Model Selection Methods for Cancer Staging and Other Disease Stratification Problems

by

Yunzhi Lin

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

(STATISTICS)

at the

UNIVERSITY OF WISCONSIN - MADISON

2012

Date of final oral examination: 11/19/2012

The dissertation is approved by the following members of the Final Oral Committee:
Richard J. Chappell, Professor, Statistics
Sijian Wang, Assistant Professor, Statistics
Menggang Yu, Associate Professor, Biostatistics and Medical Informatics
KyungMann Kim, Professor, Biostatistics and Medical Informatics
Ronald E. Gangnon, Associate Professor, Biostatistics and Medical Informatics

Acknowledgements

First and foremost I would like to thank my advisor, Professor Richard Chappell, without whose guidance this project will never be done. Throughout the past four years, with his knowledge, wisdom, and scientific attitude, Professor Chappell has not only helped me through each specific research questions, but also impacted me greatly as a researcher in general.

I also want to express my sincere gratitude to all my committee members, Professor Sijian Wang, Professor Menggang Yu, Professor Kyungmann Kim, and Professor Ronald Gangnon, whose advices and suggestions have helped me much in improving this work and turning this project to the correct direction. I would especially like to thank Professor Sijian Wang and Professor Menggang Yu, who are the major collaborators of this project. Their instructive advice and contagious enthusiasm has greatly impacted both the project and myself. I am also thankful for the Department of Statistics at the University of Wisconsin - Madison, which provides such an excellent program and help me build a solid background in both theories and applications. In addition, I would like to thank Dr. Mithat Gönen and Professor Tom Imperiale for providing the data.

Finally, I owe my deepest thanks to my parents and to my fiancé and best friend Heming Zhen for their love, care, and support throughout the years. Their love

provides my inspiration and is my driving force. A special thanks to Heming Zhen, whose mathematical insight has been of great help to me.

Abstract

The tumor-node-metastasis (TNM) staging system has been the anchor of cancer diagnosis, treatment, and prognosis for many years. For meaningful clinical use, an orderly, progressive grouping of the T and N categories into an overall staging system needs to be defined, usually with respect to a time-to-event outcome. This can be considered as a model selection problem for censored response with respect to features arranged on a partially ordered two-way grid, and the aim is to select the grouping that best classifies the patients.

This dissertation presents the effort to develop such cancer stage groupings. Two model selection methods are proposed for this task. As a first, exploratory attempt, a bootstrap model selection method is proposed by maximizing bootstrap estimates of the chosen statistical criteria. The criteria are based on prognostic ability including a landmark measure of the explained variation, the area under the ROC curve, and a concordance probability generalized from Harrell's c-index. We illustrate the utility of our method by applying it to the staging of colorectal cancer. The pros and cons of the method are discussed.

In order to overcome some of the drawbacks of the bootstrap method, a penalized regression method is proposed which resembles the lasso method. Instead of penalizing the $L1$-norm of the coefficients like lasso, in order to enforce the stage grouping we place $L1$ constraints on the differences between neighboring coefficients. The underlying mechanism is the sparsity-enforcing property of the $L1$ penalty, which forces some estimated coefficients to be the same and hence leads to stage grouping. A series of optimal groupings with different numbers of stages can be obtained by varying the tuning parameter, which gives a tree-like structure offering a visual aid on how the groupings are progressively made. We hence call the proposed method the lasso tree. Again, we illustrate the utility of our method by applying it to the staging of colorectal cancer. Simulation studies are carried out to examine the finite sample performance of the selection procedure. We demonstrate that the lasso tree is able to give the right grouping with moderate sample size, is stable with regard to changes in the data, and is not affected by random censoring. Furthermore, with slight modification of the penalties and proper choice of regularization parameters, we show that the lasso tree grouping procedure is consistent; namely, the estimator is root-n consistent and gives the correct grouping asymptotically.

The lasso tree methodology has general appeal to cancers and other diseases that use aggregate risk scores based on risk factors. With proper modification, it is applied to the risk stratification of colorectal cancer. To facilitate the efficiency of colorectal cancer screening, there is a need to stratify risk for colorectal cancer among the 90% of U.S. residents who are considered “average risk”. Logistic regression is traditionally used to estimate the risk of advanced colorectal neoplasia. However, logistic regression may be prone to overfitting and instability in variable selection. Since most of the risk factors have several categories, it is tempting to collapse these categories into fewer risk groups. In light of these considerations, a modification of the lasso tree, a penalized logistic regression method, is proposed which automatically and simultaneously selects variables, groups categories, and estimates their coefficients, by penalizing the $L1$-norm of both the coefficients and their differences. It encourages sparsity in the categories, i.e. grouping of the categories, and sparsity in the variables, i.e. variable selection. The method is applied to a recently completed large cohort study of colorectal cancer. The important variables are selected, with close categories simultaneously grouped, by penalized regression models with and without the interactions terms. The models are validated with 10-fold cross-validation. The ROC curves of the penalized regression models dominate the ROC curve of naive logistic regressions, indicating a superior discriminative performance.

Contents

Acknowledgements **i**

Abstract **iii**

List of Figures **ix**

List of Tables **xii**

1 Introduction **1**

1.1 The Cancer Staging Problem 1
1.2 Motivating Example: Colorectal Cancer 4
1.3 Scope of this Dissertation 8

2 The Bootstrap Selection Method **10**

2.1 Introduction 10
2.2 The Bootstrap Method 12
2.3 Criteria for Assessing Staging Systems 15
2.3.1 Landmark Measures 15
2.3.2 Concordance Probability 17
2.4 Application to Colorectal Cancer 21
2.4.1 MSKCC Data: Bootstrap Selection 21
2.4.2 MSKCC Data: Confidence Intervals 27
2.4.3 MSKCC Data: Majority Vote Rule 27
2.4.4 MSKCC Data: Cross-Validation 28
2.4.5 The SEER Data Example 29
2.5 Discussion and Conclusions 34

3 Penalized Regression Method: The Lasso Tree **38**

3.1 Introduction 38
3.2 The Lasso Tree 41
3.2.1 A Lasso-type Selection Procedure for Survival Outcomes . . . 41

3.2.2 Computational Approach 43
3.2.3 The Tuning Parameter and The Lasso Tree 45
3.3 Application to Colorectal Cancer 46
3.3.1 MSKCC Data: The Lasso Tree 47
3.3.2 MSKCC Data: Incorporating Information from the AJCC System 51
3.3.3 The SEER Data Example 53
3.4 Simulation Studies 58
3.4.1 Ability to Select the Correct Grouping 59
3.4.2 Estimation Stability 62
3.4.3 The Partial Ordering Constraint 62
3.5 Asymptotic Theory of the Lasso Tree 64
3.5.1 Consistency of the Grouping Procedure 64
3.5.2 Adaptive Lasso Tree 66
3.5.3 Two-dimensional Case 68
3.6 Discussion and Conclusions 70

4 Risk Stratification by Penalized Logistic Regression 74
4.1 Introduction 74
4.2 The Advanced Colorectal Neoplasia Data 78
4.3 Penalized Logistic Regression 80
4.3.1 A Modification of the Lasso Tree 80
4.3.2 Computational Approach 83
4.3.3 Inclusion of Two-way Interactions 84
4.4 Data Analysis and Results 87
4.4.1 Fitted Models 87
4.4.2 Model Validation 93
4.5 Discussion and Conclusions 95

5 Conclusion and Future Work 98
5.1 Cancer Staging and The Lasso Tree 99
5.2 CRC and Penalized Regression Methods 102

A Proofs of Consistency for the Lasso Tree 106
A.1 Reparameterization 107
A.2 Proof of Theorem 1 110
A.3 Proof of Theorem 2 114

Bibliography **120**

List of Figures

1.1 Schematic showing the AJCC 6th edition staging system for colorectal cancer. 3

1.2 Cancer-survival of colorectal cancer patients at Memorial Sloan-Kettering Cancer Center (MSKCC), by the 6th edition AJCC staging system: (A) three main stages; (B) including sub stages. 5

1.3 Survival of colorectal cancer patients identified in the Surveillance, Epidemiology, and End Results (SEER) national cancer registry, by the 6th edition AJCC staging system: (A) three main stages; (B) including sub stages. 7

2.1 Schematic showing staging systems selected by bootstrap based on the MSKCC data and the AJCC 6th edition staging system. VAR: explained variation; AUC: area under the ROC curve; K: concordance probability. 22

2.2 Cancer-specific survival of colorectal cancer patients by the selected staging systems (MSKCC data). Left panel: 3-stage systems; right panel: 6-stage systems. 24

2.3 Confidence intervals for the criteria: the top-ranked systems based on the MSKCC data and the AJCC (red). Left panel: 3-stage systems; right panel: 6-stage systems. 25

2.4 Majority voted systems. Grey bars show the % time each system is ranked from #1 to #10, and the systems are ordered by their % time ranked #1. The x-axis represents the index of the candidate systems. Left panel: 3-stage systems; right panel: 6-stage systems. 26

2.5 Schematic showing staging systems selected by bootstrap based on the SEER data and the AJCC 6th edition staging system. VAR: explained variation; AUC: area under the ROC curve; K: concordance probability. 30

2.6 Cancer-specific survival of colorectal cancer patients by the selected staging systems (SEER data). Left panel: 3-stage systems; right panel: 6-stage systems. 33

2.7 Kaplan-Meier survival curves based on the 260th ranked 3-stage system out of 388 eligible systems (SEER data). 34

3.1 The lasso tree based on the MSKCC data: coefficient estimates and BIC for the colorectal cancer example, as a function of $\log(\lambda)$. The dotted line represents the staging for $\log(\hat{\lambda}) = 1.66$, selected by minimizing BIC. K represents the number of groups. 48

3.2 Schematic showing the staging schemes for M0 colorectal cancer by (a) AJCC 6th edition staging, (b) the lasso tree selected staging, and (c) the lasso tree selected staging incorporating information from the AJCC. 49

3.3 Survivals of colorectal cancer patients by (a) AJCC staging, (b) the lasso tree selected staging, and (c) the lasso tree selected staging incorporating information from the AJCC. 50

3.4 The AJCC 6th edition staging scheme for colorectal cancer: big arrows link cells in the same substage; small arrows link cells in the same main stage but different substages. Heavier weights are imposed on differences represented by the arrows to reflect the AJCC staging, with the big arrows having the heavier weight. 51

3.5 The lasso tree based on the SEER data: coefficient estimates and BIC for the colorectal cancer example, as a function of $\log(\lambda)$. The dotted line represents the staging for $\log(\hat{\lambda}) = 2.54$, selected by minimizing BIC. K represents the number of groups. 54

3.6 Schematic showing the optimal staging schemes selected by lasso tree based on the SEER data: (a) the 9-stage system selected by BIC, (b) the selected 6-stage system, and (c) the selected 3-stage system. . . . 55

3.7 Survivals of colorectal cancer patients by (a) AJCC staging, (b) the lasso tree selected staging, and (c) the lasso tree selected staging incorporating information from the AJCC. 58

3.8 Selection proportions based on 1000 bootstrap samples from the colorectal cancer data set. The x-axis indexes the staging systems selected most often by the lasso tree, and the groupings are ordered by the percentage of time they are selected by the lasso tree. K represents the pre-specified number of groups. Conc = the concordance criterion. . 63

4.1 Coefficient estimates for men for the six selected variables under the main effect penalized logistic regression model as a function of $\log(\lambda)$. The dotted line represents the value of $\log(\lambda)$ that minimized the BIC. 91

4.2 Receiver-operating characteristic (ROC) curves for the risk prediction models: male subjects. 95

4.3 Receiver-operating characteristic (ROC) curves for the risk prediction models: female subjects. 96

A.1 Schematic showing two examples of reparameterizing the β's. Red arrows are reparameterizion with the inter-column and intra-column differences; and black arrows are reparameterizion with the inter-row and intra-row differences. 108
A.2 Schematic showing a third example of reparameterizing the β's. . . . 110
A.3 Schematic showing an example of parameterization (in thick arrows) where $\theta_{j,k}$ (the thick red arrow) has a positive contribution to all the linear combinations $\nu(\theta)$ (thin blue arrows) in which it is involved. . 119

List of Tables

1.1 The TNM staging system for colorectal cancer 2

1.2 Estimated cancer-specific 10-year survivals/hazard ratios by TNM classifications (sample size): colorectal cancer patients at Memorial Sloan-Kettering Cancer Center (MSKCC). 5

1.3 Estimated 5-year survivals/hazard ratios by TNM classifications (sample size): colorectal cancer patients identified in the Surveillance, Epidemiology, and End Results (SEER) national cancer registry. 7

2.1 Number of eligible staging systems given k. 13

2.2 Selected systems and the AJCC: the estimated criteria and their standard errors (MSKCC data). 23

2.3 Selected systems and the AJCC: the estimated criteria and their standard errors (SEER data). 31

3.1 The AJCC and the lasso tree selected systems: the estimated criteria and their standard errors. VAR: explained variation; AUC: area under the ROC curve; Concordance: the probability of concordance. 49

3.2 Selection probabilities based on 1000 simulations (sample size = 1000) 61

3.3 Selection probabilities based on 1000 simulations 64

4.1 Summary of variables in the advanced colorectal neoplasia data set . 86

4.2 Estimated coefficients for men. LR = logistic regression; PLR-1 = penalized logistic regression with only main effects; PLR-2 = penalized logistic regression with main effects and their two-way interactions. . 88

4.3 Estimated coefficients for women. LR = logistic regression; PLR-1 = penalized logistic regression with only main effects; PLR-2 = penalized logistic regression with main effects and their two-way interactions. . 89

4.4 Areas under ROC curves for the risk prediction models. The p-values are for comparing the penalized models to the naive logistic regression. 93

Chapter 1

Introduction

1.1 The Cancer Staging Problem

The development of accurate prognostic classification schemes is of great interest and concern in many areas of clinical research. In oncology, much effort has been made to define a cancer classification scheme that can facilitate prognosis, provide a basis for making treatment or other clinical decisions, and identify homogeneous groups of patients for clinical trials [1–3]. Among various classification schemes, the tumor-node-metastasis (TNM) staging system is widely used because of its simplicity and prognostic ability.

The basis of TNM staging is the anatomic extent of disease. It has three components: T for primary tumor, N for lymph nodes, and M for distant metastasis.

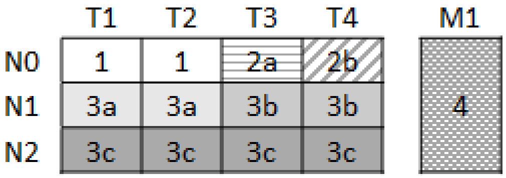
TNM staging is periodically updated. Using its 6th edition, in the case of colorectal cancer which we will use as an example in this dissertation, there are 4 categories of T, 3 of N, and 2 of M [1]. Details of the categories are provided in Table 1.1. These TNM categories jointly define 24 distinct groups, which are unwieldy for meaningful clinical use [4]. Therefore, the American Joint Committee on Cancer (AJCC) and International Union against Cancer (UICC) defined an orderly, progressive grouping of the TNM categories which reduces the system to fewer stages (4 main stages and 7 sub stages under the 6th edition [1], see Figure 1.1). Alternative grouping schemes have also been proposed by other authors.

Table 1.1: The TNM staging system for colorectal cancer

T : Primary Tumor	
T1	Tumor invades submucosa
T2	Tumor invades muscularis propria
T3	Tumor invades into pericolorectal tissues
T4	Tumor directly invades or is adherent to other organs
N : Lymph Nodes	
N0	No regional lymph node metastasis
N1	Metastasis in 1-3 regional lymph nodes
N2	Metastasis in 4 or more regional lymph nodes
M : Distant Metastasis	
M0	No distant metastasis
M1	Distant metastasis

The value and usefulness of these TNM staging systems are, however, very much debated [5]. Some of the main concerns are as follows. First, the AJCC system, as well as other proposed systems based on TNM, represent only few of the numerous possible

Figure 1.1: Schematic showing the AJCC 6th edition staging system for colorectal cancer.



	T1	T2	T3	T4	M1
N0	1	1	2a	2b	4
N1	3a	3a	3b	3b	
N2	3c	3c	3c	3c	

combinations of the T, N and M categories. Yet they are defined without systematic empirical investigation. By systematic, we mean the extensive division of the table in Figure 1.1 into all possible staging systems. There is also a lack of commonly accepted statistical methods for developing staging systems. Although a vast literature in the medical community exists on cancer staging systems, they are solely focused on evaluating and comparing existing proposals of TNM groupings [6, 7]. There have been few literature reports of statistical techniques for developing cancer stage groupings. Begg and others considered the problem of comparing alternative stage groupings where three non model-based statistical criteria were applied and compared [8]. Hothorn and Zeileis used maximally selected log rank statistics to select the optimal two-class partition determined by the T and N categories of rectal cancer patients [9]. The maximal selection, however, are unstable with regard to small changes in the data [10]. These critiques apply to staging of all types of cancers.

Here and below, M1 patients will be omitted and relegated, as they usually are, to separate consideration. The reasons for this are two-fold: 1) M1 cancers are considered systemic diseases, as opposed to M0, which is considered localized; and 2) M1 has historically been the strong indicator of poor prognosis for almost all cancers. Our goal in this dissertation is to answer the question: does the AJCC staging scheme outperform other possible T and N combinations in prognosis? If not, what is the best system out of all possible T and N combinations?

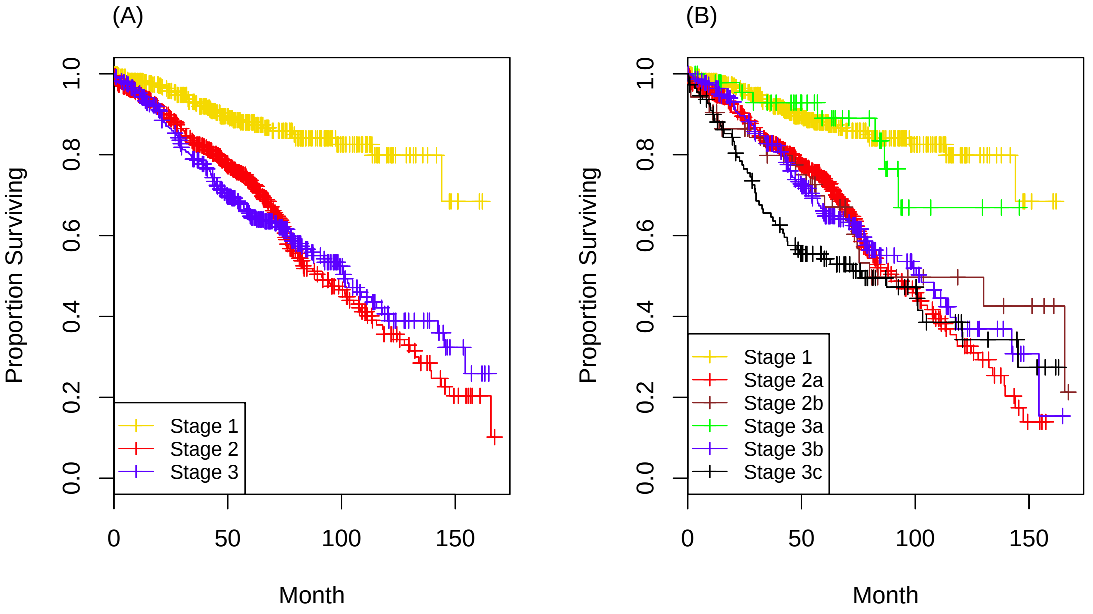
1.2 Motivating Example: Colorectal Cancer

Our motivating data example is the de-identified database of 1,326 patients with non-metastatic colorectal cancer treated at Memorial Sloan-Kettering Cancer Center (MSKCC) between January 1, 1990 and December 27, 2000 [11]. All patients are diagnosed with AJCC stage 1 to 3c disease (6th edition). The primary outcome used in the analysis is cancer-specific survival (only deaths attributable to recurrent cancer were counted as events). Of the 1,326 patients, 379 died by end of follow-up and the median survival was 115 months. Median follow-up time was 61.4 months. Table 1.2 presents the sample size, hazard ratio, and 10-year survival for each cell in the T×N table. With a couple of exceptions, apparently due to small sample sizes, there is a strong upward trend in risk with increasing T and N involvement. However, we observe a relatively poor separation of the Kaplan-Meier survival curves under the AJCC 6th edition staging system (Figure 1.2), which indicates that there might be room for improvement for the current AJCC system.

