5-7-2011

Some Topics in Roc Curves Analysis

Xin Huang
Georgia State University

SOME TOPICS IN ROC CURVES ANALYSIS

by

XIN HUANG

Under the Direction of Dr. Yixin Fang

ABSTRACT

The receiver operating characteristic (ROC) curves is a popular tool for evaluating continuous diagnostic tests. The traditional definition of ROC curves incorporates implicitly the idea of "hard" thresholding, which also results in the empirical curves being step functions. The first topic is to introduce a novel definition of soft ROC curves, which incorporates the idea of "soft" thresholding. The softness of a soft ROC curve is controlled by a regularization parameter that can be selected suitably by a cross-validation procedure. A byproduct of the soft ROC curves is that the corresponding empirical curves are smooth.

The second topic is on combination of several diagnostic tests to achieve better diagnostic accuracy. We consider the optimal linear combination that maximizes the area under the receiver operating characteristic curve (AUC); the estimates of the combination's coefficients can be obtained via a non-parametric procedure. However, for estimating the AUC associated with the estimated coefficients, the apparent estimation by re-substitution is too optimistic. To adjust for the upward bias, several methods are proposed. Among them the cross-validation approach is especially advocated, and an approximated cross-validation is developed to reduce the computational cost. Furthermore, these proposed methods can be applied for variable selection to select important diagnostic tests.

However, the above best-subset variable selection method is not practical when the number of diagnostic tests is large. The third topic is to further develop a LASSO-type procedure for variable

selection. To solve the non-convex maximization problem in the proposed procedure, an efficient algorithm is developed based on soft ROC curves, difference convex programming, and coordinate descent algorithm.

INDEX WORDS: Area under curve, Coordinate descent, Cross-validation, Difference convex programming, Diagnostic test, Over-fitting, Regularization, ROC curve, Thresholding, Variable selection

SOME TOPICS IN ROC CURVES ANALYSIS

by

XIN HUANG

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2011

SOME TOPICS IN ROC CURVES ANALYSIS

by

XIN HUANG

Chair:	Dr. Yixin Fang
Committee:	Dr. Man Jin
	Dr. Gengsheng Qin
	Dr. Yichuan Zhao

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2011

This dissertation is dedicated to my dear parents,

and all my dear friends.

ACKNOWLEDGEMENTS

There are so many to thank during my long journey in creating such a delicate dissertation. First and foremost I would like to thank Dr. Yixin Fang, my doctoral advisor, who is always willing to walk with me as I struggled against various research difficulties. I shall never forgot the afternoon call years ago, when he expressed his willingness to take a chance with me. Since then, my life has changed. His words of encouragement, quiet urgings and careful reading of all my writing will never be forgotten, and will be cherished during my whole course of life. In that same vein, I want to thank my doctoral committees – Dr. Man Jin, Dr. Gengsheng Qin, and Dr. Yichuan Zhao who agreed to come with me in thinking through my research topics. Without their support, I could not have done what I was able to do.

My parents have all been encouraging. They funded me initially for my graduate study, and are always being optimistic under any circumstance. I would never have moved forward without their mentally support. They gave me the courage to risk being my creative self wherever that may lead me.

My special thank goes to my girl friend, Shan Luo. Her presence helped me to make the transition from where I came to where I am. We took care of each other, and supported each other down to the loving tears we shed on the difficulties and happiness during our daily life.

Many people on the faculty and staff from Georgia State University assisted and encouraged me in various ways during my course of studies. I am especially grateful to professors from Math. & Stat. Department: Yu-Sheng Hsu, Marina Arav, Guantao Chen, Jun Han, Zhongshan Li, Jiawei Liu, Valerie Miller, Yuanhui Xiao, Draga Vidakovic, and Xu Zhang for all that they have taught me. My thanks must also go to professors from Institute of Public Health: Karen Gieseker, Ike Okosun, John Steward, and Sheryl Strasser, as well as professors from Department of Applied Linguistics: Sharon Cavusgil and John Stowe, who died from heart attack. I was also greatly inspired by Prof. Gengsheng Qin, for whom served as my Master thesis advisor for two years.

My graduate studies would not have been the same without the social and academic challenges and diversions provided by all my student-colleagues in the Math. & Stat. Department. I need to thank especially Haci Akcin, Binghuan Wang, Yanhong Wang, Hanfang Yang, Haochuan Zhou, Meng Zhao, and Kun Zhao.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . . . **v**

LIST OF FIGURES . . . . . **x**

LIST OF TABLES . . . . . **xi**

LIST OF ABBREVIATIONS . . . . . **xii**

Chapter 1 INTRODUCTION **1**

1.1 Soft-thresholding 1

1.2 Combinations of multiple diagnostic tests 3

1.3 AUC-LASSO 4

1.4 Brief summary 5

Chapter 2 SOFT ROC CURVES **6**

2.1 Soft ROC curves 6

2.1.1 Two-sided indecisive functions 7

2.1.2 One-sided indecisive functions 9

2.2 Selection of regularization parameter 10

2.2.1 Method based on softness 11

2.2.2 Method based on cross-validation 12

Chapter 3 COMBINING DIAGNOSTIC TESTS **14**

3.1 Properties of $\widehat{\boldsymbol{\beta}}$ 14

3.2 Estimates of the AUC associated with $\widehat{\boldsymbol{\beta}}$ 15

3.2.1 Cross-validation 15
3.2.2 Bootstrap methods 16
3.2.3 Sigmoid function smoothing 18
3.2.4 Approximated cross-validation for variable selection 19
3.2.5 Comparison with some LASSO-type procedures 20

Chapter 4 AUC-LASSO 21

4.1 Motivation 21
4.2 Review of related computational methods 22
4.2.1 Penalized likelihood methods and coordinate descent algorithms 22
4.2.2 Difference convex programming 23
4.3 AUC-LASSO 24

Chapter 5 NUMERICAL STUDIES 27

5.1 Numerical studies for soft ROC curves 27
5.1.1 Simulation study 27
5.1.2 Pancreatic cancer serum biomarkers example 28
5.2 Numerical studies for optimal combinations of diagnostic tests 29
5.2.1 Simulation studies 29
5.2.2 Pancreatic cancer serum biomarkers example 35
5.2.3 Wisconsin breast cancer study 36
5.2.4 Pima Indians diabetes study 37

Chapter 6 DISCUSSION 39

REFERENCES 41

APPENDICES 45

Appendix A SOME PROOFS AND CALCULATIONS FOR CHAPTER 2 **45**

A.1 Proof of Theorem 2.1 45

A.2 Calculation for order 0 two-sided indecisive function 45

A.3 Calculation for order 1 two-sided indecisive function 46

A.4 Calculation for order ∞ two-sided (Sigmoid) indecisive function 46

A.5 Calculation for order 0 one-sided indecisive function 47

A.6 Calculation for order 1 one-sided indecisive function 47

Appendix B PROOFS FOR CHAPTER 3 . . . . . **48**

B.1 Investigation of the GLM assumption 48

B.2 The proof of Proposition 3.1 48

B.3 The proof of Proposition 3.2 50

B.4 The derivation of the approximated cross-validation 50

LIST OF FIGURES

Figure 2.1 Two-sided I_δ and their corresponding K_δ 8

Figure 2.2 One-sided I_δ and their corresponding K_δ 9

Figure 2.3 Plots of δ versus mean difference μ for some given α 12

Figure 4.1 The DC representation for one-sided order 1 K_δ 25

Figure 5.1 Comparison of mean square errors of AUC_δ and AUC_0 28

Figure 5.2 ROC curves for Pancreatic Cancer Serum Biomarkers Example: CA-125 30

Figure 5.3 ROC curves for Pancreatic Cancer Serum Biomarkers Example: CA-19-9 31

Figure 5.4 Plot of $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$ versus λ 33

Figure 5.5 Plot of estimated AUC's versus the subsets 34

Figure 5.6 Plot of $\mathrm{AUC}_\lambda^{(\mathrm{ACV})}$ versus the subset size 37

LIST OF TABLES

Table 5.1 Performance of CV 29

Table 5.2 Different methods for estimating AUC($\widehat{\boldsymbol{\beta}}$) 32

Table 5.3 Example 5.2.2 – Pancreatic cancer serum biomarkers 35

Table 5.4 Example 5.2.4 – Pima Indians diabetes 38

LIST OF ABBREVIATIONS

- AMSE - average mean squared error
- AUC - area under the ROC curve
- BIC - bayesian information criterion
- CV - cross-validation
- DC - differenced convex
- FPR - false positive rate
- TIC - Takeuchi information criterion
- TPR - true positive rate
- GLM - generalized linear model
- LASSO - least absolute shrinkage and selection operator
- LARS - least angle regression
- LDA - linear discriminant analysis
- ROC - receiver operating characteristic
- SCAD - smoothly clipped absolute deviation
- SVM - suppor vector machine

Chapter 1

INTRODUCTION

1.1 Soft-thresholding

Laboratory diagnostic tests are one of the most important components in modern medical practice. The receiver operating characteristic (ROC) curves, the plots of true-positive rate against false-positive rate, are popular tools for evaluating continuous diagnostic tests; see, for example Pepe (2003) and Zhou et al. (2002). However, the traditional definition of ROC curves incorporates implicitly the idea of "hard" thresholding. To be specific, let T be the outcome of a continuous diagnostic test and D be the disease status. Given a threshold c, the hard-thresholding scheme (H) defines a subject as diseased ($\widehat{D} = 1$) if the test result $T = t$ exceeds c, and as non-diseased ($\widehat{D} = 0$) otherwise. It thus results in a binary classifier

$$\text{(H)} \qquad I(t-c) = \begin{cases} 1, & t - c \geq 0, \\ 0, & t - c < 0. \end{cases}$$

The ROC curve is then a graphical plot of true positives, $E\{I(T-c)|D=1\}$, versus false positives, $E\{I(T-c)|D=0\}$, for $-\infty < c < \infty$. It can be expressed as

$$R(p) = 1 - G[F^{-1}(1-p)], \quad 0 < p < 1,$$

where $F(\cdot)$ and $G(\cdot)$ are the distributions of T, given $D = 0$ and $D = 1$, respectively.

Unfortunately, there are some disadvantages in the above hard-thresholding scheme. First, it is too strict since the true disease status of a subject is hard to detect if the test result is close to the specified threshold. Second, the discontinuity of the binary classifier results in the corresponding

estimated ROC curve being a step function, while the underlying ROC curve is likely to be smooth. Finally, due to the discontinuity in the step function, the variability of the estimated ROC curve becomes large.

To overcome these disadvantages, we consider the following soft-thresholding scheme (S):

$$\text{(S)} \qquad I_\delta(t-c) = \begin{cases} 1, & t-c \geq \delta, \\ ?, & -\delta < t-c < \delta, \\ 0, & t-c < -\delta, \end{cases}$$

where the value ? is between 0 and 1 and will be discussed in the next section, and δ is a regularization parameter controlling the softness. In particular, when $\delta = 0$, the soft-thresholding simply becomes the hard-thresholding. The rationale for this scheme is that if the test result is close to the given thresholding c, then we may be indecisive about the status of the disease. Hence, we refer to $I_\delta(\cdot)$ as the indecisive function. Different indecisive functions will result in different soft ROC curves.

The idea used here is similar in principle to the one used in designing randomization tests to achieve a given significance level in hypothesis testing (Lehmann, 1997). The indecisive function has been considered in the literature of ROC. Liu et al. (2009) and Liu and Tan (2008) used an S-type function to approximate the indicator function for the empirical False Positive Rate (FPR) and True Positive Rate (TPR). Huang et al. (2010), Wang et al. (2007), and Ma and Huang (2005, 2007) used the sigmoid function to approximate the indicator function for the empirical area under the ROC curve (AUC).

Instead of looking for an approximation, in our work, we examine the definition of ROC curves directly and introduce the soft ROC curves based on the soft-thresholding. More importantly, we build a bridge between the approximation of an ROC curve and the approximation of its AUC. Moreover, continuity of the proposed soft ROC curves is a promising byproduct, although it is not our primary goal. We should point out that in the literature of ROC, many authors have discussed

methods to smooth ROC curves. For example, Zou et al. (1997) proposed a non-parametric estimator from kernel estimates of the distribution functions F and G. Peng and Zhou (2004) proposed a local linear regression for the ROC curve, while Ren et al. (2004) proposed a penalized spline linear mixed-effects model.

1.2 Combinations of multiple diagnostic tests

The ROC curve is a graphical tool for evaluating the discriminatory accuracy of diagnostic tests. Meanwhile, the AUC is a popular one-number summary index of the discriminatory accuracy; the closer to one it is, the more accurate the test is. When several diagnostic tests are available, one can combine them to achieve better diagnostic accuracy. Let $\mathbf{T} = (T_1, \cdots, T_p)^T$ be p diagnostic tests that yield continuous measurements. Assume that all tests are performed on m non-diseased subjects, yielding testing outcomes $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^T$, $i = 1, \cdots, m$, i.i.d with $\mathbf{X}$ of distribution $F(\mathbf{X})$, and on n diseased subjects, yielding outcomes $\mathbf{Y}_j = (Y_{j1}, \cdots, Y_{jp})^T$, $j = 1, \cdots, n$, i.i.d with $\mathbf{Y}$ of distribution $G(\mathbf{Y})$. We are interested in seeking a linear combination of the diagnostic tests such that the combined score achieves the maximum AUC over all the possible linear combinations. Thus, as Bamber (1975), we are interested in the following coefficient vector,

$$\boldsymbol{\beta}_0 = \arg\max_{\boldsymbol{\beta} \in \mathbf{B}} P(\boldsymbol{\beta}^T \mathbf{Y} > \boldsymbol{\beta}^T \mathbf{X}), \tag{1.1}$$

where $\mathbf{B} = \{\boldsymbol{\beta} \in \mathcal{R}^p : ||\boldsymbol{\beta}|| = 1\}$. A nonparametric estimate of $\boldsymbol{\beta}_0$ can be obtained via the following maximum-rank procedure,

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta} \in \mathbf{B}} \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I(\boldsymbol{\beta}^T \mathbf{Y}_j > \boldsymbol{\beta}^T \mathbf{X}_i), \tag{1.2}$$

where $I(\cdot)$ is the indicator function. This procedure has been well-studied in the literature of ROC curves. For example, Su and Liu (1993) discussed the optimal linear combination under the multiple-normal assumption; Pepe and Tompson (2000), Pepe et al. (2006), and Ma and Huang

(2005) discussed this procedure under the generalized linear model (GLM) assumption. However, besides re-evaluating the assumptions for the properties of $\widehat{\boldsymbol{\beta}}$, we are concerned with estimates for the AUC associated with $\widehat{\boldsymbol{\beta}}$,

$$\mathrm{AUC}(\widehat{\boldsymbol{\beta}}) = P(\widehat{\boldsymbol{\beta}}^T\mathbf{Y} > \widehat{\boldsymbol{\beta}}^T\mathbf{X}), \tag{1.3}$$

where $(\mathbf{X}, \mathbf{Y})$ are future observations independent of the observed data used for obtaining $\widehat{\boldsymbol{\beta}}$, and the probability is taken over $(\mathbf{X}, \mathbf{Y})$ conditionally on the observed data. These estimates are of great interests; after the combined test is obtained, we are interested in assessing its discriminatory accuracy.

