The

partial likelihood (3.2) at time t can be expressed as

$$L(\boldsymbol{\beta}, t) = \prod_{i=1}^{n} \prod_{0 \le u \le t} \left[\frac{\mathbf{Y}_i(u) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i)}{\sum_{j=1}^{n} \mathbf{Y}_j(u) \exp(\boldsymbol{\beta}^T \mathbf{Z}_j)} \right]^{\Delta \mathbf{N}_i(u)} \tag{3.5}$$

with $L(\boldsymbol{\beta}, \infty)$ equal to (3.2). The logarithm of (3.5) is

$$\ell(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^t \left\{ \boldsymbol{\beta}^T \mathbf{Z}_i - \log \left[\sum_{j=1}^{n} \mathbf{Y}_j(u) \exp(\boldsymbol{\beta}^T \mathbf{Z}_j) \right] \right\} d\mathbf{N}_i(u). \tag{3.6}$$

Note that $\mathbf{Z}_i$ is a random covariate vector of the i-th individual:

$$\mathbf{Z}_i = \begin{pmatrix} Z_{i1} \\ \vdots \\ Z_{ik} \end{pmatrix}.$$

We assume that $\mathbf{Z}_i$'s are not time dependent. It can be shown that if not all of the observations are censored, the log partial likelihood, $\ell(\boldsymbol{\beta}, t)$, is a strictly concave function of $\boldsymbol{\beta}$, and the maximum partial likelihood estimator exists uniquely.

Similar to the approach in the linear regression of Section 2.1, a regularized estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ based on $\ell(\boldsymbol{\beta}, t)$ is obtained as

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmax}}\ \ell(\boldsymbol{\beta}, t) \text{ subject to the constraint } P(\boldsymbol{\beta}) \le c, \tag{3.7}$$

where $P(\boldsymbol{\beta})$ is a penalty function of $\boldsymbol{\beta}$ and c is some known nonnegative constant. As in Section 2.1, the function $P(\boldsymbol{\beta})$ is assumed to be zero at the origin. It is positive and differentiable for $\boldsymbol{\beta} \neq 0$. In terms of the Lagrange multiplier $\lambda \ge 0$, (3.7) is equivalent to

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmax}} \left\{ \ell(\boldsymbol{\beta}, t) - \lambda P(\boldsymbol{\beta}) \right\}.$$

Note that $\hat{\boldsymbol{\beta}}$ here is a maximizer instead of a minimizer as we have used in Section 2.1. Therefore we use the minus sign "$-$" in the penalized term in order to keep the tuning parameter λ nonnegative as in Section 2.1.

In the following, we shall use a more general form of penalty $P_\lambda(\boldsymbol{\beta})$ instead of $\lambda P(\boldsymbol{\beta})$ as discussed in the previous chapter. In this chapter we have censored data and the log partial likelihood $\ell(\boldsymbol{\beta}, t)$ is nonlinear in $\boldsymbol{\beta}$ which differ from the linear regression $X\boldsymbol{\beta}$ and uncensored data studied in Chapter 2.

3.2 Asymptotic Properties of Estimators

Assume that $\hat{\boldsymbol{\beta}}$ is a local maximizer of the regularized log partial likelihood,

$$Q(\boldsymbol{\beta}, t) = \ell(\boldsymbol{\beta}, t) - P_\lambda(\boldsymbol{\beta}), \tag{3.8}$$

in a neighborhood $\mathcal{B}$ of the true $\boldsymbol{\beta}_0$, where $\ell(\boldsymbol{\beta}, t)$ is given by (3.6). We shall prove that given the covariate $\mathbf{Z} = \mathbf{z}$, $\hat{\boldsymbol{\beta}}$ is conditionally consistent and asymptotically normal as the sample size n goes to infinity.

To our knowledge, there is little work in the literature on the asymptotic properties of $\hat{\boldsymbol{\beta}}$ with a nonnegative constant λ. Oracle properties of $\hat{\boldsymbol{\beta}}$ under the sparsity model and in the case of λ depending on n have been studied by Fan and Li (2002) and Zhang and Lu (2007) among others. Therefore, in the following Sections 3.2.2, 3.2.3 and 3.2.4, we shall establish consistency and asymptotic normality of the estimator $\hat{\boldsymbol{\beta}}$ for both constant λ in Theorem 3.1 and 3.2 and for λ depending on n in Theorem 3.3 and 3.4. Section 3.2.1 shows preliminaries to the above theorems. Finally, in Section 3.2.5, we shall present the oracle properties of $\hat{\boldsymbol{\beta}}$ and the theorem of Fan and Li (2002) in Theorem 3.5.

3.2.1 Preliminaries for Establishing Asymptotic Properties of Estimators

Under some regularity conditions on the model, the asymptotic properties of the estimators can be obtained by investigating the asymptotic behavior of the first two partial derivatives of $Q(\boldsymbol{\beta}, t)$ (in (3.8)). Consider the partial derivative of Q,

$$\frac{\partial Q(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}} = \frac{\partial \ell(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}} - \frac{\partial P_\lambda(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}. \tag{3.9}$$

The first term on the right hand side is the score vector of the log partial likelihood function (3.6) which can be written in terms of the counting processes $\mathbf{N}_i = I[T_i \leq t, T_i \leq C_i]$ (see (3.3)), for $i = 1, \ldots, n$, as follows

$$\frac{\partial \ell(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \int_0^t \left[\mathbf{Z}_i - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, u)}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, u)} \right] d\mathbf{N}_i(u), \tag{3.10}$$

where

$$\mathbf{S}^{(0)}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i), \tag{3.11}$$

$$\mathbf{S}^{(1)}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Y}_i(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i). \tag{3.12}$$

Recall that $\mathbf{Y}_i(t) = I[T_i \geq t, C_i \geq t]$ (see (3.4)). We shall also need the second partial derivatives of $\mathbf{S}^{(0)}(\boldsymbol{\beta}, t)$:

$$\mathbf{S}^{(2)}(\boldsymbol{\beta}, t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Z}_i^T \mathbf{Y}_i(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}_i). \tag{3.13}$$

Note that $\mathbf{S}^{(0)}$ is a one-dimensional random variable, $\mathbf{S}^{(1)}$ is a k-dimensional column random vector and $\mathbf{S}^{(2)}$ is a $k \times k$ random matrix. These are notations used in Andersen and Gill (1982).

Andersen and Gill (1982) established the asymptotic consistency and normality of the maximum partial likelihood estimator of $\boldsymbol{\beta}$. We shall use their method of proof to establish asymptotic properties for the regularized estimator $\hat{\boldsymbol{\beta}}$ from the penalized likelihood $Q(\boldsymbol{\beta}, t)$.

Throughout this thesis, the notation $\xrightarrow{a.s.}$ denotes convergence almost surely, $\xrightarrow{P}$ denotes convergence in probability, and $\xrightarrow{D}$ denotes convergence in distribution. These limits are taken as $n \to \infty$ unless stated otherwise.

The following four conditions are used to establish the asymptotic results.

A. (Finite interval). Let τ be such that $\int_0^\tau h_0(t)dt < \infty$.

B. (Asymptotic stability). For $\mathbf{S}^{(0)}, \mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$, there exists a neighborhood $\mathcal{B}$ of the true parameter $\boldsymbol{\beta}_0$ and non random scalar, vector and matrix functions $s^{(0)}, s^{(1)}$ and $s^{(2)}$ defined on $\mathcal{B} \times [0, \tau]$ such that for $j = 0, 1, 2$,

$$\sup_{t \in [0,\tau], \beta \in \mathcal{B}} \|\mathbf{S}^{(j)}(\boldsymbol{\beta}, t) - s^{(j)}(\boldsymbol{\beta}, t)\| \xrightarrow{P} 0.$$

C. (Lindeberg condition). There exists $\gamma > 0$ such that

$$n^{-1/2} \sup_{1 \le i \le n, 0 \le t \le \tau} |\mathbf{Z}_i| \mathbf{Y}_i(t) I\{\boldsymbol{\beta}_0^T \mathbf{Z}_i > -\gamma |\mathbf{Z}_i|\} \xrightarrow{P} 0.$$

D. (Asymptotic regularity conditions). Let $\mathcal{B}, s^{(0)}, s^{(1)}$ and $s^{(2)}$ be as defined in Condition B. For all $\boldsymbol{\beta} \in \mathcal{B}$ and $t \in [0, \tau]$:

(1) The derivatives $s^{(1)}(\boldsymbol{\beta}, t) = \partial s^{(0)}(\boldsymbol{\beta}, t)/\partial \boldsymbol{\beta}$ and $s^{(2)}(\boldsymbol{\beta}, t) = \partial s^{(1)}(\boldsymbol{\beta}, t)/\partial \boldsymbol{\beta}$ exist.

(2) $s^{(j)}$ are bounded on $\mathcal{B} \times [0, \tau]$ for $j = 0, 1, 2$, and $s^{(0)}$ is away from zero.

(3) The family of functions $s^{(j)}(\cdot, t)$ is equicontinuous at $\boldsymbol{\beta}_0$ for $j = 0, 1, 2$. That is, given $\varepsilon > 0$, there is a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}_0$ such that

$$\sup_{\beta \in \mathcal{B}} \|s^{(j)}(\boldsymbol{\beta}, t) - s^{(j)}(\boldsymbol{\beta}_0, t)\| < \varepsilon \text{ for all } t.$$

(4) The matrix

$$\Sigma(\boldsymbol{\beta}_0, t) = \int_0^t v(\boldsymbol{\beta}_0, u) s^{(0)}(\boldsymbol{\beta}_0, u) h_0(u) du \tag{3.14}$$

is positive definite, where $v = (s^{(2)}/s^{(0)}) - (s^{(1)}/s^{(0)})(s^{(1)}/s^{(0)})^T$.

Remark: These are the conditions used in Andersen and Gill (1982). They are a variant of standard conditions used in asymptotic investigations known as local asymptotic normal conditions introduced by Le Cam (1960) (LAN). See Le Cam and Yang (2000). Condition B permits the replacement of $\boldsymbol{\beta}$ by a random vector used the proof of convergence of the information matrix $\mathcal{I}(\boldsymbol{\beta}, t)$. (See equations (3.23) and (3.27) below.) □

We shall use the following three lemmas. Lemma 3.1 and Lemma 3.2 are given by Andersen and Gill (1982). Lemma 3.3 is similar to Corollary II.2. of Andersen and Gill (1982), but they omit its proof. We shall give a proof of Lemma 3.3 below. Let

$$f(\boldsymbol{\beta}, \boldsymbol{\beta}_0, t) = \int_0^t \left\{ (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T s^{(1)}(\boldsymbol{\beta}_0, u) - \log \left[\frac{s^{(0)}(\boldsymbol{\beta}, u)}{s^{(0)}(\boldsymbol{\beta}_0, u)} \right] s^{(0)}(\boldsymbol{\beta}_0, u) \right\} h_0(u) du. \tag{3.15}$$

For all $t \in [0, \tau]$, the function $f(\boldsymbol{\beta}, \boldsymbol{\beta}_0, t)$ is nonrandom. It can be shown that $f(\boldsymbol{\beta}, \boldsymbol{\beta}_0, t)$ has a unique maximum at $\boldsymbol{\beta}_0$.

Lemma 3.1 *For any $t \in [0, \tau]$ and $\boldsymbol{\beta} \in \mathcal{B}$, the neighborhood of $\boldsymbol{\beta}_0$ defined in Condition B, under Conditions A–C and when the true parameter is $\boldsymbol{\beta}_0$,*

$$\frac{1}{n}\left[\ell(\boldsymbol{\beta}, t) - \ell(\boldsymbol{\beta}_0, t)\right] \xrightarrow{P} f(\boldsymbol{\beta}, \boldsymbol{\beta}_0, t) \ \text{ as } n \to \infty.$$

Lemma 3.2 *(Preservation of Concavity). Let E be an open convex set in R^k. Let $F_1, F_2, \ldots,$ be a sequence of random concave functions on E such that, for every $x \in E$, $F_n(x) \xrightarrow{P} F(x)$ as $n \to \infty$, where F is a real-valued function on E. Then F is also concave and for all compact $A \subset E$*

$$\sup_{x \in A} |F_n(x) - F(x)| \xrightarrow{P} 0 \ \text{ as } n \to \infty.$$

Lemma 3.3 *Let E be an open convex set in R^k. Let $F, F_1, F_2, \ldots,$ be random continuous functions on E such that for all compact $A \subset E$, $\sup_{x \in A} |F_n(x) - F(x)| \xrightarrow{P} 0$ as $n \to \infty$. Suppose that F has a unique maximum at $\hat{x} \in E$, and for any compact set A containing $\hat{x}$, there exists an $\hat{x}_n \in A$ maximizing F_n. Then $\hat{x}_n \xrightarrow{P} \hat{x}$ as $n \to \infty$.*

Note that F_n is assumed to be concave in Lemma 3.2 while it is only assumed to be continuous in Lemma 3.3.

Proof of Lemma 3.3: By hypothesis, considering any compact set $A \subset E$, for every subsequence $\{n_m\}$, there exists a further subsequence $\{n_{m_k}\} \subset \{n_m\}$ such that $\sup_{x \in A} |F_{n_{m_k}}(x) - F(x)| \xrightarrow{a.s.} 0$ as $k \to \infty$. We first show that along the sub-subsequence $\{n_{m_k}\}$, Lemma 3.3 is true. Then we extend the result to the original sequence $\{n\}$.

Consider any compact set A containing $\hat{x}$ in E. Then, for every fixed n_m, the continuous function F_{n_m} has a maximum $\hat{x}_{n_m} \in A$. By compactness of A, for every $\{\hat{x}_{n_m}\}$, there exists a subsequence $\{\hat{x}_{n_{m_k}}\} \subset \{\hat{x}_{n_m}\}$ such that $\hat{x}_{n_{m_k}} \xrightarrow{a.s.} \hat{y} \in A$.

We will show that $F(\hat{y}) \geq F(y)$ for every $y \in A$. Note that

$$F_{n_{m_k}}(\hat{x}_{n_{m_k}}) \geq F_{n_{m_k}}(y) \ \ \forall y \in A. \tag{3.16}$$

For every $\varepsilon > 0$ and k sufficiently large, we have

$$|F_{n_{m_k}}(\hat{x}_{n_{m_k}}) - F(\hat{y})| \leq |F_{n_{m_k}}(\hat{x}_{n_{m_k}}) - F(\hat{x}_{n_{m_k}})| + |F(\hat{x}_{n_{m_k}}) - F(\hat{y})| < 2\varepsilon.$$

$|F_{n_{m_k}}(\hat{x}_{n_{m_k}}) - F(\hat{x}_{n_{m_k}})| < \varepsilon$ follows by the convergence hypothesis and $|F(\hat{x}_{n_{m_k}}) - F(\hat{y})| < \varepsilon$ by continuity and our selected sequence $\{\hat{x}_{n_{m_k}}\}$.

It follows that

$$\begin{aligned} F(\hat{y}) &\geq F_{n_{m_k}}(\hat{x}_{n_{m_k}}) - 2\varepsilon, \\ F(\hat{y}) &\geq F_{n_{m_k}}(y) - 2\varepsilon, \ \text{ by (3.16)} \\ F(\hat{y}) &\geq F(y) - 2\varepsilon, \ \text{ by taking limit as } k \to \infty. \end{aligned}$$

This inequality is true for every $\varepsilon > 0$. Hence

$$F(\hat{y}) \geq F(y) \ \ \forall y \in A.$$

We have shown that the subsequence $\{\hat{x}_{n_{m_k}}\}$ converges almost surely to the limit $\hat{y}$, and that $\hat{y}$ maximizes F on A. By hypothesis, F has a unique maximum, so we conclude $\hat{y} = \hat{x}$. Then $\hat{x}_{n_{m_k}} \xrightarrow{a.s.} \hat{x}$ as $k \to \infty$.

Since for every subsequence $\{n_m\}$, there exists a further subsequence $\{n_{m_k}\}$ such that $\hat{x}_{n_{m_k}} \xrightarrow{a.s.} \hat{x}$, we have $\hat{x}_n \xrightarrow{P} \hat{x}$ as $n \to \infty$. $\square$

3.2.2 Consistency of $\hat{\boldsymbol{\beta}}$ with a nonnegative constant λ

We now establish the consistency of $\hat{\boldsymbol{\beta}}$, a local maximizer of the regularized log partial likelihood

$$Q(\boldsymbol{\beta}, t) = \ell(\boldsymbol{\beta}, t) - P_\lambda(\boldsymbol{\beta}), \tag{3.17}$$

where $\ell(\boldsymbol{\beta}, t)$ is given by (3.6).