Table 1.2: Estimated cancer-specific 10-year survivals/hazard ratios by TNM classifications (sample size): colorectal cancer patients at Memorial Sloan-Kettering Cancer Center (MSKCC).

	T1	T2	T3	T4
N0	0.87/1.00 (213)	0.73/2.44 (209)	0.33/5.63 (468)	0.50/4.99 (53)
N1	0.83/0.93 (14)	0.57/2.54 (34)	0.36/5.56 (197)	0.58/6.00 (27)
N2	0.50/4.37 (3)	1.00/0.00 (5)	0.33/8.00 (81)	0.43/10.27 (22)

Figure 1.2: Cancer-survival of colorectal cancer patients at Memorial Sloan-Kettering Cancer Center (MSKCC), by the 6th edition AJCC staging system: (A) three main stages; (B) including sub stages.



As some committee members pointed out, the MSKCC data is small and might not be representative of the United States population. Some of the categories simply have too few samples to be considered representative, such as the categories T1N2

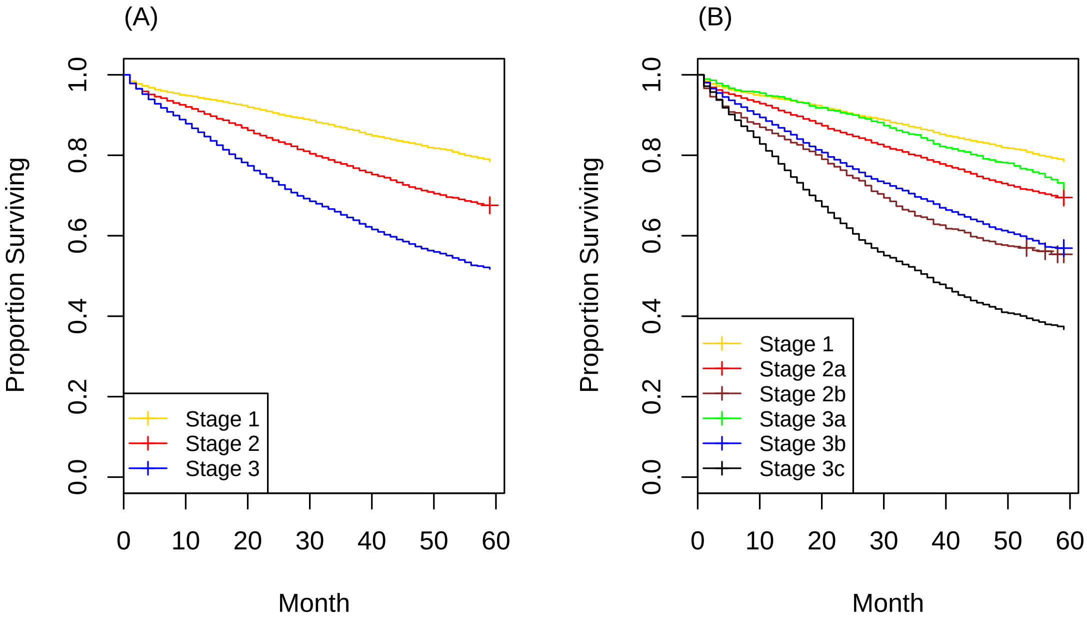
(3 samples) and T2N2 (5 samples). Therefore, a second, much larger data set is obtained and analyzed to illustrate the utility of our methods. We use data from the Surveillance, Epidemiology, and End Results (SEER) program, a large national cancer registry that collects patient records from multiple sites across the United States [12]. This national program includes 12 regional registries that cover approximately 14% of the U.S. population. The database was designed to reflect the overall characteristics of the U.S. population, including the diverse array of racial and/or ethnic groups, geographic locations, and types of cities and states [13].

For the purpose of illustration, we identify and evaluate 17,297 patients diagnosed with colon adenocarcinoma in the SEER national cancer registry the year of 2000 (January 1, 2000 - December 31, 2000). Mean age (± standard deviation) for the cohort is 68.7 ± 13.1 years. Females represent 49.9% of the group, and the overall racial and/or ethnic distribution is 84.2% whites, 8.4% blacks, and 7.4% other. All patients are diagnosed with AJCC stage 1 to 3c disease (6th edition). The primary outcome is cancer-specific survival. Of the 17,297 patients, 5,700 (33.0%) died by end of follow-up and the median follow-up time was 54 months. Overall 5-year colon cancer-specific survival for the cohort was 64.8%. Table 1.3 presents the sample size, hazard ratio, and 5-year survival for each cell in the T×N table. Unlike the MSKCC data, the SEER data, with ample samples, show a strong upward trend in risk with increasing T and N involvement with no exceptions. There is also a substantial separation of the Kaplan-Meier survival curves under the 3-stage AJCC system (Figure 1.3 (A)). However, when the sub-stages are considered, the survivals are again not in strict order with the stages (Figure 1.3 (B)).

Table 1.3: Estimated 5-year survivals/hazard ratios by TNM classifications (sample size): colorectal cancer patients identified in the Surveillance, Epidemiology, and End Results (SEER) national cancer registry.

	T1	T2	T3	T4
N0	0.81/1.00 (2048)	0.76/1.37 (2438)	0.70/1.88 (5773)	0.55/3.21 (920)
N1	0.77/1.34 (181)	0.70/1.58 (477)	0.60/2.63 (2855)	0.42/4.76 (555)
N2	0.73/1.70 (30)	0.62/2.54 (110)	0.38/4.82 (1471)	0.22/8.80 (439)

Figure 1.3: Survival of colorectal cancer patients identified in the Surveillance, Epidemiology, and End Results (SEER) national cancer registry, by the 6th edition AJCC staging system: (A) three main stages; (B) including sub stages.



1.3 Scope of this Dissertation

Searching for the best TNM grouping posed a challenging statistical problem. In this dissertation, we reframe the cancer staging problem into a model selection problem for censored response with respect to features arranged on a partially ordered two-way grid. The aim is to select the grouping that best classifies the patients. Two model selection methods are proposed for this task: 1) a bootstrap selection method, and 2) a $L1$ penalized regression method. We illustrate the utility of both methods by applying them to the staging of colorectal cancer. Considered a more promising method, the penalized regression method is further studied through simulations and its theoretical properties are developed. Extension of the penalized regression method to other applications is also studied.

The structure of the dissertation is as follows. The bootstrap selection method is described in Chapter 2, where we also introduce the three statistical criteria for evaluating cancer staging systems and illustrate the utility of the bootstrap method on the staging of colorectal cancer. In Chapter 3, we introduce the penalized regression method, which we call the lasso tree, and apply it to the colorectal cancer example. Simulation studies are carried out to examine the finite sample performance of the lasso tree selection procedure. The asymptotic theory of the lasso tree is also provided. Finally in Chapter 4, we turn to an extended application of the penalized regression method: the development of risk stratification models for colorectal cancer. Using this example, we illustrate how the penalized regression method can be modified to meet different modeling requirements and have applications to a wide range of disease areas and scientific questions.

Chapter 2

The Bootstrap Selection Method

2.1 Introduction

In this chapter, a bootstrap model selection method is proposed for the task of cancer staging based on the following considerations. First, not all T and N combinations are eligible staging systems. As both categories are ordinal, only those combinations are eligible which are ordered in T given N and vice versa. A search algorithm that satisfies this partial ordering rule is needed for generating all eligible staging systems. Second, the best staging systems can be simply defined as the ones that optimize the selection criterion chosen. Ideally, an external validation with a new population is desirable before determining the best system. In the absence of independently collected data, bootstrapping could be used as an alternative to provide replicate data sets for validating the selection [14]. Hence a bootstrap resampling strategy is proposed to estimate the optimal staging system, and to provide inference procedures (e.g. confidence intervals).

Selection criteria need to be identified that quantify the prognostic ability of candidate staging systems. A common approach for model development based on censored survival data is through the use of Cox proportional hazards model. Whereas the partial likelihood function as a statistical criterion is informative for looking at magnitude of effect, in certain clinical situations it might not be the most desirable option. It might be difficult to interpret for a non-statistician. Furthermore, since our problem is centered on evaluating prognostic classification schemes, which are inherently fully categorical and hence model-free, measures that check goodness-of-fit or that address model selection are less suitable for the task at hand. In view of these considerations, we elect to use measures that directly assess the prognostic ability of the staging systems. Several measures and ad hoc methods have been proposed for assessing prognostic ability; detailed reviews of these measures have been given by Schemper and Stare [15] and by Graf et al. [16], among others. In this paper, we elect to use the three criteria proposed by Begg et al. [8] and adapt them for comparison with our search algorithm: the explained variation for a specified "landmark" time, the area under the ROC curve for a landmark, and a concordance probability generalized from Harrell's c-index.

The rest of this chapter is structured as follows. In Section 2.2 we describe the proposed bootstrap selection method. The criteria for finding the optimal staging system are explained in Section 2.3. The method is then illustrated on the colorectal cancer example in Section 2.4. Discussions and conclusions are given in Section 2.5.

2.2 The Bootstrap Method

To identify the best staging system, we propose a search algorithm that scans through all eligible possibilities. In general, suppose the T descriptor has p categories, the N descriptor has q categories, and a k-stage system is desirable. The problem can be described by borrowing the framework of an outcome-oriented cutpoint selection problem for a censored response partitioned with respect to two ordered categorical covariates and their interaction. That is, we aim to estimate the "best" $k-1$ partition lines (cutpoints) that classify a partially ordered $p \times q$ table into k ordinal groups.

Calculating the number of all eligible partitions falls into the general mathematical problem of compositions of a grid graph [17], yet an analytical solution is not available for the general case. Numerical solutions can be obtained through computerized enumeration for small k, p, and q values, and they are given in Table 2.1 for small k's with $p = 4$ and $q = 3$, relevant to our colorectal cancer example.

TABLE 2.1: Number of eligible staging systems given k.

number of stage k	2	3	4	5	6
number of eligible systems	33	388	2,362	8,671	20,707

A value $n_{\min}$ is pre-specified for the minimum size of a stage, for example, $n_{\min}$ equals 5% of the sample size. Any system violating the $n_{\min}$ criterion will be dropped, and the remaining are the candidate systems. Let S denote the set of candidate systems, and let T_s denote the selection criterion value (see technical discussions in Section 2.3) for candidate system $s \in S$. The maximally selected criterion is

$$T_{\max} = \max_s T_s. \tag{2.1}$$

The maximally selected TN combination s^* is defined to be the one for which the maximum is attained, that is, for which the value of statistical criterion T_{s^*} equals $T_{\max}$ given k.

Our task differs from the usual cutpoint estimation problem which utilizes the maximally selected statistics. Under the maximally selected tests, the null hypothesis of interest is the independence between the covariate (to be dichotomized) and the response, and the estimation of a cutpoint comes after the rejection of the null hypothesis. This null hypothesis is irrelevant in our case as the prognostic ability of the T and N categories is well established and assumed to hold. Our inquiry takes one step further to ask, is the maximally selected TN combination s^* truly the optimal staging system for the population?

A bootstrap model selection strategy is therefore applied to estimate the optimal staging system. B bootstrap samples of size n (where n is the original sample size) are drawn with replacement from the original data. Denoting the bootstrap replication of $\hat{T}_s$ by $\hat{T}_s^b$, the bootstrap estimated criterion for candidate system s is given by

$$\tilde{T}_s = \frac{1}{B} \sum_{b=1}^{B} \hat{T}_s^b. \tag{2.2}$$

The bootstrap estimate of the best staging system $\tilde{s}^*$ is defined as the system that maximizes $\tilde{T}_s$.

There are two reasons that lead us to choose the bootstrap procedure:

1. The bootstrap method provides inference procedures (e.g. confidence intervals) for not only the optimal selected but all candidate systems, which enables us to examine the relative performance of any staging systems of interest and allows flexibility in the decision making process for clinical researchers and practitioners. In the analysis in Section 2.4, the standard bootstrap variance estimate was employed to construct the variance estimates of the measures for each candidate system, and the confidence intervals are also produced by bootstrapping.

2. The bootstrap selection procedure can easily adopt any measure of prognostic ability. In the next part of this chapter, three such measures will be introduced.

Note that a complete search through all eligible partitions for all bootstrap samples is, however, thwarted by combinational explosion. To overcome this problem, we can first compute the criterion for each eligible system using the complete data, and then only include the top m systems, say $m = 200$, (and the currently used staging system proposed by AJCC) as the "finalists" for the bootstrap selection procedure.

2.3 Criteria for Assessing Staging Systems

In this section, we discuss the three measures / criteria we choose for assessing the prognostic power of candidate staging systems.

2.3.1 Landmark Measures

An appealing way to simplify the analysis of survival data is to use a "landmark" time-point, such as 5-year or 10-year survival, and deal with only the censored binary outcome. This is frequently used in medical investigations. Here we elect to use the two landmark measures described by Begg et al. [8], the explained variation and the area under the ROC curve. Let θ_i, $i = 1, \ldots, c$, denote the probabilities of survival at the chosen landmark time or each of the c categories in the staging system, and p_i denote the prevalence of the stage categories. Let $\mu = \sum p_i\theta_i$ represent the unconditional mean outcome, and ν_i be the variance of θ_i. Then the estimated proportion of explained variation $\hat{\pi}$ is given by

$$\hat{\pi} = \frac{\sum \hat{p}_i\hat{\theta}_i^2 - (\sum \hat{p}_i\hat{\theta}_i)^2 - \sum \hat{p}_i\hat{\nu}_i}{(\sum \hat{p}_i\hat{\theta}_i)(1 - \sum \hat{p}_i\hat{\theta}_i)} \ , \tag{2.3}$$

and the area under the ROC curve $\hat{A}$ is estimated as

$$\hat{A} = \sum_{i=1}^{c} \frac{\hat{p}_i(1-\hat{\theta}_i)}{2\hat{\mu}(1-\hat{\mu})} \left\{ 2\sum_{j=1}^{i-1} \hat{p}_j\hat{\theta}_j + \hat{p}_i\hat{\theta}_i \right\} \tag{2.4}$$

where $\{\hat{\theta}_i\}$ are the Kaplan-Meier [18] estimates of the survival probabilities at the landmark time, $\{\hat{p}_i\}$ are the observed relative frequencies of the staging categories, and $\{\hat{\nu}_i\}$ are the variances of the observed values of $\{\hat{\theta}_i\}$ obtained from the Greenwood formula.

Using landmark times are less efficient statistically than using the entire survival distribution but provides for easier communication of results. In fact certain landmark times have become standards of reporting in various cancers such as 5 and 10 years in localized colorectal cancer. We include these measures also because in some situations landmark survival analysis can be more desirable than using the full survival. These include comparisons in which proportionality is obviously violated (e.g., when one stage is usually treated with a therapy which has a substantial immediate failure rate and another stage's failures tend to occur later) or those in which a landmark analysis is preferred for scientific reasons. An example of the latter might be a childhood cancer in which life extension is less relevant than the cure rate, and so a landmark measure such as 5-year survival could be used to stage these patients as a surrogate for cure.

2.3.2 Concordance Probability

Harrell et al. [19, 20] proposed the c-index as a way of estimating the concordance probability for survival data. It is defined as the probability that, for a randomly selected pair of participants, the person who fails first has the worse prognosis as predicted by the model. A limitation of Harrell's c-index is that it only takes into account usable pairs of subjects, at least one of whom has suffered the event. Begg et al. proposed an improved estimator of concordance which is adapted to account for all pairs of observations, including those for which the ordering of the survival times cannot be determined with certainty [8]. It requires the estimation of the probability of concordance for each pair of subjects and thus is computationally intensive for large sample sizes, particularly when bootstrapping. It also assumes that if the patient with the shorter censored value lives as long as the observed censored survival time in the paired patient, the remaining conditional probability of concordance is 1/2. As a result there is likely to be a conservative bias in the concordance estimator in the presence of high censoring rates [8].

Here we develop an estimator of the concordance probability under a classification scheme. Similar to Begg's approach, the new method utilizes the Kaplan-Meier estimates to evaluate the probabilities. Let K be the probability of concordance. For two patients randomly selected with stage (class) and survival time denoted by (S_1, T_1) and (S_2, T_2),

$$K = P\{(S_1 > S_2,\ T_1 < T_2) \text{ or } (S_1 < S_2,\ T_1 > T_2)\}. \tag{2.5}$$

Here we assume the survival time is inherently continuous although there could be ties in observed survival times. If $S1 = S2$, then the most common approach is to consider it equivalent to $S1 > S2$ with probability 1/2 and to $S1 < S2$ with probability 1/2. Thus (2.5) can be written as

$$K = 2P(S_1 > S_2,\ T_1 < T_2) + P(S_1 = S_2,\ T_1 < T_2). \tag{2.6}$$

Letting $S_1 = j$ and $S_2 = i$, $1 \le i < j \le k$, the first part of (2.6) can be estimated as

$$\begin{aligned}\hat{P}(S_1 > S_2,\ T_1 < T_2) &= \hat{P}(T_1 < T_2 | S_1 > S_2)\hat{P}(S_1 < S_2) \\ &= \sum\sum_{j>i} \hat{P}(T_1 < T_2 | j, i)\hat{P}(j, i) \\ &= \sum\sum_{j>i} \hat{P}(T_1 < T_2 | j, i)\frac{N_j N_i}{N(N-1)}\end{aligned} \tag{2.7}$$

where N_i, N_j are the sample sizes of stages i and j, respectively, and N is the total sample size.