The simplest estimate of $\mathrm{AUC}(\widehat{\boldsymbol{\beta}})$ is the apparent estimate by re-substitution,

$$\overline{\mathrm{AUC}}(\widehat{\boldsymbol{\beta}}) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I(\widehat{\boldsymbol{\beta}}^T\mathbf{Y}_j > \widehat{\boldsymbol{\beta}}^T\mathbf{X}_i). \tag{1.4}$$

Obviously, this estimate is too optimistic; see e.g., Efron (1983). Copas and Corbett (2002) addressed the overfitting problem when combining tests via logistic regression, and suggested a shrinkage correction for the AUC. Here we are concerned with approaches to adjusting for the upward bias of $\overline{\mathrm{AUC}}(\widehat{\boldsymbol{\beta}})$.

After the question of how to adjust the upward bias of $\overline{\mathrm{AUC}}(\widehat{\boldsymbol{\beta}})$ is answered, one can perform a best subset selection procedure to select important diagnostic tests: the adjusted AUC estimates are calculated for all possible subsets of diagnostic tests, and the subset with the largest adjusted AUC estimate is chosen as the important diagnostic tests.

1.3 AUC-LASSO

When the number of tests p is increasing, the number of subsets is increasing dramatically as $2^p - 1$, and therefore the best subset selection procedure is not practical as p increases. Moreover, the best subset variable selection method is lack of stability as analyzed (Breiman, 1996). The LASSO procedure is among the most popular methods for variable selection. It shrinks the coefficients

by imposing a L1 penalty and produces a kind of continuous subset selection. More discussion on the comparison of LASSO and best-subset selection methods can be found in Hastie et al. (2009, p.69). Therefore, we try to impose the LASSO penalty into (1.2), named AUC-LASSO, and develop corresponding efficient algorithms. We will use the soft ROC curve to approximate the non-convex objective function in (1.2), and use the difference convex programming and coordinate descent algorithms to solve the global optima of AUC-LASSO.

1.4 Brief summary

The remainder of this dissertation is organized as follows. In Chapter 2, we define the soft ROC curve, and derive some of its properties. We also propose methods to choose the regularization parameter δ. In Chapter 3, we study the optimal combinations of diagnostic tests. We re-investigate the properties of the estimated linear combination coefficient, and show the uniqueness of $\boldsymbol{\beta}_0$ and the consistency of $\widehat{\boldsymbol{\beta}}$ under some mild assumptions. We also propose several methods for estimating the AUC associated with $\widehat{\boldsymbol{\beta}}$, while all these methods can be applied as variable selection criteria. In Chapter 4, we review several related state-of-the-art computational methods for regularized likelihood methods and non-convex optimization, and propose an algorithm to solve the AUC-LASSO problem. In Chapter 5, we examine the propose methods through some simulation studies and real examples. Finally, some discussion is made in Chapter 6. All technical details are relegated to the Appendix.

Chapter 2

SOFT ROC CURVES

This chapter is organized as follows. In Section 2.1, we define the soft ROC curve, and derive some of its properties. In Section 2.2, we propose methods to choose the regularization parameter δ. All technical details are relegated to the Appendix A.

2.1 Soft ROC curves

When the indecisive function I_δ is applied with threshold c, the true positive probability equals $E\{I_\delta(T-c)|D=1\}$ and the false positive probability equals $E\{I_\delta(T-c)|D=0\}$. We can therefore define soft ROC curves as follows.

DEFINITION 2.1: *A plot of true positives, $E\{I_\delta(T-c)|D=1\}$, versus false positives, $E\{I_\delta(T-c)|D=0\}$, for all possible values of c, is called the soft ROC curve with respect to indecisive function I_δ.*

Assume that a test is performed on m non-diseased subjects, yielding testing outcomes X_i, and on n diseased subjects, yielding outcomes Y_j. Then, an empirical estimate of the soft ROC curve w.r.t. I_δ is

$$\widehat{R}_\delta(p) = 1 - \widehat{G}_\delta[\widehat{F}_\delta^{-1}(1-p)], \quad p \in (0,1), \tag{2.1}$$

where $\widehat{G}_\delta(c) = \frac{1}{n}\sum_{j=1}^{n} I_\delta(Y_j - c)$ and $\widehat{F}_\delta(c) = \frac{1}{m}\sum_{i=1}^{m} I_\delta(X_i - c)$. The area under the soft ROC curve w.r.t. I_δ (denoted by AUC_δ) is derived in the following theorem, and its proof is presented in the Appendix A.

THEOREM 2.1: *For the soft ROC curve w.r.t. to indecisive function $I_\delta(\cdot)$, we have*

$$AUC_\delta = E\{K_\delta(Y-X)\},$$

where $X \sim F(\cdot)$, $Y \sim G(\cdot)$, $K_\delta(Y - X) = \int_{-\infty}^{\infty} I_\delta(Y - c)\dot{I}_\delta(X - c)dc$, *and* $\dot{I}_\delta$ *is the derivative of* I_δ.

Thus, from Theorem 2.1, we see that an unbiased estimate of AUC_δ is given by

$$\widehat{\mathrm{AUC}}_\delta = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} K_\delta(Y_j - X_i). \tag{2.2}$$

2.1.1 Two-sided indecisive functions

We can categorize indecisive functions and soft ROC curves into one-sided and two-sided according to the following definition.

DEFINITION 2.2: *If* $I_\delta(t - c) = 0$ *for* $t < c$, I_δ *and the corresponding soft ROC curve are said to be one-sided. Otherwise, they are said to be two-sided.*

We now present some examples of indecisive functions I_δ and their corresponding K_δ, which are all displayed in Figure 2.1. The corresponding detailed calculations are presented in the Appendix A.

EXAMPLE 2.1: *Order 0 two-sided indecisive function is given by*

$$I_\delta(t - c) = \frac{1}{2}\mathbf{1}\{-\delta \le t - c < \delta\} + \mathbf{1}\{t - c \ge \delta\},$$

where $\mathbf{1}\{\cdot\}$ *is an indicator function. This implies that the disease status is totally indecisive when* t *is within* δ *of threshold* c. *The corresponding* K_δ *is*

$$K_\delta(s) = \frac{1}{4}\mathbf{1}\{-2\delta \le s < 0\} + \frac{3}{4}\mathbf{1}\{0 \le s < 2\delta\} + \mathbf{1}\{s \ge 2\delta\}.$$

EXAMPLE 2.2: *Order 1 two-sided indecisive function is given by*

$$I_\delta(t - c) = \left[\frac{1}{2} + \frac{1}{2\delta}(t - c)\right]\mathbf{1}\{-\delta \le t - c < \delta\} + \mathbf{1}\{t - c \ge \delta\}.$$

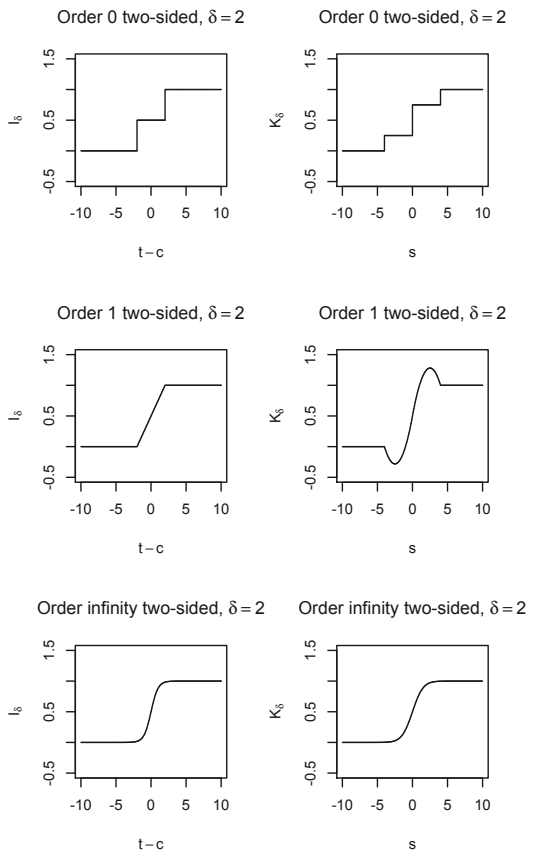


Figure 2.1. Two-sided I_δ and their corresponding K_δ

This implies that the positive probability is linear in $t - c$ when t is within δ of threshold c. The corresponding K_δ is

$$K_\delta(s) = \left[\frac{1}{2} + \frac{5s}{4\delta} - sign(s)\frac{s^2}{2\delta^2}\right] \mathbf{1}\{-2\delta \leq s < 2\delta\} + \mathbf{1}\{s \geq 2\delta\},$$

where $sign(\cdot)$ is the sign function. This K_δ takes on a strange form (see Figure 2.1) and so the order 1 two-sided indecisive function above may not be a good choice in practice.

EXAMPLE 2.3: *Order ∞ two-sided (Sigmoid) indecisive function is given by*

$$I_\delta(t - c) = \frac{1}{1 + e^{-\delta(t-c)}}.$$

An appealing property of the sigmoid function is that it has infinite derivatives. The corresponding K_δ *is*

$$K_\delta(s) = -e^{\delta s}\left[\frac{1}{1-e^{\delta s}} + \frac{1}{(1-e^{\delta s})^2}\delta s\right].$$

2.1.2 One-sided indecisive functions

In this subsection, we present two examples of one-sided indecisive functions I_δ and their corresponding K_δ, which are displayed in Figure 2.2. The indecisive functions are similar to the ones in Examples 2.1 and 2.2, but the order 1 one-sided K_δ takes on a reasonable form, unlike its two-sided counterpart. The corresponding detailed calculations are presented in the Appendix A.

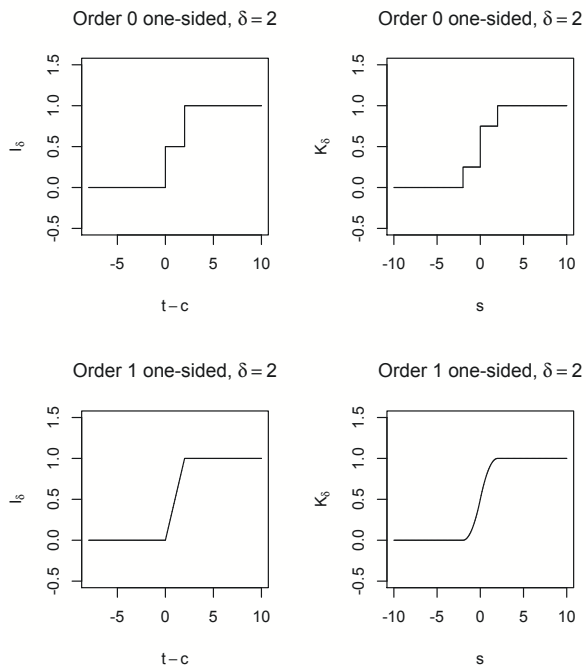


Figure 2.2. One-sided I_δ and their corresponding K_δ

EXAMPLE 2.4: *Order 0 one-sided indecisive function is given by*

$$I_\delta(t-c) = \frac{1}{2}\mathbf{1}\{0 \le t-c < \delta\} + \mathbf{1}\{t-c \ge \delta\},$$

This implies that the disease status is totally indecisive when t is in the interval $[c, c+\delta)$. The corresponding K_δ is

$$K_\delta(s) = \frac{1}{4}\mathbf{1}\{-\delta \le s < 0\} + \frac{3}{4}\mathbf{1}\{0 \le s < \delta\} + \mathbf{1}\{s \ge \delta\}.$$

EXAMPLE 2.5: *Order 1 one-sided indecisive function is given by*

$$I_\delta(t-c) = \frac{1}{\delta}(t-c)\mathbf{1}\{0 \le t-c < \delta\} + \mathbf{1}\{t-c \ge \delta\}.$$

This implies that the positive probability is linear in $t-c$ when t is in the interval $[c, c+\delta)$. The corresponding K_δ is

$$K_\delta(s) = \left[\frac{1}{2} + \frac{s}{\delta} - sign(s)\frac{s^2}{2\delta^2}\right]\mathbf{1}\{-\delta \le s < \delta\} + \mathbf{1}\{s \ge \delta\}.$$

Surprisingly, the minor change in this indecisive function from its two-sided counterpart results in a big change in the corresponding K_δ, and K_δ has a continuous derivative. In what follows, we will focus on this indecisive function. Of course, the procedures developed here for this indecisive function can also be applied to other indecisive functions.

2.2 Selection of regularization parameter

2.2.1 Method based on softness

The regularization parameter δ controls the softness of a soft ROC curve. The bigger the δ is, the softer the ROC curve is. When δ is taken as zero, it becomes the traditional ROC curve as mentioned earlier. Hence, it is important to select an appropriate regularization parameter δ. First, we define the softness of a soft ROC curve as follows.

DEFINITION 2.3: *For a soft ROC curve with regularization parameter δ, the softness is defined as*

$$\alpha = 1 - \frac{P(Y - X > \delta)}{P(Y - X > 0)},$$

where $X \sim F(\cdot)$ and $Y \sim G(\cdot)$. The hardness is then naturally defined as $1 - \alpha$.

The softness α controls the smoothness of the empirical soft ROC curve estimated from (2.1). For example, if the order 1 one-sided indecisive function is used, the softness ranges from 0 (when $\delta = 0$ in which case the soft ROC curve becomes a step function) to 1 (when $\delta = \infty$ in which case the soft ROC curve becomes a diagonal line). As mentioned before, the idea of soft-thresholding is similar to the one used in designing randomization tests in hypothesis testing. In this regard, the softness defined above is analogous to significance level in the setting of randomization tests.

Figure 2.3 shows the plots of δ versus the differences of means of diseased and non-diseased populations for some choices of α. Here, we have denoted $\mu = E\{Y\} - E\{X\}$ and have assumed that the two populations are normal with unit standard deviation.