Theorem 3.1 *(Consistency of $\hat{\boldsymbol{\beta}}$ with a nonnegative constant λ). Assume that Conditions A–D hold. Let $\mathcal{B}$ denote a neighborhood of the true parameter $\boldsymbol{\beta}_0$ satisfying Condition B. Assume that $\hat{\boldsymbol{\beta}}$ is the local maximizer of $Q(\boldsymbol{\beta}, t)$ in $\mathcal{B}$ for a given nonnegative tuning parameter λ. Then under the true parameter $\boldsymbol{\beta}_0$,*

$$\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0 \ \ \textit{as } n \to \infty.$$

Proof: Consider the difference of the regularized log partial likelihoods at $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$ in (3.17):

$$Q(\boldsymbol{\beta}, t) - Q(\boldsymbol{\beta}_0, t) = [\ell(\boldsymbol{\beta}, t) - \ell(\boldsymbol{\beta}_0, t)] - [P_\lambda(\boldsymbol{\beta}) - P_\lambda(\boldsymbol{\beta}_0)]. \tag{3.18}$$

For any compact set A containing $\boldsymbol{\beta}_0$ in $\mathcal{B}$,

$$\begin{aligned} &\sup_{\beta \in A} |[Q(\boldsymbol{\beta}, t) - Q(\boldsymbol{\beta}_0, t)]/n - f(\boldsymbol{\beta}, \boldsymbol{\beta}_0, t)| \\ &\quad \leq \sup_{\beta \in A} |[\ell(\boldsymbol{\beta}, t) - \ell(\boldsymbol{\beta}_0, t)]/n - f(\boldsymbol{\beta}, \boldsymbol{\beta}_0, t)| + \sup_{\beta \in A} |P_\lambda(\boldsymbol{\beta}) - P_\lambda(\boldsymbol{\beta}_0)|/n, \end{aligned} \tag{3.19}$$

where $f(\boldsymbol{\beta}, \boldsymbol{\beta}_0, t)$ is defined in (3.15). By Lemma 3.1 and Lemma 3.2, the first term of (3.19) converges in probability to zero since $[\ell(\boldsymbol{\beta}, t) - \ell(\boldsymbol{\beta}_0, t)]/n$ is a concave function of $\boldsymbol{\beta}$. The second term converges to zero since $\sup_{\beta \in A} |P_\lambda(\boldsymbol{\beta}) - P_\lambda(\boldsymbol{\beta}_0)|$ is

bounded on A. Let $F_n(\boldsymbol{\beta}) = [\ell(\boldsymbol{\beta}, t) - \ell(\boldsymbol{\beta}_0, t)]/n$ and $F(\boldsymbol{\beta}) = f(\boldsymbol{\beta}, \boldsymbol{\beta}_0, t)$. Since $f(\boldsymbol{\beta}, \boldsymbol{\beta}_0, t)$ has a unique maximum at $\boldsymbol{\beta}_0$, it follows by Lemma 3.3 that $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$. □

Remark: The proof shows that our penalty function $P_\lambda(\boldsymbol{\beta})$ does not play a significant role in determining the consistency of $\hat{\boldsymbol{\beta}}$.

Example 3.1 *Let $\hat{\boldsymbol{\beta}}$ be a L_p regularized estimator with $p \geq 1$, that is, $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{k} |\beta_j|^p$. Then $\hat{\boldsymbol{\beta}}$ is consistent.*

3.2.3 Asymptotic Normality of $\hat{\boldsymbol{\beta}}$ with a nonnegative constant λ

The first order Taylor expansion of $\partial Q(\boldsymbol{\beta}, t)/\partial \boldsymbol{\beta}$ at the true value $\boldsymbol{\beta}_0$ yields

$$\frac{\partial}{\partial \beta} Q(\boldsymbol{\beta}, t) = \frac{\partial}{\partial \beta} Q(\boldsymbol{\beta}_0, t) + \frac{\partial^2}{\partial \beta \partial \beta^T} Q(\boldsymbol{\beta}^*, t)(\boldsymbol{\beta} - \boldsymbol{\beta}_0), \tag{3.20}$$

where $\boldsymbol{\beta}^*$ is on the line segment between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$. Note that

$$Q(\boldsymbol{\beta}, t) = \ell(\boldsymbol{\beta}, t) - P_\lambda(\boldsymbol{\beta}) \tag{3.21}$$

(see (3.17)). Recall that the score vector process of $\ell(\boldsymbol{\beta}, t)$ (see (3.10)) is defined as

$$\begin{aligned} \frac{\partial}{\partial \beta} \ell(\boldsymbol{\beta}, t) &= \sum_{i=1}^{n} \int_0^t \left[\mathbf{Z}_i - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, u)}{\mathbf{S}^{(0)}(\boldsymbol{\beta}, u)} \right] d\mathbf{N}_i(u) \\ &\equiv \mathbf{U}(\boldsymbol{\beta}, t). \end{aligned} \tag{3.22}$$

The negative partial derivative of $\mathbf{U}(\boldsymbol{\beta}, t)$ gives the so-called "observed" information matrix (although it depends on the unknown $\boldsymbol{\beta}$):

$$\mathcal{I}(\boldsymbol{\beta}, t) = -\frac{\partial}{\partial \beta} \mathbf{U}(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^t \mathbf{V}(\boldsymbol{\beta}, u) d\mathbf{N}_i(u), \tag{3.23}$$

where

$$\mathbf{V} = (\mathbf{S}^{(2)}/\mathbf{S}^{(0)}) - (\mathbf{S}^{(1)}/\mathbf{S}^{(0)})(\mathbf{S}^{(1)}/\mathbf{S}^{(0)})^T,$$

and $\mathbf{S}^{(0)}$, $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ are defined in (3.11), (3.12) and (3.13), respectively.

Put

$$\mathbf{B}_\lambda(\boldsymbol{\beta}) = \left(\frac{\partial}{\partial \beta_1} P_\lambda(\boldsymbol{\beta}), \ldots, \frac{\partial}{\partial \beta_k} P_\lambda(\boldsymbol{\beta}) \right)^T. \tag{3.24}$$

If the first partial derivative of $P_\lambda(\boldsymbol{\beta})$ does not exist at $\beta_i = 0$, $i = 1, \ldots, k$, we set $\mathbf{B}_\lambda(\boldsymbol{\beta})$ equal to zero. Let

$$H_\lambda(\boldsymbol{\beta}) = - \begin{pmatrix} \frac{\partial^2}{\partial \beta_1^2} P_\lambda(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial \beta_1^2} P_\lambda(\boldsymbol{\beta}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \beta_k^2} P_\lambda(\boldsymbol{\beta}) & \cdots & \frac{\partial^2}{\partial \beta_k^2} P_\lambda(\boldsymbol{\beta}) \end{pmatrix}. \tag{3.25}$$

If the second partial derivative of $P_\lambda(\boldsymbol{\beta})$ does not exist at $\beta_i = 0$ or $\beta_j = 0$, $i, j = 1, \ldots, k$, we set it equal to zero.

Then from (3.20) and (3.21), we have

$$\frac{\partial}{\partial \boldsymbol{\beta}} Q(\boldsymbol{\beta}, t) = \mathbf{U}(\boldsymbol{\beta}_0, t) - \mathbf{B}_\lambda(\boldsymbol{\beta}_0) - \left[\mathcal{I}(\boldsymbol{\beta}^*, t) - H_\lambda(\boldsymbol{\beta}^*) \right] (\boldsymbol{\beta} - \boldsymbol{\beta}_0). \tag{3.26}$$

Since $\partial Q(\hat{\boldsymbol{\beta}}, t)/\partial \boldsymbol{\beta} = 0$, (3.26) can be written as

$$\frac{1}{\sqrt{n}} \left[\mathbf{U}(\boldsymbol{\beta}_0, t) - \mathbf{B}_\lambda(\boldsymbol{\beta}_0) \right] = \frac{1}{n} \left[\mathcal{I}(\boldsymbol{\beta}^*, t) - H_\lambda(\boldsymbol{\beta}^*) \right] \sqrt{n} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{3.27}$$

where $\boldsymbol{\beta}^*$ is on the line segment between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. Therefore $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converges to multivariate normal if the left hand side of (3.27) converges in distribution to a multivariate normal and $\left[\mathcal{I}(\boldsymbol{\beta}^*, t) - H_\lambda(\boldsymbol{\beta}^*) \right] / n$ converges in probability to a nonsingular nonrandom matrix.

Theorem 3.2 *(Asymptotic Normality of $\hat{\boldsymbol{\beta}}$ with a nonnegative constant λ). Let the assumptions in Theorem 3.1 hold. Assume that $\Sigma(\boldsymbol{\beta}_0, t)$, defined by (3.14), satisfies*

Condition D(4) and that each entry of $H_\lambda(\boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$. Then

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{D} N(\mathbf{0}, \Sigma(\boldsymbol{\beta}_0, t)^{-1}) \ \text{ as } n \to \infty. \tag{3.28}$$

Proof: By (3.27),

$$\frac{1}{\sqrt{n}}\left[\mathbf{U}(\boldsymbol{\beta}_0, t) - \mathbf{B}_\lambda(\boldsymbol{\beta}_0)\right] = \frac{1}{n}\left[\mathcal{I}(\boldsymbol{\beta}^*, t) - H_\lambda(\boldsymbol{\beta}^*)\right]\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

By Theorem 3.2 of Andersen and Gill (1982), under Conditions A–D, $\mathbf{U}(\boldsymbol{\beta}_0, t)/\sqrt{n} \xrightarrow{D}$ $N(0, \Sigma(\boldsymbol{\beta}_0, t))$ and $\mathcal{I}(\boldsymbol{\beta}^*, t)/n \xrightarrow{P} \Sigma(\boldsymbol{\beta}_0, t)$ as $n \to \infty$. Because $\boldsymbol{\beta}^*$ is on the line segment between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, and $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$ as $n \to \infty$ (see Theorem 3.1), for any $\varepsilon > 0$ and any $\delta > 0$, there exists a value N such that

$$P(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\| > \delta) \leq P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| > \delta) < \varepsilon,$$

for all $n \geq N$. Therefore, $\boldsymbol{\beta}^* \xrightarrow{P} \boldsymbol{\beta}_0$. Then $H_\lambda(\boldsymbol{\beta}^*) \xrightarrow{P} H_\lambda(\boldsymbol{\beta}_0)$ as $n \to \infty$. Since $\mathbf{B}_\lambda(\boldsymbol{\beta}_0)$ and $H_\lambda(\boldsymbol{\beta}_0)$ are constants, $\mathbf{B}_\lambda(\boldsymbol{\beta}_0)/\sqrt{n} \xrightarrow{P} 0$ and $H_\lambda(\boldsymbol{\beta}^*)/n \xrightarrow{P} 0$ as $n \to \infty$. Then (3.28) holds by Slutsky's theorem. □

3.2.4 Consistency and Asymptotic Normality of $\hat{\boldsymbol{\beta}}$ with λ_n

We modify our theorems of consistency (Theorem 3.1) and asymptotic normality (Theorem 3.2) for the tuning parameter λ depending on n, to be denoted by λ_n. Then three examples of regularized estimators satisfying the consistency and asymptotic normality properties are given.

Theorem 3.3 *(Consistency of $\hat{\boldsymbol{\beta}}$ with λ_n). Assume that Conditions A–D are satisfied. Assume that $\hat{\boldsymbol{\beta}}$ is a local maximizer of $Q(\boldsymbol{\beta}, t) = \ell(\boldsymbol{\beta}, t) - n\sum_{j=1}^{k} p_{\lambda_n}(\beta_j)$*

in a neighborhood $\mathcal{B}$ of the true parameter $\boldsymbol{\beta}_0$ satisfying Condition B, and λ_n is a nonnegative value depending on the sample size n. If $p_{\lambda_n}(\beta_j) = o(1)$ as $n \to \infty$ uniformly in $\mathcal{B}$ for all $j = 1, \ldots, k$, then

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0 \ \text{ as } n \to \infty.$$

Proof: Under the assumptions on the penalty function $P_\lambda(\boldsymbol{\beta}) = n\sum_{j=1}^{k} p_{\lambda_n}(\beta_j)$, with λ depending on n, it is sufficient to show that $\sup_{\beta \in A} |P_\lambda(\boldsymbol{\beta}) - P_\lambda(\boldsymbol{\beta}_0)|/n$ in (3.19) converges to zero for any compact set $A \subset \mathcal{B}$. For any compact subset A of $\mathcal{B}$, we have

$$\sup_{\beta \in A} \left| n\sum_{j=1}^{k} p_{\lambda_n}(\beta_j) - n\sum_{j=1}^{k} p_{\lambda_n}(\beta_{j0}) \right| /n \le \sum_{j=1}^{k} \left[\sup_{\beta \in A} |p_{\lambda_n}(\beta_j)| + |p_{\lambda_n}(\beta_{j0})| \right] \longrightarrow 0$$

as $n \to \infty$. The convergence to zero follows from the hypothesis that $p_{\lambda_n}(\beta_j) = o(1)$ uniformly in $\mathcal{B}$ for all $j = 1, \ldots, k$. Then following a proof similar to that of Theorem 3.1, we conclude that $\hat{\boldsymbol{\beta}}$ is consistent. □

Theorem 3.4 *(Asymptotic Normality of $\hat{\boldsymbol{\beta}}$ with λ_n). Let the assumptions in Theorem 3.3 hold. Suppose that $\Sigma(\boldsymbol{\beta}_0, t)$ satisfies Condition D(4). Then*

$$\sqrt{n}\left[\Sigma(\boldsymbol{\beta}_0, t) - \frac{1}{n}H_{\lambda_n}(\boldsymbol{\beta}_0)\right]\left\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \left[\Sigma(\boldsymbol{\beta}_0, t) - \frac{1}{n}H_{\lambda_n}(\boldsymbol{\beta}_0)\right]^{-1}\left[\frac{1}{n}\mathbf{B}_{\lambda_n}(\boldsymbol{\beta}_0)\right]\right\} \tag{3.29}$$

converges to a normal distribution $N(\mathbf{0}, \Sigma(\boldsymbol{\beta}_0, t))$ as $n \to \infty$, where $\mathbf{B}_{\lambda_n}$ and H_{λ_n} are defined as in (3.24) and (3.25) with λ replaced by λ_n.

If $\mathbf{B}_{\lambda_n}(\boldsymbol{\beta})/\sqrt{n}$ converges to a vector of k functions $\mathbf{b}(\boldsymbol{\beta})$ and $-H_{\lambda_n}(\boldsymbol{\beta})/n$ converges to a $k \times k$ matrix $\Sigma_\lambda(\boldsymbol{\beta})$ componentwise as $n \to \infty$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mu(\boldsymbol{\beta}_0), \Sigma^*(\boldsymbol{\beta}_0, t)) \ \text{ as } n \to \infty, \tag{3.30}$$

where

$$\mu(\boldsymbol{\beta}_0) = -\left[\Sigma(\boldsymbol{\beta}_0, t) + \Sigma_\lambda(\boldsymbol{\beta}_0)\right]^{-1} \mathbf{b}(\boldsymbol{\beta}_0),$$

$$\Sigma^*(\boldsymbol{\beta}_0, t) = \left[\Sigma(\boldsymbol{\beta}_0, t) + \Sigma_\lambda(\boldsymbol{\beta}_0)\right]^{-1} \Sigma(\boldsymbol{\beta}_0, t) \left[\Sigma(\boldsymbol{\beta}_0, t) + \Sigma_\lambda(\boldsymbol{\beta}_0)\right]^{-1}.$$

Proof: From (3.27), $\mathbf{U}(\boldsymbol{\beta}_0, t)/\sqrt{n}$ is equal to

$$\sqrt{n}\left[\frac{1}{n}\mathcal{I}(\boldsymbol{\beta}^*, t) - \frac{1}{n}H_{\lambda_n}(\boldsymbol{\beta}^*)\right]\left\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \left[\frac{1}{n}\mathcal{I}(\boldsymbol{\beta}^*, t) - \frac{1}{n}H_{\lambda_n}(\boldsymbol{\beta}^*)\right]^{-1}\left[\frac{1}{n}\mathbf{B}_{\lambda_n}(\boldsymbol{\beta}_0)\right]\right\}. \tag{3.31}$$

By Andersen and Gill (1982) and Theorem 3.3, $\mathbf{U}(\boldsymbol{\beta}_0, t)/\sqrt{n} \xrightarrow{D} N(0, \Sigma(\boldsymbol{\beta}_0, t))$, $\mathcal{I}(\boldsymbol{\beta}^*, t)/n \xrightarrow{P} \Sigma(\boldsymbol{\beta}_0, t)$ and $H_{\lambda_n}(\boldsymbol{\beta}^*) \xrightarrow{P} H_{\lambda_n}(\boldsymbol{\beta}_0)$. Therefore (3.29) converges to $N(0, \Sigma(\boldsymbol{\beta}_0, t))$ by Slutsky's theorem.

Moreover, (3.29) is equivalent to

$$\sqrt{n}\left[\Sigma(\boldsymbol{\beta}_0, t) - \frac{1}{n}H_{\lambda_n}(\boldsymbol{\beta}_0)\right](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{1}{\sqrt{n}}\mathbf{B}_{\lambda_n}(\boldsymbol{\beta}_0).$$

Since $-H_{\lambda_n}(\boldsymbol{\beta}_0)/n \to \Sigma_\lambda(\boldsymbol{\beta}_0)$ pointwise and $\mathbf{B}_{\lambda_n}(\boldsymbol{\beta}_0)/\sqrt{n} \to \mathbf{b}(\boldsymbol{\beta}_0)$ pointwise, we have (3.30) by Slutsky's theorem. □

The following are examples of regularized estimators with their consistency and asymptotic normality properties.

Example 3.2 *(Lasso). Suppose* $P_\lambda(\boldsymbol{\beta}) = n\lambda_n \sum_{j=1}^k |\beta_j|$. *If* $\lambda_n \to 0$, *then* $\hat{\boldsymbol{\beta}}$ *is consistent,* $\mathbf{B}_{\lambda_n}(\boldsymbol{\beta}) = n\lambda_n sign(\boldsymbol{\beta})$ *and* $H_{\lambda_n}(\boldsymbol{\beta}) = \mathbf{0}_{k\times k}$. *If* $\sqrt{n}\lambda_n$ *converges to a constant* C, *then* $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ *converges to a multivariate normal distribution with mean* $-C\Sigma(\boldsymbol{\beta}_0, t)^{-1} sign(\boldsymbol{\beta}_0)$ *and covariance matrix* $\Sigma(\boldsymbol{\beta}_0, t)^{-1}$.

Example 3.3 *(Adaptive Lasso). Suppose* $P_\lambda(\boldsymbol{\beta}) = n\lambda_n \sum_{j=1}^k |\beta_j|/\tilde{\beta}_j^2$, *where* $\tilde{\beta}_j$ *is the OLS estimator of* β_j. *If* $\lambda_n \to 0$, *then* $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$ *as* $n \to \infty$, $\mathbf{B}_{\lambda_n}(\boldsymbol{\beta}) =$

$n\lambda_n sign(\boldsymbol{\beta}) diag(1/\tilde{\beta}_1^2, \ldots, 1/\tilde{\beta}_k^2)$ *and* $H_{\lambda_n}(\boldsymbol{\beta}) = \mathbf{0}_{k\times k}$. *If* $\sqrt{n}\lambda_n$ *converges to a constant* C, *then* $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ *converges to a multivariate normal distribution with mean* $-C\Sigma(\boldsymbol{\beta}_0, t)^{-1} sign(\boldsymbol{\beta}_0) diag(1/\beta_{10}^2, \ldots, 1/\beta_{k0}^2)$ *and covariance matrix* $\Sigma(\boldsymbol{\beta}_0, t)^{-1}$.