Given i and j, and the last event time in all groups denoted by $t_{\max}$, we have

$$P(T_1 < T_2) = P(T_1 < T_2,\ T_1 \le t_{\max}) + P(T_1 < T_2,\ T_1 > t_{\max}). \tag{2.8}$$

When at least one event occurred,

$$\begin{aligned}P(T_1 < T_2,\ T_1 \le t_{\max}) &= \int_0^\infty dt_2 \int_0^{t_2} f_1(t_1) f_2(t_2) dt_1 \\ &= \sum_{t \in \{t_j\}} [S_j(t^-) - S_j(t)] S_i(t)\end{aligned} \tag{2.9}$$

where S_i and S_j can be estimated by the Kaplan-Meier survival estimators in stage i and j, and $\{t_j\}$ are the observed event times in stage j.

In the case when both observations are censored,

$$P(T_1 < T_2,\ T_1 > t_{\max}) = S_1(t_{\max}) S_2(t_{\max}) P(T_1 < T_2 | T_1, T_2 > t_{\max}). \tag{2.10}$$

The conditional probability $P(T_1 < T_2|T_1, T_2 > t_{\max})$ is not estimable, but can be conservatively assumed to be 1/2 as in Begg et al., or assumed to be equal to the overall concordance $P(T_1 < T_2)$. The latter is adopted in our method. That is,

$$\hat{P}(T_1 < T_2) = \frac{\sum_{t \in \{t\}_j} [\hat{S}_j(t^-) - \hat{S}_j(t)]\hat{S}_i(t)}{1 - \hat{S}_1(t_{\max})\hat{S}_2(t_{\max})} \ . \tag{2.11}$$

Similarly the second part of (2.6) can be estimated as

$$\hat{P}(S_1 = S_2,\ T_1 < T_2) = \frac{1}{2}\sum_i \frac{N_i(N_i - 1)}{N(N-1)} \tag{2.12}$$

and the overall concordance estimator is given by

$$\hat{K} = 2\sum\sum_{j>i}\left\{\frac{N_j N_i}{N(N-1)}\frac{\sum_{t \in \{t\}_j} [\hat{S}_j(t^-) - \hat{S}_j(t)]\hat{S}_i(t)}{1 - \hat{S}_1(t_{\max})\hat{S}_2(t_{\max})}\right\} + \frac{1}{2}\sum_i \frac{N_i(N_i - 1)}{N(N-1)} \ . \tag{2.13}$$

The new estimator improves upon Harrell's c-index, particularly in the presence of a large amount of censoring, by including comparisons between censored individuals. It is also much faster to implement than Begg's method. The statistic suffers from the usual criticism applied to concordance statistics; that is, they look only at the ranks of individuals and thus might be insensitive to small model improvements. Using survival times, however, often requires parametric modeling and alternative measures that are sensitive to small changes can also be sensitive to model choice.

Using ranks can also be a benefit in that K is robust to outlying observations.

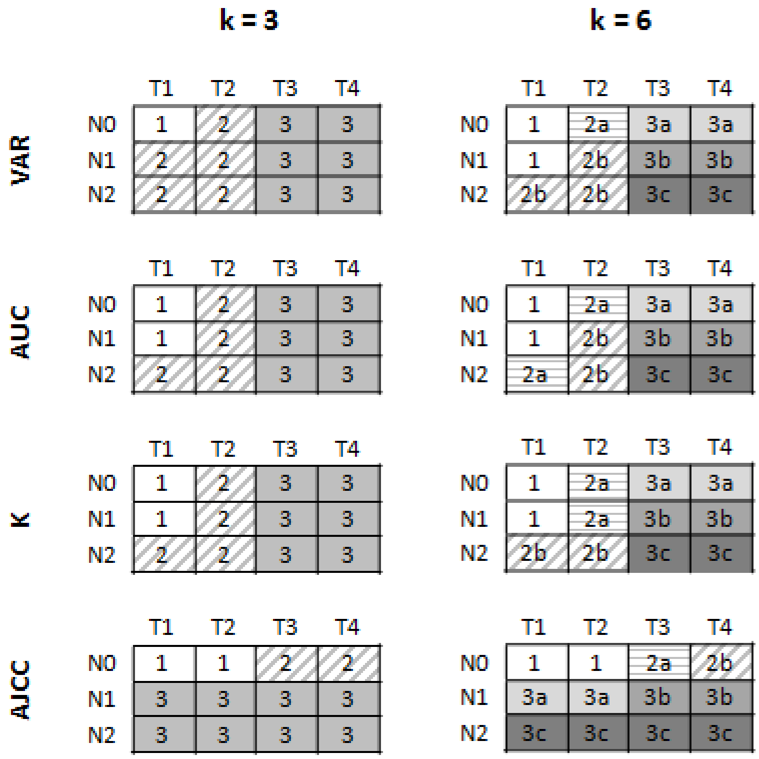
2.4 Application to Colorectal Cancer

We illustrated the utility of the proposed bootstrap method by applying it to the staging of colorectal cancer. The number of stages is given as $k = 3$ and $k = 6$, corresponding to numbers of main and sub-stages in the 6th edition AJCC staging system. For the percent explained variation and the area under the curve measures, we tried both landmark times of 5 years and 10 years, based on the median follow-up time, and the results are very similar. We hence report here only the results from the 5-year landmark analysis. We use the MSKCC data as the primary example and base our major presentation on them, including the bootstrap selection, the inferences, and the validations, in Sections 2.4.1 to 2.4.4. In Section 2.4.5 we describe our experiment with the SEER data.

2.4.1 MSKCC Data: Bootstrap Selection

The staging systems selected by maximizing the bootstrap estimates of each of the criteria described in Section 2.3, given $k = 3$ and $k = 6$, respectively, are presented in Figure 2.1, as well as the AJCC system for comparison. The systems selected by the three criteria are similar to each other and quite different from the AJCC system. Unlike the AJCC which separates stage 3 horizontally at N1, the bootstrap selected systems all classify groups primarily by the T categories (vertically). This is consistent with what we observe in Table 1.2, where the estimated 10-year survivals are much lower and the hazard ratios are much greater in categories T3 and T4.

FIGURE 2.1: Schematic showing staging systems selected by bootstrap based on the MSKCC data and the AJCC 6th edition staging system. VAR: explained variation; AUC: area under the ROC curve; K: concordance probability.



VAR, k = 3

	T1	T2	T3	T4
N0	1	2	3	3
N1	2	2	3	3
N2	2	2	3	3

VAR, k = 6

	T1	T2	T3	T4
N0	1	2a	3a	3a
N1	1	2b	3b	3b
N2	2b	2b	3c	3c

AUC, k = 3

	T1	T2	T3	T4
N0	1	2	3	3
N1	1	2	3	3
N2	2	2	3	3

AUC, k = 6

	T1	T2	T3	T4
N0	1	2a	3a	3a
N1	1	2b	3b	3b
N2	2a	2b	3c	3c

K, k = 3

	T1	T2	T3	T4
N0	1	2	3	3
N1	1	2	3	3
N2	2	2	3	3

K, k = 6

	T1	T2	T3	T4
N0	1	2a	3a	3a
N1	1	2a	3b	3b
N2	2b	2b	3c	3c

AJCC, k = 3

	T1	T2	T3	T4
N0	1	1	2	2
N1	3	3	3	3
N2	3	3	3	3

AJCC, k = 6

	T1	T2	T3	T4
N0	1	1	2a	2b
N1	3a	3a	3b	3b
N2	3c	3c	3c	3c

Let A1 and A2 denote the 3-stage systems selected by explained variation, and by area under the curve and concordance probability, respectively, and let B1, B2, and B3 denote the 6-stage systems selected by the three criteria, respectively. Table

Table 2.2: Selected systems and the AJCC: the estimated criteria and their standard errors (MSKCC data).

		Criteria (SE)		
	System	VAR	AUC	K
k = 3	A1	0.684 (0.010)	0.705 (0.011)	0.662 (0.008)
	A2	0.684 (0.010)	0.705 (0.012)	0.663 (0.007)
	AJCC	0.627 (0.008)	0.643 (0.011)	0.622 (0.008)
k = 6	B1	0.688 (0.009)	0.708 (0.011)	0.667 (0.008)
	B2	0.688 (0.009)	0.709 (0.011)	0.666 (0.008)
	B3	0.687 (0.009)	0.708 (0.011)	0.667 (0.007)
	AJCC	0.642 (0.012)	0.660 (0.013)	0.623 (0.008)

2.2 shows the estimated value of the three criteria for these selected systems and the AJCC. The bootstrap selected systems are very similar with respect to all three criteria, which is not surprising given that the systems highly resemble each other. The prognostic power increases minimally as the number of stages increase from 3 to 6, indicating there is not much to gain by adding more sub-stages. In addition, of course, a 3-stage system is easier to use than a 6-stage one. The AJCC system is inferior to the selected ones in all cases.

Kaplan-Meier survival curves for the selected staging systems are displayed in Figure 2.2. All five systems show a substantial degree of prognostic separation and a clear advantage over the AJCC system in Figure 1.2. There is considerable overlap of survival curves in the right panel because of the larger number of stages, which again raise the question whether 6 distinct stages are too many.

FIGURE 2.2: Cancer-specific survival of colorectal cancer patients by the selected staging systems (MSKCC data). Left panel: 3-stage systems; right panel: 6-stage systems.

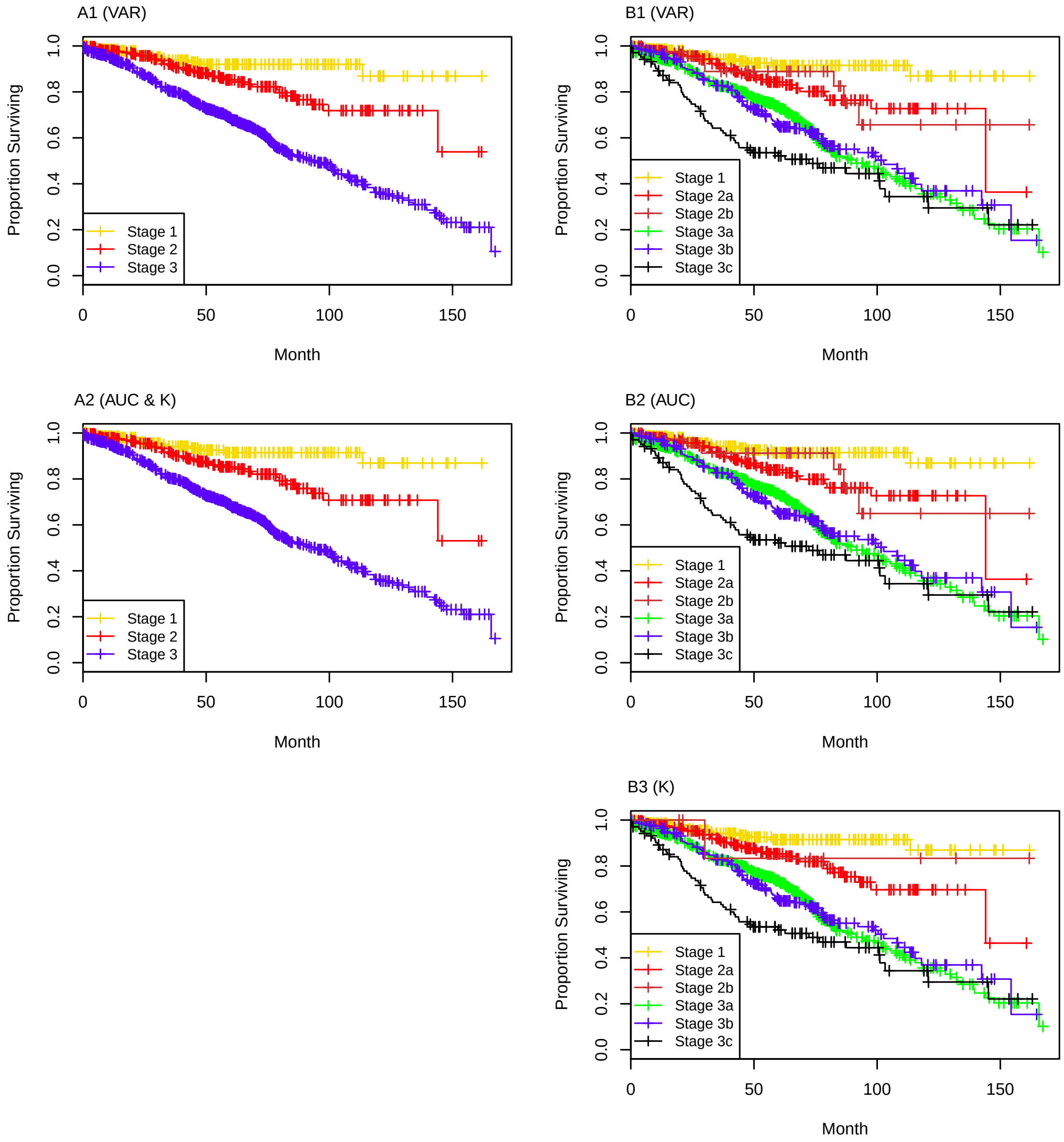


Figure 2.3: Confidence intervals for the criteria: the top-ranked systems based on the MSKCC data and the AJCC (red). Left panel: 3-stage systems; right panel: 6-stage systems.

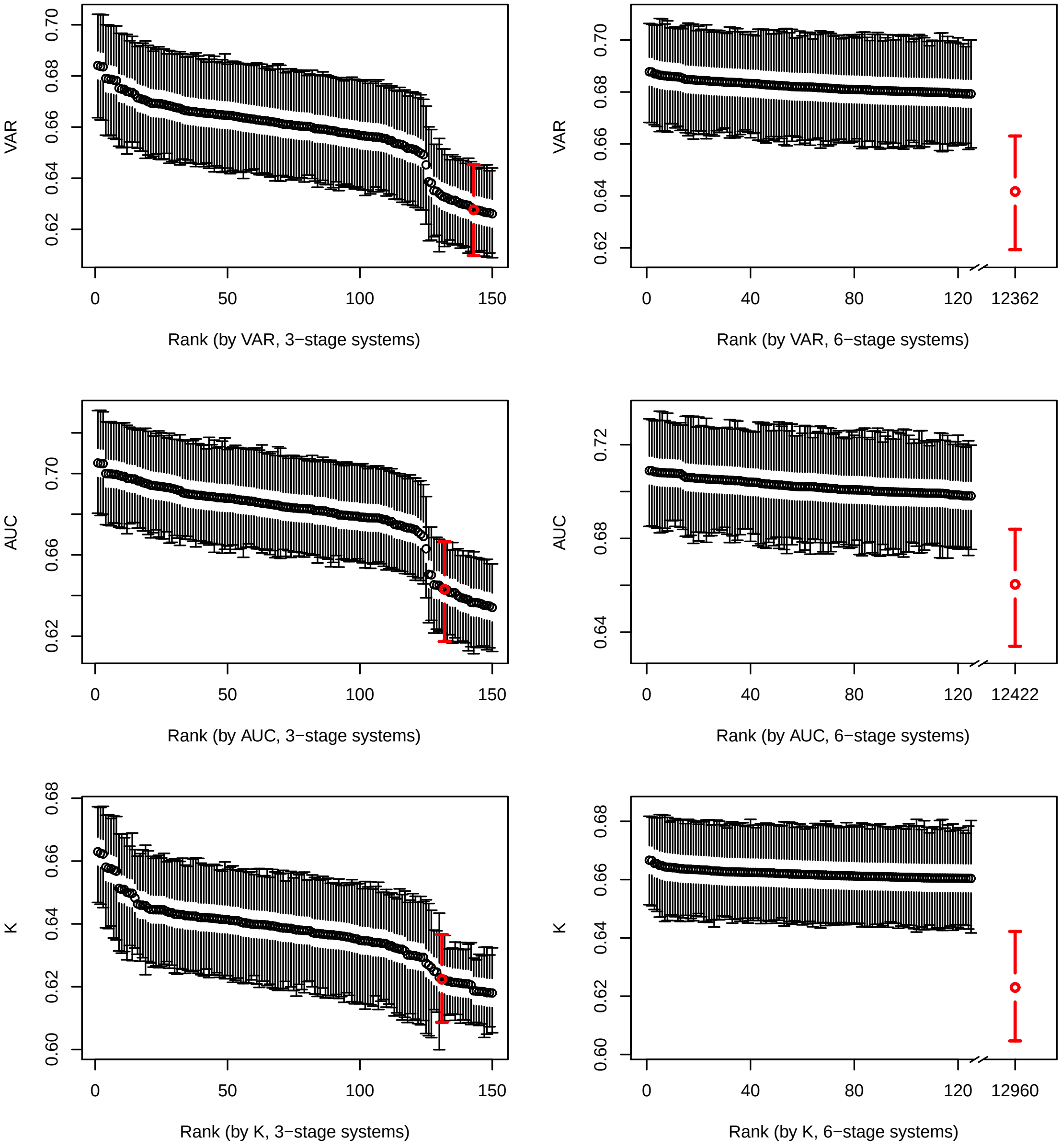
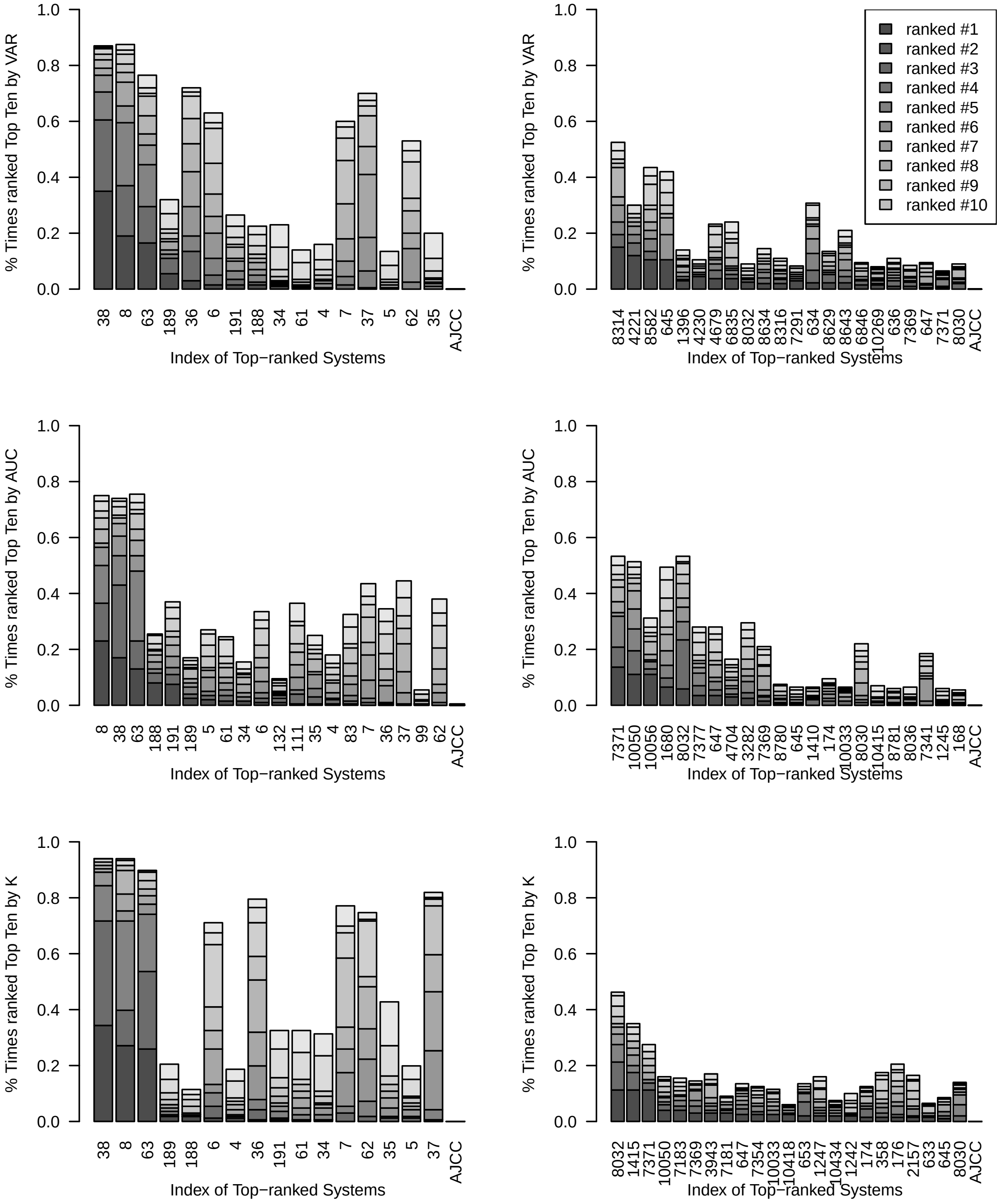


FIGURE 2.4: Majority voted systems. Grey bars show the % time each system is ranked from #1 to #10, and the systems are ordered by their % time ranked #1. The x-axis represents the index of the candidate systems. Left panel: 3-stage systems; right panel: 6-stage systems.