Evidently, a non-parametric estimate of softness is given by

$$\alpha = 1 - \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} I(Y_j - X_i - \delta)}{\sum_{i=1}^{m} \sum_{j=1}^{n} I(Y_j - X_i)}. \tag{2.3}$$

So, for a pre-specified α, we can choose a regularization parameter δ; but, the determination of α is quite subjective. Recall that the same issue is present in hypothesis testing wherein the significance level is usually taken to be 5%. From the simulation study we have carried out, we would suggest

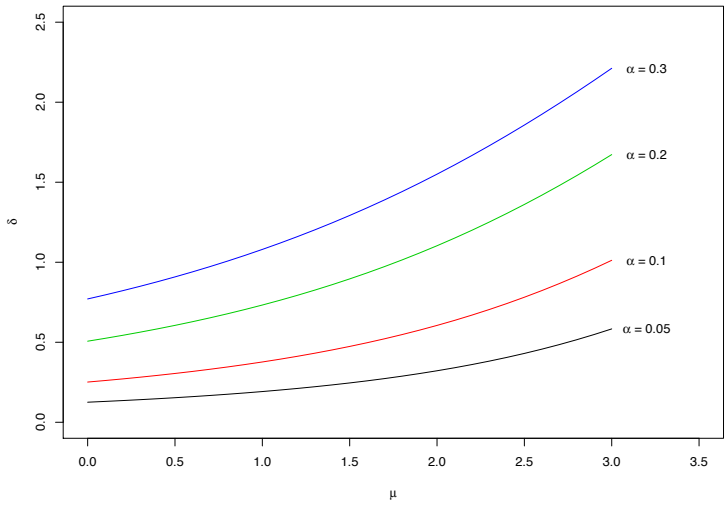


Figure 2.3. Plots of δ versus mean difference μ for some given α

considering softness between 0.1 and 0.3. In the next subsection, we propose a cross-validation procedure for selecting an appropriate δ without pre-fixing α.

2.2.2 Method based on cross-validation

In this subsection, we propose a cross-validation (CV) procedure for selecting δ by minimizing the average mean squared error (AMSE) (Ren et al., 2004),

$$\text{AMSE}(\delta) = E\left\{\frac{1}{K}\sum_{k=1}^{K}\left[\hat{R}_\delta(p_k) - R(p_k)\right]^2\right\}, \tag{2.4}$$

where p_k is in a fine grid of $(0,1)$, $k = 1, \cdots, K$.

For this purpose, we randomly split the sample into two parts, or we randomly split the diseased and non-diseased samples into two parts each. For each random split, we treat one part as a training sample and the other as a validation sample. Based on the training sample, we construct the soft ROC curve and obtain the estimate $\widehat{R}_\delta^{(1)}$, and based on the validation sample, we construct the regular ROC curve and obtain the estimate $\widehat{R}^{(2)}$. By repeating this random split many times, we

obtain the following cross-validation estimate of the AMSE:

$$\mathrm{CV}_\delta = \frac{1}{H}\frac{1}{K}\sum_{h=1}^{H}\sum_{k=1}^{K}\left[\widehat{R}_\delta^{(1,h)}(p_k) - \widehat{R}^{(2,h)}(p_k)\right]^2, \tag{2.5}$$

where H is the number of random splits. Then, δ is chosen as the one that minimizes CV_δ in (2.5).

The split ratio (training/validation) can be chosen to be either 1:1 or 2:1. From our simulation study, we observed that the results are not sensitive to the split ratio. Such an idea of cross-validation has been considered by many authors including Bickel and Levina (2008).

Chapter 3

COMBINING DIAGNOSTIC TESTS

This chapter is organized as follows. In Section 3.1, we show the uniqueness of $\boldsymbol{\beta}_0$ and the consistency of $\widehat{\boldsymbol{\beta}}$ under some mild assumptions. In Section 3.2, we propose several methods for estimating the AUC associated with $\widehat{\boldsymbol{\beta}}$. All these methods can be applied as variable selection criteria. All the technical proofs are provided in Appendix B.

3.1 Properties of $\widehat{\boldsymbol{\beta}}$

Let $\mathbf{Z} = \mathbf{Y} - \mathbf{X} = (Z_1, \cdots, Z_p)^T$ with distribution $H(\mathbf{Z})$. To examine the properties of $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$, in this section, we assume the support of H is not contained in any proper linear subspace of $\mathcal{R}^p$. We also assume the marginal density of Z_k satisfies $f_{Z_k}(Z_k = 0) \neq 0, k = 1, \cdots, p$. Pepe et al. (2006) studied the properties of $\widehat{\boldsymbol{\beta}}$ under the GLM assumption,

$$P(D = 1|\mathbf{T}) = h(\boldsymbol{\beta}_0^T \mathbf{T}), \tag{3.1}$$

where D is the indicator of the disease status with one being diseased and zero being non-diseased, and h is an increasing link function. By arguments in Appendix B.1, the GLM assumption cannot encompass the setting where $\mathbf{X}$ and $\mathbf{Y}$ are following multivariate normal distributions with different covariance matrices. In this section, we re-investigate the properties of $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$ without the GLM assumption. First of all, we show that $\boldsymbol{\beta}_0$ is unique under some mild assumption. The proof is in Appendix B.2.

PROPOSITION 3.1: *If for any $\boldsymbol{\beta} \neq \boldsymbol{\gamma} \in \mathbf{B}$,*

$$E^2(\boldsymbol{\beta}^T \mathbf{Z}|\boldsymbol{\gamma}^T \mathbf{Z} = 0) + E^2(\boldsymbol{\gamma}^T \mathbf{Z}|\boldsymbol{\beta}^T \mathbf{Z} = 0) \neq 0,$$

then $\boldsymbol{\beta}_0$ is unique.

Unfortunately, the assumption made in Proposition 3.1 cannot imply the GLM assumption. But at least it provides an alternative assumption to investigate the properties of $\boldsymbol{\beta}_0$. If $\boldsymbol{\beta}_0$ is unique, then many good asymptotic properties of $\widehat{\boldsymbol{\beta}}$ can be derived. For example, the consistency of $\widehat{\boldsymbol{\beta}}$ is stated in Proposition 3.1, with the proof in Appendix B.3.

PROPOSITION 3.2: *If $\boldsymbol{\beta}_0$ is unique, then $\widehat{\boldsymbol{\beta}} \xrightarrow{a.s.} \boldsymbol{\beta}_0$, as $m, n \longrightarrow \infty$.*

To illustrate the assumption made in Proposition 3.1, we examine the setting where $\mathbf{X}$ and $\mathbf{Y}$ follow multivariate normal distributions with different covariance matrices respectively. In this setting, for any given $\boldsymbol{\beta} \neq \boldsymbol{\gamma}$, $V^{(1)} = \boldsymbol{\beta}^T\mathbf{Z} \sim N(\mu_1, \sigma_1^2)$ and $V^{(2)} = \boldsymbol{\gamma}^T\mathbf{Z} \sim N(\mu_2, \sigma_2^2)$. Then $Corr(V^{(1)}, V^{(2)}) = \rho \neq 1$, and $V^{(1)}$ and $V^{(2)}$ can be expressed as $V^{(2)} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(V^{(1)} - \mu_1) + \epsilon$ and $V^{(1)} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(V^{(2)} - \mu_2) + e$, where $\epsilon \sim N(0, (1-\rho^2)\sigma_2^2)$ and $e \sim N(0, (1-\rho^2)\sigma_1^2)$. If $E(V^{(2)}|V^{(1)} = 0) = 0$ and $E(V^{(1)}|V^{(2)} = 0) = 0$, then $\frac{\mu_2}{\sigma_2} = \rho\frac{\mu_1}{\sigma_1}$ and $\frac{\mu_1}{\sigma_1} = \rho\frac{\mu_2}{\sigma_2}$, and then $\rho = 1$, which leads to a contradiction. Therefore, in this setting, the assumption made in Proposition 3.1 is satisfied.

3.2 Estimates of the AUC associated with $\widehat{\boldsymbol{\beta}}$

This section provides several methods for estimating the AUC associated with $\widehat{\boldsymbol{\beta}}$, AUC($\widehat{\boldsymbol{\beta}}$). Note that AUC($\widehat{\boldsymbol{\beta}}$) is the counterpart of the prediction error in linear regression, and therefore the apparent estimate, $\overline{\text{AUC}}(\widehat{\boldsymbol{\beta}})$, which is the counterpart of the training error in linear regression, overestimates it.

3.2.1 Cross-validation

As pointed out by Hastie et al. (2009, p.241), the simplest and most widely used method for estimating the prediction error is cross-validation. Therefore, we consider cross-validation method

for estimating AUC($\widehat{\boldsymbol{\beta}}$) first. The leave-one-pair-out cross-validation estimate is

$$\mathrm{AUC}^{(\mathrm{CV})} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I(\widehat{\boldsymbol{\beta}}^{(-ij)T} \mathbf{Y}_j > \widehat{\boldsymbol{\beta}}^{(-ij)T} \mathbf{X}_i), \tag{3.2}$$

where $\widehat{\boldsymbol{\beta}}^{(-ij)}$ is the solution to procedure (1.2) without pair (i, j).

If many diagnostic tests are available and some of them are redundant, we might want to seek a subset of diagnostic tests based on which the combined test has the largest AUC. This becomes a variable selection problem, because including more redundant diagnostic tests leads to finding a combination with which the AUC is farther away from the maximum AUC, AUC($\boldsymbol{\beta}_0$). The variable selection based on $\mathrm{AUC}^{(\mathrm{CV})}$ can be done as follows. At each subset of diagnostic tests (there are $2^p - 1$ possible subsets), $\mathrm{AUC}^{(\mathrm{CV})}$ is calculated, and then the subset with the largest $\mathrm{AUC}^{(\mathrm{CV})}$ is selected as the "best" subset.

For the purpose of variable selection, alternatives such as the five-fold cross-validation and the ten-fold cross-validation can be applied in stead of the leave-one-pair-out cross-validation; e.g., Hastie et al. (2009, p.242). These alternatives are simpler and more efficient. However, in this paper, estimating the AUC associated with $\widehat{\boldsymbol{\beta}}$ is the main concern, and the leave-one-out cross-validation is a nearly unbiased estimate.

3.2.2 Bootstrap methods

The cross-validation estimate is nearly unbiased, but has relatively large variance, because of the discontinuity of the indicator function in (3.2). In order to reduce the variance of the cross-validation, we can apply the bootstrap smoothing introduced by Efron (1983). Let $\mathbf{X}^{*(i)b}$ be the bth bootstrap sample from the empirical distribution on $\mathbf{X}^{(i)}$, the training set of $\mathbf{X}$ without the ith observation, and $\mathbf{Y}^{*(j)b}$ be the bth bootstrap sample from the empirical distribution on $\mathbf{Y}^{(j)}$, the training set of $\mathbf{Y}$ without the jth observation. Then the leave-one-pair-out bootstrap cross-

validation estimate of AUC($\widehat{\boldsymbol{\beta}}$) is defined as

$$\mathrm{AUC}^{(\mathrm{BT})} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{B}\sum_{b=1}^{B} I(\widehat{\boldsymbol{\beta}}^{*b(-ij)T}\mathbf{Y}_j > \widehat{\boldsymbol{\beta}}^{*b(-ij)T}\mathbf{X}_i), \tag{3.3}$$

where $\widehat{\boldsymbol{\beta}}^{*b(-ij)}$ is the solution to procedure (1.2) based on the bth bootstrap samples from $\mathbf{X}^{(i)}$ and $\mathbf{Y}^{(j)}$.

As discussed in Efron (1983), the bootstrap cross-validation estimate has the training-set-size bias. Because the average number of distinct observations in each bootstrap sample is about $0.632(m+n)$, the "0.632 estimate" proposed in Efron (1983) is designed to alleviate this bias. It is defined as

$$\mathrm{AUC}^{(.632)} = 0.368\overline{\mathrm{AUC}} + 0.632\mathrm{AUC}^{(\mathrm{BT})}. \tag{3.4}$$

The 0.632 estimate works well in light fitting situations, but can break down in overfit ones (Breiman et al., 1984). To improve the 0.632 estimate, Efron and Tibshirani (1997) proposed the "0.632+ estimate", defined by

$$\mathrm{AUC}^{(.632+)} = (1-\widehat{\omega})\overline{\mathrm{AUC}} + \widehat{\omega}\mathrm{AUC}^{(\mathrm{BT})}, \tag{3.5}$$

where $\widehat{\omega} = \frac{0.632}{1-0.368\widehat{R}}$. Here $\widehat{R}$ is called the relative overfitting rate, and in this framework,

$$\widehat{R} = \frac{\overline{\mathrm{AUC}} - \mathrm{AUC}^{(\mathrm{BT})}}{\overline{\mathrm{AUC}} - \gamma},$$

where γ is called the no-information error (i.e. the prediction rate if the input and class labels were independent), and in this setting, $\gamma = 0.5$.

3.2.3 Sigmoid function smoothing

Another way to reduce the variance of the cross-validation estimate is by smoothing the indicator function in (3.2). As Ma and Huang (2005; 2007), a reasonable choice of smoothing function is sigmoid function, $g_\lambda(u) = \frac{1}{1+e^{-\lambda u}}$, where λ is large enough. A by-product of this smoothing function is the computational convenience. Let $\mathbf{Z}_{ij} = \mathbf{Y}_j - \mathbf{X}_i$, and

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\max_{\boldsymbol{\beta}\in\mathbf{B}} \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} g_\lambda(\boldsymbol{\beta}^T \mathbf{Z}_{ij}). \tag{3.6}$$

Ma and Huang (2007) developed the asymptotic properties of $\widehat{\boldsymbol{\beta}}_\lambda$ under the GLM assumption. We can also verify the results under the assumption made in Section 3.1.

In this subsection, we are interested in estimating the AUC associated with $\widehat{\boldsymbol{\beta}}_\lambda$,

$$\mathrm{AUC}(\widehat{\boldsymbol{\beta}}_\lambda) = P(\widehat{\boldsymbol{\beta}}_\lambda^T \mathbf{Y} > \widehat{\boldsymbol{\beta}}_\lambda^T \mathbf{X}). \tag{3.7}$$

The cross-validation estimate of $\mathrm{AUC}(\widehat{\boldsymbol{\beta}}_\lambda)$ is

$$\mathrm{AUC}_\lambda^{(\mathrm{CV})} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I(\widehat{\boldsymbol{\beta}}_\lambda^{(-ij)T} \mathbf{Y}_j > \widehat{\boldsymbol{\beta}}_\lambda^{(-ij)T} \mathbf{X}_i), \tag{3.8}$$

where $\widehat{\boldsymbol{\beta}}_\lambda^{(-ij)}$ is the solution to procedure (3.6) without pair (i, j).