Example 3.4 *(Ridge). Suppose* $P_\lambda(\boldsymbol{\beta}) = n\lambda_n \sum_{j=1}^k \beta_j^2$. *If* $\lambda_n \to 0$, *then* $\hat{\boldsymbol{\beta}}$ *is consistent,* $\mathbf{B}_{\lambda_n}(\boldsymbol{\beta}) = 2n\lambda_n\boldsymbol{\beta}$ *and* $H_\lambda(\boldsymbol{\beta}) = -2n\lambda_n\mathbf{I}_{k\times k}$. *If* $\sqrt{n}\lambda_n$ *converges to a constant* C, *then* $\mathbf{B}_{\lambda_n}(\boldsymbol{\beta})/\sqrt{n} \to 2C\boldsymbol{\beta}$ *and* $-H_{\lambda_n}(\boldsymbol{\beta})/n = 2\lambda_n\mathbf{I}_{k\times k} \to \mathbf{0}_{k\times k}$. *Hence,* $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ *converges to a normal distribution with mean* $-2C\Sigma(\boldsymbol{\beta}_0, t)^{-1}\boldsymbol{\beta}_0$ *and covariance matrix* $\Sigma(\boldsymbol{\beta}_0, t)^{-1}$.

3.2.5 Oracle Properties of $\hat{\boldsymbol{\beta}}$ with λ_n

In variable selection and estimation, oracle properties of estimators in the regularized regression model are studied in the literature. See Zou (2006) and references therein. Oracle properties are some asymptotically optimal properties which can be conveniently described as follows. Suppose we know a priori that s components (for $s < k$) of the true parameter $\boldsymbol{\beta}_0$ are nonzero while the remaining $(k-s)$ components are zero. Rearrange these components so that

$$\boldsymbol{\beta}_0 = \begin{pmatrix} \boldsymbol{\beta}_{10} \\ \\ \boldsymbol{\beta}_{20} \end{pmatrix}, \tag{3.32}$$

where $\boldsymbol{\beta}_{10} = (\beta_{10}, \ldots, \beta_{s0})^T$ with $\beta_{i0} \neq 0$ for $i = 1, \ldots, s$, and $\boldsymbol{\beta}_{20} = (\beta_{(s+1)0}, \ldots, \beta_{k0})^T$ with $\beta_{j0} = 0$ for $j = s+1, \ldots, k$. That is, the true model depends only on a relatively small subset of the components of $\boldsymbol{\beta}_0$ while the other components are zero. This is called a sparsity condition. Sparsity also refers to, with probability tending

to one, the estimator for $\boldsymbol{\beta}_{20}$ is zero. We shall consider the case where the tuning parameter depends on n, to be denoted by λ_n, and the penalty function is

$$P_{\lambda_n}(\boldsymbol{\beta}) = n \sum_{j=1}^{k} p_{\lambda_n}(\beta_j), \tag{3.33}$$

where $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(\beta_j) \geq 0$ for all j. Under conditions on λ_n to be specified later, the regularized estimator

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$$

of $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$ has oracle properties.

Following Fan and Li (2002), we say an estimation procedure Δ that produces the estimator $\hat{\boldsymbol{\beta}}(\Delta)$ has oracle properties if it is a consistent variable selection, that is,

(a) With probability tending to one, the procedure Δ gives *correct identification of* $\boldsymbol{\beta}_{20}$, i.e. $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$,

and $\boldsymbol{\beta}_{10}$ can be estimated as well as if the correct sub-model were known in advance. In the present case, it means $\hat{\boldsymbol{\beta}}_1$ has an optimal estimation rate, in the sense of

(b) *Asymptotic normality*, i.e. the distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})$ converges to a normal distribution with mean zero and covariance matrix $\mathcal{I}_1^{-1}(\boldsymbol{\beta}_{10}, \mathbf{0})$, where $\mathcal{I}_1(\boldsymbol{\beta}_{10}, \mathbf{0})$ is the Fisher information for $\boldsymbol{\beta}_1$, knowing $\boldsymbol{\beta}_{20} = 0$.

We begin with the presentation of the results of Fan and Li (Theorem 3.2, 2002) who established oracle properties of the local maximizer $\hat{\boldsymbol{\beta}}$ under certain conditions

and with a proper choice of regularization parameter λ_n. Their theorem is restated in this thesis as Theorem 3.5.

Theorem 3.5 *(Oracle properties, Fan and Li 2002). Assume that Conditions A–D of Section 3.2 are satisfied. Let the sparsity condition (3.32) hold. Consider the penalty function* $P_{\lambda_n}(\boldsymbol{\beta}) = n\sum_{j=1}^{k} p_{\lambda_n}(\beta_j)$ *given in (3.33). Let* p'_{λ_n} *and* p''_{λ_n} *denote the first and the second derivatives of* p_{λ_n} *with respect to* β_j*. Assume that* $a_n = \max\{|p'_{\lambda_n}(|\beta_{j0}|)| : j = 1, \ldots, s\} = O(1/\sqrt{n})$ *and* $b_n = \max\{|p''_{\lambda_n}(|\beta_{j0}|)| : j = 1, \ldots, s\} = o(1)$ *as* $n \to \infty$*. Suppose (a)* $\lambda_n \to 0$*, (b)* $\sqrt{n}\lambda_n \to \infty$ *as* $n \to \infty$*, and (c)* $\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$.

Let $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$ *be the local maximizer of* $Q(\boldsymbol{\beta}, t) = \ell(\boldsymbol{\beta}, t) - P_{\lambda_n}(\boldsymbol{\beta})$ *in the neighborhood* $\mathcal{B}(\boldsymbol{\beta}_0, C/\sqrt{n})$*, where* C *is a given positive constant. Then with probability tending to one,*

(i) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$,

(ii)

$$\sqrt{n}\left[\Sigma_1(\boldsymbol{\beta}_0, t) + \Sigma_{\lambda_n}\right]\left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + \left[\Sigma_1(\boldsymbol{\beta}_0, t) + \Sigma_{\lambda_n}\right]^{-1}\mathbf{b}_{\lambda_n}\right\} \tag{3.34}$$

converges to a normal distribution $N(\mathbf{0}, \Sigma_1(\boldsymbol{\beta}_0, t))$ *as* $n \to \infty$*, where* $\Sigma_1(\boldsymbol{\beta}_0, t)$ *is the principal* $s \times s$ *submatrix of* $\Sigma(\boldsymbol{\beta}_0, t)$ *defined in (3.14),*

$$\Sigma_{\lambda_n} = diag\left(p''_{\lambda_n}(|\beta_{10}|), \ldots, p''_{\lambda_n}(|\beta_{s0}|\right),$$

and

$$\mathbf{b}_{\lambda_n} = \left(p'_{\lambda_n}(|\beta_{10}|)sign(\beta_{10}), \ldots, p'_{\lambda_n}(|\beta_{s0}|)sign(\beta_{s0})\right).$$

Remark: We find the proof given in Fan and Li (2002) is not transparent. The repeated use of bounded in probability $O_p(\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|/\sqrt{n})$, where $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|$ is nonrandom, adds to the confusion.

3.3 Generalized Cross-Validation Criterion

We have shown how to estimate the tuning parameter λ in the regularized linear regression by using the GCV criterion in Section 2.3.2. Recall that the GCV function is given by

$$\text{GCV}(\lambda) = \frac{\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{n[1 - tr(A)/n]^2}. \tag{3.35}$$

However, the GCV is not an easily implementable tool for the proportional hazards model. In the linear regression model (2.1) of Chapter 2, the random variable $\mathbf{y}_i$ is explicitly modeled by

$$\mathbf{y}_i = \sum_{j=1}^{k} x_{ij}\beta_j + \mathbf{e}_i \;\; i = 1, \ldots, n.$$

Once the estimator $\hat{\boldsymbol{\beta}}_\lambda^{-i}$ is calculated, the estimate $\hat{\mathbf{y}}_i = X_i\hat{\boldsymbol{\beta}}_\lambda^{-i}$ of $\mathbf{y}_i$ can be determined immediately as it is needed in GCV(λ) (see (3.35)). On the contrary, in the proportional hazards model, we model the survival probability or equivalently the hazard rate of a patient, but not the actual survival time (T) of a patient which is the required input variable in GCV(λ) with T_i taking the role of $\mathbf{y}_i$. In theory, we could obtain an estimate $\hat{T}_i$ of T_i by sampling from an estimated survival time distribution of the proportional hazards model. However, such a task would introduce additional sampling error and computational inefficiency. Moreover, the linear assumption (2.18) is not satisfied. These problems exist even if we assume that all

survival data are uncensored. The other reason is that the linear assumption (2.18) is not satisfied. In other words, we can estimate the hazard rate that a patient will die at time t, so that we can compute $\sum_{i=1}^{n}(h(t)_i - \hat{h}(t)_i)^2$ while considering the hazard rate $h(t)_i$ as $\mathbf{y}_i$, but unfortunately there is no evidence that there exists an $n \times n$ matrix A such that $(\hat{h}(t)_1, \ldots, \hat{h}(t)_n)^T = A(h(t)_1, \ldots, h(t)_n)^T$.

A traditional way to apply the GCV criterion to the regularized proportional hazards model is to replace the sum of squares errors by minus log partial likelihood, that is,

$$\mathrm{GCV}(\lambda) = \frac{-\ell(\hat{\boldsymbol{\beta}}, t)}{n[1 - tr(A)/n]^2}. \tag{3.36}$$

Tibshirani (1997) assumed that the proportional hazards model can be simplified as a generalized linear model and proposed that $\hat{\boldsymbol{\beta}} \approx (\mathbf{Z}^T\mathbf{D}\mathbf{Z} + \lambda\mathbf{W})^{-1}\mathbf{Z}^T\mathbf{D}\mathbf{Z}\boldsymbol{\beta}$, where $\mathbf{D}$ is a diagonal matrix with the same diagonal elements as $-\nabla^2\ell(\hat{\boldsymbol{\beta}}, t)$ and $\mathbf{W}$ is the diagonal matrix of $|\hat{\boldsymbol{\beta}}|^{-1}$ for the Lasso penalty. Therefore,

$$A_1(\lambda) = \mathbf{Z}(\mathbf{Z}^T\mathbf{D}\mathbf{Z} + \lambda\mathbf{W})^{-1}\mathbf{Z}^T\mathbf{D}. \tag{3.37}$$

Fan and Li (2002) claimed that an approximate linear relationship between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ can be derived from the iterative Newton-Raphson algorithm and defined

$$A_2(\lambda) = \left[\nabla^2\ell(\hat{\boldsymbol{\beta}}, t) + \Sigma_\lambda(\hat{\boldsymbol{\beta}})\right]^{-1} \nabla^2\ell(\hat{\boldsymbol{\beta}}, t) \tag{3.38}$$

where $\Sigma_\lambda(\hat{\boldsymbol{\beta}}) = \mathrm{diag}\left(\partial P_\lambda(|\hat{\boldsymbol{\beta}}|)/\partial|\boldsymbol{\beta}|\right)\mathbf{W}$, $P_\lambda(\cdot)$ is the penalty function and $\mathbf{W}$ is defined as in (3.37).

The purpose of introducing the GCV criterion is to reduce the intensive computation of cross-validation. Its success in the linear regression model depends

critically on a linear relationship between the true parameter and its estimator. In the proportional hazards model, however, the necessary linear relationship can only be obtained by an approximation since the log partial likelihood $\ell(\boldsymbol{\beta}, t)$ or the regularized estimator $\hat{\boldsymbol{\beta}}$ is not a linear function of $\boldsymbol{\beta}$ at all. Different approximations can result in different linear relations like (3.37) and (3.38), and then make the GCV score vary.

Therefore, to avoid the linear assumption between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ is a strong motivation for us to propose a nonparametric method called AUC criterion in the following chapter.

Chapter 4

Area under Receiver Operating Characteristic Curve Criterion

In this chapter, we present the receiver operating characteristic (ROC) curve, define the area under a ROC curve (AUC) as a measure of diagnostic performance of the estimated proportional hazards model. We propose a method of selecting the tuning parameter λ by maximizing the AUC.

4.1 The ROC and AUC

Consider a population Ω of individuals and a particular disease that affects some of the individuals in the population, for example, individuals with a certain type of cancer. There are a variety of clinical tests for diagnosis of the disease. The ROC is a widely used method for evaluating the performance of a diagnostic test. A good test would have high probability of true positive diagnosis and low probability of false positive. In this thesis, we shall use the ROC in a different way. Here the ROC will be used to determine the tuning parameter λ in selecting variables (or covariates) that are most relevant to the disease under study.

We shall begin with the definition of ROC. Let D denote the true disease status of an individual in the population with $D = 1$ indicating the presence of the disease and $D = 0$, the absence of the disease. Consider a particular diagnostic test and let W represent the measurement used in the diagnostic test. We assume that W is a

real-valued random variable with a continuous distribution. Specifying a threshold value w, the event $[W > w]$ indicates that the diagnostic test is positive while the event $[W \leq w]$ indicates that the test is negative. For any specified threshold w, define the true positive probability and the false positive probability, respectively, by

$$\mathrm{TP}(w) = P(W > w|D = 1), \tag{4.1}$$

$$\mathrm{FP}(w) = P(W > w|D = 0). \tag{4.2}$$

The ROC curve (see Figure 4.1) is defined as the path in the first quadrant obtained by connecting all the pairs $(\mathrm{FP}(w), \mathrm{TP}(w))$ as w runs through the entire range of the threshold.

Conceptually, let us divide the population Ω into two subsets Ω_1 and Ω_0 of individuals with and without the said disease. We introduce two independent random variables W_1 and W_0 that carry the conditional distributions (4.1) and (4.2) respectively. Then W_1 is the diagnostic measurement of a randomly selected individual from Ω_1, and W_0 is the measurement on an individual randomly selected from Ω_0. Therefore, the diagnostic measurement W of a randomly selected individual from Ω has the following probability distribution

$$\begin{aligned} P(W > w) &= P(W > w|D = 1)P(D = 1) + P(W > w|D = 0)P(D = 0) \\ &= P(W_1 > w)P(D = 1) + P(W_0 > w)P(D = 0). \end{aligned}$$

Let S_1 and S_0 denote the survival functions of W_1 and W_0, respectively, that is,

$$S_1(w) = P(W_1 > w) = P(W > w|D = 1),$$

$$S_0(w) = P(W_0 > w) = P(W > w|D = 0).$$

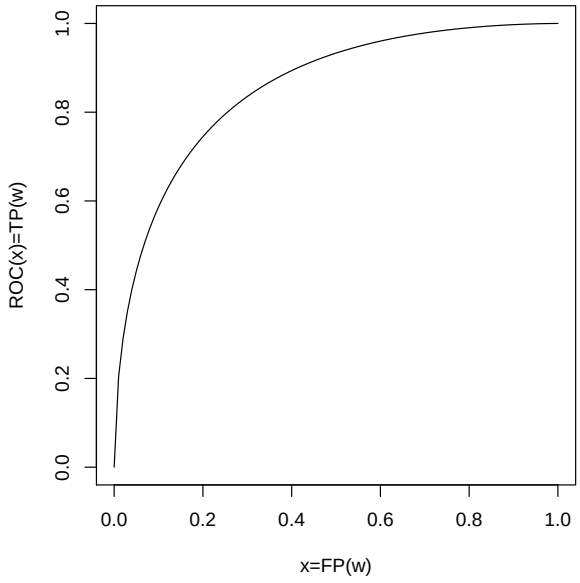

Figure 4.1: An ROC curve.

We change variable from w to x by putting $x = S_0(w)$ and consider the inverse function of S_0 defined by

$$S_0^{-1}(x) = \inf\{w : S_0(w) \geq x\}.$$

Then the ROC curve is

$$\mathrm{ROC}(x) = S_1(S_0^{-1}(x)), \quad x \in [0, 1]. \tag{4.3}$$

Because $S_0^{-1}(x)$ is nonincreasing in x, $\mathrm{ROC}(x)$ is a nondecreasing function of x. See the illustration in Figure 4.1.

The ROC curve is a fundamental statistical tool for measuring the diagnostic accuracy of a clinical test. Figure 4.2 gives an illustration of the use of ROC. If, for

some threshold w^*, $\mathrm{TP}(w^*) = 1$ and $\mathrm{FP}(w^*) = 0$, then we have a perfect test which can distinguish a diseased individual from a healthy one. Thus $\mathrm{TP}(w) = 1$ for all $w \leq w^*$, and $\mathrm{FP}(w) = 0$ for all $w \geq w^*$. The corresponding ROC curve is the left and upper borders of the unit square in the first quadrant, the curve A in Figure 4.2. Various tests (with different measurements W) can be compared visually by using their corresponding ROC curves because regardless of the scales of different W's (say blood sugar or blood pressure), the W's have been converted to $X = S_0(W)$.

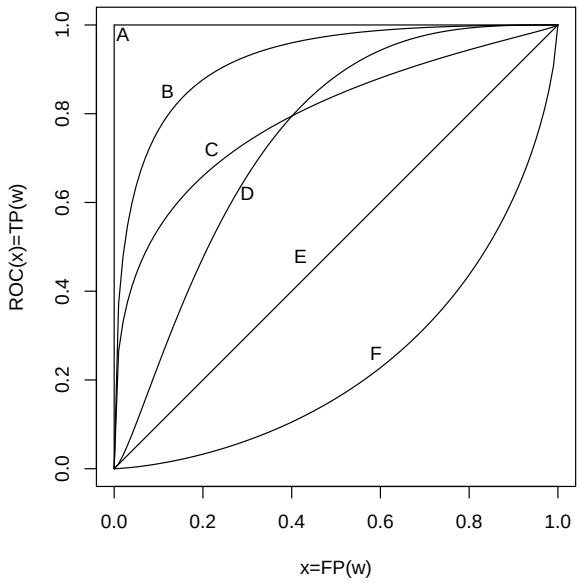

Figure 4.2: ROC curves. Curve A is a perfect curve. Curve B is better than curves C, D, E and F. Curve E is non-informative. Curve F is the worst among these six curves.