2.4.2 MSKCC Data: Confidence Intervals

The bootstrap selection provides inference procedures for not only the optimal selected, but all candidate systems. Figure 2.3 shows the confidence intervals of each of the criteria for the top-ranked systems and the AJCC. The systems are ordered by their rankings with regard to the bootstrap estimated criteria. The top systems are in fact very close in terms of prognostic power, especially for 6-stage systems where the top 100 systems are virtually identical due to the fact that the systems are only slightly different from one another in their definition. It is hence difficult to select one best system, but it allows flexibility in the decision making process for clinical researchers, who might incorporate both statistical evidence and medical insight into their considerations. Among the 388 3-stage systems (only the top 150 shown), the AJCC ranks around 135 (35%), and it ranks around 12500 (60%) among the 20707 6-stage systems (the top 125 shown). Again, the AJCC demonstrates clearly lower prognostic power than the top systems.

2.4.3 MSKCC Data: Majority Vote Rule

Instead of choosing the system to maximize the bootstrap estimated criteria, the optimal staging system can also be decided by a simple majority vote rule. For each bootstrap replication, the candidate systems are ranked with respect to the criterion, and the systems ranked top 10 at least 5% of the time are defined to be "good" staging systems. We show these "good" systems in Figure 2.4. There is an obvious differential effect for 3-stage systems; the top-3 ranked systems are consistent under all criteria and enjoy a majority of votes. This is not the case for 6-stage systems whose votes are very widely spread.

The x-axis represents the index of the candidate systems. Among the 3-stage systems, candidates #38, #8, and #63 are constantly selected as the top systems, showing a clear advantage over other candidates. Candidates #38 and #8 are in fact systems A1 and A2, respectively, in Section 5.1. The top-ranked 6-stage systems, #8314, #7371, and #8032, correspond to systems B1, B2, and B3, respectively. The AJCC system is included but it is never ranked top 10.

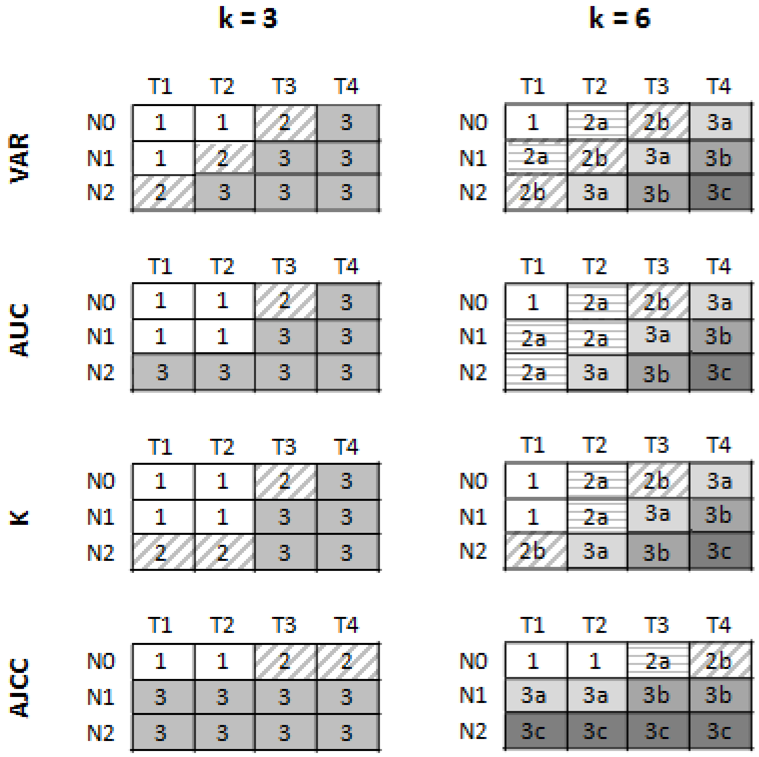
2.4.4 MSKCC Data: Cross-Validation

Table 2.2 compares maximally selected values with the one given by fixed model, the AJCC, without adjusting for the maximization process. This could result in maximization bias that elevates the performance of the bootstrap selected systems, a problem well known in the context of model selection or cutpoint selection [21–23]. One approach for correcting the maximization bias is the use of cross-validation [21]. Here we use 10-fold cross-validation to reevaluate the bootstrap procedure. Basically, the data are randomly split into ten parts of similar size. Ten times we use 9/10 of the data for selection and each time apply the selected system to the omitted 1/10th of the data. Once the procedure is complete, all patients in the sample have been assigned to a stage. We then compute the estimated criteria under this staging assignment. For example, here we use the bootstrap selection procedure with the concordance criterion and set the desired number of stages to be 3. With the cross-validation adjustment, the estimated criteria are 0.669 (VAR), 0.688 (AUC), and 0.650 (K), which are still substantially superior to the estimates for AJCC in Table 2.2, indicating there is much room for improvement in the current system.

2.4.5 The SEER Data Example

As a tertiary cancer center, the Memorial Sloan-Kettering Cancer Center does not typically have a representative sample of all cancer patients. In fact, as one of the top cancer centers in the U.S., the MSKCC is very likely to see a different patient population from the general colon cancer population. Analyses of data that are more representative of the U.S. population are needed before a valid recommendation could be made about improving cancer staging systems. Given the limited scope of this dissertation, we will only present here a brief example using the national population-based data from the SEER cancer registry. More detailed explorations using such national databases are very desirable in future publications in statistical journals and medical journal as well.

Figure 2.5: Schematic showing staging systems selected by bootstrap based on the SEER data and the AJCC 6th edition staging system. VAR: explained variation; AUC: area under the ROC curve; K: concordance probability.



k = 3

VAR

	T1	T2	T3	T4
N0	1	1	2	3
N1	1	2	3	3
N2	2	3	3	3

AUC

	T1	T2	T3	T4
N0	1	1	2	3
N1	1	1	3	3
N2	3	3	3	3

K

	T1	T2	T3	T4
N0	1	1	2	3
N1	1	1	3	3
N2	2	2	3	3

AJCC

	T1	T2	T3	T4
N0	1	1	2	2
N1	3	3	3	3
N2	3	3	3	3

k = 6

VAR

	T1	T2	T3	T4
N0	1	2a	2b	3a
N1	2a	2b	3a	3b
N2	2b	3a	3b	3c

AUC

	T1	T2	T3	T4
N0	1	2a	2b	3a
N1	2a	2a	3a	3b
N2	2a	3a	3b	3c

K

	T1	T2	T3	T4
N0	1	2a	2b	3a
N1	1	2a	3a	3b
N2	2b	3a	3b	3c

AJCC

	T1	T2	T3	T4
N0	1	1	2a	2b
N1	3a	3a	3b	3b
N2	3c	3c	3c	3c

The bootstrap selected optimal systems based on the SEER data, as expected, are slightly different from those based on the MSKCC data due to the inherent differences between these two populations. Figure 2.5 displays the selected systems under each statistical criterion, given $k = 3$ and $k = 6$, respectively, as well as the AJCC system for comparison. Again, the systems selected by the three criteria are similar to each other and quite different from the AJCC system. Unlike the AJCC system or the optimal systems selected based on the MSKCC data which separate stages in "straight lines", either horizontally or vertically, the optimal systems obtained from the SEER data partition the T×N table into more irregular shapes. Many of the stages lie along the diagonals in the T×N rectangular. This demonstrates the flexibility of the bootstrap selection method; any “eligible” combinations of the T and N categories that satisfy the partial ordering constrains can be a potential optimal system. Other than the partial ordering constraints we explicitly and intentionally avoid starting the process with judgments about which groups should be combined. On the other hand, these more irregularly configured systems might raise a concern as to whether they are clinically practical or useful, in which case it is necessary to seek the input from clinical and medical experts.

Table 2.3: Selected systems and the AJCC: the estimated criteria and their standard errors (SEER data).

		Criteria (SE)		
	System	VAR	AUC	K
k = 3	C1	0.634 (0.001)	0.665 (0.002)	0.612 (0.002)
	C2	0.633 (0.001)	0.666 (0.002)	0.613 (0.002)
	C3	0.634 (0.002)	0.666 (0.002)	0.613 (0.002)
	AJCC	0.625 (0.002)	0.655 (0.001)	0.604 (0.002)
k = 6	D1	0.646 (0.002)	0.677 (0.002)	0.624 (0.002)
	D2	0.645 (0.001)	0.677 (0.002)	0.624 (0.002)
	D3	0.646 (0.002)	0.676 (0.001)	0.625 (0.002)
	AJCC	0.635 (0.002)	0.664 (0.002)	0.614 (0.002)

Let C1, C2 and C3 denote the 3-stage systems selected by explained variation, by area under the curve, and by concordance probability, respectively, and let D1, D2, and D3 denote the 6-stage systems selected by the three criteria, respectively. Table

2.3 shows the estimated values of the three criteria for these selected systems and for the AJCC. The bootstrap selected systems are very similar with respect to all three criteria, which is not surprising given that the systems highly resemble each other. The prognostic power increases slightly as the number of stages increases from 3 to 6. The AJCC system is inferior to the selected ones in all cases.

Kaplan-Meier survival curves for the selected staging systems are displayed in Figure 2.6. All six systems show a substantial degree of prognostic separation. When 3-stage systems are of concern, the advantage of the selected systems over the AJCC (Figure 1.3) is not visibly apparent, despite the fact that the AJCC ranks around 100th (26%) amongst all 388 candidate systems. In fact, for SEER data, most of the eligible 3-stage combinations give similarly good separations in survivals. For instance, the system ranks 260 out of all 388 eligible 3-stage systems has the survival separation as shown in Figure 2.7. The reason lies in the fact that the partial ordering rule already guarantees that any eligible system is well ordered in terms of disease severity, and that the large sample size of the SEER data ensures that any grouping would have sufficient sample size in each stage (in a 3-stage system). In other words, randomness is largely removed. The same cannot be said for systems with more stages, for example, 6-stage systems where randomness could be preserved in stages with small numbers of subjects. The gain in prognostic separation is considerable for 6-stage systems as shown in Figure 2.6 (right panel).

Figure 2.6: Cancer-specific survival of colorectal cancer patients by the selected staging systems (SEER data). Left panel: 3-stage systems; right panel: 6-stage systems.

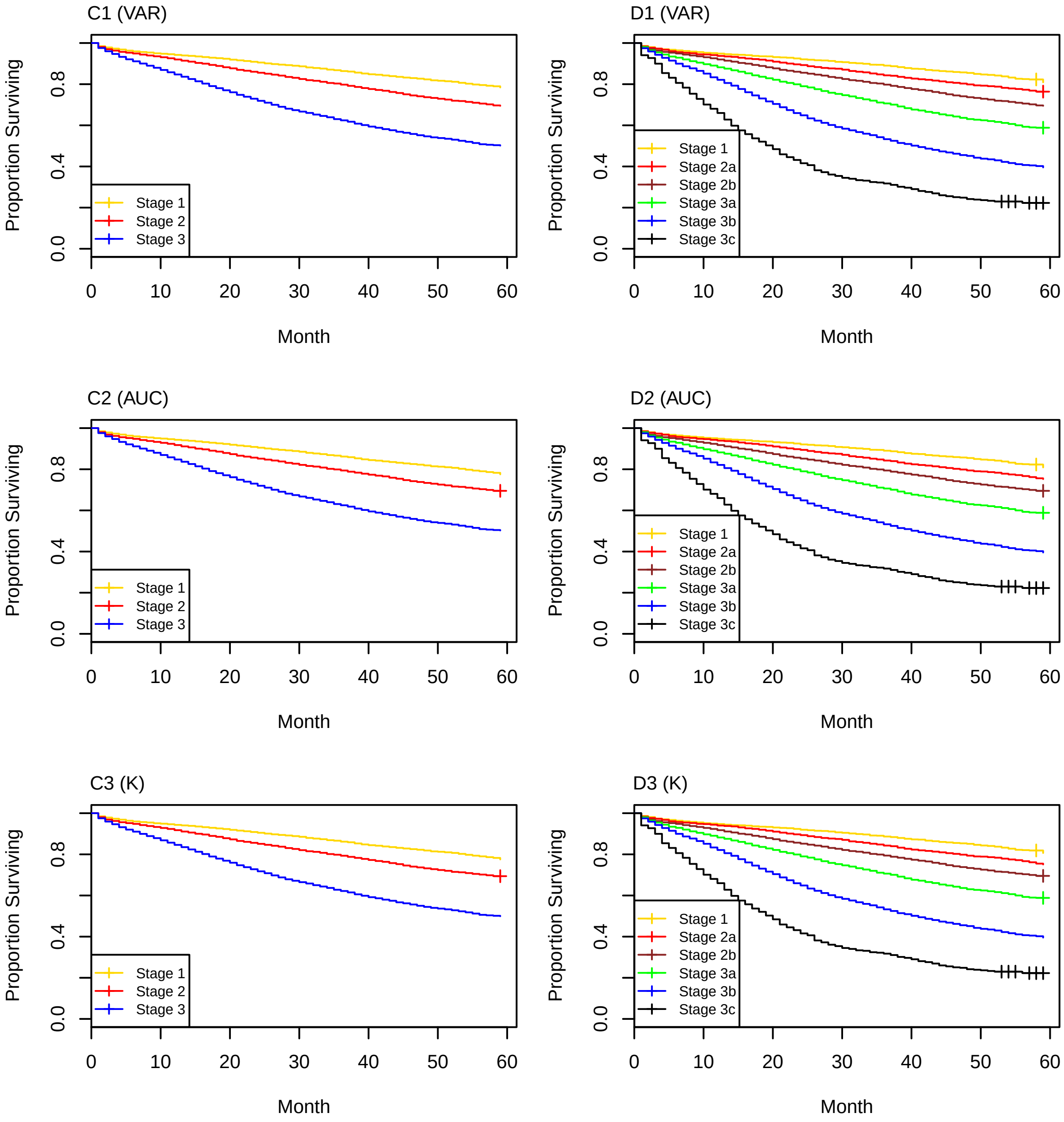
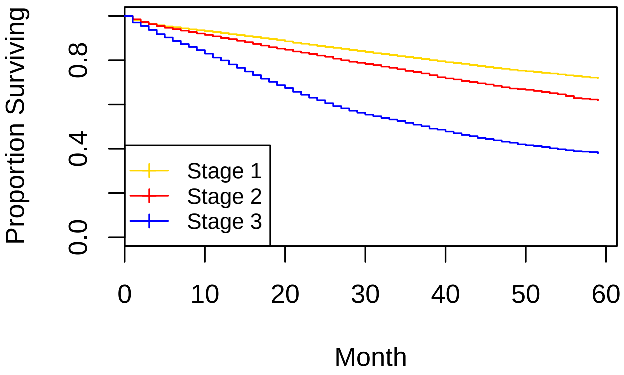


Figure 2.7: Kaplan-Meier survival curves based on the 260th ranked 3-stage system out of 388 eligible systems (SEER data).



2.5 Discussion and Conclusions

An accurate staging system is crucial for predicting patient outcome and guiding treatment strategy. For decades investigators have developed and refined stage groupings using a combination of medical knowledge and observational studies, yet there appears to be no well established statistical method for objectively incorporating quantitative evidence into this process. In this chapter, we have proposed a systematic selection method for the development of cancer staging systems, and illustrated the utility of this method by applying it to the staging of colorectal cancer. The staging systems selected by the three criteria are similar to each other while quite different from and superior to the current AJCC system, indicating there might be room for improvement in selecting it.

Our analysis of the colorectal cancer data has provided some insight into the prognostic power of the TNM staging system. For example, the selected systems based on the MSKCC data (A1, A2, B1, B2, and B3) are virtually identical in their prognostic accuracy regardless of which of the three evaluative measures is used. The selected 6-stage systems are a further division of the 3-stage systems with no apparent improvement in separating the survivals. The findings from the SEER data suggest similar conclusions. Thus, it might be reasonable to favor a more parsimonious system as urged in Gönen and Weiser [4]. Naturally, the choice between systems with three, six, or other numbers of stages involves a variety of considerations. The solution is an essentially medical one which combines issues of treatment regimen distinctions, diagnostic ease, and clinical practice. We recommend that an analyst give medical researchers several staging systems in a range of practical sizes along with their performance score as in Table 2.2 to allow them to compare the systems' prognostic capabilities.

We use bootstraps to provide bias-corrected estimates of performance for the staging systems. This addresses the internal validity which is a prerequisite for external validity yet does not guarantee it. External validity of a prognostic system can be established by being tested and found accurate across increasingly diverse settings. The selected systems should be tested across multiple independent investigators, geographic sites, and follow-up periods for accuracy and generalizability. The use of population-based datasets is important in establishing a staging system that is useful for the general patient population.

All final systems obtained from the MSKCC data suggest that the most essential information is contained in the contrast between the tumor invading through the muscularis propria (T3 and T4) and otherwise (T1 and T2). This is in sharp contrast to AJCC where the primary distinction is between node-positive (N1 and N2) and node-negative (N0) cancers. On the other hand, the findings from the SEER data suggest otherwise: the groupings primarily lie along the diagonals of the T×N table. The differences in the results come from the inherent differences between the two populations, which again prompts the need for external validations across independent populations.