For any given λ, $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$ is a nearly unbiased estimate of $\mathrm{AUC}(\widehat{\boldsymbol{\beta}}_\lambda)$. Smaller λ results in bigger bias but smaller variance of $\widehat{\boldsymbol{\beta}}_\lambda$ as an estimate of $\boldsymbol{\beta}_0$. An appropriate λ can be decided at the full set of diagnostic tests as follows. Over a fine grid of λ, $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$'s are calculated, and then the value of λ achieving the largest $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$ is selected. As the simulations show, the curve of $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$ increases as λ increases, achieves its peak at some λ, and then follows a slow turn down. Ma and Huang (2005) mentioned a rule of thumb for choosing the tuning parameter, and found that the results are not sensitive to λ as long as λ is large enough.

3.2.4 Approximated cross-validation for variable selection

For the purpose of variable selection, at each subset of diagnostic tests, $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$ is calculated, and then the subset achieving the largest $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$ is selected as the "best" subset. This method is straightforward, but the computation is intensive. For example, to find the "best" subset from p tests, we need to conduct the maximization procedure $(2^p-1)mn$ times. To reduce the computation cost, we develop an approximated cross-validation method. By this method, at each subset, we only need to perform the maximization procedure once.

First, to speed up the above variable selection process, we can approximate $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$ by

$$\widehat{\mathrm{AUC}}_\lambda^{(\mathrm{CV})} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} g_\lambda(\widehat{\boldsymbol{\beta}}_\lambda^{(-ij)T}\mathbf{Z}_{ij}), \tag{3.9}$$

Then, by Taylor expansion, the above expression can be further approximated by

$$\mathrm{AUC}_\lambda^{(\mathrm{ACV})} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} g_\lambda(\widehat{\boldsymbol{\beta}}_\lambda^{T}\mathbf{Z}_{ij}) + \mathrm{trace}(\widehat{L}_2^{-1}\widehat{L}_1), \tag{3.10}$$

where

$$\widehat{L}_1 = \frac{1}{m^2}\sum_{i=1}^{m}\bar{D}_{i\cdot}\bar{D}_{i\cdot}^T + \frac{1}{n^2}\sum_{j=1}^{n}\bar{D}_{\cdot j}\bar{D}_{\cdot j}^T - \frac{1}{m^2n^2}\sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij}D_{ij}^T,$$

$$\widehat{L}_2 = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{Z}_{ij}\ddot{g}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda^{T}\mathbf{Z}_{ij})\mathbf{Z}_{ij}^T,$$

with $\dot{g}_\lambda(u)$ and $\ddot{g}_\lambda(u)$ being the first two derivatives of $g_\lambda(u)$, $D_{ij} = \mathbf{Z}_{ij}\dot{g}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda^{T}\mathbf{Z}_{ij})$, $\bar{D}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} D_{ij}$, and $\bar{D}_{\cdot j} = \frac{1}{m}\sum_{i=1}^{m} D_{ij}$.

The derivation of $\mathrm{AUC}_\lambda^{(\mathrm{ACV})}$ follows the arguments in Stone (1977); see Appendix B.4 for details. Actually, the expression of $\mathrm{AUC}_\lambda^{(\mathrm{ACV})}$ is similar to the Takeuchi Information Criterion (TIC) proposed by Takeuchi (1976). For programming, it might be interesting to note that $\dot{g}_\lambda(u)$ can be expressed as $\lambda g_\lambda(u)[1-g_\lambda(u)]$.

3.2.5 Comparison with some LASSO-type procedures

Many methods for variable selection are proposed in the literature of classification; among them, the LASSO-type procedures are the most popular. However, for classification, the variable selection procedures are examined based on the misclassification rate, rather than on the compromise between the sensitivity and the specificity through the ROC curve. We propose several variable selection procedures with the focus on the AUC. In the three real examples analyzed in the numerical studies section, we compare our procedures, which are based on the AUC, with the sLDA and the logistic-LASSO, but not with the DALASS because of its instability (personal communication).

In the next chapter, we will work on directly applying the LASSO constrain on (1.2), and propose an AUC-LASSO method using the soft ROC approximation.

Chapter 4

AUC-LASSO

This chapter is organized as follows. The motivation is discussed in Section 4.1. A detailed review of related computational methods is presented in Section 4.2. And the main algorithm is proposed in Section 4.3.

4.1 Motivation

The approximated cross-validation method proposed in Chapter 3 successfully adjusted the upward bias of apparent AUC estimate, and can be used as a criteria to select important diagnostic tests. It dramatically reduces the computation compared to the cross-validation procedure. However, the approximated cross-validation method is still based on the best subset selection procedure. And it becomes impractical as the number of diagnostic tests increases. The LASSO-type procedure are the most popular methods for variable selection, performed by regularizing the coefficient via L1 norm. Because the nature of the L1 constrain, it tends to produce some coefficients to be 0 and hence produce a "continuous" variable selection procedure. Many LASSO-type procedures are proposed in the literature of classification. For instance, Friedman et al. (2010) incorporated the coordinate descent algorithm with logistic regression (Logistic-LASSO); Trendafilov and Jolliffe (2007) proposed the discriminant analysis via the LASSO (DALASS); and Wu et al. (2009) proposed the sparse linear discriminant analysis (sLDA). However, for classification, the variable selection procedures are examined based on the misclassification rate, rather than on the compromise between the sensitivity and the specificity through the ROC curve. We are concerned with

developing algorithms to solve the AUC-LASSO problem:

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbf{B}} \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I(\boldsymbol{\beta}^T \mathbf{Y}_j > \boldsymbol{\beta}^T \mathbf{X}_i) + \rho \sum_{j=1}^{p} |\beta_i|. \tag{4.1}$$

The challenge of this problem is that the objective function is not convex, and there is generally no efficient method to compute the global optima for high-dimensional non-convex optimization problem. Note that the constrain of coefficient $\boldsymbol{\beta}$ on the unit circle in (4.1) is mainly due to the identifiability. To achieve the same purpose, instead of restricting $\boldsymbol{\beta}$ on the unit circle, similar to Ma and Huang (2005), we assume the first diagnostic test is the anchor, and set $\beta_{(1)} = 1$. When dealing with the AUC-LASSO problem, we still use $\boldsymbol{\beta}$ to denote the coefficients $(1, \beta_{(2)}, \ldots, \beta_{(p)})^T$. Ma and Huang (2005) proposed a method of how to choose the anchor diagnostic test via the adjusted t-statistic. Thus, our problem becomes

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I(\boldsymbol{\beta}^T \mathbf{Y}_j > \boldsymbol{\beta}^T \mathbf{X}_i) + \rho \sum_{j=1}^{p} |\beta_i|. \tag{4.2}$$

However, the objective function in (4.2) is still non-convex. To develop an efficient algorithm, we employ the idea of the soft ROC curves to approximate the empirical objective function. To find an approximate solution for the non-convex problem, we first decompose the approximated objective function into a difference of two convex functions, then apply the difference convex programming (An and Tao, 1997) to find the global optima.

4.2 Review of related computational methods

4.2.1 Penalized likelihood methods and coordinate descent algorithms

Penalized likelihood methods are among the most popular approaches for automatically and simultaneously variable selection. However, optimizing the penalized likelihood can be a challenging task. The penalized optimization problems include ridge regression (Hoerl and Kennard, 1970),

bridge regression (Frank and Friedman, 1993), LASSO (Tibshirani, 1996), SVM (Vapnik, 1998), SCAD (Fan and Li, 2001), group LASSO (Yuan and Lin, 2006) and elastic net (Zou and Hastie, 2005). For LASSO, Efron et al. (2004) exploited the fact that the coefficient profiles are piecewise linear, and proposed the LARS algorithm for computing the entire solution path for the linear regression models, which attains the same computational cost as ordinary least-square fit. Furthermore, to characterize the class of problems where piecewise linear solution path exists, Rosset and Zhu (2007) pointed out that the objective function has to be piecewise quadratic and the penalty has to be piecewise linear.

Friedman et al. (2007) utilized the cyclical coordinate descent methods as an alternative to produce the entire solution path without verifying the existence of piecewise linearity. And Friedman et al. (2010) extended this algorithm to the generalized linear models with elastic-net penalties. This algorithm solves the solutions along an entire path of values for the regularization parameters, while the current estimates are used as warm starts. The coordinate descent algorithm is proven to be efficient for high dimensional problems. To implement the coordinate descent algorithms, one only needs to iteratively solve a sequence of univariate problems (the coefficients except the current one are fixed) with "partial residuals" as the response variable until convergence of all the estimates.

4.2.2 Difference convex programming

An and Tao (1997) studied the computation of global optima when an objective function $h(\omega)$ has a Differenced Convex (DC) representation: $h(\omega) = h_1(\omega) - h_2(\omega)$, where $h_1(\omega)$ and $h_2(\omega)$ are convex functions. Liu et al. (2005) developed computational tools of ψ-learning by first decomposing the non-convex objective function into difference of two convex functions, then applying the DC algorithm to solve the global optimization problem. Li and Yu (2009) utilized the DC algorithm and coordinate descent algorithms to solve the solution path for robust and sparse bridge regression. To implement the DC programming, one needs to construct a sequence of subproblems, which replacing $h_2(\omega)$ by its affine minorization function $h_2(\omega^0) + < \omega - \omega^{(0)}, \nabla h_2(\omega^{(0)}) >$ and solve them iteratively,

where $\nabla h_2(\omega^{(0)})$ is the subgradient of $h_2(\omega)$ at $\omega^{(0)}$ with respect to ω, and $< \cdot, \cdot >$ is the notation for inner product. Thus, after some calculation, given the solution for the $(i-1)$th subproblem, the ith subproblem solves

$$\omega^{(i)} = \arg\max_{\omega} h_1(\omega) - < \omega, \nabla h_2(\omega^{(i-1)}) > .$$

4.3 AUC-LASSO

For our problem, in order to decompose the objective function, we incorporate the order 1 one-sided indecisive function proposed in chapter 1 to approximate the empirical AUC function in (4.2), and the problem becomes

$$\widehat{\boldsymbol{\beta}}_\delta = \arg\max_{\boldsymbol{\beta}} \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} K_\delta(\boldsymbol{\beta}^T \mathbf{Y}_j > \boldsymbol{\beta}^T \mathbf{X}_i) + \rho \sum_{j=1}^{p} |\beta_j|, \tag{4.3}$$

where

$$K_\delta(s) = \left[\frac{1}{2} + \frac{s}{\delta} - \text{sign}(s)\frac{s^2}{2\delta^2}\right] I(-\delta \le s < \delta) + I(s \ge \delta),$$

and $\delta > 0$ is a regularization parameter, which can be selected by a cross-validation procedure as proposed in chapter 1. It can be shown that, $K_\delta(s)$ has the following DC representation:

$$K_\delta(s) = K_\delta^{(1)}(s) - K_\delta^{(2)}(s),$$

where

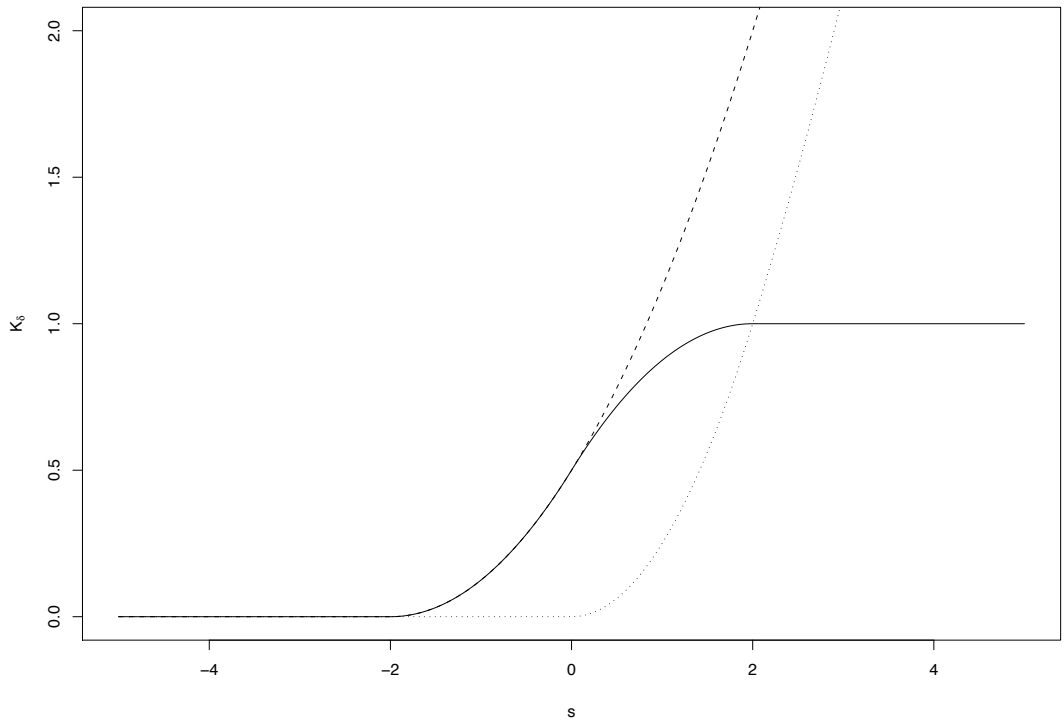
$$K_\delta^{(1)}(s) = (\frac{1}{2} + \frac{t}{\delta} + \frac{t^2}{2\delta^2}) I(t \ge -\delta),$$

and

$$K_\delta^{(2)}(s) = \frac{t^2}{\delta^2} I(0 \le t < \delta) + (-\frac{1}{2} + \frac{t}{\delta} + \frac{t^2}{2\delta^2}) I(t \ge \delta).$$

Figure 4.1 shows the plot of $K_\delta^{(1)}(s)$, $K_\delta^{(2)}(s)$ and their difference $K_\delta(s)$.

Figure 4.1. The DC representation for one-sided order 1 K_δ



Therefore, the problem (4.3) can be decomposed into:

$$\widehat{\boldsymbol{\beta}}_\delta = \arg\max_{\boldsymbol{\beta}} S_1(\boldsymbol{\beta}) - S_2(\boldsymbol{\beta}), \tag{4.4}$$

where

$$S_1(\boldsymbol{\beta}) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} K_\delta^{(1)}(\boldsymbol{\beta}^T\mathbf{Z}_{ij}) + \rho\sum_{j=1}^{p}|\beta_j|,$$

$$S_2(\boldsymbol{\beta}) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} K_\delta^{(2)}(\boldsymbol{\beta}^T\mathbf{Z}_{ij}),$$

and $\mathbf{Z}_{ij} = \mathbf{Y}_{ij} - \mathbf{X}_{ij}$. Thus, to solve the problem (4.3) is equivalent to solve the DC problem (4.4).