Better curves are closer to the upper left hand corner. For example, curve A is the best. Curve B is the second best. Curves C and D cannot be ordered because

they cross each other, but both of them are better than curves E and F. Curve E has a 50% chance of giving false and true positive result. We say this diagnosis is uninformative. Any curve lying below the non-informative curve is a bad diagnostic test because for any threshold value w, the probability that a false positive is larger than a true positive. Therefore curve F is the worst among all the curves.

It is possible for two ROC curves to cross each one another, like curves C and D, which shows that neither of the two diagnostic tests is necessarily better then the other. Because of this problem, in order to rank the performance of various diagnostic tests, sometimes the area under the ROC curve (AUC) is used.

The AUC is defined by

$$\text{AUC} = \int_0^1 \text{ROC}(x)dx. \tag{4.4}$$

Obviously, because W_1 and W_0 are independent,

$$\begin{aligned} \text{AUC} &= \int_{\infty}^{-\infty} S_1(w)dS_0(w) \\ &= \int_{-\infty}^{\infty} P(W_1 > w)dP(W_0 \leq w) \\ &= P(W_1 > W_0). \end{aligned}$$

The value of AUC has been used to compare the performance of diagnostic tests. The larger the AUC, the better the diagnostic test. See Figure 4.2. The AUC value of curve A is one, the AUC of curve E is one-half, the AUCs of curves above E is between one-half and one, and the AUCs of curves below E are less than one-half. Thus if the AUC value of curve C is larger than the one of curve D, we say that test C is better than test D.

4.2 ROC and AUC for Proportional Hazards Model

In this section, we give the definition of the ROC and the AUC in terms of the proportional hazards model. We then study estimation problems using the AUC. In particular, an AUC criterion will be used for selecting the tuning parameter λ in the regularized log partial likelihood (3.8).

4.2.1 Definition

Let T be the survival time of an individual. For a fixed time u, let the indicator $I[T \leq u]$ play the role of our binary outcome D. That is, using the terminology in Section 4.1, we refer to individuals whose survival times are less or equal to u as "diseased" and those who survive longer than u as "healthy" ones or "disease-free". With a judicious choice of u, we divide the population into two groups: those who died by time u as Ω_1 and those who survived beyond u as Ω_0. An example of this is the standard practice of using the five-year survival time as a measure of an individual's success with a cancer treatment. Thus $D = 1$ if a patient dies within five years and $D = 0$ if s/he lives beyond five years. Given a known covariate vector $\mathbf{Z}$, we use a linear combination of regression coefficients $\boldsymbol{\beta}$ and the covariate, $\boldsymbol{\beta}^T\mathbf{Z}$, as the measurement W of a diagnostic test.

The proportional hazards model gives the survival probability of an individual:

$$P(T > u) = \exp\left\{-\int_0^u h_0(t)dt \exp(\boldsymbol{\beta}^T\mathbf{Z})\right\}. \tag{4.5}$$

In this thesis, the proposed variable selection criterion for the proportional hazards model will be applied to patients with squamous cell cancer in the head and neck

region. There are many potential covariates $\mathbf{Z}$ for this disease (see Table F.1). A goal of this thesis is to use our proposed statistical method to look for a subset of covariates as good predictors of the survival time T.

In the application of the ROC and the AUC, we take $W = \boldsymbol{\beta}^T\mathbf{Z}$ as the measurement of the diagnostic test. Then the true positive and false positive probabilities of (4.1) and (4.2) are given, respectively, by

$$\mathrm{TP}(w,u) = P(\boldsymbol{\beta}^T\mathbf{Z} > w | T \le u), \tag{4.6}$$

$$\mathrm{FP}(w,u) = P(\boldsymbol{\beta}^T\mathbf{Z} > w | T > u). \tag{4.7}$$

Now TP and FP are functions of w and u. According to (4.3), the ROC function is

$$\mathrm{ROC}(x,u) = \mathrm{TP}(\mathrm{FP}^{-1}(x,u),u) \tag{4.8}$$

for all $x \in [0,1]$ and $u \in [0,\infty)$, where $\mathrm{FP}^{-1}(x,u) = \inf\{w : \mathrm{FP}(w,u) \ge x\}$. The AUC is a function of u:

$$\mathrm{AUC}(u) = \int_0^1 \mathrm{ROC}(x,u)dx. \tag{4.9}$$

Suppose that $\mathbf{Z}_1$ is the covariate vector of a randomly selected individual from the population Ω_1 of patients who have died by time u, and $\mathbf{Z}_0$ is that of a randomly selected individual from the population Ω_0 of patients whose survival times are longer than u. By Equation (4.9),

$$\mathrm{AUC}(u) = P(\boldsymbol{\beta}^T\mathbf{Z}_1 > \boldsymbol{\beta}^T\mathbf{Z}_0). \tag{4.10}$$

As an example, the value of $\mathrm{AUC}(u) = 0.9$ means that there is a 90% probability that a randomly selected individual from population Ω_0 will have a value of $\boldsymbol{\beta}^T\mathbf{Z}$ larger than that of a randomly selected individual from population Ω_1.

4.2.2 Estimation

Let $(\tilde{T}_i, \delta_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, be an i.i.d. sample of n random vectors, where $\tilde{T}_i$ is the observed survival time of the i-th individual whose censoring indicator and covariate vector are denoted by δ_i and $\mathbf{Z}_i$ as defined in Section 3.1. The first step of the estimation problem is to select an optimal regularized parameter, λ, for the regularized log partial likelihood (3.8). To use our proposed AUC criterion for determining λ, it is necessary to estimate the true and the false positive probabilities (TP and FP). From these estimates of TP and FP, estimates of the ROC and AUC will be derived.

The problems of estimating TP, FP, ROC and AUC have been studied in the literature; see, for example, Heagerty and Zheng (2005) and references therein. Their linear predictor $W = \boldsymbol{\beta}^T\mathbf{Z}$ as the diagnostic measurement is the same as the one used in this thesis. Their true positive probability is given by

$$\text{TP} = P(\boldsymbol{\beta}^T\mathbf{Z} \geq w | T < t). \tag{4.11}$$

If T follows the proportional hazard distribution with the same $\boldsymbol{\beta}^T\mathbf{Z}$ in the hazard, then except for the nuisance hazard $h_0(t)$, the diagnostic measurement $\boldsymbol{\beta}^T\mathbf{Z}$ has no ability to distinguish a diseased individual from a disease-free individual. This may be the reason for Heagerty and Zheng to introduce an additional parameter γ in the proportional hazard model as

$$h(t) = h_0(t) \exp(\gamma\boldsymbol{\beta}^T\mathbf{Z}) \tag{4.12}$$

However, introducing the additional parameter γ would make the parameter $\boldsymbol{\beta}$ non-identifiable in the estimation.

Our purpose of estimating TP, FP, ROC and AUC is to determine λ. In Appendix A, we use simulation to illustrate the problem of using (4.11) with T having the proportional hazard distribution. Simulation shows that the AUC criterion using the $\widehat{\mathrm{AUC}}(u,\lambda)$ (A.4) has no power of discrimination. This means that the regularized estimator $\hat{\boldsymbol{\beta}}$ selected by the AUC was always the same as the MLE $\tilde{\boldsymbol{\beta}}$ obtained by maximizing the log partial likelihood without a penalty function (i.e. $\lambda = 0$).

The procedure we use is the following. Given a time u, divide the sample as follows. Consider the subset of samples $\{i : \tilde{T}_i \leq u, \delta_i = 1\}$ from the "diseased" population Ω_1, and the subset $\{i : \tilde{T}_i > u\}$ from the "healthy" population Ω_0. As for the set $\{i : \tilde{T}_i \leq u, \delta_i = 0\}$, there is no information as to which population it belongs.

Based on the known "diseased" and "healthy" subsets, $\{i : \tilde{T}_i \leq u, \delta_i = 1\}$ and $\{i : \tilde{T}_i > u\}$, we estimate the true positive probability (4.6) and false positive probability (4.7), respectively, by their empirical distribution functions:

$$\widehat{\mathrm{TP}}(w,u) = \sum_{i=1}^{n} \delta_i I[\boldsymbol{\beta}^T \mathbf{Z}_i \geq w, \tilde{T}_i \leq u]/n_1, \tag{4.13}$$

$$\widehat{\mathrm{FP}}(w,u) = \sum_{i=1}^{n} I[\boldsymbol{\beta}^T \mathbf{Z}_i \geq w, \tilde{T}_i > u]/n_2, \tag{4.14}$$

where $n_1 = \sum_{i=1}^{n} \delta_i I[\tilde{T}_i \leq u]$ and $n_2 = \sum_{i=1}^{n} I[\tilde{T}_i > u]$. The ROC function (4.8) can be estimated by

$$\widehat{\mathrm{ROC}}(x,u) = \widehat{\mathrm{TP}}(\widehat{\mathrm{FP}}^{-1}(x,u),u), \tag{4.15}$$

where $\widehat{\mathrm{FP}}^{-1}(x,u) = \inf\{w : \widehat{\mathrm{FP}}(w,u) \geq x\}$. The estimator of the AUC function

(4.9) is given by

$$\widehat{\mathrm{AUC}}(u) = \int_0^1 \widehat{\mathrm{ROC}}(x, u)dx. \tag{4.16}$$

If $\boldsymbol{\beta}$ is known, then $\widehat{\mathrm{AUC}}(u)$ can be written as a U statistic of the form

$$\sum_{j=1}^{n}\sum_{i\neq j}\delta_i \left\{ I[\boldsymbol{\beta}^T\mathbf{Z}_i > \boldsymbol{\beta}^T\mathbf{Z}_j, \tilde{T}_i \leq u, \tilde{T}_j > u] + \frac{1}{2} I[\boldsymbol{\beta}^T\mathbf{Z}_i = \boldsymbol{\beta}^T\mathbf{Z}_j, \tilde{T}_i \leq u, \tilde{T}_j > u] \right\} / n_1 n_2. \tag{4.17}$$

In our case, the unknown $\boldsymbol{\beta}$'s in (4.13), (4.14) and (4.17) are replaced by the regularized estimator $\hat{\boldsymbol{\beta}}_\lambda$'s with a fixed tuning parameter λ. To emphasize the role of λ, we shall use the notation $\widehat{\mathrm{AUC}}(u, \lambda)$ instead of $\widehat{\mathrm{AUC}}(u)$ to indicate it is a function of u and λ.

The next section describes an AUC criterion for selecting the tuning parameter λ in the regularized proportional hazards model.

4.3 AUC Criterion for Determining λ

Suppose Θ_λ is a set of possible tuning parameters λ. Given a λ, we can obtain an estimator $\hat{\boldsymbol{\beta}}_\lambda$ by maximizing the regularized log partial likelihood $\{\ell(\boldsymbol{\beta}, t) - P_\lambda(\boldsymbol{\beta})\}$ in Equation (3.8). We carry out the regulation estimation for each $\lambda \in \Theta_\lambda$ and compute the estimated AUC value (4.17) for the estimator $\hat{\boldsymbol{\beta}}_\lambda$. If there exists a λ^* that maximizes (4.17), we say $\hat{\boldsymbol{\beta}}_{\lambda^*}$ is the best regularized estimator by the AUC criterion. That is, given a specific time u and a fixed domain Θ_λ of λ, we consider $\hat{\boldsymbol{\beta}}_{\lambda^*}$ as the selected regularized estimator by the AUC criterion if

$$\lambda^* = \underset{\lambda\in\Theta_\lambda}{\operatorname{argmax}}\, \widehat{\mathrm{AUC}}(u, \lambda).$$

The choice of u depends on the survival time we are interested in. For example, "five years" is a popular choice for measuring the success of a cancer treatment. Note that u cannot be so small or so large that the diseased population Ω_1 or the healthy population Ω_0 are empty sets.

One may take $\Theta_\lambda = \{\lambda : \lambda \geq 0\}$ as the set of all possible choices of λ, but that would demand intensive computations. Often due to different research purposes, it is possible to limit Θ_λ to a subset of $\{\lambda : \lambda \geq 0\}$. For example, let $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\beta}_{\lambda,1}, \ldots, \hat{\beta}_{\lambda,k})$. We may choose $\Theta_\lambda^\kappa = \{\lambda \geq 0 : \sum_{j=1}^k I[\hat{\beta}_{\lambda,j} \neq 0] \leq \kappa\} \subset \{\lambda : \lambda \geq 0\}$. This would guarantee that the dimension of selected estimator $\hat{\boldsymbol{\beta}}_{\lambda_\kappa^*}$ is at most κ, where $\lambda_\kappa^* = \underset{\lambda \in \Theta_\lambda^\kappa}{\operatorname{argmax}} \widehat{\mathrm{AUC}}(u, \lambda)$. We can also choose the subset $\Theta_\lambda^\alpha = \{\lambda \geq 0 : \mathrm{AUC}(u, \lambda) > \alpha\}$. Then the selected $\hat{\boldsymbol{\beta}}_{\lambda_\alpha^*}$ guarantees its resulting AUC value larger than α, where $\lambda_\alpha^* = \underset{\lambda \in \Theta_\lambda^\alpha}{\operatorname{argmax}} \widehat{\mathrm{AUC}}(u, \lambda)$.

Chapter 5

Computational Study

In this chapter, we develop a computational algorithm based on the R software to find the regularized estimator $\hat{\boldsymbol{\beta}}$. Using simulated data, we compare the AUC criterion with the GCV criterion in three scenarios.

5.1 The Algorithm

The k-dimensional regularized estimator $\hat{\boldsymbol{\beta}}$ is obtained by maximizing the regularized log partial likelihood

$$Q(\boldsymbol{\beta}, t) = \ell(\boldsymbol{\beta}, t) - P_\lambda(\boldsymbol{\beta}), \tag{5.1}$$

where $\ell(\boldsymbol{\beta}, t)$ is the log partial likelihood given in (3.6). The Newton-Raphson method is used for solving $\hat{\boldsymbol{\beta}}$ iteratively.

Taylor's expansion of $\ell(\boldsymbol{\beta}, t)$ at the true value $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0k})^T$ with a linear quadratic approximation of $P_\lambda(\boldsymbol{\beta})$ yields

$$\begin{aligned} \hat{Q}(\boldsymbol{\beta}, t|\boldsymbol{\beta}_0) &= \Big[\ell(\boldsymbol{\beta}_0, t) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \nabla \ell(\boldsymbol{\beta}_0, t) \\ &\quad + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \nabla^2 \ell(\boldsymbol{\beta}_0, t)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\Big] - \boldsymbol{\beta}^T D_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}, \end{aligned} \tag{5.2}$$

where $\nabla \ell(\boldsymbol{\beta}_0, t) = \partial \ell(\boldsymbol{\beta}_0, t)/\partial \boldsymbol{\beta}$ is a gradient vector, $\nabla^2 \ell(\boldsymbol{\beta}_0, t) = \partial^2 \ell(\boldsymbol{\beta}_0, t)/\partial \boldsymbol{\beta}\boldsymbol{\beta}^T$ is a Hessian matrix, and $D_\lambda(\boldsymbol{\beta}_0)$ is a $k \times k$ matrix such that $\boldsymbol{\beta}^T D_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}$ is a quadratic approximation of $P_\lambda(\boldsymbol{\beta})$. For example, if $P_\lambda(\boldsymbol{\beta})$ is a Lasso penalty function (Equation

(2.9) with $p = 1$), we take $D_\lambda(\boldsymbol{\beta}_0)$ as a diagonal matrix whose j-th main diagonal entry is $\lambda/|\beta_{0j}|$ if $\beta_{0j} \neq 0$ and zero if otherwise. This is obtained by approximating

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{k} |\beta_j| \approx \lambda \sum_{j=1}^{k} \beta_j^2/|\beta_{0j}| \text{ for } \beta_j \approx \beta_{0j}.$$

This application was used by Tibshirani (1997), and followed by many others to reduce the computational burden. There are other quadratic approximations, see e.g. Fan and Li (2002).

Let $\boldsymbol{\beta}_\lambda^{(0)}$ denote the selected initial value. Then $\boldsymbol{\beta}_\lambda^{(1)} = \underset{\beta}{\text{argmax}}\, \hat{Q}(\boldsymbol{\beta}, t|\boldsymbol{\beta}_\lambda^{(0)})$. The optimization of (5.1) can be updated by iteration, for $i \geq 1$,

$$\begin{aligned}
\boldsymbol{\beta}_\lambda^{(i+1)} &= \underset{\beta}{\text{argmax}}\, \hat{Q}(\boldsymbol{\beta}, t|\boldsymbol{\beta}_\lambda^{(i)}) \\
&= \underset{\beta}{\text{argmax}} \Big\{ \boldsymbol{\beta}^T \left[\nabla \ell(\boldsymbol{\beta}_\lambda^{(i)}, t) - \nabla^2 \ell(\boldsymbol{\beta}_\lambda^{(i)}, t) \boldsymbol{\beta}_\lambda^{(i)} \right] \\
&\qquad\qquad + \frac{1}{2}\, \boldsymbol{\beta}^T \left[\nabla^2 \ell(\boldsymbol{\beta}_\lambda^{(i)}, t) - 2D_\lambda(\boldsymbol{\beta}_\lambda^{(i)}) \right] \boldsymbol{\beta} \Big\} \\
&= -\left[\nabla^2 \ell(\boldsymbol{\beta}_\lambda^{(i)}, t) - 2D_\lambda(\boldsymbol{\beta}_\lambda^{(i)}) \right]^{-1} \left[\nabla \ell(\boldsymbol{\beta}_\lambda^{(i)}, t) - \nabla^2 \ell(\boldsymbol{\beta}_\lambda^{(i)}, t) \right]. \qquad (5.3)
\end{aligned}$$

We set the stopping criterion as

$$\|\boldsymbol{\beta}_\lambda^{(i+1)} - \boldsymbol{\beta}_\lambda^{(i)}\| < 10^{-7}. \qquad (5.4)$$

That is, when (5.4) is satisfied, we set $\hat{\boldsymbol{\beta}}_\lambda = \boldsymbol{\beta}_\lambda^{(i+1)}$. See detailed R programming codes in Appendix B.1 and B.2.