As we mentioned, the relatively irregularly configured systems obtained from the SEER data could raise concerns as to whether they are clinically practical or useful. Because cancer staging has been as much about anatomic interpretation as it is about accurate prognosis, a staging system that is prognostically optimal is unlikely to be adopted if it does not respect the anatomic extent of disease. Staging systems are most useful when they are both prognostically optimal and anatomically interpretable. If such anatomical interpretability requires a more "regular" configuration, we can always identify, from the confidence interval plots or the bar plots, other near-top systems that satisfy such requirement, and a compromise can be reach between statistical evaluation and medical preference and common sense. Another reasonable proposal would be to use clinically sensible combinations as a way to constrain which groupings are allowed from the beginning rather than in the end. After all, there is nothing in our selection method which forbids certain ways of grouping or combination being pre-specified.

TNM staging is applicable to virtually any type of solid tumor hence, although we used colorectal cancer as illustration, our methodology has general appeal to other cancers and other diseases that use aggregate risk scores based on ordinal (or ordinalized) risk factors, such as the ATP III score for high-blood cholesterol that can benefit from optimal aggregation [24]. Our methodology is applicable in principle to binary outcomes as well.

Chapter 3

Penalized Regression Method: The Lasso Tree

3.1 Introduction

The bootstrap selection method for cancer staging suffers from the following drawbacks: 1) the number of stages needs to be pre-specified; 2) a complete search through all eligible partitions is thwarted by combinational explosion when the desired number of stages is large; and 3) the procedure, essentially a form of best subset model selection, could be unstable with regard to small changes in the data [10].

An intuitive approach that could speed up the computation is to use tree-based methods such as recursive partitioning [25]. However, these tree-based methods do not capture all types of T and N combinations. At each split, trees must have full / complete separation with respect to one variable. That is, in the T×N table, a tree method will split fully along all columns or rows conditional on the existing splits. In other words, the grouping that results from a tree must have partitions in straight lines. Yet the true staging system might have a different configuration. For instance, the newly published AJCC 7th edition has a partition along the diagonal [2], which can not be achieved by a tree method. Hence, a more flexible and less computationally intensive method is needed for estimating cancer stage groupings.

In this chapter, we propose a $L1$ penalized regression method that satisfies these requirements. Specifically, the development of cancer stage groupings can be considered a model selection problem for a censored response grouped with respect to features arranged on a partially ordered two-way grid (the T×N table), with the T and N categories partial ordered. An attractive way to reduce the time complexity of an exhaustive search is to introduce an $L1$ penalty in a regression model. In order to yield the grouping effect, we constrain the differences of coefficients that are one unit apart in both directions to be small. To be specific, we require

$$\sum |\beta_{j,k} - \beta_{j,k-1}| + \sum |\beta_{j,k} - \beta_{j-1,k}| \leq s \tag{3.1}$$

where $\beta_{j,k}$ is the coefficient for the cell with T $= j$ and N $= k$, and $s > 0$ is a pre-specified tuning parameter. The constraint leads to some of the estimated coefficients being exactly the same, which provides the desired stage grouping. An attractive feature of this method is that a series of optimal groupings with different numbers of groups can be obtained as a function of the tuning parameter s. This gives a tree-like structure for partitioning the T×N table, and thus offers doctors and medical researchers a visual aid on how the groupings are made progressively and a freedom to choose among different numbers of stages. We use the term "lasso tree" for the proposed method.

The continuous constraint function shrinks the difference of coefficients toward zero continuously and is expect to result in a more stable stage grouping than that provided by best subset selection [26]. This strategy has been proved to be appropriate in other statistical problems such as fused lasso [27, 28]. Unlike the fused lasso, the lasso tree only focuses on sparsity in the differences of the coefficients but not the coefficients themselves. It also takes into account the partial ordering characteristic of the T and N categories.

The structure of this chapter is as follows. In Section 3.2 we describe the lasso tree and the algorithm for obtaining the estimates. The method is illustrated on the colorectal cancer data example in Section 3.3, where we also show that the proposed method can incorporate information from the AJCC or other sources by posing different weights on the penalty terms. Simulation studies comparing the lasso tree with the best subset selection methods are presented in Section 3.4. Section 3.5 establishes the asymptotic properties of the penalized likelihood estimators. Some discussion is given in Section 3.6.

3.2 The Lasso Tree

3.2.1 A Lasso-type Selection Procedure for Survival Outcomes

In general, suppose the T descriptor has p categories and the N descriptor has q categories. The data for n subjects are of the form $(y_1, \delta_1, X_1), \ldots, (y_n, \delta_n, X_n)$, with δ_i describing whether y_i is a survival time ($\delta_i = 1$) or a censoring time ($\delta_i = 0$) and X_i denoting the vector of covariates for the i^{th} individual. Under the Cox proportional hazards model, the T×N table can be seen as a categorical covariate with $p \times q$ levels and the model can be written as

$$\lambda(t) = \lambda_0(t)\text{exp}(\beta^T X) \tag{3.2}$$

where $\beta = \{\beta_{j,k}\}, j = 1, ..., p, k = 1, ..., q,$ are the regression coefficients for cells in the T×N table and $\lambda_0(t)$ is an unspecified baseline hazard function. Equation (3.2) can be solved through maximizing the partial likelihood function

$$L(\beta) = \prod_{r \in D} \frac{\exp(\beta^T X_r)}{\sum_{j \in R_r} \exp(\beta^T X_j)} \tag{3.3}$$

where D is the set of indices of the events and R_r denotes the set of indices of the individuals at risk at time $t_r - 0$.

The grouping problem can be addressed by borrowing the framework of a lasso-type model selection problem; instead of estimating β such that some of its components are exactly 0 as in the usual implementation of the lasso, we aim to estimate β such that some of the solution coefficients are exactly the same. Hence, instead of constraints on the coefficients we pose constraints on the differences between neighboring coefficients. For the stage grouping problem specifically, β is ordered in both T and N directions such that $\beta_{j,1} \leq \ldots \leq \beta_{j,q}$ and $\beta_{1,k} \leq \ldots \leq \beta_{p,k}$, $j = 1, ..., p$, $k = 1, ..., q$, and this partial ordering constraint is also applied. We propose to estimate β as follows

$$\begin{aligned} \hat{\beta} = \operatorname{argmax} \ell(\beta), \quad \text{subject to} \quad & \sum_{j=1}^{p} \sum_{k=2}^{q} |\beta_{j,k} - \beta_{j,k-1}| + \sum_{j=2}^{p} \sum_{k=1}^{q} |\beta_{j,k} - \beta_{j-1,k}| \leq s \\ \text{and} \quad & \beta_{j,1} \leq \ldots \leq \beta_{j,q} \ \text{ and } \ \beta_{1,k} \leq \ldots \leq \beta_{p,k}, \ \ j = 1, ..., p, \ k = 1, ..., q \end{aligned} \tag{3.4}$$

or, equivalently,

$$\hat{\beta} = \text{argmin}\left\{-\ell(\beta) + \lambda(\sum_{j=1}^{p}\sum_{k=2}^{q}|\beta_{j,k} - \beta_{j,k-1}| + \sum_{j=2}^{p}\sum_{k=1}^{q}|\beta_{j,k} - \beta_{j-1,k}|)\right\}$$
$$\text{subject to} \quad \beta_{j,1} \leq \ldots \leq \beta_{j,q} \;\text{ and }\; \beta_{1,k} \leq \ldots \leq \beta_{p,k}, \quad j = 1, ..., p,\; k = 1, ..., q \tag{3.5}$$

where $\ell(\beta) = \text{log}L(\beta)$ and $s, \lambda > 0$ are tuning parameters. The underlying mechanism is the sparsity-enforcing property of the $L1$ penalty, which is expected to give a reduced number of unique $\beta_{j,k}$ values that represent different groups.

3.2.2 Computational Approach

If each neighboring difference is penalized equivalently, as in (3.4) and (3.5), because of the ordering constraint, the absolute values can be dropped and the objective function can be simplified as

$$\min_{\beta}\left\{-\ell(\beta) + \lambda\Big(-\sum_{j=1}^{p-1}\beta_{j,1} - \sum_{k=1}^{q-1}\beta_{1,k} + \sum_{j=2}^{p}\beta_{j,q} + \sum_{k=2}^{q}\beta_{p,k}\Big)\right\}$$
$$\text{subject to} \quad \beta_{j,1} \leq \ldots \leq \beta_{j,q} \;\text{ and }\; \beta_{1,k} \leq \ldots \leq \beta_{p,k}, \quad j = 1, ..., p,\; k = 1, ..., q. \tag{3.6}$$

Note that only the coefficients of the "boundary cells" in the T×N table are taken into account in (3.6). Yet this is mathematically equivalent to (3.4) and (3.5) and will give the same estimates as (3.4) and (3.5).

Tibshirani gave an iterative procedure to solve the $L1$ penalized Cox proportional hazards model by expressing the usual Newton-Raphson update as an iterative reweighted least squares (IRLS) step and then replacing the weighted least squares step by a constrained weighted least squares procedure [26]. Since our problem does not involve high-dimensional data, this procedure is quite adequate for computing its estimates. Define $\eta = X\beta$, $u = \partial\ell/\partial\eta$, $A = -\partial^2\ell/\partial\eta\eta^T$, and $z = \eta + A^{-1}u$. Denote $P_\lambda(\beta) = \lambda(-\sum_{j=1}^{p-1}\beta_{j,1} - \sum_{k=1}^{q-1}\beta_{1,k} + \sum_{j=2}^{p}\beta_{j,q} + \sum_{k=2}^{q}\beta_{p,k})$. The iterative procedure is as follows:

1. Fix λ and initialize $\hat{\beta} = 0$.

2. Compute η, u, A and z based on the current value of $\hat{\beta}$.

3. Minimize $(z - X\beta)^T A(z - X\beta) + P_\lambda(\beta)$ subject to $\beta_{j,1} \leq \ldots \leq \beta_{j,q}$ and $\beta_{1,k} \leq \ldots \leq \beta_{p,k}$.

4. Repeat Steps 2 and 3 until convergence of $\hat{\beta}$.

With the absolute values dropped, the minimization in Step 3 is a simple quadratic program with linear inequality constraints. It requires $O(n^2)$ computations with A being a full matrix. To speed up the computation, one can replace A with a diagonal matrix D with the diagonal entries of A [26]. The procedure typically runs between p and q iterations, and converges quickly based on our empirical experience.

3.2.3 The Tuning Parameter and The Lasso Tree

The estimates from (3.5) depend on the tuning parameter λ. When $\lambda = 0$, the solution is the usual Cox model estimate. As λ increases the absolute differences between neighboring coefficients go to 0 successively, corresponding to the successive grouping of the cells, until all cells are in one group. This is similar to pruning a tree bottom-up and hence we call the proposed method the lasso tree. The estimated coefficients from the lasso tree fit can be displayed as a function of the tuning parameter λ; an example is given in Section 3.3. "Warm starts" are used to efficiently compute the path of solutions over a grid of values for λ: starting at a solution for the previous λ, the solution for the next λ can be found relatively quickly.

We propose to use the Bayesian Information Criterion (BIC) [29]

$$\mathrm{BIC}(\lambda) = -2\ell(\hat{\beta}_\lambda) + k_\lambda \mathrm{ln}(n) \tag{3.7}$$

to select the tuning parameter λ, where $\ell(\hat{\beta}_\lambda)$ is the log-partial likelihood for the constrained fit with λ, and k_λ is the degrees of freedom in the model. Here we estimate k_λ by the number of unique parameters (the number of groups identified). Intuitively, the BIC inflates the negative log-partial likelihood by a penalty term proportional to the effective number of parameters. The BIC is calculated over a grid of values of λ which are uniformly distributed on the log-scale from 0 (yielding

$p \times q$ stages) to some big number (reducing to a single stage), and the value $\hat{\lambda}$ yielding the lowest estimated BIC is selected.

Other popular methods for tuning parameter selection include the Akaike Information Criterion (AIC) [30], cross-validation (CV), and generalized cross-validation (GCV) [31]. It is known that the BIC is consistent for model selection while the CV, GCV and AIC are not [32]. Since our primary focus is model selection rather than prediction we elect to use BIC as the tuning parameter selection criterion in this paper. In fact, simulation studies have shown that the GCV statistic is inferior to the BIC in terms of selecting the correct grouping. In addition, AIC tends to select less sparse groupings compared to BIC.

3.3 Application to Colorectal Cancer

In this section, we illustrate the utility of the lasso tree method by applying it to the staging of colorectal cancer. Again, we use the MSKCC data as the primary example and base our major presentation on them, including the lasso tree selection and the incorporation of prior information, in Section 3.3.1 and 3.3.2. In Section 3.3.3 we describe our experiment with the SEER data.

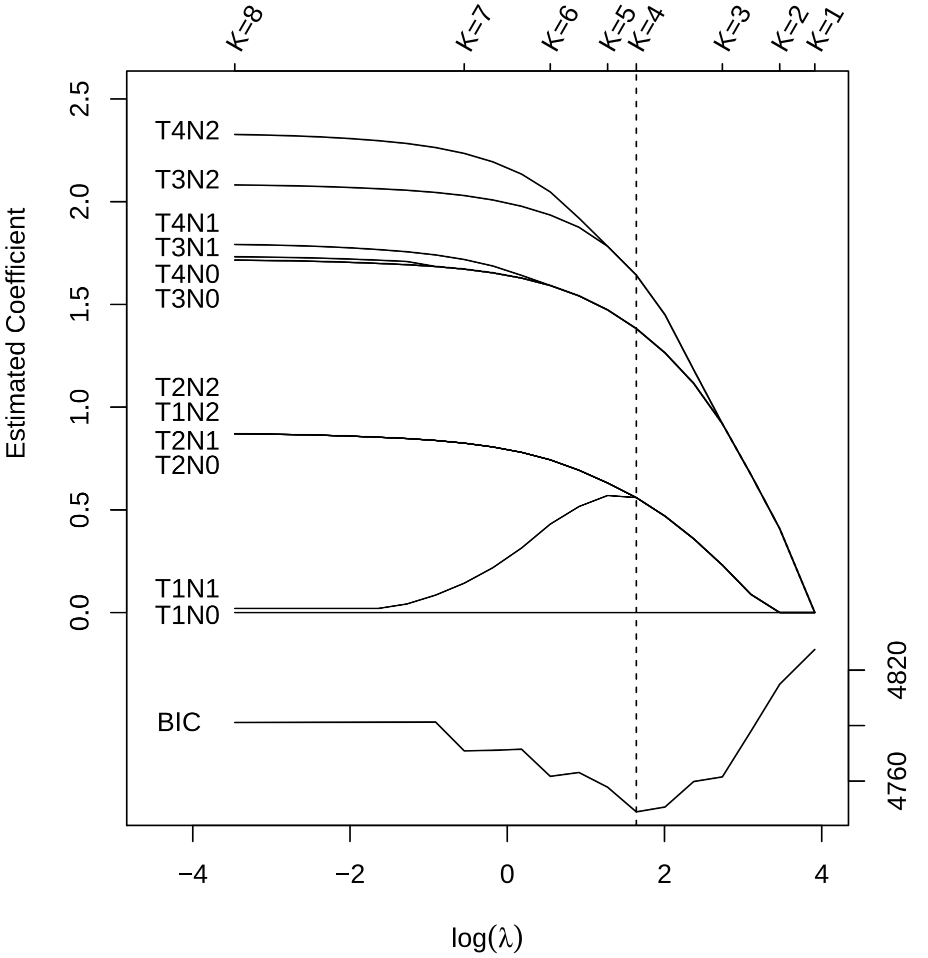
3.3.1 MSKCC Data: The Lasso Tree

Figure 3.1 shows the estimated coefficients from the lasso tree fit as a function of the log tuning parameter $\log(\lambda)$. Labels on top of the graph show when the groupings occur. The cell T1N0 is set to be the reference group. There is a left-right tree structure and the cells merge successively in a roughly monotone fashion. The monotonicity is with respect to the grouping process, where a higher level grouping always contains the lower ones as subsets.

The tree structure starts with 8 groups instead of 12 because some of the cells are forced to merge by the partial ordering constraints. Specifically, the unconstrained coefficients in cells T2N0, T2N1, T1N2 and T2N2 are not in line with the natural ordering and hence these cells are aggregated into one group at the very beginning. The same can be said for cells T3N0 and T4N0. Interestingly, T1N1 starts very close to T1N0 but separates from it until it merges with a higher stage. Note that since only 1 patient died among the 14 patients in T1N1, this behavior of $\hat{\beta}_{1,1}$ can be attributed to the large variation within this subcategory.

The vertical dotted line is drawn at 1.66, the value of $\log(\lambda)$ that minimizes the BIC. The grouping selected by BIC has 4 stages (Figure 3.2(b)), with estimated hazard ratios $\exp(\hat{\beta}) = (1.00, 1.75, 3.96, 5.19)$. Unlike the AJCC grouping (Figure 3.2(a)) which divides stage 3 horizontally at N1, the lasso tree selected grouping classifies stages primarily by the T categories (vertically), which is in fact very similar

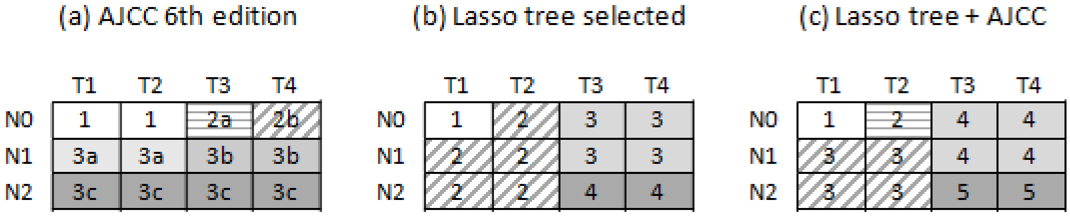
FIGURE 3.1: The lasso tree based on the MSKCC data: coefficient estimates and BIC for the colorectal cancer example, as a function of $\log(\lambda)$. The dotted line represents the staging for $\log(\hat{\lambda}) = 1.66$, selected by minimizing BIC. K represents the number of groups.



to the systems selected by the bootstrap method. The Kaplan-Meier survival curves for the AJCC and the lasso tree selected staging are displayed in Figure 3.3(a-b). Clearly, the lasso tree selected staging scheme gives a better separation of survivals than does the AJCC staging system. This result again indicates that there could be much room for improvement for the current AJCC staging.

To formally evaluate the AJCC and the lasso tree selected staging systems, we use the three criteria for cancer staging proposed in Chapter 2: 1) explained variation; 2) area under the ROC curve; and 3) the probability of concordance of stage and survival. Table 3.1 shows for the two systems the estimated values of the three criteria. The estimates of standard errors are obtained by bootstrapping. As expected, the lasso tree selected system has substantially greater prognostic power as measured by all three criteria than the AJCC system.

FIGURE 3.2: Schematic showing the staging schemes for M0 colorectal cancer by (a) AJCC 6th edition staging, (b) the lasso tree selected staging, and (c) the lasso tree selected staging incorporating information from the AJCC.