With the aforementioned DC representation, the ith subproblem of DC algorithm in this problem is

$$\boldsymbol{\beta}^{(i)} = \arg\max_{\boldsymbol{\beta}} S_1(\boldsymbol{\beta}) - < \boldsymbol{\beta}, \nabla S_2(\boldsymbol{\beta}^{(i-1)}) >, \tag{4.5}$$

and the subgradient of $S_2(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ at $\boldsymbol{\beta}^{(0)}$ is

$$\begin{aligned} \nabla S_2(\boldsymbol{\beta}^{(0)}) &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{2\mathbf{Z}_{ij}\mathbf{Z}_{ij}^T\boldsymbol{\beta}}{\delta^2} I(0 \le \boldsymbol{\beta}^T\mathbf{Z}_{ij} < \delta) \\ &\quad + \left(\frac{\mathbf{Z}_{ij}}{\delta} + \frac{\mathbf{Z}_{ij}\mathbf{Z}_{ij}^T\boldsymbol{\beta}}{\delta^2}\right) I(\boldsymbol{\beta}^T\mathbf{Z}_{ij} \ge \delta). \end{aligned}$$

Thus, for a given ρ, the algorithm can be implemented such that in each iteration of the DC programming, the coordinate descent algorithm is applied to solve (4.5), until the convergence of the estimate. Similar to Friedman et al. (2007), this algorithm can be repeatedly used to solve the solutions of (4.5) along an entire path of values for the regularization parameters, while the current estimates are used as warm starts.

Chapter 5

NUMERICAL STUDIES

5.1 Numerical studies for soft ROC curves

In this section, the proposed methods in Chapter 2 are examined through a Monte Carlo simulation study and a real example.

5.1.1 Simulation study

First, in order to examine the method of choosing δ with pre-specified softness α, 500 datasets were generated from normal distributions with unit standard deviation and mean difference between two populations being $\mu_y - \mu_x = 1.5$. In each dataset, there were $n = 100$ diseased subjects and $m = 100$ non-diseased subjects, and $\widehat{\delta}$ was calculated from (2.3) with $\alpha = 0.05$, 0.1, or 0.2. Then, we compared the two square errors, $(\widehat{\mathrm{AUC}}_{\widehat{\delta}} - \mathrm{AUC}_0)^2$ and $(\widehat{\mathrm{AUC}}_0 - \mathrm{AUC}_0)^2$, where $\widehat{\mathrm{AUC}}_{\widehat{\delta}}$ is calculated from (2.2) with $\widehat{\delta}$, AUC_0 is the true AUC, and $\widehat{\mathrm{AUC}}_0$ is the hard-thresholding empirical estimate of AUC_0. Figure 5.1 shows side-by-side box-plots of these two square errors, and we observe that $\widehat{\mathrm{AUC}}_{\delta}$ has smaller MSE than $\widehat{\mathrm{AUC}}_0$, especially for the choice of $\alpha = 0.2$.

Next, we investigate the performance of the CV procedure through two simulation studies. In the first study, data were generated from two normal distributions with unit standard deviation and means being one of the following cases: (i) $(\mu_y, \mu_x) = (1, 0)$, (ii) $(\mu_y, \mu_x) = (1.5, 0)$, (iii) $(\mu_y, \mu_x) = (2, 0)$, and (iv) $(\mu_y, \mu_x) = (2.5, 0)$. The sample sizes were taken as $m = n = 50$ or $m = n = 100$, and the split ratio was set as 2:1. In the second study, settings were the same except that the data were generated from two double exponential distributions. For each simulation study,

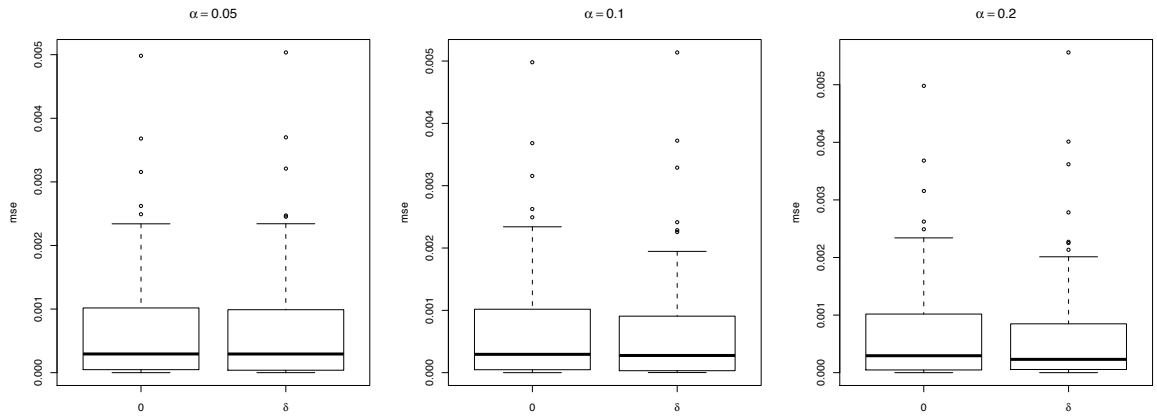


Figure 5.1. Comparison of mean square errors of AUC_δ and AUC_0

300 replications were performed to calculate the efficiency measure

$$\frac{\mathrm{AMSE}(\widehat{\delta}^{cv})}{\mathrm{AMSE}(0)}, \tag{5.1}$$

and the efficacy measure,

$$\frac{\mathrm{AMSE}(\widehat{\delta}^{cv})}{\min_\delta \mathrm{AMSE}(\delta)}, \tag{5.2}$$

where $\widehat{\delta}^{cv}$ is the δ chosen by the CV procedure. The simulation results so obtained are summarized in Table 5.1. These results show that all efficiencies are less than 1, while efficacies are all close to 1, which indicates the optimality of $\widehat{\delta}^{cv}$. From this table, we also observe that $\widehat{\delta}^{cv}$ is decreasing when the difference $\mu_y - \mu_x$ increases. In fact, when Y and X are well-distinguished, the indecisive interval vanishes.

5.1.2 Pancreatic cancer serum biomarkers example

The dataset comes from a case-control study at Mayo Clinic which included 90 patients with pancreatic cancer and 51 subjects with pancreatitis. These data were originally analyzed by Wieand et al. (1989). Two continuous positive scale serum biomarkers were available to diagnose a patient

Table 5.1. Performance of CV

		$m = n = 50$			$m = n = 100$		
Distribution	(μ_y, μ_x)	Efficiency	Efficacy	$\widehat{\delta}^{cv}$	Efficiency	Efficacy	$\widehat{\delta}^{cv}$
Normal	(1, 0)	0.9611	1.1599	1.670	0.9607	1.0972	1.628
	(1.5, 0)	0.9792	1.1875	1.172	0.9703	1.0903	1.030
	(2, 0)	0.9956	1.1224	0.676	0.9852	1.0453	0.648
	(2.5, 0)	0.9956	1.0781	0.414	0.9472	1.0526	0.364
Double	(1, 0)	0.9601	1.1905	1.490	0.9597	1.1895	1.404
Exponential	(1.5, 0)	0.9743	1.1579	1.437	0.9571	1.1012	1.320
	(2, 0)	0.9687	1.1761	1.263	0.9474	1.1080	1.184
	(2.5, 0)	0.9895	1.1017	1.098	0.9768	1.1005	1.013

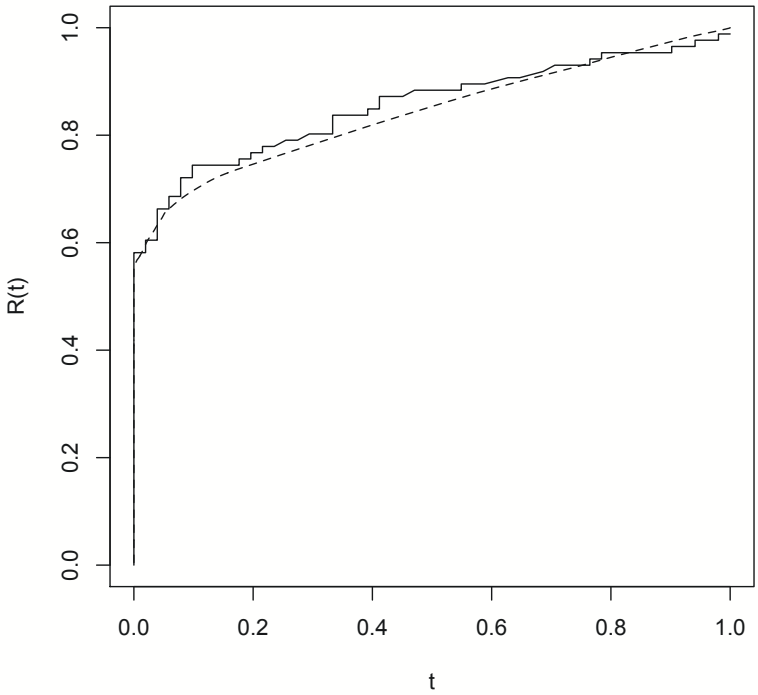
with pancreatic cancer: CA-125, a cancer antigen, and CA-19-9, a carbohydrate antigen. We applied the CV method to select regularization parameters for CA-125 and CA-19-9, which turn out to be 0.04 and 0.115, respectively. The corresponding ROC and soft ROC curves are displayed in Figure 5.2 and Figure 5.3. From these figures, we observe that the AUC is considerably higher for CA-125.

5.2 Numerical studies for optimal combinations of diagnostic tests

In this section, the proposed methods in Chapter 3 are illustrated by some Monte Carlo simulation studies and applications to three real examples.

5.2.1 Simulation studies

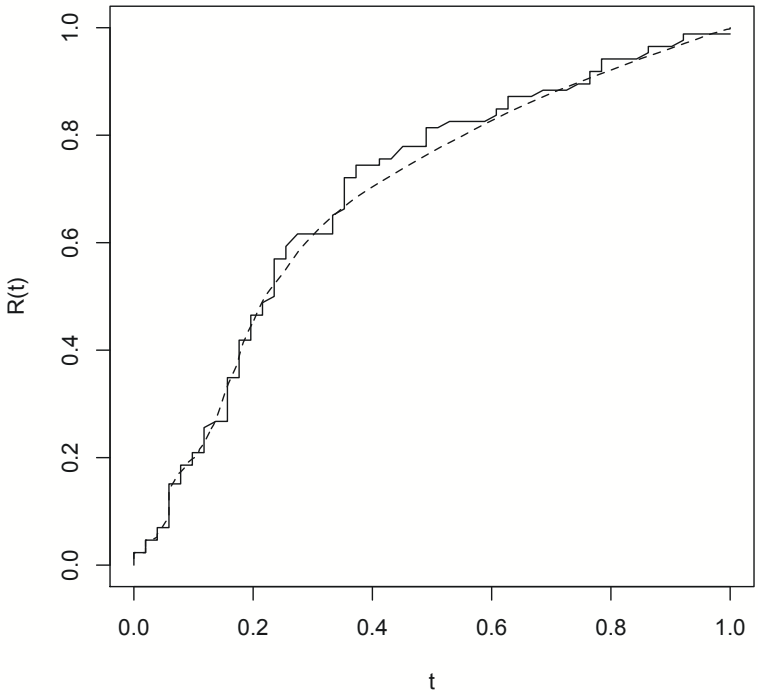
First, we examine the performances of different methods, $\overline{\mathrm{AUC}}(\widehat{\boldsymbol{\beta}})$, $\mathrm{AUC}^{(\mathrm{CV})}$, $\mathrm{AUC}^{(\mathrm{BT})}$, $\mathrm{AUC}^{(.632)}$, and $\mathrm{AUC}^{(.632+)}$ as estimates of the AUC associated $\widehat{\boldsymbol{\beta}}$, $\mathrm{AUC}(\widehat{\boldsymbol{\beta}})$, under two groups of settings. Assume that in each simulated sample there are $m = 50$ non-diseased subjects and $n = 50$ diseased subjects, and there are $p = 2, 3$, or 4 diagnostic tests. Denote the means of testing outcomes in non-diseased and diseased, respectively, by μ_x and μ_y. In the first group of settings, testing



Solid line: empirical ROC curve; Dotted line: soft ROC curve.

Figure 5.2. ROC curves for Pancreatic Cancer Serum Biomarkers Example: CA-125

outcomes are generated from two multivariate normal distributions with means being one of the following three cases: (i) $p = 2$, $\mu_x = (0, 0)^T$, $\mu_y = (1, 1)^T$; (ii) $p = 3$, $\mu_x = (0, 0, 0)^T$, $\mu_y = (1, 1, 0)^T$; or (iii) $p = 4$, $\mu_x = (0, 0, 0, 0)^T$, $\mu_y = (1, 1, 0, 0)^T$. The variance of each testing outcome is set as $\sigma^2 = 1$ and the correlation between any two testing outcomes is set as $\rho = 0.3$. In the second group of settings, testing outcomes are generated from two multivariate exponential distributions with means being one of the following three cases: (i) $p = 2$, $\mu_x = (1, 1)^T$, $\mu_y = (5, 5)^T$; (ii) $p = 3$, $\mu_x = (1, 1, 1)^T$, $\mu_y = (5, 5, 1)^T$; (iii) $p = 4$, $\mu_x = (1, 1, 1, 1)^T$, $\mu_y = (5, 5, 1, 1)^T$. To generate multivariate exponential random variables, we use the generator proposed by Marshall et al. (1967) and the correlation between any two testing outcomes is set as $\rho = 0.3$. For each setting, 500 repetitions are generated.



Solid line: empirical ROC curve; Dotted line: soft ROC curve.

Figure 5.3. ROC curves for Pancreatic Cancer Serum Biomarkers Example: CA-19-9

The above simulation results are summarized in Table 5.2. For each method, the average, the variance, and the mean square error over 500 repetitions are reported. We find that $\overline{\mathrm{AUC}}(\widehat{\boldsymbol{\beta}})$ is always biased upward. Among those bias-adjusted estimates, $\mathrm{AUC}^{(\mathrm{BT})}$, $\mathrm{AUC}^{(.632)}$, and $\mathrm{AUC}^{(.632+)}$ have relatively lower variance. However, $\mathrm{AUC}^{(.632)}$ and $\mathrm{AUC}^{(.632+)}$ are slightly biased upward. By contrast, $\mathrm{AUC}^{(\mathrm{CV})}$ and $\mathrm{AUC}^{(\mathrm{BT})}$ are almost unbiased. Although the $\mathrm{AUC}^{(\mathrm{BT})}$ is of slightly lower variance, because of its computational intensity, we advocate that $\mathrm{AUC}^{(\mathrm{CV})}$ is a good estimate of $\mathrm{AUC}(\widehat{\boldsymbol{\beta}})$.