Our next task is to select the "best" tuning parameter $\hat{\lambda}$ and set $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ as our optimal regularized estimator of $\boldsymbol{\beta}$. We shall use the AUC criterion discussed in Section 4.3 and also a commonly used criterion based on GCV given in Equation (3.36) of Section 3.3. The results for $\hat{\lambda}$ will be compared for both criteria. Equation

(3.36) is

$$\widehat{\mathrm{GCV}}(\lambda) = \frac{-\ell(\hat{\boldsymbol{\beta}}_\lambda, t)}{n[1 - tr(\hat{A})/n]^2}, \tag{5.5}$$

where $\hat{A} = [\nabla^2\ell(\hat{\boldsymbol{\beta}}_\lambda, t) + D_\lambda(\hat{\boldsymbol{\beta}}_\lambda)]^{-1}\nabla^2\ell(\hat{\boldsymbol{\beta}}_\lambda, t)$ (see Equation (3.38)). Given a domain Θ_λ of the tuning parameter λ, the "best" selected regularized estimator by the GCV criterion is $\hat{\boldsymbol{\beta}}_{\hat{\lambda}_{GCV}}$ where

$$\hat{\lambda}_{GCV} = \underset{\lambda\in\Theta_\lambda}{\operatorname{argmin}}\, \widehat{\mathrm{GCV}}(\lambda). \tag{5.6}$$

For our AUC criterion introduced in Section 4.3, the "best" selected regularized estimator will be $\hat{\boldsymbol{\beta}}_{\hat{\lambda}_{AUC}}$ where

$$\hat{\lambda}_{AUC} = \underset{\lambda\in\Theta_\lambda}{\operatorname{argmin}}\, \widehat{\mathrm{AUC}}(u, \lambda) \tag{5.7}$$

and u is a specified time. Detailed R programming codes for calculating the AUC and the GCV are given in Appendix B.3 and B.4, respectively.

The iterative procedures are conducted in the following steps:

1. Choose a λ from its domain Θ_λ. If Θ_λ is an interval, for example, $\lambda \in [0, 10]$, choose $\lambda_j, j = 1, \ldots, J$, where $0 = \lambda_1 < \lambda_2 < \ldots < \lambda_J = 10$, and J is a large number.

2. Let the initial vector $\boldsymbol{\beta}_\lambda^{(0)}$ be a zero vector $(0, \ldots, 0)^T$, and initially set $i = 0$.

3. Compute $\nabla\ell(\boldsymbol{\beta}_\lambda^{(i)}, t)$, $\nabla^2\ell(\boldsymbol{\beta}_\lambda^{(i)}, t)$ and $D_\lambda(\boldsymbol{\beta}_\lambda^{(i)})$ based on $\boldsymbol{\beta}_\lambda^{(i)}$.

4. Solve (5.3). Its solution is $\boldsymbol{\beta}_\lambda^{(i+1)}$. Let $\boldsymbol{\beta}_\lambda^{(i+1)}$ be the new $\boldsymbol{\beta}_\lambda^{(i)}$ in step 3.

5. Repeat steps 3 and 4 until (5.4) is achieved. Let the result $\boldsymbol{\beta}_\lambda^{(i+1)}$ be the estimator $\hat{\boldsymbol{\beta}}_\lambda$.

6. Use $\hat{\boldsymbol{\beta}}_\lambda$ in step 5 to compute (a) $\widehat{\text{AUC}}(u, \lambda)$ for a specific time u (4.9) or (b) $\widehat{\text{GCV}}(\lambda)$ (5.5).

7. Repeat step 1 through 6 until all possible λ values are used.

8. Obtain (a) $\hat{\boldsymbol{\beta}}_{\hat{\lambda}_{AUC}}$ (5.7) and (b) $\hat{\boldsymbol{\beta}}_{\hat{\lambda}_{GCV}}$ (5.6).

In the following we use simulated data to calculate the Lasso estimator $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ with both the AUC and the GCV criteria, and to compare these two criteria.

5.2 Comparison of the AUC and GCV Criteria

Three different scenarios are used to simulate censored survival data from the proportional hazards model

$$h(t|\mathbf{Z} = \mathbf{z}) = \exp(\boldsymbol{\beta}_0^T \mathbf{z}). \tag{5.8}$$

Note that this is the model (3.1) with $h_0(t) = 1$. Appendix C gives details of how simulated censored survival data are obtained. Appendix B.5 gives the R programming codes. The Lasso penalty

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{k} |\beta_j|$$

is used in the regularized log partial likelihood (5.1).

Using simulated data and the algorithm of Section 5.1, we calculate the regularized estimator $\hat{\boldsymbol{\beta}}_{\hat{\lambda}} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)^T$ based on either the AUC or the GCV criteria. In the AUC criterion, we set the specific time u of $\widehat{\text{AUC}}(u, \lambda)$ to be the median of simulated survival times, u_m.

Scenario I: The true parameter is $\boldsymbol{\beta}_0 = (2, 0, 3, 0, 0, 0, 0, 1)^T$ with dimension $k = 8$. The covariate vector $\mathbf{Z}$ in (5.8) follows a multivariate standard normal distribution with mean $\mu_{\mathbf{Z}}$ $k \times 1$ zero vector and covariance matrix $\Sigma_{\mathbf{Z}}$ $k \times k$ identity matrix.

In this scenario, we generate 500 data sets ($r = 500$) of 100 observations ($n = 100$), each exactly according to the above specifications. All samples use the same covariate vectors, $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$. The estimator $\hat{\boldsymbol{\beta}}_{\hat{\lambda}} = (\hat{\beta}_1, \ldots, \hat{\beta}_8)^T$ is calculated for each data set according to two different criteria, AUC and GCV, and two different censoring proportions, 10% censoring and 30% censoring.

Table 5.1 shows the sample mean of 500 regularized estimators $\hat{\beta}_j$'s,

$$\bar{\hat{\beta}}_j = \frac{1}{r} \sum_{m=1}^{r} \hat{\beta}_{jm}, \quad j = 1, \ldots, k, \tag{5.9}$$

rounded to one decimal place, where $r = 500$, $k = 8$ and $\hat{\beta}_{jm}$ is the j-th component of the regularized estimator from the m-th simulated data set. Note that the sample means, $\bar{\hat{\beta}}_j, j = 2, 4, 5, 6, 7$, turned out to be equal to the true parameter value (i.e. β_{02}, β_{04}, β_{05}, β_{06}, β_{07}) of zero. In the case of nonzero true parameters (i.e. β_{01}, β_{03}, β_{08}), the $\bar{\hat{\beta}}_j, j = 1, 3, 8$, are smaller than their corresponding true values (2,3,1), which indicate the shrinkage effect. Both the AUC and the GCV criteria select the nonzero parameters correctly. However, the sample means determined by AUC are closer to the true parameter values than those determined by GCV.

Table 5.1 also provides the sample standard deviation of the estimated regression coefficient,

$$s_j(\hat{\beta}_j) = \sqrt{\frac{1}{r-1} \sum_{m=1}^{r} (\hat{\beta}_{jm} - \bar{\hat{\beta}}_j)^2}, \quad j = 1, \ldots, k, \tag{5.10}$$

Table 5.1: The sample means ($\bar{\hat{\beta}}_j$, $j = 1, \ldots, 8$) of 500 $\hat{\beta}_j$'s selected by the AUC or the GCV criteria with 10% and 30% censored survival data under scenario I. The numbers in parentheses are sample standard deviations (s_j).

	10% Censoring		30% Censoring	
$\boldsymbol{\beta}_0$ (True value)	AUC	GCV	AUC	GCV
$\beta_{01} = 2$	1.8 (0.31)	1.7 (0.25)	1.8 (0.35)	1.7 (0.27)
$\beta_{02} = 0$	0.0 (0.09)	0.0 (0.08)	0.0 (0.12)	0.0 (0.10)
$\beta_{03} = 3$	2.7 (0.43)	2.6 (0.35)	2.7 (0.48)	2.6 (0.37)
$\beta_{04} = 0$	0.0 (0.12)	0.0 (0.09)	0.0 (0.12)	0.0 (0.08)
$\beta_{05} = 0$	0.0 (0.10)	0.0 (0.07)	0.0 (0.11)	0.0 (0.09)
$\beta_{06} = 0$	0.0 (0.11)	0.0 (0.09)	0.0 (0.12)	0.0 (0.09)
$\beta_{07} = 0$	0.0 (0.11)	0.0 (0.08)	0.0 (0.12)	0.0 (0.09)
$\beta_{08} = 1$	0.9 (0.21)	0.8 (0.17)	0.9 (0.24)	0.8 (0.20)

Table 5.2: The estimated coefficient of variation of $\hat{\beta}_j$'s with nonzero sample means in Table 5.1.

	10% Censoring		30% Censoring	
	AUC	GCV	AUC	GCV
$\hat{\beta}_1$	0.010	0.009	0.012	0.009
$\hat{\beta}_3$	0.011	0.010	0.013	0.010
$\hat{\beta}_8$	0.010	0.009	0.011	0.010

which are rounded to two decimal places. Note that all of the s_j's of $\hat{\beta}_j$'s selected by the AUC are larger than the s_j's of $\hat{\beta}_j$'s selected by the GCV.

As a measure of the relative variability of the estimator, we look at the estimated coefficient of variation,

$$\frac{\mathrm{SE}(\bar{\hat{\beta}}_j)}{\bar{\hat{\beta}}_j} = \frac{\sqrt{s_j^2/r}}{\bar{\hat{\beta}}_j}, \quad \text{for } \bar{\hat{\beta}}_j \neq 0. \tag{5.11}$$

The estimated coefficients of variation (5.11) for nonzero $\bar{\hat{\beta}}_j, j = 1, 3, 8$, are given in Table 5.2. It shows that in both 10% and 30% censored data, the $\hat{\beta}_j$'s selected by AUC have larger relative variability than those selected by GCV. However, their differences are small ($\leq 0.3\%$).

In the following, we consider two other scenarios:

Scenario II: Same design as scenario I except that the covariance matrix $\Sigma_{\mathbf{Z}}$ of the covariate vector $\mathbf{Z}$ in (5.8) is given by $Cov(Z_i, Z_j) = 0.5^{|i-j|}$ for all $i, j = 1, \ldots, k$, so that the covariates are dependent. The sample size is $n = 100$, and the number of data sets is $r = 500$.

Scenario III: Same design as scenario I except that the true parameter is $\boldsymbol{\beta}_0 = (\underbrace{0, \ldots, 0}_{5}, \underbrace{1, \ldots, 1}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{1, \ldots, 1}_{5})^T$ with dimension $k = 20$. The sample size and the number of data sets are the same as in scenario II ($n = 100$ and $r = 500$).

The sample means $\bar{\hat{\beta}}_j$'s and the sample standard deviations s_j's obtained from scenario II and III are given in Table 5.3 and Table 5.4, respectively. The estimated coefficients of variation (5.11) for nonzero $\bar{\hat{\beta}}_j$'s in Table 5.3 and Table 5.4 are shown in Table 5.5 and Table 5.6.

Regardless of the dependence of covariates and the increase of dimension of

the true paramenter, we obtained similar results to those obtained in scenario I:

(1) When the true coefficient β_{0j} is zero, the sample mean of the regularized estimator, $\bar{\hat{\beta}}_j$, is zero. When the true coefficient β_{0j} is not zero, the $\bar{\hat{\beta}}_j$ is not, either. That is, the regularized estimator can eliminate insignificant variables but contain the significant ones.

(2) When the true coefficient β_{0j} is not zero, $\bar{\hat{\beta}}_j$ is smaller than the true value, which gives a shrinkage of significant variable. Under two different criteria, the $\bar{\hat{\beta}}_j$ obtained by AUC is closer to the true coefficient value than that obtained by GCV.

(3) Although the sample standard deviation s_j's of $\hat{\beta}_j$'s selected by the AUC are larger than the s_j's of $\hat{\beta}_j$'s selected by the GCV, the difference of the relative variability of the regularized estimator selected by these two criteria is less or equal to 0.4% when the size of data set is 500.

Table 5.3: The sample means ($\bar{\hat{\beta}}_j$, $j = 1, \ldots, 8$) of 500 $\hat{\beta}_j$'s selected by the AUC or the GCV criteria with 10% and 30% censored survival data under scenario II. The numbers in parentheses are sample standard deviations (s_j).

	10% Censoring		30% Censoring	
$\boldsymbol{\beta}_0$ (True value)	AUC	GCV	AUC	GCV
$\beta_{01} = 2$	1.8 (0.34)	1.7 (0.24)	1.8 (0.38)	1.7 (0.25)
$\beta_{02} = 0$	0.0 (0.16)	0.0 (0.11)	0.0 (0.17)	0.0 (0.11)
$\beta_{03} = 3$	2.8 (0.48)	2.6 (0.33)	2.8 (0.54)	2.6 (0.35)
$\beta_{04} = 0$	0.0 (0.13)	0.0 (0.09)	0.0 (0.16)	0.0 (0.10)
$\beta_{05} = 0$	0.0 (0.14)	0.0 (0.09)	0.0 (0.14)	0.0 (0.08)
$\beta_{06} = 0$	0.0 (0.15)	0.0 (0.10)	0.0 (0.17)	0.0 (0.09)
$\beta_{07} = 0$	0.0 (0.15)	0.0 (0.10)	0.0 (0.16)	0.0 (0.10)
$\beta_{08} = 1$	0.9 (0.23)	0.8 (0.16)	0.9 (0.24)	0.8 (0.19)

Table 5.4: The sample means ($\bar{\hat{\beta}}_j$, $j = 1, \ldots, 20$) of 500 $\hat{\beta}_j$'s selected by the AUC or the GCV criteria with 10% and 30% censored survival data under scenario III. The numbers in parentheses are sample standard deviations (s_j).

	10% Censoring		30% Censoring	
$\boldsymbol{\beta}_0$ (True value)	AUC	GCV	AUC	GCV
$\beta_{0,1} = 0$	0.0 (0.13)	0.0 (0.09)	0.0 (0.13)	0.0 (0.09)
$\beta_{0,2} = 0$	0.0 (0.12)	0.0 (0.08)	0.0 (0.14)	0.0 (0.09)
$\beta_{0,3} = 0$	0.0 (0.13)	0.0 (0.08)	0.0 (0.13)	0.0 (0.09)
$\beta_{0,4} = 0$	0.0 (0.11)	0.0 (0.08)	0.0 (0.13)	0.0 (0.09)
$\beta_{0,5} = 0$	0.0 (0.13)	0.0 (0.10)	0.0 (0.14)	0.0 (0.09)
$\beta_{0,6} = 1$	0.9 (0.27)	0.8 (0.19)	0.9 (0.29)	0.8 (0.20)
$\beta_{0,7} = 1$	0.9 (0.25)	0.8 (0.18)	0.9 (0.29)	0.8 (0.20)
$\beta_{0,8} = 1$	0.9 (0.27)	0.8 (0.18)	0.9 (0.28)	0.8 (0.20)
$\beta_{0,9} = 1$	0.9 (0.27)	0.8 (0.20)	0.9 (0.30)	0.8 (0.23)
$\beta_{0,10} = 1$	0.9 (0.27)	0.8 (0.20)	0.9 (0.31)	0.8 (0.22)
$\beta_{0,11} = 0$	0.0 (0.12)	0.0 (0.09)	0.0 (0.14)	0.0 (0.10)
$\beta_{0,12} = 0$	0.0 (0.12)	0.0 (0.08)	0.0 (0.14)	0.0 (0.09)
$\beta_{0,13} = 0$	0.0 (0.12)	0.0 (0.08)	0.0 (0.14)	0.0 (0.10)
$\beta_{0,14} = 0$	0.0 (0.13)	0.0 (0.09)	0.0 (0.13)	0.0 (0.10)
$\beta_{0,15} = 0$	0.0 (0.13)	0.0 (0.09)	0.0 (0.13)	0.0 (0.09)
$\beta_{0,16} = 1$	0.9 (0.26)	0.8 (0.19)	0.9 (0.29)	0.8 (0.22)
$\beta_{0,17} = 1$	0.9 (0.28)	0.8 (0.19)	0.9 (0.29)	0.8 (0.21)
$\beta_{0,18} = 1$	0.9 (0.25)	0.8 (0.19)	0.9 (0.31)	0.8 (0.23)
$\beta_{0,19} = 1$	0.9 (0.27)	0.8 (0.20)	0.9 (0.29)	0.8 (0.21)
$\beta_{0,20} = 1$	0.9 (0.26)	0.8 (0.19)	0.9 (0.30)	0.8 (0.22)

Table 5.5: The estimated coefficient of variation of $\hat{\beta}_j$'s with nonzero sample means in Table 5.3.

	10% Censoring		30% Censoring	
	AUC	GCV	AUC	GCV
$\hat{\beta}_1$	0.011	0.008	0.013	0.009
$\hat{\beta}_3$	0.013	0.009	0.014	0.010
$\hat{\beta}_8$	0.011	0.008	0.011	0.010

Table 5.6: The estimated coefficient of variation of $\hat{\beta}_j$'s with nonzero sample means in Table 5.4.