(a) AJCC 6th edition

	T1	T2	T3	T4
N0	1	1	2a	2b
N1	3a	3a	3b	3b
N2	3c	3c	3c	3c

(b) Lasso tree selected

	T1	T2	T3	T4
N0	1	2	3	3
N1	2	2	3	3
N2	2	2	4	4

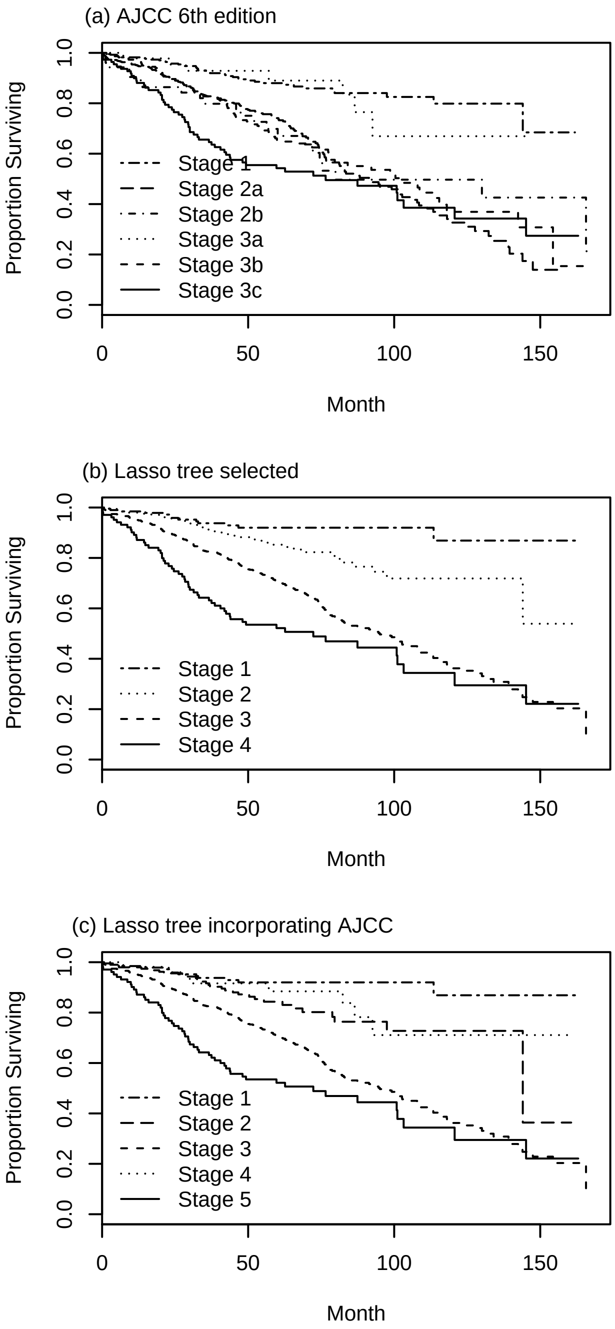
(c) Lasso tree + AJCC

	T1	T2	T3	T4
N0	1	2	4	4
N1	3	3	4	4
N2	3	3	5	5

TABLE 3.1: The AJCC and the lasso tree selected systems: the estimated criteria and their standard errors. VAR: explained variation; AUC: area under the ROC curve; Concordance: the probability of concordance.

	Criteria (SE)		
System	VAR	AUC	Concordance
AJCC	0.642 (0.012)	0.660 (0.013)	0.623 (0.008)
Lasso tree selected	0.686 (0.009)	0.706 (0.011)	0.664 (0.008)

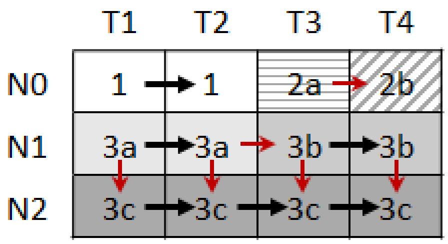
Figure 3.3: Survivals of colorectal cancer patients by (a) AJCC staging, (b) the lasso tree selected staging, and (c) the lasso tree selected staging incorporating information from the AJCC.



3.3.2 MSKCC Data: Incorporating Information from the AJCC System

The AJCC staging system has been seen as the most concerted effort to design a universally acceptable staging system and, since its introduction, has been used in clinical practice throughout the world. We recognize that suggesting a redefinition of the AJCC grouping scheme can be very difficult and comes with a cost regarding our future ability to make comparisons to past experience. On the other hand, the AJCC staging scheme has been developed using a combination of medical knowledge and observational studies, and hence could contain valuable information on prognostic separation of cancer patients. It is reasonable to incorporate this information when developing new systems, which can be done by modifying the penalty terms to reflect it.

FIGURE 3.4: The AJCC 6th edition staging scheme for colorectal cancer: big arrows link cells in the same substage; small arrows link cells in the same main stage but different substages. Heavier weights are imposed on differences represented by the arrows to reflect the AJCC staging, with the big arrows having the heavier weight.



	T1	T2	T3	T4
N0	1	1	2a	2b
N1	3a	3a	3b	3b
N2	3c	3c	3c	3c

We include the information from AJCC in the regression model by posing a heavier penalty on the differences between cells that are in the same stage according to AJCC, in other words, the differences that are zero under AJCC. More specifically, the 6th edition AJCC on colorectal cancer has three main stages and six substages. The differences that are zero under the substage grouping are $\beta_{1,2} - \beta_{1,1}$, $\beta_{2,2} - \beta_{2,1}$, $\beta_{2,4} - \beta_{2,3}$, $\beta_{3,2} - \beta_{3,1}$, $\beta_{3,3} - \beta_{3,2}$, and $\beta_{3,4} - \beta_{3,3}$ (big arrows in Figure 3.4); and the additional differences that are zero under the main stage grouping are $\beta_{1,4} - \beta_{1,3}$, $\beta_{2,3} - \beta_{2,2}$, $\beta_{3,1} - \beta_{2,1}$, $\beta_{3,2} - \beta_{2,2}$, $\beta_{3,3} - \beta_{2,3}$, and $\beta_{3,4} - \beta_{2,4}$ (small arrows in Figure 3.4). Different weights are imposed on these two sets of differences to reflect these two levels of grouping; cells in the same substage are expected to be closer than cells in the same main stage but different substages. Penalty terms $|\beta_{j,k} - \beta_{j,k-1}|$ (or $|\beta_{j,k} - \beta_{j-1,k}|$) in (3.5) corresponding to the first set of differences are replaced by $w_1|\beta_{j,k} - \beta_{j,k-1}|$ (or $w_1|\beta_{j,k} - \beta_{j-1,k}|$), and penalty terms corresponding to the second set of differences are replaced by $w_2|\beta_{j,k} - \beta_{j,k-1}|$ (or $w_2|\beta_{j,k} - \beta_{j-1,k}|$), where $w_1 > w_2 > 1$. The remaining differences have weights of 1. These cells (linked by arrows in Figure 3.4) are hence forced to aggregate more aggressively than the rest, leading to a staging system that might look more like the AJCC.

We apply this modified modeling on the colorectal cancer data with two choices of w_i's: 1) $w_1 = 4$, $w_2 = 2$; and 2) $w_1 = 10$, $w_2 = 5$. The optimal groupings selected by the BIC are identical under both choices of w_i's with slightly different hazard ratio estimates. The selected scheme has 5 groups (Figure 3.2(c)), with $\exp(\hat{\beta}) = (1.00,$

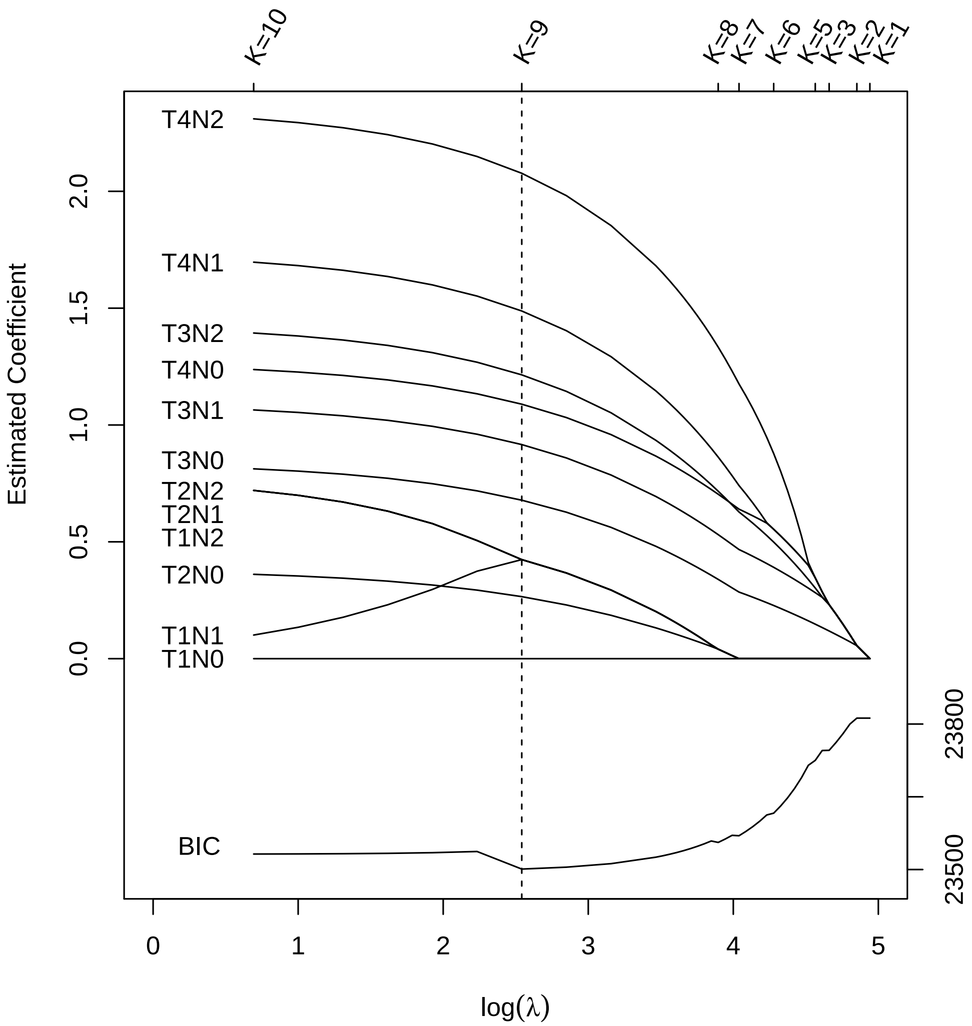
1.83, 1.99, 4.28, 5.66) when $w = (4, 2)$ and $\exp(\hat{\beta}) = (1.00,\ 1.71,\ 2.23,\ 3.98,\ 5.37)$ when $w = (10, 5)$. This grouping is in fact almost identical to the one selected by the lasso tree alone, except that the cell T2N0 is now separated as a single stage. The information contained in the AJCC scheme is overwhelmed by the information contained in the data. Figure 3.3(c) shows the survival curves under this 5-stage system. As a further division of the previous 4-stage scheme, the 5-stage scheme seems to offer no apparent improvement in separating the survivals. Thus, it might be reasonable to favor a more parsimonious system as urged by Gönen and Weiser [4].

3.3.3 The SEER Data Example

In this section, we describe our brief experiment with the SEER data. We apply the lasso tree method to the SEER data by searching through a range of λ's. The estimated coefficients from the lasso tree fit as a function of the log tuning parameter $\log(\lambda)$ are shown in Figure 3.5, as well as the BIC as a function of $\log(\lambda)$. Again, there is a left-right tree structure and the cells merge successively in a roughly monotone fashion.

With the SEER data, the tree structure starts with 10 groups as the cells T1N2, T2N1, and T2N2 are forced to merge at the very beginning by the partial ordering constraints. The vertical dashed line is drawn at 2.54, the value of $\log(\lambda)$ that

Figure 3.5: The lasso tree based on the SEER data: coefficient estimates and BIC for the colorectal cancer example, as a function of $\log(\lambda)$. The dotted line represents the staging for $\log(\hat{\lambda}) = 2.54$, selected by minimizing BIC. K represents the number of groups.

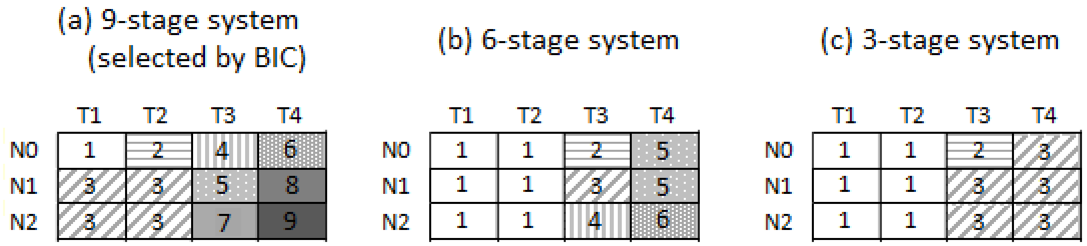


minimizes the BIC. Unlike the sparse, 4-stage grouping selected based on the MSKCC data, the grouping selected by BIC here has as many as 9 stages. A schematic of this optimally selected grouping is shown in Figure 3.6(a): every single cell in the T×N table forms a unique stage, with the exception of the cells T1N1, T1N2, T2N1, and

T2N2. The corresponding estimated hazard ratios are exp(β) = (1.00, 1.30, 1.53, 1.97, 2.50, 2.97, 3.37, 4.43, 7.98). The BIC increases as the cells are further grouped.

The very large sample size of the SEER data has contributed to this less sparse grouping. The only considerable randomness in this data lies in these few small cells: T1N1, T1N2, T2N1, and T2N2. Once these four cells merge, the unexplained variation in the model is largely removed and an optimal staging system is reached judged by the BIC. A more sparse system might be reached by using a criterion that impose a larger penalty on the number of explanatory variables than the BIC.

FIGURE 3.6: Schematic showing the optimal staging schemes selected by lasso tree based on the SEER data: (a) the 9-stage system selected by BIC, (b) the selected 6-stage system, and (c) the selected 3-stage system.



(a) 9-stage system (selected by BIC)

	T1	T2	T3	T4
N0	1	2	4	6
N1	3	3	5	8
N2	3	3	7	9

(b) 6-stage system

	T1	T2	T3	T4
N0	1	1	2	5
N1	1	1	3	5
N2	1	1	4	6

(c) 3-stage system

	T1	T2	T3	T4
N0	1	1	2	3
N1	1	1	3	3
N2	1	1	3	3

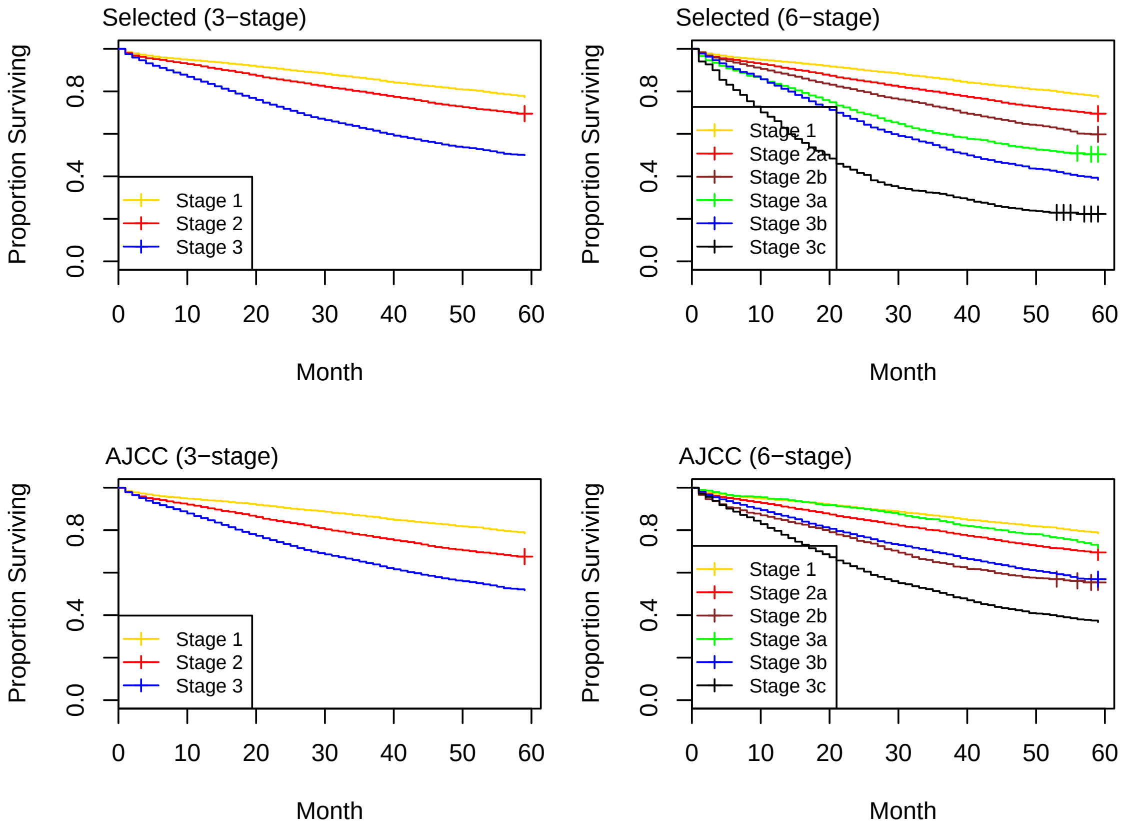
Despite the selection of a 9-stage system by the BIC, we can always choose other optimal systems given by the lasso tree if a more sparse system is desired. Figure 3.6 shows two such systems with 6 stages and 3 stages, respectively. Similar to those selected based on the MSKCC data, these selected groupings classify stages primarily by the T categories (vertically). However, they are quite different from the systems selected by bootstrap based on the same data (Figure 2.5), on which more discussions will be given shortly.

Similar to what we found previously with the MSKCC data, the category of T1N1 starts close to T1N0 but separates from it until it merges with a higher stage (the group of T1N2, T2N1, and T2N2). Again, this is due to the relatively small sample size of this category. With 181 subjects (compared to the average cell sample size of 1,442), this behavior of $\hat{\beta}_{1,1}$ can be attributed to the large variation within this category. When λ increases, heavier penalties will drag $\hat{\beta}_{1,1}$ towards $\hat{\beta}_{1,2}$ and $\hat{\beta}_{2,1}$, the coefficients in its neighboring cells. Although there is also penalty on the difference between T1N1 and T1N0, this effect is outweighed by the joint effect from T1N2 and T2N1.