Secondly, we provide an example on how to choose the tuning parameter λ in $\mathrm{AUC}_{\lambda}^{(\mathrm{CV})}$. Assume that there are $m = 50$ non-diseased subjects and $n = 50$ diseased subjects, and $p = 6$ diagnostic tests generated from two multivariate normal distributions with means being $\mu_x = (0, 0, 0, 0, 0, 0)^T$

Table 5.2. Different methods for estimating $\mathrm{AUC}(\widehat{\boldsymbol{\beta}})$

		Normal			Exponential		
Setting	Method	AVE	VAR	MSE	AVE	VAR	MSE
$p = 2$	$\mathrm{AUC}(\widehat{\boldsymbol{\beta}})$	0.8059	–	–	0.9055	–	–
	$\overline{\mathrm{AUC}}(\widehat{\boldsymbol{\beta}})$	0.8207	0.0025	0.0027	0.9224	0.0010	0.0013
	$\mathrm{AUC}^{(\mathrm{CV})}$	0.8095	0.0027	0.0027	0.9163	0.0010	0.0012
	$\mathrm{AUC}^{(\mathrm{BT})}$	0.8131	0.0025	0.0025	0.9158	0.0010	0.0011
	$\mathrm{AUC}^{(.632)}$	0.8159	0.0025	0.0026	0.9182	0.0010	0.0012
	$\mathrm{AUC}^{(.632+)}$	0.8159	0.0025	0.0026	0.9182	0.0010	0.0012
$p = 3$	$\mathrm{AUC}(\widehat{\boldsymbol{\beta}})$	0.8034	–	–	0.8987	–	–
	$\overline{\mathrm{AUC}}(\widehat{\boldsymbol{\beta}})$	0.8249	0.0026	0.0042	0.9276	0.0015	0.0024
	$\mathrm{AUC}^{(\mathrm{CV})}$	0.8028	0.0029	0.0033	0.9127	0.0018	0.0021
	$\mathrm{AUC}^{(\mathrm{BT})}$	0.8082	0.0027	0.0033	0.9135	0.0017	0.0020
	$\mathrm{AUC}^{(.632)}$	0.8143	0.0026	0.0036	0.9187	0.0016	0.0021
	$\mathrm{AUC}^{(.632+)}$	0.8141	0.0026	0.0036	0.9186	0.0016	0.0021
$p = 4$	$\mathrm{AUC}(\widehat{\boldsymbol{\beta}})$	0.8021	–	–	0.8968	–	–
	$\overline{\mathrm{AUC}}(\widehat{\boldsymbol{\beta}})$	0.8249	0.0027	0.0045	0.9293	0.0011	0.0022
	$\mathrm{AUC}^{(\mathrm{CV})}$	0.8019	0.0027	0.0029	0.9044	0.0016	0.0016
	$\mathrm{AUC}^{(\mathrm{BT})}$	0.8059	0.0025	0.0030	0.9079	0.0012	0.0013
	$\mathrm{AUC}^{(.632)}$	0.8129	0.0025	0.0035	0.9158	0.0011	0.0015
	$\mathrm{AUC}^{(.632+)}$	0.8127	0.0025	0.0035	0.9155	0.0011	0.0015

and $\mu_y = (1.5, 1.5, 1.5, 0, 0, 0)^T$. The variance of each testing outcome is set as $\sigma^2 = 1$ and the correlation between any two outcomes is set as $\rho = 0.3$. To obtain $\widehat{\boldsymbol{\beta}}_\lambda$ and $\widehat{\boldsymbol{\beta}}_\lambda^{(-ij)}$ we apply the Lagrange multipliers and Quasi-Newton method. Figure 5.4 shows the plot of $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$ against the choice of λ. As λ increases, the curve increases, achieves its peak around $\lambda = 8$, and then follows a slow turn down. We examine many other settings; because the results are similar, they are not reported here.

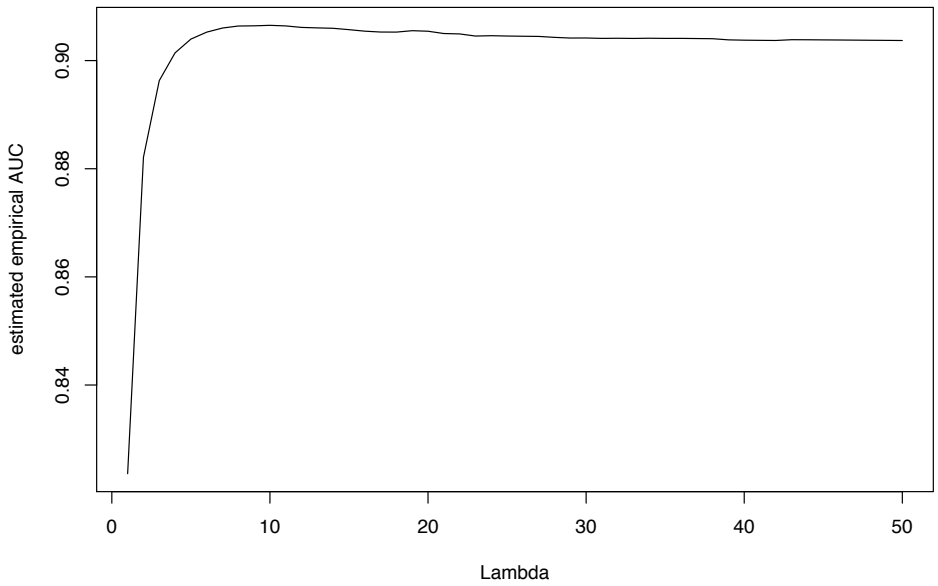
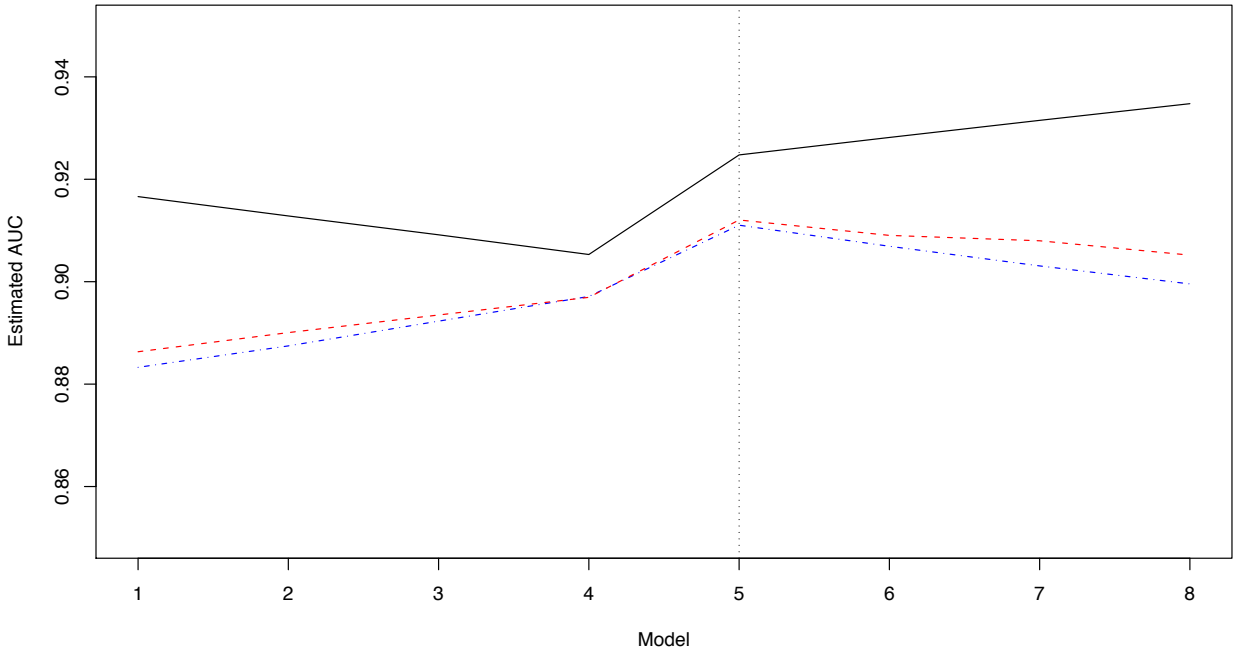


Figure 5.4. Plot of $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$ versus λ

Thirdly, we investigate the performances of $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$ and $\mathrm{AUC}_\lambda^{(\mathrm{ACV})}$ as objective functions for variable selection. We consider the same example in the previous paragraph and generate 500 repetitions. We use $\lambda = 8$ and consider their performances at the following eight subsets: $S_1 = \{1, 2, 4, 5, 6\}$, $S_2 = \{1, 2, 4, 5\}$, $S_3 = \{1, 2, 4\}$, $S_4 = \{1, 2\}$, $S_5 = \{1, 2, 3\}$, $S_6 = \{1, 2, 3, 4\}$, $S_7 = \{1, 2, 3, 4, 5\}$, and $S_8 = \{1, 2, 3, 4, 5, 6\}$. In particular, $S_5 = \{1, 2, 3\}$ is the true best subset, the one including only those important tests. The results are summarized in Figure 5.5. We find

that $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$ and $\mathrm{AUC}_\lambda^{(\mathrm{ACV})}$ are very close to each other, and both of them achieve their maximum at the true best subset S_5. Therefore, both methods can be used as variable selection criteria. However, the apparent estimate by re-substitution, $\overline{\mathrm{AUC}}(\widehat{\boldsymbol{\beta}}_\lambda)$, always increases when more tests are included, and therefore it cannot be used as a variable selection criterion.



Solid line: $\overline{\mathrm{AUC}}(\widehat{\boldsymbol{\beta}})$; Dashed line: $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$; Dotted line: $\mathrm{AUC}_\lambda^{(\mathrm{ACV})}$

Figure 5.5. Plot of estimated AUC's versus the subsets

Lastly, to further examine the performances of $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$ and $\mathrm{AUC}_\lambda^{(\mathrm{ACV})}$ as variable selection criteria to select the true best subset, we consider the same example in the previous paragraph and generate 500 repetitions. The selected subsets are divided into the following four categories: (i) selecting correctly the best subset S_5; (ii) missing at least one important test in S_5; (iii) including only one redundant test; (iv) including two or three more redundant tests. The simulation results show that: based on $\mathrm{AUC}_\lambda^{(\mathrm{CV})}$, about 53.6% of all the times the variable selection procedure selects the true best subset S_5, about 17.6% selects one subset in category (ii), about 12.4% select one

Table 5.3. Example 5.2.2 – Pancreatic cancer serum biomarkers

$\overline{\text{AUC}}(\widehat{\boldsymbol{\beta}}_\lambda)$	$\text{AUC}_\lambda^{(\text{ACV})}$	V_1	V_2	V_1^2	V_2^2	$V_1 * V_2$
0.7151	0.7051	0.8405	0.4859	-0.0908	0.0062	-0.2220
0.7136	0.7100*	0.8426	0.4856	–	-0.005	-0.2329
0.7136	0.7026	0.8386	0.4873	–	–	-0.2435
0.7096	0.6979	0.8592	0.5117	–	–	–
0.6865	0.6865	1	–	–	–	–

* the selected "best" subset

subset in category (iii), and about 16.4% select one subset in category (iv); based on $\text{AUC}_\lambda^{(\text{ACV})}$, about 42.6% of all the times the variable selection procedure selects the true best subset S_5, about 15.8% selects one subset in category (ii), about 13.5% select one subset in category (iii), and about 28.1% select one subset in category (iv). Therefore, we find that the two criteria are similar to each other, although $\text{AUC}_\lambda^{(\text{ACV})}$ is slightly more conservative.

5.2.2 Pancreatic cancer serum biomarkers example

We revisit the pancreatic cancer serum biomarkers example in section 5.1.2. Recall that two continuous positive scale serum biomarkers were available to diagnose a patient with pancreatic cancer: CA-125 (V_1), a cancer antigen, and CA-19-9 (V_2), a carbohydrate antigen. In this example, we also included their interaction and the quadratic terms. We search all the possible subsets to find the subset with the largest AUC. Table 5.3 shows this selection procedure, where each row shows the results for the subset that has the largest $\text{AUC}_\lambda^{(\text{ACV})}$ values among those subsets of equal size. For each subset, $\overline{\text{AUC}}(\widehat{\boldsymbol{\beta}})$ and $\text{AUC}_\lambda^{(\text{ACV})}$ are shown along with the linear combination coefficient vector $\widehat{\boldsymbol{\beta}}$. It shows that, the selected "best" subset includes CA-125, CA-19-9, CA-19-9 square, and the interaction.

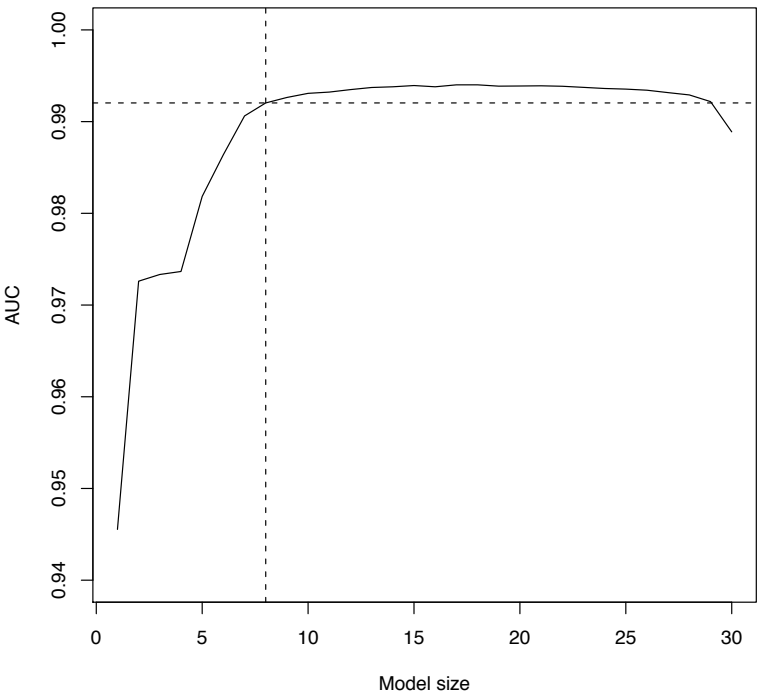
We also compare this result with the ones from the Logistic-LASSO and the sLDA. Here (and in the following two real examples) the Logistic-LASSO applies the C_p criterion and the sLDA

applies the BIC criterion to select the tuning parameter. When they are applied to this dataset, both procedures result in selecting the full set as the "best" subset, which includes both biomarkers as well as their interaction and the quadratic terms. The estimated coefficients from the Logistic-LASSO are 0.7900, 0.5922, -0.1043, -0.0579 and -0.1044; while from the sLDA are 0.8854, 0.4406, -0.1007, -0.0322 and -0.1036. However, as discussed in Section 3.4.1, they cannot provide the estimate of $\mathrm{AUC}(\widehat{\boldsymbol{\beta}})$ at the same time.