	10% Censoring		30% Censoring	
	AUC	GCV	AUC	GCV
$\hat{\beta}_6$	0.013	0.010	0.014	0.010
$\hat{\beta}_7$	0.012	0.009	0.014	0.010
$\hat{\beta}_8$	0.013	0.009	0.013	0.010
$\hat{\beta}_9$	0.013	0.010	0.014	0.012
$\hat{\beta}_{10}$	0.013	0.010	0.015	0.011
$\hat{\beta}_{16}$	0.012	0.010	0.014	0.011
$\hat{\beta}_{17}$	0.013	0.010	0.014	0.011
$\hat{\beta}_{18}$	0.012	0.010	0.015	0.012
$\hat{\beta}_{19}$	0.013	0.010	0.014	0.011
$\hat{\beta}_{20}$	0.012	0.010	0.014	0.011

Chapter 6

Application

6.1 Data: Survival Times of Squamous Cell Carcinoma

We acquired the clinical data from trial 9501 of the Radiation Therapy Oncology Group (RTOG) tumor bank. Between September, 1995, and April, 2000, 459 patients who had resectable squamous cell carcinoma of the head and neck region were enrolled in a randomized trial. Of these, 142 patients had tissue biopsy available for immunohistochemistry (IHC) by pathologists at the University of Maryland Greenebaum Cancer Center. Discarding missing or incomplete observations, there are 122 patients left for the study.

Basic information on patients was collected before they received the treatment. It indicates age, gender, primary tumor site, Karnofsky performance status (KPS), TN staging of tumor and smoking history. Patients younger than eighteen were excluded from entering the trial. The age of the 122 patients ranges from 31 to 79 with sample mean 55.48 and sample standard deviation 9.79. There were 106 males and 16 females. Six different primary tumor sites were identified: oral cavity, oropharynx, hypopharynx, supraglottic larynx, glottic larynx and subglottic larynx. KPS is a medical index for classifying patients' functional impairment. It is a measure to determine whether a patient can receive chemotherapy and dose adjustment. The KPS ranges from 0 to 100 with 0 indicating the death status

and 100 indicating the normal status. Patients whose KPS is larger than 60 were eligible to participate in the clinical trial. The T and N stages are descriptors of how much the cancer has spread, where T takes into account the size of a primary tumor and N represents regional lymph node involvement, following the American Joint Commission (AJC) staging system. Detailed classifications of KPS, T stage and N stage are shown in Appendix D and E. The smoking history of a patient was dichotomized: whether the patient has ever used cigarettes and whether the patient is currently using cigarettes within 6 months.

The 122 patients who entered into the trial were assigned at random to one of two treatments within eight weeks after their surgery had been performed. Treatment one is radiation treatment (RT) alone, once a day, five days a week for six weeks. Treatment two is chemotherapy (CT) of the drug cisplatin given to a patient every three weeks on days 1, 22 and 43 concurrent with radiotherapy following the same protocol as in treatment one. The follow-up assessments were reported starting at the third week after the end of six-week treatment, then every three months during the first year following treatment, then every six months for the next two years, and annually after the third year. Among the 122 cases, progression-free survival times, the length of time during and after treatment in which a patient does not get worse, were observed from 0.01 to 9.19 years with mean 2.71 years and median 1.40 years, and 41 (34%) cases were censored which occurred throughout the follow-up.

After surgery, patients' tumor biopsy samples were stored in the RTOG tumor bank for further examination. For each patient in our study, four genetic markers

were given scores ranging from 0 to 3 through IHC staining, a technique for visually identifying antigens or proteins in tissue sections by means of antigen-antibody interactions. These genetic markers were B-cell lymphoma 2 (Bcl2), glutathione S-transferase π (GSTπ), protein 53 (p53) and thymidylate synthase (TS). Appendix F lists the detailed information of 122 patients, and Table 6.1 gives a summary.

Table 6.1: A summary of patients' information of head and neck cancer data. (Total number of patients is 122.)

	Variable	Number of patients (%)
1	Age	
	31-40	9 (07%)
	41-50	28 (23%)
	51-60	45 (37%)
	61-70	34 (28%)
	71-79	6 (05%)
2	Gender	
	Female	16 (13%)
	Male	106 (87%)
3	Treatment	
	Only RT	56 (46%)
	RT + CT	66 (54%)
4	Primary tumor site	
	Oral cavity	25 (20%)
	Oropharynx	55 (45%)
	Hypopharynx	16 (13%)
	Supraglottic larynx	20 (16%)
	Glottic larynx	4 (03%)
	Subglottic larynx	2 (02%)

Continued...

	Variable	Number of patients (%)
5	KPS	
	60	1 (01%)
	70	21 (17%)
	80	31 (25%)
	90	52 (43%)
	100	17 (14%)
6	T stage	
	T1	15 (12%)
	T2	30 (25%)
	T3	37 (30%)
	T4	40 (33%)
7	N stage	
	N0	2 (02%)
	N1	1 (01%)
	N2a	8 (07%)
	N2b	88 (72%)
	N2c	22 (18%)
	N3	1 (01%)
	Whether ever smoked	
	No	5 (04%)
	Yes	117 (96%)
8	Whether smoking in the recent 6 months	
	No	45 (37%)
	Yes	77 (63%)
	Four genetic markers	Mean (Standard deviation)
9	Bcl2	1.10 (0.87)
10	GSTπ	1.94 (0.88)
11	p53	1.50 (1.18)
12	TS	1.46 (0.62)

Based on the method developed, we select significant variables that influence patients' progression-free survival time. Before doing the analysis, it is necessary to make some adjustment for sparse data. Note that, in the categories of primary tumor site, only 3% of sites are in the glottic larynx and 2% in the subglottic larynx. In anatomy, since supraglottic, glottic and subglottic larynx are subdivisions of the larynx, we decided to combine the data from these three subdivisions and call the combined category "larynx". The combination increased the number of patients to twenty six which would help to reduce the sample error in analysis. Also, in the smoking category, all but 5 patients have smoked in the past. Because fewer than 5% never smoked and in an effort to maintain reasonable sample sizes in each category, we decided not to consider the variable "whether patients ever smoked". Otherwise, the sample sizes would be too small for analysis which may result in unstable estimation.

6.2 The Lasso Analysis

In this section, we use the regularized log partial likelihood of the proportional hazards model,

$$Q(\boldsymbol{\beta}, t) = \ell(\boldsymbol{\beta}, t) - P_\lambda(\boldsymbol{\beta}), \tag{6.1}$$

to analyze the head and neck cancer data presented in Section 6.1. The goal is to study the effect of explanatory variables (regressors) on patients' progression-free survival time and to find significant explanatory variables that affect the survival

time. We use the Lasso penalty function,

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{k} |\beta_j|.$$

The regularized parameter λ is selected by both the AUC and the GCV criteria. Each criterion yields a set of estimates, $\hat{\beta}_j$'s. In using the AUC criterion, we consider three different times, 1.4 years (the sample median of patients' progression-free survival times), 2.71 years (the sample mean of patients' progression-free survival times) and 5 years (a popular choice of progression-free survival times for measuring the success of a cancer treatment). These chosen times (u) are used to separate patients into two groups. The group of patients who survival beyond time u corresponds to the "healthy" group with $D = 0$, while the group that died before time u corresponds to $D = 1$ (a "diseased" patient). This is discussed in Section 4.2.1. For easy reference, we call the three AUC criteria "criterion 1", "criterion 2" and "criterion 3" for $u = 1.4$, 2.71 and 5 years, respectively. The GCV criterion will be called "criterion 4".

In the data set, the primary site can be one of four categories, oral cavity, oropharynx, hypopharynx and larynx. We use a dummy variable P1 with values 1 and 0 to indicate if a patient's primary tumor had been found in oral cavity or not. Similarly, P2 and P3 are dummy variables indicating the presence or absence of the primary tumors in oropharynx and hypopharynx, respectively. If P1, P2 and P3 are all zero, then the primary tumor is in larynx. Therefore, the total number of explanatory variables in the proportional hazards model is fourteen (i.e. $k = 14$). The Lasso estimators of the coefficients β's with the corresponding standard errors

are given in Table 6.2.

Table 6.2: The results of Lasso estimation for the head and neck cancer data.

Estimated coefficient	AUC(u) criterion			GCV criterion
	u=1.4 yrs (Criterion 1)	u=2.71 yrs (Criterion 2)	u=5 yrs (Criterion 3)	(Criterion 4)
$\hat{\beta}_1$ (Age)	0.00	0.00	0.00	0.00
$\hat{\beta}_2$ (Gender)	-0.15	-0.02	-0.04	-0.14
$\hat{\beta}_3$ (Treatment)	-0.01	-0.14	-0.01	-0.15
$\hat{\beta}_4$ (P1)	0.01	0.00	0.00	0.00
$\hat{\beta}_5$ (P2)	0.00	-0.02	-0.01	-0.17
$\hat{\beta}_6$ (P3)	0.00	0.00	0.07	0.00
$\hat{\beta}_7$ (KPS)	-0.03	-0.03	-0.03	-0.03
$\hat{\beta}_8$ (T stage)	0.07	0.03	0.06	0.03
$\hat{\beta}_9$ (N stage)	0.04	0.01	0.03	0.01
$\hat{\beta}_{10}$ (Smoking)	0.24	0.02	0.18	0.01
$\hat{\beta}_{11}$ (Bcl2)	-0.29	-0.24	-0.27	-0.22
$\hat{\beta}_{12}$ (GSTπ)	0.25	0.17	0.22	0.14
$\hat{\beta}_{13}$ (p53)	-0.02	-0.01	-0.01	-0.01
$\hat{\beta}_{14}$ (TS)	-0.34	-0.18	-0.30	-0.16

Any estimate $\hat{\beta}_j$ with $|\hat{\beta}_j| \leq 0.05$ is regarded insignificant. That is, $\hat{\beta}_j$'s influence in predicting patients' survival time in the model will be ignored. According to this definition, five explanatory variables, age, P1, KPS, N stage and p53 are regarded as insignificant ones while the three genetic markers, Bcl2, GSTπ and TS, are classified as significant explanatory variables for all four criteria. There is no

uniformity by these criteria in the gender, T stage, P2 and smoking. Gender, T stage, smoking are significant in criterion 1, but become insignificant in criterion 2. P3, T stage, Smoking, are significant in criterion 3 while gender, treatment, P2 are significant in criterion 4. The choice of $|\hat{\beta}_j| \leq 0.05$ is arbitrary. The discussion above would be somewhat different if a different criterion were used.

Comparing the AUC and the GCV criteria, we see a reasonable consistency in the values of the estimates and the signs of the estimates. The signs are identical in all criteria. The performance of the both criteria is about the same. One advantage of the AUC criterion is its time-dependence. One can trace the changes in these explanatory variables over time. A remarkable result is that these explanatory variables do not change significantly over time. It is also interesting to note that age has no effect in predicting a patient's survival time.

Because of the shrinkage effect as discussed in Section 5.2, it is possible that the regularized estimators, $\hat{\beta}_j$, deemed significant explanatory variables may underestimate their true values β_j. To study the effect of shrinkage, we carried out the estimation of these significant β_j's by maximizing the log partial likelihood (6.1) without a penalty function (i.e. $P_\lambda(\boldsymbol{\beta}) = 0$). The results are given in Table 6.3 which shows that only the estimate of P2 in criterion 4 has its value less than that in Table 6.2. All other estimates are larger than the corresponding estimates in Table 6.2. From Table 6.3, we can calculate the increasing (or decreasing) hazard for each one unit increases in the variable and predict the probability of patients' progression-free survival times.

Table 6.3: Estimates of significant predictors, $\hat{\beta}_j$'s, obtained by maximizing the log partial likelihood without penalty. Standard errors are given in parentheses. Significant β_j's are determined in Table 6.2. Dash – indicates the insignificance of β_j.

Estimated coefficient	AUC(u) Selection			GCV Selection
	u=1.4 yrs (Criterion 1)	u=2.71 yrs (Criterion 2)	u=5 yrs (Criterion 3)	(Criterion 4)
$\hat{\beta}_1$ (Age)	–	–	–	–
$\hat{\beta}_2$ (Gender)	-0.24 (0.34)	–	–	-0.30 (0.34)
$\hat{\beta}_3$ (Treatment)	–	-0.14 (0.23)	–	-0.17 (0.23)
$\hat{\beta}_4$ (P1)	–	–	–	–
$\hat{\beta}_5$ (P2)	–	–	–	-0.13 (0.23)
$\hat{\beta}_6$ (P3)	–	–	0.07 (0.33)	–
$\hat{\beta}_7$ (KPS)	–	–	–	–
$\hat{\beta}_8$ (T stage)	0.13 (0.12)	–	0.13 (0.12)	–
$\hat{\beta}_9$ (N stage)	–	–	–	–
$\hat{\beta}_{10}$ (Smoking)	0.55 (0.25)	–	0.54 (0.45)	–
$\hat{\beta}_{11}$ (Bcl2)	-0.35 (0.14)	-0.38 (0.15)	-0.35 (0.15)	-0.36 (0.15)
$\hat{\beta}_{12}$ (GSTπ)	0.27 (0.14)	0.25 (0.14)	0.25 (0.14)	0.27 (0.14)
$\hat{\beta}_{13}$ (p53)	–	–	–	–
$\hat{\beta}_{14}$ (TS)	-0.43 (0.19)	-0.29 (0.19)	-0.41 (0.20)	-0.29 (0.19)

Chapter 7

Summary and Conclusion

The main goal of this thesis has been the development of a sound statistical procedure for selecting significant variables that can accurately predict a patient's survival time. We developed such a statistical procedure and a computational algorithm for the proportional hazards model and right-censored survival times. Our procedure is built upon the regularized variable selection method, the Lasso, which first introduced by Tibshirani (1996) for the linear regression model and later for the proportional hazards model (Tibshirani, 1997). This regularization procedure is computationally intensive. The success of this method depends on not only having optimal statistical properties but also a good choice of the regularization parameter λ or the tuning variable.

Chapters 2 and 3 contain mainly the literature review relevant to this thesis. In the review, we proved a few lemmas that are not available to us in the literature. The generalized cross-validation (GCV) criterion have been used in the literature for selecting λ. In this thesis, a new method of determining λ, called the area under the ROC curve (AUC) criterion, is proposed for the proportional hazards model. This is given in Chapter 4. The application of Chapter 4 is provided in Chapter 5. The superiority of the AUC criterion over the GCV lies in its interpretation of the survival data. The GCV criterion is at least computationally more suitable for

linear regression model than for the proportional hazards model. This is due to the fact that the linear regression model models the observation $\mathbf{Y}$ directly in a linear form $X\boldsymbol{\beta}$ plus error while the proportional hazards models the hazards or survival probability resulting in a non-linear structure. The comparison of the AUC and the GCV criteria were carried out in Chapter 5. Using the Lasso penalty as an example, our simulation results show that the performance of variable selection and shrinkage of significant variables by AUC criterion is similar to that of the traditional GCV criterion, but the magnitude of shrinkage is different. The AUC criterion is a function of the survival time. Using the AUC criterion, variable selection can be made time-dependent as it has been observed that some of the significant covariates change over time. An iterative algorithm based on the Newton-Raphson method was developed for computation in R.

In this thesis, we established the consistency and the asymptotic normality of regularized estimator of the regression coefficient in the proportional hazards model for a fixed λ. In variable selection, another kind of consistency needs to be addressed, namely the variable selection consistency. It requires the procedure to select the right subset of regression coefficients if in truth only this subset has all the nonzero coefficients. This is a part of the oracle properties of a statistical procedure. We have reviewed some of the literature on the oracle properties with the tuning parameter depending on the sample size n. Fan and Li (2002) proved that that "the oracle properties hold" with probability tending to one as n tends to infinity. This is much weaker than requiring that the estimates of insignificant parameters are zero. We will study this problem in the future. It is worth noting that in our

simulations, both our AUC criterion and the traditional GCV criterion identified the true zero regression coefficients which exhibit the oracle property. When the true coefficients are not zero, then the AUC criterion produced estimates closer to the true parameters than that of the GCV.

As an illustration, we applied the method developed in this thesis to a set of survival data of patients who had squamous cell cancer of the head and neck. The results show the three genetic markers, Bcl2, GSTπ and TS are significant variables while the other variables, age, primary tumor site, Karnofsky performance status, N stage and genetic marker p53 are insignificant.