This particular behaver of $\hat{\beta}_{1,1}$ in fact reveals an interesting, and probably useful, feature of the lasso tree method. In our method, the penalties / constraints are imposed upon the differences between neighboring coefficients, where"neighbors" are defined as adjacent cells in the rows or in the columns. This hence does not include adjacent cells in the diagonal direction. Therefore, rather than merging with a cell next to them along the diagonal, cells are much more likely to be grouped with their row or column neighbors. This implies that the resulting systems will most likely have stages configured in rectangular shapes. In this particular example, this feature leads to the merging of T1N1 with its two adjacent cells, T1N2 and T2N1 (and hence also T2N2), which results in the rectangular group of the four cells combined (stage 3 in Figure 3.6(a)), rather than the merging of T1N1 and T1N0 which would result in a lower triangle group of T1N2, T2N1, and T2N2. This feature also explains the difference between the systems selected by the bootstrap method and the lasso tree method. The bootstrap method, by exhaustively searching through all eligible systems and judging them solely by their prognostic abilities, might select systems that have irregular configurations like those presented in Figure 2.5. While the lasso tree method, by using constraints on immediate row and column neighbors, would have selected systems as in Figure 3.6 that have more regular, rectangular configurations. Both approaches have pros and cons, and we believe the lasso tree method, by producing more regularly configured, and hence more interpretable, systems, will prove useful in many situations.

The Kaplan-Meier survival curves for the 3-stage and 6-stage lasso tree selected systems (we suspect the 9-stage system too unwieldy for clinical use and hence do not include it here) and the AJCC are displayed in Figure 3.7. Again, the lasso tree selected 6-stage scheme gives a better separation of survivals than does the 6-stage AJCC staging system, while the difference in prognostic ability is not substantial when 3-stage systems are of concern.

FIGURE 3.7: Survivals of colorectal cancer patients by (a) AJCC staging, (b) the lasso tree selected staging, and (c) the lasso tree selected staging incorporating information from the AJCC.



3.4 Simulation Studies

In this section we present simulation studies to investigate the finite sample properties of the lasso tree. The performance of the lasso tree and the existing approaches for cancer staging will be compared from two aspects: the ability to select the correct grouping and robustness with regard to changes in the data. We also investigate the role of the partial ordering constraint on the lasso tree.

3.4.1 Ability to Select the Correct Grouping

A hypothetical 4-stage system is used to generate the data, whose sample distribution in the T×N table is chosen to be representative of the real colorectal cancer data. The sample size is set to be 1000, also representative of the colorectal cancer data. Based on the "true" grouping, we generate survival times from two models: (A) the exponential model $\log\lambda = x'\beta$; and (B) the log-normal model $\log T = -x'\beta + \sigma W$, where W has a standard normal distribution and $\sigma = 0.8$. The coefficients are set to be $\beta = \log(1, 2, 4, 8)$ for a moderate effect and $\beta = \log(1.0, 1.2, 1.5, 2.0)$ for a small effect. Censoring times are generated from a Unif(0, τ) distribution, where τ is chosen to produce 40% and 80% censoring.

For comparison, the bootstrap and best subset methods are included and the criteria for selection are chosen to be the area under ROC curve (AUC) for 10-year survival and the probability of concordance of grouping and survival (Concordance). Other selection criteria that quantify the prognostics ability of candidate groupings, such as the partial likelihood, BIC, etc., can also be use under the bootstrap and best subset methods. Yet they perform unfavorably as compared to AUC and Concordance, and hence are dropped from the simulation. Note that both these two approaches assume the true number of groups K is known or pre-specified.

Table 3.2 reports the empirical probabilities (based on 1000 simulations) of selecting the correct 4-stage system. Two significant digits are shown based on the maximum Monte Carlo standard error ($\sqrt{0.5 * 0.5/1000} = 0.016$). For the lasso tree, we report selection probabilities both when assuming the true number of groups K = 4 is known and when K is unknown and estimated by the BIC. The empirical results show that the lasso tree is able to select the correct grouping, especially when the true number of groups K is pre-specified; the probability of selecting the true grouping when fixing K = 4 is over 70% in all scenarios studied. When the BIC is used to select the grouping, the successful rate is slightly lower (around 65%). However, it is worthwhile to emphasize that most of the remaining groupings selected by the BIC are in fact nested in the true 4-stage grouping, and so their errors involve falsely splitting stages rather than erroneously combining them. Moreover, the grouping selected by the BIC is very robust to the degree of censoring and the effect size. The bootstrap and best subset methods give less satisfactory results, especially the best subset selection methods which show very low probabilities of selecting the correct grouping even though the true number of groups K is pre-specified.

Table 3.2: Selection probabilities based on 1000 simulations (sample size = 1000)

		Method for Selection					
		Lasso Tree		Bootstrap		Best Subset	
True model	% Censored	K known	K unknown	AUC	Concordance	AUC	Concordance
A*(1,2,4,8)	40%	0.94	0.68	0.57	0.62	0.40	0.33
	80%	0.88	0.65	0.38	0.43	0.23	0.14
A*(1.0, 1.5, 2.5, 4.0)	40%	0.80	0.68	0.44	0.48	0.22	0.14
	80%	0.72	0.62	0.32	0.31	0.13	0.09
B*(1,2,4,8)	40%	0.99	0.74	0.81	0.82	0.78	0.77
	80%	0.95	0.73	0.67	0.67	0.62	0.63
B*(1.0, 1.5, 2.5, 4.0)	40%	0.94	0.74	0.56	0.58	0.53	0.52
	80%	0.89	0.74	0.51	0.50	0.46	0.44

* A: exponential model; B: log-normal model.

3.4.2 Estimation Stability

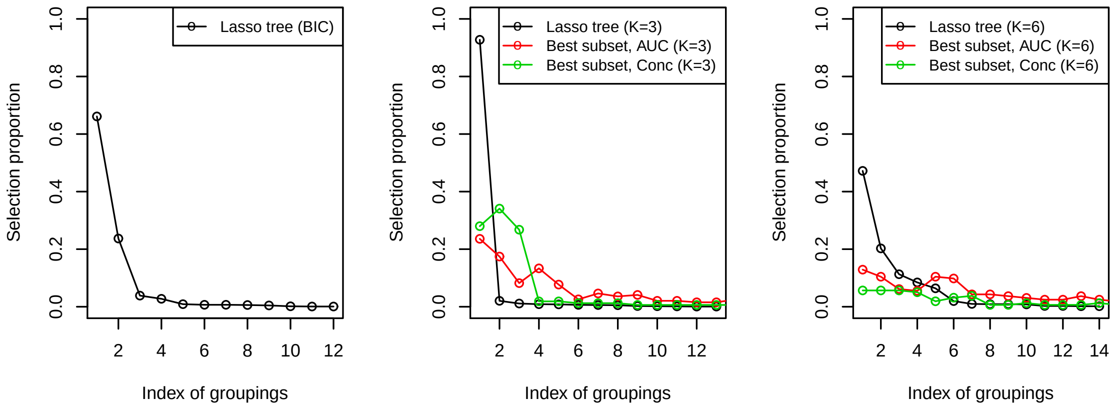
Bootstrap samples are drawn from the colorectal cancer data set with replacement to evaluate the estimation stability of the lasso tree with respect to changes in the data. For comparison, we include the best subset selection and assume the number of groups to be K = 3 and K= 6, corresponding to the two AJCC 6th edition groupings, respectively. The lasso tree selection is made accordingly and, additionally, by estimating K with the BIC. The empirical selection proportions based on 1000 bootstrap samples are shown in Figure 3.8. The algorithm is stable if it selects one grouping large proportion of time.

As expected, the lasso tree gives most stable results, with one dominant staging when fixing K = 3 or estimating K with BIC, and a few dominant groupings when K = 6. The estimate is more stable towards the top of the lasso tree (when K is small). The best subset selection gives most unstable results, especially when K is large.

3.4.3 The Partial Ordering Constraint

The partial ordering constraint $\beta_{j,1} \leq \ldots \leq \beta_{j,q}$ and $\beta_{1,k} \leq \ldots \leq \beta_{p,k}$, $j = 1, ..., p$, $k = 1, ..., q$, is included in (3.5) to assure the estimates are ordered in both T and N. One could expect to relax this constraint when the ordinal trends in T and N are apparent. We are interested in whether formulation (3.5) will give similar results

Figure 3.8: Selection proportions based on 1000 bootstrap samples from the colorectal cancer data set. The x-axis indexes the staging systems selected most often by the lasso tree, and the groupings are ordered by the percentage of time they are selected by the lasso tree. K represents the pre-specified number of groups. Conc = the concordance criterion.



with and without the ordering constraint. This is equivalent to asking if the reformulation in (3.6) (when the ordering constraint is present) gives results similar to formulation (3.5) without the ordering constraint. Here we investigate the role of the partial ordering by removing it from the lasso tree and rerunning the simulations in Section 3.4.1. The results for the exponential model are shown in Table 3.3. The selection probabilities are not much affected even for small effects, indicating that the two formulations are almost equivalent when the ordinal trends in T and N are apparent.

Table 3.3: Selection probabilities based on 1000 simulations

		Lasso Tree			
		With constraint		Without constraint	
True model	% Censored	K known	K unknown	K known	K unknown
A*(1,2,4,8)	40%	0.94	0.68	0.94	0.67
	80%	0.88	0.65	0.86	0.64
A*(1.0, 1.5, 2.5, 4.0)	40%	0.80	0.68	0.78	0.65
	80%	0.72	0.62	0.70	0.60

* A: exponential model

3.5 Asymptotic Theory of the Lasso Tree

3.5.1 Consistency of the Grouping Procedure

Fan and Li [33, 34] established the sampling property and oracle property for a class of variable selection procedures via nonconcave penalized likelihood. With proper choice of regularization parameters, they showed that the smoothly clipped absolute deviation (SCAD) penalty perform as well as the oracle procedure in variable selection; namely, they work as well as if the correct submodel were known. However, the lasso estimator does not possess the oracle properties as the lasso shrinkage produces biased estimates for the large coefficients.

The lasso tree solves a different problem rather than selecting non-zero variables: the grouping of parameters. We hence define its "oracle property", or the *consistency* of the grouping procedure, accordingly. Denote by $\beta^0 \in \mathbb{R}^p$ the true value of β. There exists a partition $\{\mathcal{G}_1^0, \mathcal{G}_2^0, ..., \mathcal{G}_{K^0}^0\}$ of $\{1, 2, ..., p\}$ that groups the values of β^0 into K^0 groups, and a vector $\mu^0 \in \mathbb{R}^{K^0}$ such that the true values of β can be written as

$$\beta^0 = \sum_{k=1}^{K^0} \mu_k^0 1_{\mathcal{G}_k^0}, \tag{3.8}$$

where $1_{\mathcal{G}}$ is the indicator function of the set $\mathcal{G} \subseteq \{1, 2, ..., p\}$, i.e. the p-dimensional vector whose j-th coordinate is 1 if $j \in \mathcal{G}$ and 0 otherwise. Similarly, let $\{\hat{\mathcal{G}}_1, \hat{\mathcal{G}}_2, ..., \hat{\mathcal{G}}_{\hat{K}}\}$ be the partition for the lasso tree estimate $\hat{\beta}$, $\hat{K}$ the estimated number of groups, and $\hat{\mu} \in \mathbb{R}^{\hat{K}}$ the estimated vector of group values. Following the language of Fan and Li [33], we say the lasso tree grouping procedure is *consistent* if $\hat{\beta}$ has the following properties asymptotically:

1. Is root-n consistent;

2. Identifies the correct grouping, $\{\mathcal{G}_1^0, \mathcal{G}_2^0, ..., \mathcal{G}_{K^0}^0\} = \{\hat{\mathcal{G}}_1, \hat{\mathcal{G}}_2, ..., \hat{\mathcal{G}}_{\hat{K}}\}$.

In this section we will show that an improved, adaptive lasso tree grouping procedure possesses the *consistency* properties with a proper choice of the regularization parameter.

3.5.2 Adaptive Lasso Tree

To explore the asymptotic behavior of the lasso tree, we first consider the one-dimensional situation (e.g. only T or N is of interest). The penalized likelihood estimator is

$$\hat{\beta} = \operatorname{argmin} \left\{ -\ell_x(\beta) + n\lambda_n \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \right\}, \tag{3.9}$$

where the β's may be unconstrained or subject to some ordering constraints (e.g. $\beta_1 \leq \ldots \leq \beta_p$ in the cancer staging case). It is easily noticed that the model can be reparameterized by writing $\theta = \{\theta_j\} = (\beta_1,\ \beta_2 - \beta_1,\ \ldots,\ \beta_p - \beta_{p-1})^T$ and

$$\hat{\theta} = \operatorname{argmin} \left\{ -\ell_z(\theta) + n\lambda_n \sum_{j=2}^{p} |\theta_j| \right\}, \tag{3.10}$$

where $z = \{z_{ij}\}$ with $z_{ij} = \sum_{k=j}^{p} x_{ij}$. Hence the grouping problem has been transformed into a regular lasso problem, and proving its consistency properties is equivalent to proving the oracle properties for a regular lasso, which, according to Fan and Li [34], does not hold. Next, we improve the lasso tree method in order to achieve the consistency properties.

To improve the lasso tree method, we apply the adaptive idea which has been used by various authors (Wang, Li and Tsai [35]; Zou [36]; Zhang and Lu [37]) that penalizes different coefficients differently. Specifically, we consider the weighted $\mathcal{L}1$

penalty,

$$\operatorname{argmin}\left\{-\ell_x(\beta)+n\lambda_n\sum_{j=2}^{p}w_j|\beta_j-\beta_{j-1}|\right\}, \tag{3.11}$$

where the positive weights $w=(w_2,...,w_p)^T$ are chosen adaptively by data. We elect to use $w_j=1/|\tilde{\beta}_j-\tilde{\beta}_{j-1}|$, where $\tilde{\beta}=(\tilde{\beta}_1,...,\tilde{\beta}_p)^T$ is the maximum likelihood estimator of β. This way, small penalties are imposed on large differences and large penalties on small differences, which avoids excessive penalties on large differences. By reparameterization, the problem is transformed into

$$\operatorname{argmin}\left\{-\ell_z(\theta)+n\lambda_n\sum_{j=2}^{p}|\theta_j|/|\tilde{\theta}_j|\right\}, \tag{3.12}$$

where $\tilde{\theta}$ is the maximizer of the log partial likelihood $\ell_z(\theta)$. This is equivalent to the adaptive lasso which has been shown to enjoy the *oracle* property [36, 37]. Therefore, the adaptive lasso tree grouping procedure in (3.11) also enjoys the *consistency* properties as defined in Section 3.5.1. Note that any consistent estimators of β can be used, and $\tilde{\beta}$ is just a convenient choice [36, 37].

If the β's are ordered as in the cancer staging example, i.e. $\beta_1\leq\ldots\leq\beta_p$, then (3.12) would subject to $\theta_j\geq 0,\ j=1,...,p$ and the absolute-signs can be dropped for the θ's. The *consistency* properties can be proved following the same argument in Zhang and Lu [37] with no extra difficulty.

3.5.3 Two-dimensional Case

Now let's consider the two-dimensional situation as in (3.5). Similarly, we propose the adaptive lasso tree with adaptively weighted $\mathcal{L}1$ penalties as follows

$$\hat{\beta} = \text{argmin}\left\{-\ell_x(\beta) + n\lambda_n \sum_{j=1}^{p}\sum_{k=2}^{q} \frac{|\beta_{j,k} - \beta_{j,k-1}|}{|\tilde{\beta}_{j,k} - \tilde{\beta}_{j,k-1}|} + n\lambda_n \sum_{j=2}^{p}\sum_{k=1}^{q} \frac{|\beta_{j,k} - \beta_{j-1,k}|}{|\tilde{\beta}_{j,k} - \tilde{\beta}_{j-1,k}|}\right\} \tag{3.13}$$

where $\tilde{\beta} = (\tilde{\beta}_{1,1}, ..., \tilde{\beta}_{j,k})^T$ is the maximum likelihood estimator of β. We will use steps similar to the proofs of Zhang and Lu [37] to establish the *consistency* properties of the adaptive lasso tree for our grouping problem, under the Cox model with general settings. It will be shown that the adaptive lasso tree estimator converges at rate $O_p(n^{-1/2})$ and gives the correct grouping asymptotically. In this section, we only state the theoretical results. The proofs will be given in the Appendix.

The following theorem shows that $\hat{\beta}$ is root-n consistent if $\lambda_n \to 0$ at an appropriate rate. With no extra difficulty, the root-n consistency property can be generalized to cases where the β's are naturally ordered such that $\beta_{j,1} \leq \ldots \leq \beta_{j,q}$ and $\beta_{1,k} \leq \ldots \leq \beta_{p,k},\ j = 1, \ldots, p,\ k = 1, \ldots, q$.

Theorem 3.1. *Assume that $(x_1, T_1, C_1), \ldots, (x_n, T_n, C_n)$ are independently and identically distributed according to the population (x, T, C), and that T_i and C_i are independent given x_i. If $\sqrt{n}\lambda_n = O_p(1)$, then the two-dimensional adaptive lasso tree estimator satisfies $\|\hat{\beta} - \beta^0\| = O_p(n^{-1/2})$.*

Next we show that, when the β's are naturally ordered as in the cancer staging example, i.e. $\beta_{j,1} \leq \ldots \leq \beta_{j,q}$ and $\beta_{1,k} \leq \ldots \leq \beta_{p,k}$, $j = 1, \ldots, p$, $k = 1, \ldots, q$, and the penalized likelihood estimator (3.13) is obtained under this ordering constraint, the adaptive lasso tree estimator has the oracle property as defined in Section 3.5.1 if λ_n is chosen properly.

Theorem 3.2. *Assume that the β's are naturally ordered such that $\beta_{j,1} \leq \ldots \leq \beta_{j,q}$ and $\beta_{1,k} \leq \ldots \leq \beta_{p,k}$, $j = 1, \ldots, p$, $k = 1, \ldots, q$, and that the adaptive lasso tree estimator is obtained under this ordering constraint. If $\sqrt{n}\lambda_n \to 0$ and $n\lambda_n \to \infty$, then under the condition of Theorem 1, with probability tending to 1, the root-n adaptive lasso tree estimator $\hat{\beta}$ must identify the correct grouping: $\{\mathcal{G}_1^0, \mathcal{G}_2^0, ..., \mathcal{G}_{K^0}^0\} = \{\hat{\mathcal{G}}_1, \hat{\mathcal{G}}_2, ..., \hat{\mathcal{G}}_{\hat{K}}\}$.*

The two theorems together establish the consistency of the adaptive lasso tree grouping procedure. With proper choice of regularization parameter λ_n and adaptive weights w (the inverse of any root-n consistent estimator of β_0), the penalized likelihood estimators possess the consistency properties under some mild regularity conditions. In practice, data-driven methods, such as BIC (Section 3.2.3) and GCV, are employed to select λ_n. For regular lasso with linear estimators (in terms of response variable), asymptotic optimal properties of choice of λ_n have been studied in series of papers by Wahba [38] and Li [39] and references therein. With the local quadratic approximations in Section 3.2.2, the resulting estimators will be approximately locally linear. It is of interest to establish the asymptotic properties of the proposed estimators with a data-driven λ_n. Further studies on this issue are needed, but it is beyond the scope of this dissertation.