5.2.3 Wisconsin breast cancer study

The dataset comes from the Wisconsin Breast Cancer Study which included 375 patients with benign breast cancer and 212 patients with malignant breast cancer and was originally analyzed by Wolberg et al. (1995). There were 30 features served as diagnostic tests computed from a digitized image of a fine needle aspirate of the subject's breast mass. The dataset is available at http://archive.ics.uci.edu/ml/datasets.html.

Since there were 30 tests in this example, instead of searching all the possible subsets, we used the backward elimination procedure. We started with the full set, eliminated one variable each time, and kept the subset with the largest $\mathrm{AUC}_{\lambda}^{(\mathrm{ACV})}$. Repeating this until there was only one test included. Figure 5.6 shows the plot of $\mathrm{AUC}_{\lambda}^{(\mathrm{ACV})}$ versus the subset size through the backward elimination procedure. The curve shows that the $\mathrm{AUC}_{\lambda}^{(\mathrm{ACV})}$ increases as the subset size increases, then slowly deceases. Applying the one-standard-error rule (Hastie et al., 2009, p.244), we might choose the most parsimonious model within one standard error away the subset of the largest $\mathrm{AUC}_{\lambda}^{(\mathrm{ACV})}$, that is the subset including tests 7, 11, 12, 20, 22, 24, 28, and 30, with corresponding $\mathrm{AUC}_{\lambda}^{(\mathrm{ACV})} = 0.9920$. The ad-hoc one-standard-error rule is arguable, but sometimes it is used together with cross-validation procedures (Hastie et al., 2009, p.244) because they are conservative. Without the one-standard-error correction, our methods select a subset of 16 tests.



Example 5.2.3 – Wisconsin Breast Cancer Study

Figure 5.6. Plot of $\mathrm{AUC}_{\lambda}^{(\mathrm{ACV})}$ versus the subset size

The Logistic-LASSO selects the "best" subset including tests 1, 2, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 24, 25, 27, 28, 29, and 30. While the sLDA selects the "best" subset including tests 1, 2, 3, 8, 11, 21, 22, 23, 25, 27, 28, and 29.

5.2.4 Pima Indians diabetes study

The dataset comes from the Pima Indians Diabetes Study which included 268 patients with signs of diabetes and 500 patients without signs of diabetes and was originally analyzed by Smith et al. (1988). There were eight features: Number of times pregnant (V_1), Plasma glucose concentration (V_2), Diastolic blood pressure (V_3), Triceps skin fold thickness (V_4), 2-Hour serum insulin (V_5), Body

Table 5.4. Example 5.2.4 – Pima Indians diabetes

$\overline{\mathrm{AUC}}(\widehat{\boldsymbol{\beta}}_\lambda)$	$\mathrm{AUC}_\lambda^{(\mathrm{ACV})}$	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8
0.8310	0.8225	0.2572	0.7798	-0.1360	-0.0427	-0.0503	0.4515	0.2741	0.1549
0.8308	0.8234	0.2596	0.7878	-0.1434	–	-0.0670	0.4354	0.2681	0.1593
0.8303	0.8239*	0.2689	0.7829	-0.1489	–	–	0.4411	0.2658	0.1655
0.8281	0.8234	0.3394	0.7910	-0.1194	–	–	0.4219	0.2584	–
0.8262	0.8224	0.3272	0.8058	–	–	–	0.4146	0.2679	–
0.8175	0.8146	0.3367	0.8355	–	–	–	0.4343	–	–
0.8022	0.8010	–	0.8757	–	–	–	0.4828	–	–
0.7779	0.7779	–	1	–	–	–	–	–	–

* the selected "best" subset

mass index (V_6), Diabetes pedigree function (V_7), and age (V_8). This dataset is also available at http://archive.ics.uci.edu/ml/datasets.html.

The results are shown in Table 5.4, where each row shows the results for the subset that has the largest $\mathrm{AUC}_\lambda^{(\mathrm{ACV})}$ values among those subsets of equal size. For each subset, $\overline{\mathrm{AUC}}(\widehat{\boldsymbol{\beta}})$ and $\mathrm{AUC}_\lambda^{(\mathrm{ACV})}$ are shown along with the linear combination coefficient vector $\widehat{\boldsymbol{\beta}}$. It shows that the subset with the largest $\mathrm{AUC}_\lambda^{(\mathrm{ACV})}$ is the one including V_1, V_2, V_3, V_6, V_7 and V_8.

The logistic-LASSO selects the "best" subset including V_1, V_2, V_3, V_5, V_6, V_7, and V_8, with the estimated coefficients being 0.2831, 0.7693, -0.1694, -0.0860, 0.4828, 0.2118 and 0.1180. The sLDA selects the same "best" subset, with the coefficients being 0.2909, 0.7886, -0.1629, -0.0457, 0.4425, 0.2102, and 0.1578.

Chapter 6

DISCUSSION

Many authors have considered using the sigmoid function to approximate the indicator function when calculating the AUC, but without clear reasoning. In the first topic, by introducing soft ROC curves, we have provided a connection between the approximation to ROC curve and the approximation to the corresponding AUC. This explains in some way as to why we can use some function to approximate the indicator function while calculating the AUC.

The selection of the regularization parameter in a soft ROC curve is a critical issue. The application of the proposed cross-validation procedure is straightforward. Since the cross-validation is one of the most popular methods for model selection, we have examined it in the present context, by means of Monte Carlo simulation studies and a real example, and shown that it performs well. However, the consistency of the proposed cross-validation procedure remains as an open problem.

The second topic is relatively old, which considers the optimal linear combination of diagnostic tests in terms of maximizing the AUC. However, it raises two new issues. One is that we re-investigate the properties of the estimated linear combination coefficients without the GLM assumption, which is usually made in the literature. The other is that several estimates are proposed for estimating the AUC associated with the estimated combination coefficients. Here the AUC associated with the estimated combination coefficients is important, because it is the counterpart of the prediction error in linear regression models and thus it can serve as a variable selection criterion to select important diagnostic tests from many available ones.

For the later issue, we most advocate the cross-validation procedure, which works very well both as an estimate for the AUC associated with the estimated coefficients and as a variable selection criterion.

In the last topic, we work on the AUC-LASSO problem, which is challenging due to the fact that the objective function is not convex. The proposed algorithm utilizes the soft ROC curves as an approximation to the empirical objective function, and two state-of-the-art optimization algorithms: DC programming and coordinate descent as the optimization tools. As presented in section 4.3, we are still working on the last step of the proposed algorithm: use coordinate descent to iteratively solve the subproblem (4.5). Furthermore, the approximated objective function in (4.3) is piecewise quadratic, thus the problem (4.3) has a piecewise linear solution path. As discussed in section 4.2.1, with such a good property, another efficient algorithm similar to LARS can also be developed.

REFERENCES

[1] An, L. T. and Tao, P. D. (1997). Solving a class of linearly constrained indefinite quadratic problems by D.C. algorithms. *Journal of Global Optimization* **11**: 253 - 285.

[2] Bamber, D. C. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic curve graph. *Journal of Mathematical Psychology* **12,** 387–415.

[3] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrics. *Annals of Statistics.* **36,** 199-227.

[4] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth, New York.

[5] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24**: 2350 - 2383.

[6] Copas, J. B. and Corbett, P. (2002). Overestimation of the receiver operating characteristic curve for logistic regression. *Biometrika* **89,** 315–331.

[7] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78,** 316–331.

[8] Efron, B and Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association* **92,** 548–560.

[9] Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association* **99,** 619–642.

[10] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics.* **32**: 407 - 499.

[11] 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.

[12] Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**: 109 -148.

[13] Friedman, J., Hastie, T., Hoefling, H., and Tibshirani, R. (2007). Pathwise Coordinate Optimization. *The Annals of Applied Statistics.* **2**: 302 - 332.

[14] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software.* **33**(1).

[15] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning, 2nd Edition.* New York: Springer.

[16] Han, A. K. (1977). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of the Royal Statistical Society* **39,** 44–47.

[17] Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**: 55 - 67.

[18] Huang, X., Qin, G. and Fang, Y. (2010). Optimal combinations of diagnostic tests based on AUC. In press, *Biometrics.*

[19] Lehmann, E. L. (1997). *Testing Statistical Hypotheses, 2nd edition.* New York: Springer.

[20] Li, B. and Yu, Q. (2009). Robust and sparse bridge regression. *Statistics and its interface.* **2**: 481 - 491.

[21] Liu, Y., Shen, X., and Doss, H. (2005). Multicategory psi-learning and support vector machine: computational tools. *Journal of Computational and Graphical Statistics.* **14**: 219 - 236.

[22] Liu, Z. and Tan, M. (2008). ROC-based utility function maximization for feature selection and classification with applications to high-dimensional protease data. *Biometrics* **64,** 1155–1161.

[23] Liu, Z., Tan, M. and Jiang, F. (2009). Regularized F-measure maximization for feature selection and classification. *Journal of Biomedicine and Biotechnology* **2009,** 617946

[24] Ma, S. and Huang, J. (2005). Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* **21,** 4356–4362.

[25] Ma, S. and Huang, J. (2007). Combining multiple markers for classification using ROC. *Biometrics* **63,** 751–757.

[26] Marshall, A. W. and Olkin, I. (1967). A multivariate exponential distribution. *Journal of the American Statistical Association* **62,** 30–44.

[27] Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1,** 123–140.

[28] Pepe, M. S. (2004). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* New York: Oxford University Press.

[29] Pepe, M. S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **62,** 221–229.

[30] Peng, L. and Zhou, X. H. (2004). Local linear smoothing of receiver operating characteristic (ROC) curves. *Journal of Statistical Planning and Inference* **118,** 129–143.

[31] Ren, H., Zhou, X. H. and Liang, H. (2004). A flexible method for estimating the ROC curve. *Journal of Applied Statistics* **31,** 773–784.

[32] Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *Annals of Statistics.* **35**: 1012 - 1030.

[33] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., and Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Symposium on Computer Applications and Medical Care*, 261–265.

[34] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of Econometrics* **35,** 303–316.

[35] Su, J. Q. and Liu, J. S. (1993). Linear combination of multiple diagnostic markers. *Journal of the American Statistical Association* **88,** 1350–1355.

[36] Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting *Suri-Kagaku (Mathematics Sciences)* **153,** 12–18.

[37] Trendafilov, N. T. and Jolliffe, I. T. (2007). DALASS: variable selection in discriminant analysis via the LASSO. *Computational Statistics and Data Analysis* **51,** 3718–3736.

[38] Vapnik, V. (1998). *Statistical Learning Theory.* New York: John Wiley.

[39] Wang, Z., Chang, Y. I., Ying, Z., Zhu, L. and Yang, Y. (2007) A parsimonious thresholding-independent protein feature selection method through the area under receiver operating characteristic curve. *Bioinformatics* **23,** 2788–2794.

[40] Wieand, S., Gail, M. H., James, B. R., and James, K. L. (1989) A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76,** 585–592.

[41] Wolberg, W.H., Street, W.N., Heisey, D.M., and Mangasarian, O.L. (1995) Computer-derived nuclear features distinguish malignant from benign breast cytology. *Human Pathology* **26,** 792–796.

[42] Wu, M., Zhang, L., Wang, Z., Christiani, D., and Lin, X. (2009). Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/parthway and gene selection. *Bioinformatics* **25,** 1145-1151.

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

[44] Zhou, X. H., Obuchowski, N., and McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine.* New York: John Wiley & Sons.

[45] Zou, K. H., Hall, W. J. and Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic curves for continuous diagnostic tests. *Statistics in Medicine* **16,** 2143–2156.

[46] 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.

Appendix A

SOME PROOF AND CALCULATIONS FOR CHAPTER 2

A.1 Proof of Theorem 2.1

Let $S_D = E\{I_\delta(Y-c)\}$ and $S_{\bar{D}} = E\{I_\delta(X-c)\}$. From (2.1), $R_\delta(t) = S_D(S_{\bar{D}}^{-1}(t))$, $t \in (0,1)$. Thus,

$$\begin{aligned} AUC_\delta &= \int_0^1 S_D(S_{\bar{D}}^{-1}(t))dt \\ &= \int_\infty^{-\infty} S_D(c)dS_{\bar{D}}(c) \\ &= E\int_\infty^{-\infty} I_\delta(Y-c)dE\{I_\delta(X-c)\} \\ &= E\int_{-\infty}^{\infty} I_\delta(Y-c)\dot{I}_\delta(X-c)dc \\ &= E\{K_\delta(Y-X)\}, \end{aligned}$$

which completes the proof.

A.2 Calculation for order 0 two-sided indecisive function

In this case, $\dot{I}_\delta(t-c) = \frac{1}{2}\mathbf{1}\{t-c=-\delta\} + \frac{1}{2}\mathbf{1}\{t-c=\delta\}$. So, we have

$$\begin{aligned} K_\delta(Y-X) &= \int_{-\infty}^{\infty} I_\delta(Y-c)\dot{I}_\delta(X-c)dc \\ &= \int_{-\infty}^{\infty} \left(\frac{1}{2}\mathbf{1}\{Y-\delta < c < Y+\delta\} + \mathbf{1}\{c \le Y-\delta\}\right)\dot{I}_\delta(X-c)dc \\ &= \frac{1}{4}\mathbf{1}\{-2\delta \le Y-X < 0\} + \frac{3}{4}\mathbf{1}\{0 \le Y-X < 2\delta\} + \mathbf{1}\{Y-X \ge 2\delta\}. \end{aligned}$$

A.3 Calculation for order 1 two-sided indecisive function

In this case, $\dot{I}_\delta(t-c) = \frac{1}{2\delta}\mathbf{1}\{-\delta \le t-c < \delta\}$. So, we have

$$\begin{aligned}
K_\delta(Y-X) &= \int_{-\infty}^{\infty} I_\delta(Y-c)\dot{I}_\delta(X-c)dc \\
&= \int_{-\infty}^{\infty} \left[\frac{1}{4\delta} + \frac{1}{4\delta^2}(Y-c)\right] \mathbf{1}\{Y-\delta < c \le Y+\delta\}\mathbf{1}\{X-\delta < c \le X+\delta\} \\
&\quad + \frac{1}{2\delta} I(c \le Y-\delta) I(X-\delta < c \le X+\delta)dc \\
&= 0 \cdot \mathbf{1}\{Y-X < -2\delta\} + \mathbf{1}\{Y-X \ge 2\delta\} \\
&\quad + \left\{\int_{X-\delta}^{Y-\delta} \frac{1}{2\delta}dc + \int_{Y-\delta}^{X+\delta} [\frac{1}{4\delta} + \frac{1}{4\delta^2}(Y-c)]dc\right\} \mathbf{1}\{0 \le Y-X < 2\delta\} \\
&\quad + \left\{\int_{X-\delta}^{Y+\delta} \left[\frac{1}{4\delta} + \frac{1}{4\delta^2}(Y-c)\right] dc\right\} \mathbf{1}\{-2\delta \le Y-X < 0\} \\
&= \left[\frac{1}{2} + \frac{5(Y-X)}{4\delta} - \mathrm{sign}(Y-X)\frac{(Y-X)^2}{2\delta^2}\right] \mathbf{1}\{-2\delta \le Y-X < 2\delta\} \\
&\quad + \mathbf{1}\{Y-X \ge 2\delta\}.
\end{aligned}$$