Appendix A

Parametric Estimation of TP, FP, ROC and AUC

Let $(\tilde{T}_i, \delta_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, be an i.i.d. sample of n survival data, where $\tilde{T}_i$ is observed survival time of the i-th individual whose censoring indicator and the covariate vector are denoted by δ_i and $\mathbf{Z}_i$ as defined in Section 3.1. Let $g(x)$ be the probability density function (pdf) of $\boldsymbol{\beta}^T\mathbf{Z}$. With some simple algebra, the true positive probability (4.6) and the false positive probability (4.7) can be written, respectively, as

$$\begin{aligned} \mathrm{TP}(w,u) &= \frac{\int_w^\infty P(T \le u|x)g(x)\,dx}{\int_{-\infty}^\infty P(T \le u|x)g(x)\,dx} \\ \mathrm{FP}(w,u) &= \frac{\int_w^\infty P(T > u|x)g(x)\,dx}{\int_{-\infty}^\infty P(T > u|x)g(x)\,dx}. \end{aligned}$$

In the proportional hazards model (3.1), the conditional survival function of T given $\mathbf{Z} = \mathbf{z}$ is

$$P(T > u|\mathbf{z}) = \exp\left[-H_0(u)\exp(\boldsymbol{\beta}^T\mathbf{z})\right],$$

where $H_0(u) = \int_0^t h_0(t)dt$ is the cumulative hazard of $h_0(u)$ as defined in Section 3.1. Following Breslow (1972, 1974), we use $\hat{H}_0(u)$ as the estimate of $H_0(u)$:

$$\hat{H}_0(u) = \sum_{\tilde{T}_i \le u} \frac{\delta_i}{\sum_{j\in R_i}\exp(\hat{\boldsymbol{\beta}}_\lambda^T\mathbf{Z}_j)},$$

where R_i is the risk set at time $\tilde{T}_i$ defined by $R_i = \{j : \tilde{T}_j \ge \tilde{T}_i\}$, and $\hat{\boldsymbol{\beta}}_\lambda$ is the regularized estimator of $\boldsymbol{\beta}$ obtained by maximizing the regularized log partial

likelihood $\{\ell(\boldsymbol{\beta}, u) - P_\lambda(\boldsymbol{\beta})\}$ with a fixed tuning parameter λ. Therefore, the survival function $P(T > u|\boldsymbol{\beta}^T\mathbf{Z} = x)$ can be estimated by

$$\hat{S}(u|\boldsymbol{\beta}^T\mathbf{Z} = x) = \exp\left[-\hat{H}_0(u)\exp(x)\right].$$

Besides, we consider the empirical function $\sum_{i=1}^n I[\boldsymbol{\beta}^T\mathbf{Z} \le x]/n$ as the estimator of $P(\boldsymbol{\beta}^T\mathbf{Z} \le x)$. Since $\boldsymbol{\beta}$ is unknown, we replace $\boldsymbol{\beta}$ by its estimator $\hat{\boldsymbol{\beta}}_\lambda$. Let

$$\hat{P}(\boldsymbol{\beta}^T\mathbf{Z} \le x) = \frac{1}{n}\sum_{i=1}^n I[\hat{\boldsymbol{\beta}}_\lambda^T\mathbf{Z} \le x].$$

Then the true positive and false positive probabilities can be estimated, respectively, by

$$\begin{aligned}
\widehat{\mathrm{TP}}(w, u, \lambda) &= \frac{\int_w^\infty \left[1 - \hat{S}(u|\boldsymbol{\beta}^T\mathbf{Z} = x)\right] d\hat{P}(\boldsymbol{\beta}^T\mathbf{Z} \le x)}{\int_{-\infty}^\infty \left[1 - \hat{S}(u|\boldsymbol{\beta}^T\mathbf{Z} = x)\right] d\hat{P}(\boldsymbol{\beta}^T\mathbf{Z} \le x)} \\
&= \frac{\sum_{i=1}^n \left\{1 - \exp\left[-\hat{H}_0(u)\exp(\hat{\boldsymbol{\beta}}_\lambda^T\mathbf{Z}_i)\right]\right\} I[\hat{\boldsymbol{\beta}}_\lambda^T\mathbf{Z}_i > w]}{\sum_{i=1}^n \left\{1 - \exp\left[-\hat{H}_0(u)\exp(\hat{\boldsymbol{\beta}}_\lambda^T\mathbf{Z}_i)\right]\right\}}, \qquad \text{(A.1)}
\end{aligned}$$

and

$$\begin{aligned}
\widehat{\mathrm{FP}}(w, u, \lambda) &= \frac{\int_w^\infty \hat{S}(u|\boldsymbol{\beta}^T\mathbf{Z} = x) d\hat{P}(\boldsymbol{\beta}^T\mathbf{Z} \le x)}{\int_{-\infty}^\infty \hat{S}(u|\boldsymbol{\beta}^T\mathbf{Z} = x) d\hat{P}(\boldsymbol{\beta}^T\mathbf{Z} \le x)} \\
&= \frac{\sum_{i=1}^n \exp\left[-\hat{H}_0(u)\exp(\hat{\boldsymbol{\beta}}_\lambda^T\mathbf{Z}_i)\right] I[\hat{\boldsymbol{\beta}}_\lambda^T\mathbf{Z}_i > w]}{\sum_{i=1}^n \exp\left[-\hat{H}_0(u)\exp(\hat{\boldsymbol{\beta}}_\lambda^T\mathbf{Z}_i)\right]}. \qquad \text{(A.2)}
\end{aligned}$$

The ROC function (4.8) can be estimated by

$$\widehat{\mathrm{ROC}}(x, u, \lambda) = \widehat{\mathrm{TP}}(\widehat{\mathrm{FP}}^{-1}(x, u, \lambda), u, \lambda), \qquad \text{(A.3)}$$

where $\widehat{\mathrm{FP}}^{-1}(x, u, \lambda) = \inf\{w : \widehat{\mathrm{FP}}(w, u, \lambda) \ge x\}$. Then the estimator of the AUC function (4.9) is given by

$$\widehat{\mathrm{AUC}}(u, \lambda) = \int_0^1 \widehat{\mathrm{ROC}}(x, u, \lambda) dx. \qquad \text{(A.4)}$$

Appendix B

R Programming Codes

B.1 Computation of $\ell(\boldsymbol{\beta}, t)$, $\nabla\ell(\boldsymbol{\beta}, t)$ and $\nabla^2\ell(\boldsymbol{\beta}, t)$

```
# Input: y is an n-dimensional vector of observed survival times; Z is an n × k
# matrix of covariate variables; delta is an n vector of censoring status (1 for death;
# 0 for censored); beta is a k-dimensional vector of regression coefficients β.
# Output: l.like is the log partial likelihood ℓ(β, t); l.grad is the gradient vector
# ∇ℓ(β, t); l.hess is the Hessian matrix ∇²ℓ(β, t).

l.like <- function(y, Z, delta, beta){
    l <- sum(delta * Z %*% beta)
    for (i in 1:length(y)){
        l <- l-delta[i] * log(sum(exp(Z[y>=y[i],] %*% beta)))
    }
    l
}
```

```
l.grad <- function(y, Z, delta, beta){
    l <- as.vector(delta %*% Z)
    for (i in 1:length(y)){
       X <- Z[y>=y[i],]
       l <- l-delta[i] * t(exp(X %*% beta)) %*% X / sum(exp(X %*% beta))
    }
    l
}

l.hess <- function(y, Z, delta, beta){
    D <-diag(as.vector(exp(Z %*% beta)))
    temp <- matrix(rep(y,length(y)),ncol=length(y))
    IR <- (t(temp)>=temp)
    w <- IR %*% exp(Z %*% beta)
    l <- 0
    for(i in 1:length(y)){
       A <- w[i]*D-exp(Z %*% beta) %*% t(exp(Z %*% beta))
       X <- diag(IR[i,]) %*% Z
       l <- l+(delta[i]/w[i]^2) * t(X) %*% A %*% X
    }
    -l
}
```

B.2 Estimation of $\hat{\boldsymbol{\beta}}$ with Lasso Penalty

```
# Input: Definitions of y, Z and delta are the same as above; lambda is a
# tuning parameter; tolerance is a tolerant value for convergence in iterations
# (the default is 10^-7).
# Output: beta is the vector of estimated coefficients β̂.


Beta.lasso <- function(y,Z,delta,lambda,tolerance=10^-7){
    n <- dim(Z)[1]
    k <- dim(Z)[2]
    beta <- rep(0,k)
    tol <- 1
    while(tol>=tolerance){
       old.beta <- beta
       G <- l.grad(y,Z,delta,beta)
       H <- l.hess(y,Z,delta,beta)
       D <- lambda*diag(replace(beta, (1:k)[beta!=0], 1/abs(beta[beta!=0])))
       beta <- as.vector(-solve(H-2*D, t(G)-H%*%beta))
       tol <- sqrt(sum(beta-old.beta)^2)
    }
beta
}
```

B.3 Computation of $\widehat{\mathrm{AUC}}(u, \lambda)$

```
# Input: Definitions of y, Z, delta and beta are the same as B.2. u is a
# fixed time to separate data into two groups.
# Output: AUC(u, λ).

AUC.fun <- function(y, Z, delta, beta, u){
    temp <- order(y)
    M <- as.numeric(Z %*% beta)
    M.orderbyy <- M[temp]
    delta.orderbyy <- delta[temp]
    ind <- (delta.orderbyy==1 | sort(y)>u)
    newy <- sort(y)[ind]
    risky <- as.numeric(newy>u)
    n1 <- sum(risky==0)
    n <- length(newy)
    newM <- M.orderbyy[ind]
    mean.rank <- mean(rank(newM)[risky==0])
    (mean.rank - (n1+1)/2)/(n-n1)
}
```

B.4 Computation of $\widehat{\mathrm{GCV}}(\lambda)$

```
# Input: Definitions of y, Z, delta, beta and lambda are the same as B.2.

# Output: GCV value for Lasso penalty based on Fan and Li (2002).


GCV.fun <- function(y, Z, delta, beta, lambda){

    n <- length(y)

    a <- replace(rep(0,n), [round(beta,4)!=0], 1/abs(round(beta,4)))

    D <- lambda*diag(a)

    H <- l.hess(y, Z, delta, beta)

    e <- sum(diag(H %*% solve(H+D)))

    -l.like(y, Z, delta, beta)/(n*(1-e/n)^2)

}
```

B.5 Simulation of $(\tilde{T}_i, \delta_i, \mathbf{Z}_i)$

```
#Input: n is the sample size; beta is the vector of regression coefficients

#in the true model; mu and sigma are the mean vector and the covariance

#matrix of Z; r is a vector of given censoring rates.

#Output: a data frame of survival data with observed survival time T_i,

#censoring status delta_i (1 for death; 0 for censored) and covariate vector Z_i.

#Use the package "MASS".
```

```
library(MASS)

Z <- mvrnorm(n, mu, sigma)

Simu.SurvData.Z <- function(n,beta,Z,r){

    T <- as.vector(-log(runif(n))/(exp(Z %*% beta)))

    v <- r*exp(Z %*% beta)/(1-r)

    C <- rexp(n,v)

    status <- as.numeric(T<=C)

    y <- status*T+(1-status)*C

    if (n==1) Z <- t(Z)

    data.frame(time=y, status, Z)

}
```

Appendix C

Simulation of Censored Survival Data

In this thesis, we suppose that the hazard function of the survival time T is given by

$$h(t|\mathbf{Z}=\mathbf{z}) = h_0(t)\exp(\boldsymbol{\beta}_0^T\mathbf{z}),$$

where $h_0(t)=1$, $\boldsymbol{\beta}_0$ is a given k-dimensional parameter, and $\mathbf{Z}$ is a k-dimensional covariate vector randomly selected from a multivariate normal distribution with a given $k\times 1$ mean vector $\mu_{\mathbf{Z}}$ and a given $k\times k$ covariance matrix $\Sigma_{\mathbf{Z}}$. Suppose that the censoring time C follows an exponential distribution $Exp(1/\nu)$ with $\nu>0$, the observed time is $\tilde{T}=\min(T,C)$, and the censoring indicator is $\delta=I[T\le C]$. Let the censoring rate $r=P(T>C|\mathbf{Z}=\mathbf{z})$ is given. We generate a sample of size n i.i.d. censored survival data $(\tilde{T}_i,\delta_i,\mathbf{Z}_i)$, $i=1,\ldots,n$ as follows.

Step 1: Generate covariate vectors $\mathbf{Z}_i, i=1,\ldots,n$, from a multivariate normal distribution with mean $\mu_{\mathbf{Z}}$ and covariance $\Sigma_{\mathbf{Z}}$.

Step 2: Generate $U_i, i=1,\ldots,n$, from a uniform distribution $U[0,1]$, and set the survival time T_i as

$$H_0^{-1}[-\log(U_i)/\exp(\boldsymbol{\beta}_0^T\mathbf{Z}_i)] = -\log(U_i)/\exp(\boldsymbol{\beta}_0^T\mathbf{Z}_i),$$

since the survival function

$$S(t|\mathbf{z}) = \exp[-\int_0^t h(u|\mathbf{z})du] = \exp[-H_0(t)\exp(\boldsymbol{\beta}_0^T\mathbf{z})] \qquad \text{(C.1)}$$

follows a standard uniform distribution and $H_0(t) = \int_0^t h_0(u)du = t$.

Step 3: Generate censoring times C_i, $i = 1, \ldots, n$, from an exponential distribution with the scale parameter $\nu = r\exp(\boldsymbol{\beta}_0^T\mathbf{Z}_i)/(1-r)$ since we have

$$\begin{aligned} r &= P(T > C|\mathbf{Z} = \mathbf{z}) \\ &= \int_0^\infty P(T > t|\mathbf{Z} = \mathbf{z})\nu\exp(-\nu t)\,dt \\ &= \nu\int_0^\infty \exp[-t\exp(\boldsymbol{\beta}_0^T\mathbf{z})]\exp(-\nu t)\,dt, \text{ by } (C.1) \\ &= \frac{\nu}{\nu + \exp(\boldsymbol{\beta}_0^T\mathbf{z})}. \end{aligned}$$

Step 4: Set censoring indicators $\delta_i = 1$ if $T_i \leq C_i$, and $\delta_i = 0$ otherwise, for all $i = 1, \ldots, n$.

Step 5: Set observable survival times $\tilde{T}_i = \delta_i T_i + (1-\delta_i)C_i, i = 1, \ldots, n$.

Appendix B.5 shows the R programming codes of simulating censored survival data $(\tilde{T}_i, \delta_i, \mathbf{Z}_i)$.

Appendix D

Karnofsky Performance Status

100	Normal; no complaints; no evidence of disease.
90	Able to carry on normal activity; minor signs or symptoms of disease.
80	Normal activity with effort; some signs or symptoms of disease.
70	Cares for self; unable to carry on normal activity or to do work.
60	Requires occasional assistance from others but able to care for most needs.
50	Requires considerable assistance and frequent medical care.
40	Disabled; requires special care and assistance.
30	Severely disabled; hospitalization is indicated, although death not imminent.
20	Very sick; hospitalization necessary; active support treatment is necessary.
10	Moribund; fatal processes progressing rapidly.
0	Dead.

Appendix E

American Joint Commission Staging, 4th Edition

Primary Tumor (T) Stage

Oral Cavity

T1 Tumor 2 cm or less in greatest dimension.

T2 Tumor more than 2 but not more than 4 cm in greatest dimension.

T3 Tumor more than 4 cm in greatest dimension.

T4 Tumor invades adjacent structures.

Oropharynx

T1 Tumor 2 cm or less in greatest dimension.

T2 Tumor more than 2 but not more than 4 cm in greatest dimension.

T3 Tumor more than 4 cm in greatest dimension.

T4 Tumor invades adjacent structures.

Hypopharynx

T1 Tumor limited to one subsite of hypopharynx.

T2 Tumor invades more than one subsite of hypopharynx or an adjacent site, without fixation of hemilarynx.

T3 Tumor invades more than one subsite of hypopharynx or an adjacent site, with fixation of hemilarynx.

T4 Tumor invades adjacent structures.

Supraglottis Larynx

T1 Tumor limited to one subsite of supraglottis with normal mobility.

T2 Tumor invades more than one subsite of supraglottis or glottis with vocal cord morbidity.

T3 Tumor limited to larynx with vocal cord fixation and/or extension to involve postcricoid area, medical wall of pyriform sinus or pre-epiglottic tissues.

T4 Massive tumor extending beyond the larynx to involve oropharynx, soft tissue of neck or destuction of thyroid cartilage.

Glottis Larynx

T1 Tumor limited to the vocal cord(s) with normal mobility.

T2 Tumor extends to supraglottis and/or subglottis and/or with impaired vocal cord morbidity.

T3 Tumor limited to larynx with vocal cord fixation.

T4 Tumor invades through thyroid cartilage and/or extends to other tissues beyond the larynx.

Subglottis Larynx

T1 Tumor limited to the subglottis.

T2 Tumor extends to vocal cord(s) with normal or impaired mobility.

T3 Tumor limited to larynx with vocal cord fixation.

T4 Tumor invades through cricoid or thyroid cartilage and/or extends to other tissues beyond the larynx.

Nodal Involvement (N) Stage

N0	No regional lymph node metastasis.
N1	Metastasis in a single ipsilateral lymph node, 3 cm or less in greatest dimension.
N2a	Metastasis in a single ipsilateral lymph node more than 3 cm but not more than 6 cm in greatest dimension.
N2b	Metastasis in multiple ipsilateral lymph nodes, none more than 6 cm in greatest dimension.
N2c	Metastasis in bilateral or contralateral lymph nodes, none more than 6 cm in greatest dimension.
N3	Metastasis in a lymph node more than 6 cm in greatest dimension.

Appendix F

Head and Neck Cancer Data

122 observations are listed in Table F.1. The variables in the data set are:

N	Case number.
A	Age in years.
G	Gender. 0=female; 1=male.
Tr	Treatment. 0=radiotherapy; 1=radiotherapy plus concurrent chemotherapy.
P	Primary tumor site. 1=oral cavity; 2=oropharynx; 3=hypopharynx; 4=supraglottic larynx; 5=glottic larynx; 6=subglottic larynx.
K	Karnofsky performance status.
T	T stage. 1=T1; 2=T2; 3=T3; 4=T4.
N	N stage. 0=N0; 1=N1; 2=N2a; 3=N2b; 4=N2c; 5=N3.
EC	Ever used cigarettes. 0=no; 1=yes.
CC	Currently using cigarettes within 6 months. 0=no; 1=yes.
Bcl2	Score of B-cell lymphoma 2 IHC staining.
GST	Score of glutathione S-transferase π IHC staining.
p53	Score of protein 53 IHC staining.
TS	Score of thymidylate synthase IHC staining.
PFS	Progression-free survival time in years.
C	Progression-free survival status. 0=censored; 1=failed.

Table F.1: Head and Neck Cancer Data

N	A	G	Tr	P	K	T	N	EC	CC	Bcl2	GST	p53	TS	PFS	C
1	67	1	0	3	80	2	2	1	0	0.00	1.33	0.00	0.00	6.44	0
2	54	1	1	2	90	2	3	1	1	2.00	2.67	1.50	2.00	1.02	1
3	46	1	1	4	100	3	3	1	1	1.00	2.67	0.00	0.00	0.70	1
4	34	1	0	2	70	3	3	1	1	0.00	3.00	0.67	1.50	0.62	1
5	48	1	1	4	80	4	3	1	1	1.33	1.67	3.00	2.00	0.42	1
6	63	1	0	2	100	3	3	1	0	0.00	1.00	0.00	1.67	1.06	1
7	77	1	0	1	70	3	3	1	1	0.33	2.67	2.00	2.00	1.68	1
8	52	1	1	2	100	2	3	1	0	0.67	1.67	1.67	1.33	9.19	0
9	46	0	1	2	70	3	0	1	0	1.00	0.00	2.00	1.00	0.81	1
10	56	1	0	2	90	4	3	1	1	1.00	3.00	0.00	1.00	0.73	1
11	65	1	1	6	90	3	3	1	1	0.33	2.00	0.00	1.33	1.00	1
12	53	1	1	4	70	4	3	1	1	0.33	1.67	2.00	1.67	4.89	0
13	52	1	0	3	100	3	4	1	1	0.00	2.67	0.33	2.00	1.86	0
14	67	1	1	4	90	4	4	1	0	2.67	3.00	3.00	2.00	2.32	1
15	47	0	1	2	80	2	3	1	1	0.00	1.00	0.00	1.00	0.24	1
16	67	1	1	2	70	4	3	1	1	1.00	1.33	0.00	1.67	0.26	1
17	39	1	0	1	70	2	3	1	0	1.00	0.00	0.00	1.00	0.91	1
18	60	0	0	4	100	4	4	1	1	1.33	3.00	0.00	1.33	0.61	1
19	37	1	1	2	80	3	3	1	1	0.33	2.00	2.33	2.00	1.02	1
20	51	1	0	2	90	4	3	1	0	1.33	1.67	1.00	1.00	4.70	0
21	66	1	1	5	90	4	3	1	1	0.33	1.00	2.67	1.67	8.09	0
22	38	1	1	2	90	4	4	1	1	1.00	2.33	1.00	1.00	0.50	1
23	65	1	0	2	90	4	3	1	0	0.33	2.00	1.33	2.33	5.49	1
24	60	1	0	2	70	3	4	0	0	1.00	3.00	3.00	2.33	0.81	1
25	50	1	0	2	90	2	3	1	1	1.33	3.00	0.00	1.00	1.12	1

Continued. . .

N	A	G	Tr	P	K	T	N	EC	CC	Bcl2	GST	p53	TS	PFS	C
26	56	1	0	4	90	4	4	1	1	2.00	1.67	2.67	1.00	5.57	1
27	65	0	1	2	100	2	3	1	0	1.00	2.50	1.33	1.00	7.65	0
28	54	1	1	2	90	2	3	1	0	2.00	3.00	2.67	1.67	2.04	1
29	65	0	1	2	70	2	4	1	1	1.33	2.33	3.00	1.33	0.03	1
30	69	1	1	3	90	4	3	1	0	0.33	3.00	2.00	1.67	8.32	0
31	58	1	0	3	60	4	3	1	1	0.00	1.67	2.33	0.33	0.01	1
32	62	1	1	1	100	3	2	1	1	0.00	3.00	2.33	1.00	0.93	1
33	47	1	1	1	100	1	3	1	1	0.00	2.33	0.00	1.33	0.31	1
34	48	1	1	1	70	1	3	1	1	3.00	2.00	1.67	3.00	7.40	0
35	65	1	0	5	90	4	3	1	0	0.33	1.67	2.67	1.00	7.21	0
36	43	1	1	2	100	2	3	1	1	1.00	3.00	0.00	1.33	2.77	1
37	56	0	1	2	90	1	2	1	0	1.00	0.33	0.00	2.67	7.50	0
38	50	1	0	2	90	3	3	1	0	3.00	3.00	2.00	2.00	8.20	0
39	63	1	0	1	80	2	3	1	1	1.67	1.33	3.00	1.33	2.21	1
40	48	0	0	4	80	4	4	1	1	0.67	0.00	2.33	2.00	7.22	0
41	59	0	1	1	90	4	3	1	1	0.00	2.67	2.33	0.67	0.54	1
42	31	1	1	1	80	2	3	1	1	0.50	2.00	0.00	1.50	7.24	0
43	43	0	1	1	90	4	3	1	1	1.00	2.33	2.00	2.67	6.87	0
44	49	1	1	2	90	1	3	1	1	3.00	0.33	1.67	1.00	6.85	0
45	47	1	0	4	100	4	3	1	1	0.00	2.00	0.00	1.67	0.21	0
46	61	0	1	1	80	3	3	1	1	0.00	1.67	0.00	0.33	0.28	1
47	72	1	1	3	70	4	2	1	1	1.00	1.67	3.00	2.00	0.26	1
48	51	1	0	3	100	3	3	1	1	1.00	2.00	0.00	0.33	0.93	1
49	68	1	0	2	70	1	3	1	1	0.33	1.00	3.00	2.67	1.71	0
50	59	1	1	2	80	3	2	1	1	3.00	1.00	3.00	2.33	7.72	1
51	40	1	0	2	100	3	3	1	1	0.33	1.67	0.00	1.00	0.58	1
52	53	1	0	3	80	4	4	1	1	0.67	3.00	0.00	1.00	0.23	1

Continued...

N	A	G	Tr	P	K	T	N	EC	CC	Bcl2	GST	p53	TS	PFS	C
53	55	1	0	2	90	2	3	1	1	2.67	1.33	1.67	2.33	6.11	0
54	55	1	0	2	100	4	3	1	1	1.00	2.67	0.00	1.33	0.06	1
55	46	1	1	2	90	1	3	0	0	1.00	1.00	1.00	1.67	7.21	0
56	56	1	1	3	90	4	3	0	0	1.00	3.00	3.00	1.33	0.90	1
57	31	1	0	2	80	2	3	1	1	1.00	3.00	2.33	2.67	1.61	1
58	59	1	0	1	90	3	3	0	0	0.33	2.00	0.00	1.67	0.42	1
59	63	1	0	4	90	3	4	1	0	1.00	3.00	2.67	1.00	0.72	1
60	64	1	1	4	90	3	3	1	0	2.00	1.33	2.67	2.00	7.49	1
61	65	1	1	1	90	2	3	1	0	1.00	3.00	0.00	0.50	6.63	1
62	45	1	0	2	80	4	3	1	1	1.00	2.67	1.67	2.00	2.63	1
63	51	1	1	2	80	3	3	1	0	0.33	1.33	3.00	1.33	0.54	1
64	52	1	1	2	70	4	3	1	0	0.67	2.00	0.00	1.00	1.78	1
65	49	1	0	3	80	4	3	1	1	1.33	3.00	0.00	1.33	0.93	1
66	32	1	0	2	90	3	3	1	1	0.67	1.33	2.67	2.00	1.13	1
67	57	1	1	3	70	4	4	1	1	0.00	3.00	3.00	1.00	0.34	1
68	60	1	0	4	90	2	4	1	0	0.67	3.00	0.00	1.00	1.91	1
69	78	1	1	3	80	3	3	1	1	1.00	1.33	2.00	1.00	4.30	1
70	60	1	1	4	70	2	4	1	0	2.00	1.00	1.50	1.00	0.55	0
71	57	1	0	1	90	4	3	1	1	0.33	1.67	0.00	1.67	1.62	1
72	52	1	0	2	90	2	3	1	1	1.67	3.00	3.00	1.33	0.71	1
73	40	1	0	2	90	1	3	1	1	3.00	1.50	0.67	1.67	5.49	0
74	42	0	0	2	90	1	3	1	1	1.00	1.33	0.00	1.00	0.96	1
75	66	1	0	4	80	3	4	1	0	0.00	2.00	3.00	1.33	0.24	1
76	55	1	0	1	90	2	3	1	1	0.00	1.00	2.33	0.33	2.29	1
77	58	1	0	4	80	4	3	1	1	2.00	3.00	2.00	1.00	3.67	0
78	50	1	1	3	70	4	3	1	1	1.00	1.50	2.00	1.50	0.02	1
79	74	1	0	2	70	4	3	1	1	3.00	1.67	1.00	2.00	6.20	0

Continued...

N	A	G	Tr	P	K	T	N	EC	CC	Bcl2	GST	p53	TS	PFS	C
80	50	1	1	2	90	2	3	1	0	2.00	2.00	2.33	2.00	6.73	0
81	64	1	0	3	90	4	3	1	1	2.00	2.67	3.00	2.00	6.30	0
82	47	1	0	1	80	1	3	1	0	0.33	3.00	0.00	1.00	1.17	1
83	66	1	0	1	80	3	3	1	1	0.00	1.33	3.00	1.00	0.79	0
84	50	0	1	4	80	3	4	1	1	2.67	1.33	3.00	1.00	2.91	1
85	62	1	0	5	90	3	3	1	1	1.67	3.00	2.33	0.67	3.43	1
86	51	1	0	2	80	1	3	1	1	2.00	1.67	2.67	0.67	3.72	1
87	51	1	1	1	90	3	3	1	1	0.33	2.00	1.67	1.33	3.14	1
88	54	1	0	4	80	1	3	1	1	1.50	3.00	3.00	1.00	0.74	1
89	48	1	1	2	90	2	3	0	0	1.67	0.00	0.33	1.67	6.19	0
90	44	1	1	2	90	3	3	1	1	2.33	1.00	0.33	1.67	5.22	0
91	62	1	0	4	80	3	3	1	0	1.00	1.67	3.00	1.67	0.30	1
92	54	0	0	4	80	4	4	1	1	1.33	1.00	0.00	2.00	0.75	1
93	56	1	1	2	100	3	2	1	0	3.00	1.00	2.00	2.67	5.56	0
94	48	1	0	2	100	3	3	1	0	1.50	0.00	1.50	1.50	5.10	0
95	61	1	0	2	90	2	3	1	1	0.00	2.33	0.00	1.00	1.68	1
96	52	1	0	1	70	4	4	1	1	0.00	0.33	2.00	1.00	0.45	0
97	47	1	0	1	90	1	3	1	1	1.50	3.00	1.00	2.50	6.02	0
98	62	1	0	3	90	1	2	1	0	0.67	1.00	3.00	2.33	0.34	1
99	60	0	1	2	70	4	3	1	0	0.00	1.33	1.00	1.00	3.49	0
100	54	1	0	2	80	4	3	1	0	2.00	2.33	0.00	2.67	0.34	1
101	70	1	1	3	80	3	3	1	0	1.00	2.00	2.67	2.00	0.52	1
102	66	1	1	2	90	2	3	1	0	1.67	2.67	2.33	1.33	4.90	0
103	48	0	0	1	70	4	4	1	1	1.67	2.67	3.00	2.00	1.02	1
104	64	1	0	1	80	3	4	1	1	0.33	0.00	0.00	1.33	0.94	1
105	55	1	0	2	90	1	3	1	0	1.00	2.67	1.33	1.00	4.65	0
106	54	1	1	2	70	3	3	1	0	0.33	1.67	1.67	2.00	0.71	1

Continued. . .

N	A	G	Tr	P	K	T	N	EC	CC	Bcl2	GST	p53	TS	PFS	C
107	50	1	1	5	90	4	5	1	1	1.67	2.33	3.00	1.67	5.36	0
108	68	1	1	2	70	2	4	1	1	3.00	2.67	0.00	2.67	1.42	1
109	64	1	1	2	90	2	3	1	1	1.00	1.67	1.33	1.00	0.09	1
110	53	1	1	4	100	3	3	1	0	1.33	1.67	0.00	0.67	0.99	1
111	51	1	0	2	90	2	2	1	0	1.67	2.67	1.33	2.00	5.37	0
112	63	1	0	1	80	1	3	1	1	2.00	2.33	3.00	1.33	2.12	1
113	54	1	1	4	90	4	4	1	0	0.33	3.00	1.67	1.00	0.39	1
114	61	1	1	2	80	2	3	1	1	0.33	2.33	3.00	1.00	1.39	1
115	63	1	0	3	90	4	3	1	1	0.67	1.00	3.00	1.00	0.90	1
116	63	1	0	2	80	2	3	1	0	1.00	2.33	2.00	1.33	0.22	1
117	79	1	0	1	80	2	0	1	1	0.00	1.67	1.00	1.67	0.32	0
118	76	0	0	1	90	3	3	1	0	1.00	0.00	0.00	0.67	0.74	1
119	53	1	1	6	90	4	1	1	1	1.67	3.00	0.33	2.33	3.83	1
120	58	1	0	1	90	3	3	1	1	3.00	1.00	3.00	2.33	2.97	1
121	52	1	0	2	90	3	4	1	1	1.00	3.00	0.00	1.33	0.86	0
122	47	1	0	2	100	2	3	1	0	1.33	2.00	2.00	1.00	2.93	0

Bibliography

[1] Andersen, P. K. and Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study, *The Annals of Statistics,* 10, 1100-1120.

[2] Antoniadis, A. and Fan, J. (2001). Regularization of Wavelet Approximations (with discussion), *Journal of the American Statistical Association,* 96, 939-967.

[3] Breslow, N. (1972). Discussion on Professor Cox's Paper, Regression Models and Life Tables, *Journal of the Royal Statistical Society, Series B (Methodological)* 34, 216-217.

[4] Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical-Theoretic Approach.* Springer-Verlag.

[5] Chen, S. S. , Donoho, D. L. and Saunders, M. A. (2001). Atomic Decomposition by Basis Pursuit, *SIAM Review,* 43, 129-159.

[6] Cox, D. R. (1975). Partial Likelihood, *Biometrika,* 62, 269-276.

[7] 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, *Numerische Mathematik,* 31, 377-403.

[8] Donoho, D. and Johnstone, I. (1994). Ideal Spatial Adaptation via Wavelet Shrinkages, *Biometrika,* 81, 425-455.

[9] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. J. (2004). Least Angle Regression (with discussion), *The Annals of Statistics,* 32, 407-499.

[10] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association,* 96, 1348-1360.

[11] Fan, J. and Li, R. (2002). Variable Selection for Cox's Proportional Hazards Model and Frailty Model, *The Annals of Statistics,* 30, 74-99.

[12] Fleming, T. R. and Harrington, D. P. (1991). *Counting Process and Survival Analysis.* New York. John Wiley & Sons.

[13] Frank, I. E. and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools, *Technometrics,* 35, 109-135.

[14] Fu, W. (1998). Penalized Regression: The Bridge Versus the Lasso, *Journal of Computational and Graphical Statistics,* 7, 397-416.

[15] Greene, F. L., Page, D. L., Fleming I. D., Fritz, A. and Balch, C. M. (2002). *AJCC Cancer Staging Manual, 6th Edition*, Springer-Verlag, New York.

[16] Gui, J. and Li, H. (2005). Penalized Cox Regression Analysis in the High-Dimensional and Low-Sample Size Settings, with Applications to Microarray Gene Expression Data, *Bioinformatics,* 21, 3001-3008.

[17] Heagerty, P. J. and Zheng, Y. (2005). Survival Model Predictive Accuracy and ROC Curves, *Biometrics,* 61, 92-105.

[18] Knight, K. and Fu, W. (2000). Asymptotics for Lasso-Type Estimators, *The Annals of Statistics,* 28, 1356-1378.

[19] Le Cam, L. (1960). Locally Asymptotically Normal Families of Distributions. *Univ. Calif. Publ. Statist.,* 3, 27-98.

[20] Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts*, Springer-Verlag, New York.

[21] Meinshausen, N. and Bühlmann, P. (2006). High-Dimensional Graphs and Variable Selection with the Lasso, *The Annals of Statistics,* 34, 1436-1462.

[22] Mor, V., Laliberte, L., Morris, J. N. and Wiemann, M. (1984). The Karnofsky Performance Status Scale: An Examination of its Reliability and Validity in a Research Setting, *Cancer,* 53, 2002-2007.

[23] Patel, S. G. and Shah, J. P. (2005). TNM Staging of Cancers of the Head and Neck: Striving for Uniformity Among Diversity, *CA Cancer Journal for Clinicians,* 55, 242-258.

[24] Peduzzi, P. N., Hardy, R. J. and Holford, T. R. (1980). A Stepwise Variable Selection Procedure for Nonlinear Regression Models, *Biometrics,* 36, 511-516.

[25] Pepe, M. S. (2000). Receiver Operating Characteristic Methodology, *Journal of the American Statistical Association,* 95, 308-311.

[26] Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, New York.

[27] Schumaker, L., Nikitakis, N., Goloubeva, O., Tan, M., Taylor, R. and Cullen, K. J. (2008). Elevated Expression of Glutathione S-Transferase π and p53 Confers Poor Prognosis in Head and Neck Cancer Patients Treated with Chemoradiotherapy but not Radiotherapy Alone, *Clinical Cancer Research,* 14, 5877-5883.

[28] Shen, X. and Ye, J. (2002). Adaptive Model Selection, *Journal of the American Statistical Association,* 97, 210-221.

[29] Shiga, H., Heath, E. I., Rasmussen, A. A., Trock, B., Johnston, P. G., Forastiere, A. A., Langmacher, M., Baylor, A., Lee, M. and Cullen, K. J. (1999). Prognostic Value of p53, Glutathione S-Transferase π and Thymidylate Synthase for neoadjuvant Cisplatin-based Chemotherapy in Head and Neck Cancer, *Clinical Cancer Research,* 5, 4097-4104.

[30] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B,* 58, 267-288.

[31] Tibshirani, R. (1997). The Lasso Method for Variable Selection in the Cox Model, *Statistics in Medicine,* 16, 385-395.

[32] Wang, Y. (2004). Model Selection, *Handbook of Computational Statistics, Gentle, J. E., Härdle, W. and Mori, Y. (Editors)*, Springer-Verlag Berlin Heidelberg, 437-466.

[33] Xu, R. and O'Quigley, J. (2000). Proportional Hazards Estimate of the Conditional Survival Function, *Journal of the Royal Statistical Society: Series B,* 62, 667-680.

[34] Yuan, M. and Lin, Y. (2005). Efficient Empirical Bayes Variable Selection and Estimation in Linear Models, *Journal of the American Statistical Association,* 100, 1215-1225.

[35] Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's Proportional Hazards Model, *Biometrika,* 94, 691-703.

[36] Zou, H. (2006). The Adaptive Lasso and its Oracle Properties, *Journal of the American Statistical Association,* 101, 1418-1429.

[37] Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net, *Journal of the Royal Statistical Society: Series B,* 67, 301-320.