A.4 Calculation for order ∞ two-sided (Sigmoid) indecisive function

In this case, $\dot{I}_\delta(t-c) = \delta I_\delta(t-c)[1-I_\delta(t-c)]$. So, we have

$$\begin{aligned}
K_\delta(Y-X) &= \int_{-\infty}^{\infty} I_\delta(Y-c)\dot{I}_\delta(X-c)dc \\
&= \int_{-\infty}^{\infty} \frac{\delta e^{-\delta(X-c)}}{[1+e^{-\delta(Y-c)}][1+e^{-\delta(X-c)}]^2}dc \\
&= -e^{\delta(Y-X)}\left[\frac{1}{1-e^{\delta(Y-X)}} + \frac{1}{(1-e^{\delta(Y-X)})^2}\delta(Y-X)\right].
\end{aligned}$$

A.5 Calculation for order 0 one-sided indecisive function

In this case, $\dot{I}_\delta(t-c) = \frac{1}{2}\mathbf{1}\{t-c=0\} + \frac{1}{2}\mathbf{1}\{t-c=\delta\}$. So, we have

$$\begin{aligned}
K_\delta(Y-X) &= \int_{-\infty}^{\infty} I_\delta(Y-c)\dot{I}_\delta(X-c)dc \\
&= \int_{-\infty}^{\infty} \left(\frac{1}{2}\mathbf{1}\{Y-\delta<c<Y\} + \mathbf{1}\{c \le Y-\delta)\right\} \dot{I}_\delta(X-c)dc \\
&= \frac{1}{4}\mathbf{1}\{-\delta \le Y-X<0\} + \frac{3}{4}\mathbf{1}\{0 \le Y-X<\delta\} + \mathbf{1}\{Y-X \ge \delta\}.
\end{aligned}$$

A.6 Calculation for order 1 one-sided indecisive function

In this case, $\dot{I}_\delta(t-c) = \frac{1}{\delta}\mathbf{1}\{0 \le t-c<\delta\}$. So, we have

$$\begin{aligned}
K_\delta(Y-X) &= \int_{-\infty}^{\infty} I_\delta(Y-c)\dot{I}_\delta(X-c)dc \\
&= \int_{-\infty}^{\infty} \left[\frac{1}{\delta^2}(Y-c)\mathbf{1}\{Y-\delta<c \le Y\}\mathbf{1}\{X-\delta<c \le X\}\right. \\
&\quad \left. +\frac{1}{\delta}\mathbf{1}\{c \le Y-\delta\}\mathbf{1}\{X-\delta<c \le X\}\right] dc \\
&= 0\cdot\mathbf{1}\{Y-X<-\delta\} + \mathbf{1}\{Y-X \ge \delta\} \\
&\quad + \left[\int_{X-\delta}^{Y-\delta} \frac{1}{\delta}dc + \int_{Y-\delta}^{X} \frac{1}{\delta^2}(Y-c)dc\right]\mathbf{1}\{0 \le Y-X<\delta\} \\
&\quad + \left[\int_{X-\delta}^{Y} \frac{1}{\delta^2}(Y-c)dc\right]\mathbf{1}\{-\delta \le Y-X<0\} \\
&= \left[\frac{1}{2} + \frac{(Y-X)}{\delta} - \mathrm{sign}(Y-X)\frac{(Y-X)^2}{2\delta^2}\right]\mathbf{1}\{-\delta \le Y-X<\delta\} + \mathbf{1}\{Y-X \ge \delta\}.
\end{aligned}$$

Appendix B

PROOFS FOR CHAPTER 3

B.1 Investigation of the GLM assumption

Under the GLM assumption (3.1),

$$\begin{aligned}\frac{G(\mathbf{T})}{F(\mathbf{T})} &= \frac{P(T|D=1)}{P(T|D=0)} \\ &= \frac{P(D=1|T)P(D=0)}{P(D=0|T)P(D=1)} = \frac{h(\boldsymbol{\beta}_0^T\mathbf{T})}{1-h(\boldsymbol{\beta}_0^T\mathbf{T})}\cdot C,\end{aligned}$$

where C is a constant. It can be easily shown that, assuming that $\mathbf{X}$ and $\mathbf{Y}$ follow multivariate normal distributions with different covariance matrices, the above equality cannot hold.

B.2 The proof of Proposition 3.1

First, we show that, for random variables Z_1 and Z_2 of marginal densities f_{Z_1} and f_{Z_2} respectively, the following equality holds:

$$\left.\frac{\partial P(Z_1 > -Z_2\beta)}{\partial\beta}\right|_{\beta=0} = E(Z_2|Z_1=0)\cdot f_{Z_1}(Z_1=0). \tag{B.1}$$

In fact, from

$$\begin{aligned}&P(Z_1 > -Z_2\beta) = \\ &\int_{-\infty}^{\infty}\int_{-z_2\beta}^{\infty} f_{Z_1|Z_2}(z_1|Z_2=z_2)f_{Z_2}(Z_2=z_2)dz_1dz_2,\end{aligned}$$

we have

$$
\begin{aligned}
& \left.\frac{\partial P(Z_1 > -Z_2\beta)}{\partial \beta}\right|_{\beta=0} \\
= & \int_{-\infty}^{\infty} z_2 f_{Z_1|Z_2}(Z_1 = 0|Z_2 = z_2) f_{Z_2}(Z_2 = z_2) dz_2 \\
= & \int_{-\infty}^{\infty} z_2 f(Z_1 = 0, Z_2 = z_2) dz_2 \\
= & \int_{-\infty}^{\infty} z_2 f_{Z_2|Z_1=0}(Z_2 = z_2|Z_1 = 0) f_{Z_1}(Z_1 = 0) dz_2 \\
= & E(Z_2|Z_1 = 0) \cdot f_{Z_1}(Z_1 = 0).
\end{aligned}
$$

Then, for p-dim random vector $\mathbf{Z}$, assume that there are $\boldsymbol{\beta}^{(1)} \neq \boldsymbol{\beta}^{(2)} \in \mathbf{B}$ such that

$$
P(\boldsymbol{\beta}^{(1)T}\mathbf{Z} > 0) = P(\boldsymbol{\beta}^{(2)T}\mathbf{Z} > 0) = \max_{\boldsymbol{\beta}\in\mathbf{B}} P(\boldsymbol{\beta}^{T}\mathbf{Z} > 0).
$$

Let $\boldsymbol{\omega}^{(1)} = \boldsymbol{\beta}^{(1)T}\mathbf{Z}$ and $\boldsymbol{\omega}^{(2)} = \boldsymbol{\beta}^{(2)T}\mathbf{Z}$. Noting that $P(\omega^{(1)} > 0) = \max_{\boldsymbol{\beta}\in\mathbf{B}} P(\boldsymbol{\beta}^{T}\mathbf{Z} > 0) = \max_{\gamma\in\mathcal{R}} P(\{\boldsymbol{\omega}^{(1)} + \gamma\boldsymbol{\omega}^{(2)}\}/||\boldsymbol{\beta}^{(1)} + \gamma\boldsymbol{\beta}^{(2)}|| > 0) = \max_{\gamma\in\mathcal{R}} P(\boldsymbol{\omega}^{(1)} + \gamma\boldsymbol{\omega}^{(2)} > 0)$, we have

$$
\left.\frac{\partial P(\boldsymbol{\omega}^{(1)} + \gamma\boldsymbol{\omega}^{(2)} > 0)}{\partial \gamma}\right|_{\gamma=0} = 0.
$$

Similarly, we have

$$
\left.\frac{\partial P(\boldsymbol{\omega}^{(2)} + \gamma\boldsymbol{\omega}^{(1)} > 0)}{\partial \gamma}\right|_{\gamma=0} = 0.
$$

Therefore, from (B.1), we have $E(\boldsymbol{\omega}^{(2)}|\boldsymbol{\omega}^{(1)} = 0) = 0$ and $E(\boldsymbol{\omega}^{(1)}|\boldsymbol{\omega}^{(2)} = 0) = 0$. This is a contradiction to the assumption made in the proposition. Hence, $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)}$, and then the maximization solution is unique.

B.3 The proof of Proposition 3.2

Let $S_{mn}(\boldsymbol{\beta}) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I(\boldsymbol{\beta}^T\mathbf{Y}_j > \boldsymbol{\beta}^T\mathbf{X}_i)$. By the arguments in Han (1987), we have

$$S_{mn}(\boldsymbol{\beta}) \xrightarrow{a.s.} \mathrm{AUC}(\boldsymbol{\beta}), \text{ uniformaly in } \boldsymbol{\beta} \in \mathbf{B},$$

where $\mathbf{B} = \{\boldsymbol{\beta} : ||\boldsymbol{\beta}|| = 1\}$. $\mathrm{AUC}(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta} \in \mathbf{B}$ and obtains the unique maximum at $\boldsymbol{\beta}_0$. Let $\mathbf{B}_0$ be an open set in $\mathcal{R}^p$ such that $\boldsymbol{\beta}_0 \in \mathbf{B}_0$. Then $\mathbf{B}_1 = \mathbf{B} - (\mathbf{B}_0 \bigcap \mathbf{B})$ is compact and there exists $\xi = \mathrm{AUC}(\boldsymbol{\beta}_0) - \min_{\boldsymbol{\beta}\in(B_1)} \mathrm{AUC}(\boldsymbol{\beta}) > 0$. Note that from the uniformly convergent argument, there exist m_1 and n_1 such that for all $m > m_1$ and $n > n_1$,

$$|S_{mn}(\boldsymbol{\beta}) - \mathrm{AUC}(\boldsymbol{\beta})| < \frac{\xi}{2}, \text{ uniformly in } \boldsymbol{\beta} \in \mathbf{B}.$$

This implies that $\widehat{\boldsymbol{\beta}} \in \mathbf{B}_0$ for all $m > m_1$ and $n > n_1$. Then $\widehat{\boldsymbol{\beta}} \xrightarrow{\text{a.s.}} \boldsymbol{\beta}_0$.

B.4 The derivation of the approximated cross-validation

Let $Q(\boldsymbol{\beta}) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} g(\boldsymbol{\beta}^T\mathbf{Z}_{ij})$. For any i and j,

$$\begin{aligned} &Q^{'}(\widehat{\boldsymbol{\beta}}_\lambda^{(-ij)}) - \frac{1}{mn}[\sum_{j_0=1}^{n} \mathbf{Z}_{ij_0}\dot{g}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda^{(-ij)T}\mathbf{Z}_{ij_0}) \\ &+ \sum_{i_0=1}^{m} \mathbf{Z}_{i_0j}\dot{g}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda^{(-ij)T}\mathbf{Z}_{i_0j}) - \mathbf{Z}_{ij}\dot{g}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda^{(-ij)T}\mathbf{Z}_{ij})] = 0. \end{aligned}$$

By Taylor's expansion, $\widehat{\mathrm{AUC}}_\lambda^{(\mathrm{CV})}$ can be expressed as

$$\begin{aligned} &Q(\widehat{\boldsymbol{\beta}}_\lambda) + \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(\widehat{\boldsymbol{\beta}}_\lambda^{(-ij)} - \widehat{\boldsymbol{\beta}}_\lambda)^T\mathbf{Z}_{ij}\dot{g}(\widehat{\boldsymbol{\beta}}_\lambda \\ &+a_{ij}(\widehat{\boldsymbol{\beta}}_\lambda^{(-ij)} - \widehat{\boldsymbol{\beta}}_\lambda)^T\mathbf{Z}_{ij}), \end{aligned}$$

and

$$Q^{'}(\widehat{\boldsymbol{\beta}}_{\lambda}^{(-ij)}) = Q^{''}(\widehat{\boldsymbol{\beta}}_{\lambda} + b_{ij}(\widehat{\boldsymbol{\beta}}_{\lambda}^{(-ij)} - \widehat{\boldsymbol{\beta}}_{\lambda}))(\widehat{\boldsymbol{\beta}}_{\lambda}^{(-ij)} - \widehat{\boldsymbol{\beta}}_{\lambda}),$$

where $|a_{ij}| \leq 1$ and $|b_{ij}| \leq 1$, noting $Q^{'}(\widehat{\boldsymbol{\beta}}_{\lambda}) = 0$. Therefore, $\widehat{\mathrm{AUC}}_{\lambda}^{(\mathrm{CV})}$ can be expressed as

$$\begin{aligned}
&\frac{1}{m^2n^2}\sum_{i=1}^{m}\sum_{j=1}^{n}(\sum_{j_0=1}^{n}\dot{g}_{\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda}^{(-ij)T}\mathbf{Z}_{ij_0})\mathbf{Z}_{ij_0}^{T} \\
&+\sum_{i_0=1}^{m}\dot{g}_{\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda}^{(-ij)T}\mathbf{Z}_{i_0j})\mathbf{Z}_{i_0j}^{T} - \dot{g}_{\lambda}(\widehat{\boldsymbol{\beta}}_{\lambda}^{(-ij)T}\mathbf{Z}_{ij})\mathbf{Z}_{ij}^{T}) \\
&Q^{''-1}(\widehat{\boldsymbol{\beta}_{\lambda}} + b_{ij}(\widehat{\boldsymbol{\beta}}_{\lambda}^{(-ij)} - \widehat{\boldsymbol{\beta}}_{\lambda})]\mathbf{Z}_{ij}\dot{g}_{\lambda}[\widehat{\boldsymbol{\beta}}_{\lambda} \\
&+a_{ij}(\widehat{\boldsymbol{\beta}}_{\lambda}^{(-ij)} - \widehat{\boldsymbol{\beta}}_{\lambda})^{T}\mathbf{Z}_{ij}) + Q(\widehat{\boldsymbol{\beta}}_{\lambda}).
\end{aligned}$$

Together with $\widehat{\boldsymbol{\beta}}_{\lambda} \xrightarrow{P} \boldsymbol{\beta}_{0,\lambda}$ and $\widehat{\boldsymbol{\beta}}_{\lambda}^{(-ij)} \xrightarrow{P} \boldsymbol{\beta}_{0,\lambda}$, where $\boldsymbol{\beta}_{0,\lambda} = \arg\min Eg_{\lambda}(\boldsymbol{\beta}^{T}\mathbf{Z})$, the $\mathrm{AUC}_{\lambda}^{(\mathrm{ACV})}$ is derived.