3.6 Discussion and Conclusions

In this chapter, we have proposed and studied the lasso tree stage selection method for censored survival data via a penalized likelihood approach. With slight modification of the penalties and proper choice of regularization parameters, we show that the lasso tree grouping procedure is consistent; namely, the estimator is root-n consistent and gives the correct grouping asymptotically. In finite sample situations the method is shown to be effective in estimating the best staging system and its estimates stable with regard to changes in the data. The tree structure resulting from varying the tuning parameter provides a visual aid on how the groupings are made progressively and allows flexibility in the decision making process for clinical researchers and practitioners.

Our analysis of the colorectal cancer data partly confirms the findings in Chapter 2. From the MSKCC data, the selected systems (Figure 3.2(b)) is very similar to the ones selected by the bootstrap method in terms of their configuration and prognostic accuracy. Again, the selected scheme suggests that the most essential information is contained in the contrast between the tumor invading through the muscularis propria (T3 and T4) and otherwise (T1 and T2), rather than between node-positive (N1 and N2) and node-negative (N0) cancers as in AJCC. The findings from the SEER data suggest similar trend. The considerable difference between the prognostic ability of the optimal systems and the AJCC categories is concerning especially in light of the fact that optimal systems are comparable to AJCC in terms of simplicity and interpretability. We hope these findings could sparkle some discussions within the statistical and medical communities and contribute to an improved cancer staging system.

In Section 3.3.2, we tried combining prior beliefs on cancer staging - the AJCC system - into the selection process by imposing different weights on the penalty terms. Since the AJCC grouping bears almost no resemblance to the data selected grouping, the data eventually overwhelmed the prior belief in both choices of weights. The same task might be tackled from a Bayesian perspective. The prior domain knowledge can be quantified in the form of prior distributions elicited from the AJCC. More specifically, we may obtain the prior distributions of the regression coefficients by fitting the AJCC model / grouping to an independent data set, for example the SEER data. Posterior distributions of model parameters can then be obtained for the penalized Cox proportional hazards model using Bayesian methodology.

The findings from the SEER data reveal an important property of the lasso tree grouping procedure. That is, by penalizing only the differences between row and column "neighbors", the method forces the cells in the T×N table to aggregate into more rectangular groups. In other words, the resulting optimal systems will most likely be more regularly configured, and therefore simpler and more anatomically interpretable, than the systems selected by the bootstrap selection method. This could become another attractive feature of the lasso tree method, because staging systems are most useful when they are both prognostically optimal and anatomically interpretable. A staging system that is prognostically optimal is unlikely to be adopted if it does not respect the anatomic extent of disease.

In the data analysis in Section 3.3.1 we notice that the 4 and 5-stage groupings selected by the lasso tree have nearly identical BIC's. A careful examination of the analysis suggests that the 4-stage grouping is indeed more desirable because it aggregates T1N1, the second stage in the 5-stage grouping with inadequate sample size, into a larger group. Stages with low prevalence are not useful for clinical treatment decisions. In fact, one of the many criteria for a good staging system defined by Groome and others is a balanced distribution of patients across the groups [6]. This was achieved by minimizing BIC in the colorectal cancer example, yet in more general cases it might be desirable to pursue a balanced distribution of patients by modifying the penalty as follows

$$\hat{\beta} = \text{argmin}\left\{-\ell(\beta) + \lambda(\sum_{j=1}^{p}\sum_{k=2}^{q}\tau_{j,k}|\beta_{j,k} - \beta_{j,k-1}| + \sum_{j=2}^{p}\sum_{k=1}^{q}\upsilon_{j,k}|\beta_{j,k} - \beta_{j-1,k}|)\right\}$$

$$\text{subject to} \quad \beta_{j,1} \leq \ldots \leq \beta_{j,q} \text{ and } \beta_{1,k} \leq \ldots \leq \beta_{p,k}, \quad j = 1, ..., p,\ k = 1, ..., q \tag{3.14}$$

where the positive weights τ and υ are chosen to be inversely proportional to the sample size in the corresponding cells. That is, $\tau_{j,k} = 1/(n_{j,k} + n_{j,k-1})$ and $\upsilon_{j,k} = 1/(n_{j,k} + n_{j-1,k})$, where $n_{j,k}$ is the sample size in the cell with T $= j$ and N $= k$. This places a heavier penalty on cells with small sample sizes and forces them to aggregate, leading to a more balanced distribution of stage sample sizes.

Chapter 4

Risk Stratification by Penalized Logistic Regression

In this chapter, we will see one example of how the lasso tree method can be modified and extended to applications in other scientific areas. More specifically, we will see how it can be applied to risk stratification problems under logistic regression model settings.

4.1 Introduction

Risk stratification models are useful tools in medicine to support tasks such as benchmarking, identification of patients at risk, and individual clinical decision making. A number of techniques have been suggested for the development of clinical risk stratification models, including a variety of statistical methods (e.g., logistic and linear regression, discriminant analysis, and recursive partitioning), and the clinical judgment of experts [25, 40]. For predicting binary outcomes, such as mortality or the presence of disease, logistic regression has emerged as the statistical technique of choice [41].

Logistic regression is widely used to model medical problems because the methodology is well established and coefficients may have intuitive clinical interpretations. However, when a number of risk factors are presented, logistic regression may be inadequate to handle these variables including their interactions; such highly parameterized models may overfit the data and could perform poorly for prediction. Moreover, the logistic model breaks down in the face of sparse outcomes for the different categories determined by these risk factors. To identify the "important" variables in predicting the outcome, model selection methods such as stepwise deletion and subset selection are often adopted. These techniques, though practically useful, are prone to problems such as a lack of stability as analyzed, for example, by Breiman [10]. Another disadvantage of logistic regression is that, unlike classification methods such as decision trees [25], it cannot be easily converted to a set of rules, a limitation that may reduce its clinical utility.

In this chapter, we focus on scenarios where the risk factors are categorical, which is common in clinical settings. In cases where there is no *a priori* ordering expected between the categories and the outcome, a categorical covariate is modeled by the use of dummy variables. However, in many cases including the cancer staging problem, we expect the effect of category on the outcome to follow some natural ordering. For instance, the hazard ratio for the light smoker category is expected to be smaller than that for the heavy smoker category. Similar to the staging problem, when the coefficients of two neighboring categories are close in risk magnitude, it is tempting to collapse them into one risk group for easier clinical use. This, along with the above-mentioned concerns, motivates us to propose a modified logistic regression method that could automatically and simultaneously selects variables, groups categories, and estimates their coefficients.

Following the same line of thought that gave rise to the lasso tree, we attempt the double tasks of selection and grouping by using a lasso-type penalty in the usual logistic regression. Specifically, we pose constraints on neighboring coefficients such that

$$\sum_j |\beta_{j,1}| + \sum_j \sum_k |\beta_{j,k} - \beta_{j,k-1}| \leq s \tag{4.1}$$

where $\beta_{j,k}$ is the coefficient for the kth level of the jth covariate, and $s > 0$ is a pre-specified tuning parameter. These penalty terms together encourage sparsity in both variable selection and the grouping of the categories.

An attractive feature of this penalized regression method is that, by including fewer variables into the model and at the same time aggregating their categories, it produces a relatively small number of unique predicted values. These predicted values can then be directly used in decision rules for risk stratification or treatment selection. The well-known tree-based methods, being self-explanatory and easily converted to a set of rules, are theoretically applicable [25]. However, since a decision tree does not assign estimated coefficient values to the variables deemed important, the magnitude of the covariate effects could be somewhat unclear. Moreover, as decision trees use a "divide and conquer" method, they tend to perform well if a few highly relevant attributes exist, but less so if many complex interactions are present.

The penalized regression method can be easily adapted to handle two-way interactions of interest. This represents another strength of the proposed approach. For instance, for colorectal cancer which we use as our motivating example in Section 4.2, none existing epidemiology studies of this disease has explore interactions systematically.

The structure of this chapter is as follows. In Section 4.2 we describe the motivating data example of advanced colorectal neoplasia. The proposed penalized logistic regression method is described in Section 4.3 along with the computational approach and estimation of the tuning parameter. The method is then illustrated using the example of advanced colorectal neoplasia in Section 4.4. Discussion and conclusions are presented in Section 4.5.

4.2 The Advanced Colorectal Neoplasia Data

Colorectal cancer (CRC) is the second leading cause of death from cancer in the United States. This year, it is estimated that there will be 147,000 newly diagnosed cases of CRC and nearly 50,000 deaths associated with this disease [42]. Screening is an effective way to reduce cause-specific mortality. Colonoscopy is the most commonly used screening test in the U.S., promoted in cancer-prevention guidelines for people starting at age 50 because of its higher sensitivity than other less costly procedures such as stool-sample tests [43–45]. Colonoscopy allows doctors to examine the entire colon and remove abnormal tissue growths called adenomatous polyps that may progress to cancer. However, high non-adherence to colonoscopy is observed because of its risks, cost, feasibility (availability and insurance coverage), and uncertain incremental benefit over other screening tests for meaningful patient outcomes such as cancer-related morbidity and mortality [46, 47].

One reason for support of widespread colonoscopic screening is that there is no accurate and precise way to stratify risk for advanced colorectal neoplasia (CRC and advanced, adenomatous polyps) among the 90% of U.S. residents who are considered "average risk". If such stratification could be established, then a tailored screening recommendation would be both highly effective and cost-effective. For example, people in the subgroup at very low risk for advanced neoplasia could have screening deferred or performed with methods less invasive than colonoscopy; for people at high risk, colonoscopy would be considered the preferred strategy. Tailoring according to risk of advanced neoplasia could also be useful for allocating CRC screening resources.

In this chapter, we investigate such risk stratification rules for advanced neoplasia among people considered to be average-risk. We use a recently completed large cohort study funded by the National Cancer Institute of subjects undergoing first time screening colonoscopy in a variety of clinical outpatient settings. The targeted risk factors are derived from the NCI's CRC Risk Assessment tool (http://www.cancer.gov/ colorectalcancerrisk) and include a previous cancer-negative sigmoidoscopy / colonoscopy in the last 10 years, polyp history in the last 10 years, history of CRC in first-degree relatives, aspirin and non-steroidal anti-inflammatory drug (NSAID) use, cigarette smoking, body mass index (BMI), leisure-time vigorous activity, vegetable consumption, and for women, post-menopausal estrogen use. All risk factors are categorical variables, with two to four levels. The derived rules are expected to facilitate decisions about initial CRC screening.

Logistic regression models have been used in the literature to estimate the risks of CRC based on quantifiable risk factors. For instance, with a similar group of variables, Freedman et al. developed models for men and women that use logistic regression to estimate future risk for CRC [48]. Here we will illustrate that the proposed penalized logistic regression can be a better choice for developing such risk stratification tools than the usual logistic regression.

4.3 Penalized Logistic Regression

4.3.1 A Modification of the Lasso Tree

We consider a prediction problem with N cases having binary outcomes $y_1, y_2, ..., y_N$ and covariates $x_{ij}, i = 1, 2, ..., N, j = 1, 2, ..., p$. In logistic regression, the outcome y_i follows a Bernoulli probability function that takes on the value 1 with probability π_i and 0 with probability $1 - \pi_i$, where π_i varies over the observations as an inverse logistic function of the vector x_i:

$$\pi_i = \frac{1}{1 + \exp(-\beta^T x_i)} \ . \tag{4.2}$$

To estimate β, we can maximize the conditional log-likelihood

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^{N} \Big[y_i \text{log} \pi_i + (1 - y_i) \text{log}(1 - \pi_i) \Big] \\ &= \sum_{i=1}^{N} \Big[y_i \beta^T x_i - \log(1 + \exp(\beta^T x_i)) \Big] \end{aligned} \tag{4.3}$$

with respect to the regression coefficients $\beta = \{\beta_j\}, \ j = 1, 2, ..., p$. The usual iteratively reweighted least squares (IRLS) procedure is used to obtain maximum likelihood estimates of the parameters [49].

We focus on situations where the covariates are categorical, which corresponds to the advanced colorectal neoplasia study and is very common in clinical settings.

We rewrite β as $\{\beta_{j,k}\}, j = 1, 2, ..., q, \;\; k = 1, 2, ..., n_j$, where q is the number of covariates and n_j is the number of categories or levels (excluding the reference level) for covariate j. Suppose that all covariates have an *a priori* ordering. Then, without loss of generality, β is ordered such that $0 \le \beta_{j,1} \le \ldots \le \beta_{j,n_j},\; j = 1, ..., q$, with 0 being the coefficient of the reference level. The double tasks of selection and grouping can be attempted by using a lasso-type model selection technique. We propose to estimate β as follows

$$\begin{aligned} \hat{\beta} = \operatorname{argmin} & \left\{ -\ell(\beta) + \lambda \sum_{j=1}^{q} |\beta_{j,1}| + \lambda \sum_{j=1}^{q} \sum_{k=2}^{n_i} |\beta_{j,k} - \beta_{j,k-1}| \right\} \\ \text{subject to} \quad & 0 \le \beta_{j,1} \le \ldots \le \beta_{j,n_j}, \quad j = 1, \ldots, q \end{aligned} \tag{4.4}$$

where λ is the tuning parameter. The two penalty terms together encourage sparsity in the variables, i.e. variable selection, and sparsity in the categories, i.e. grouping of the categories.

The sparsity-enforcing property of the penalty results in fewer variables as well as fewer categories in the final model, leading to a relatively small number of unique predicted values. These predicted values can then be directly used as decision rules for risk stratification or for guiding a management strategy. The penalty provides a continuous model that ensures the stability of model selection. It also facilitates model stability in the presence of sparse outcome data for different categories determined by these risk factors.

Our method naturally deals with ordinal and categorical risk factors by imposing constraints. Again, with the ordering constraint, the absolute values in (4.4) can be dropped and the objective function can be simplified as

$$\begin{aligned} &\min_{\beta}\left\{-\ell(\beta)+\lambda\sum_{j=1}^{q}\beta_{j,n_j}\right\} \\ &\text{subject to}\quad 0\le\beta_{j,1}\le\ldots\le\beta_{j,n_j},\quad j=1,\ldots,q. \end{aligned} \tag{4.5}$$

Note that only the coefficient for the highest level category of each covariate is taken into account in (4.5). Yet this is mathematically equivalent to (4.4) and will give the same estimates as the original formulation. Normally, different weights are given to covariates with different numbers of levels in order to avoid excess penalty on covariates with large number of categories. This is not needed here since the penalty only involves one coefficient for each covariate.

The penalty can be easily adapted for covariates without *a priori* ordering or that are partially ordered. For covariates without *a priori* ordering, the penalty is the summation of all pairwise absolute differences (including the differences with the reference level):

$$\lambda\sum_{k=1}^{n_j}|\beta_{j,k}|+\lambda\sum_{k=2}^{n_j}\sum_{l=1}^{k-1}|\beta_{j,k}-\beta_{j,l}|. \tag{4.6}$$

Now with all pairwise absolute differences included, a smaller weight will be given to the penalty term in order to avoid excess penalty on this covariate. This is similar for covariates that are partially ordered.

4.3.2 Computational Approach

With the absolute values dropped, the penalized logistic regression in (4.5) can be solved by the usual IRLS procedure with the weighted least squares step replaced by a constrained weighted least squares procedure. Let X denote the design matrix with x_i as the ith row and $\pi = (\pi_1, ..., \pi_n)^T$, where $\pi_i = 1/(1 + e^{-\beta^T x_i})$. Denote $A = \text{diag}\big(\pi_i(1-\pi_i)\big)$, $z = X\beta + A^{-1}(y - \pi)$, and $P_\lambda(\beta) = \lambda \sum_{j=1}^{q} \beta_{j,n_j}$. Then at the k'th iteration, $\hat{\beta}^{(k)}$ is the solution of

$$\text{argmin}\Big\{(z - X\beta)^T A(z - X\beta) + P_\lambda(\beta)\Big\}, \tag{4.7}$$

where z and A are based on $\hat{\beta}^{(k-1)}$. The iterative procedure is as follows:

1. Fix λ and initialize $\hat{\beta} = 0$.
2. Compute π, A and z based on the current value of $\hat{\beta}$.
3. Minimize $(z - X\beta)^T A(z - X\beta) + P_\lambda(\beta)$ subject to $0 \leq \beta_{j,1} \leq \ldots \leq \beta_{j,n_j}$, $j = 1, ..., q$.
4. Repeat steps 2 and 3 until convergence of $\hat{\beta}$.

The minimization in step 3 can be done through a quadratic programming procedure. When "warm starts" are used for computing the path of solutions over a grid of λ's, the initial $\hat{\beta}$ in step 1 is set to be the solution for the previous λ.

When covariates without *a priori* ordering are present, their contribution to the penalty as in (4.6) can be added into $P_\lambda(\beta)$ with proper weights. The same iterative procedure is then applied. The computation can become more difficult when the absolute values remain in the penalty. In this case, the computational approach introduced by Tibshirani et al. for the fused lasso can be applied as an alternative [27].

As in the lasso tree method, estimates from (4.5) depend on the tuning parameter λ. When $\lambda = 0$, the solution is the usual logistic regression estimate. As λ increases the absolute differences between neighboring coefficients go to 0 successively, corresponding to the successive grouping and dropping of the coefficients, until all coefficients are dropped. The BIC is again used for the automatic estimation of the tuning parameter λ.

4.3.3 Inclusion of Two-way Interactions

The penalized regression method can be adapted to handle two-way interactions of interest. For simplicity, we consider a model with two categorical covariates with p and q levels (excluding the reference levels), respectively, and their interaction terms. With some abuse of notation, we denote $\theta = (\alpha_1, ..., \alpha_p, \beta_1, ..., \beta_q, \nu_{1,1}, ..., \nu_{p,q})^T$ as the model parameters, where α's and β's are regression coefficients for the two main effects and ν's are coefficients for the two-way interaction. Let X denote the design matrix with the interaction and x_i the ith row of X. The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{N} \Big[y_i \theta^T x_i - \log(1 + \exp(\theta^T x_i)) \Big].$$

To develop the penalty, we consider the interaction terms $\{\nu_{j,k}\}$, $j = 1, ..., p$, $k = 1, ..., q$, as features arranged on a two-way grid, like the T×N table. It is then very natural to constrain the differences between neighboring coefficients in both directions in the two-way grid, as well as the difference with the reference, such that

$$|\nu_{1,1}| + \sum |\nu_{j,k} - \nu_{j,k-1}| + \sum |\nu_{j,k} - \nu_{j-1,k}| \leq s. \tag{4.8}$$

Hence when both main effects are *a priori* ordered, the penalized logistic regression can be written as

$$\hat{\theta} = \text{argmin} \left\{ -\ell(\theta) + \lambda \Big(\alpha_p + \beta_q + |\nu_{1,1}| + \sum_{j=1}^{p} \sum_{k=2}^{q} |\nu_{j,k} - \nu_{j,k-1}| + \sum_{j=2}^{p} \sum_{k=1}^{q} |\nu_{j,k} - \nu_{j-1,k}| \Big) \right\}$$
$$\text{subject to} \quad 0 \leq \alpha_1 \leq \ldots \leq \alpha_p \;\; \text{and} \;\; 0 \leq \beta_1 \leq \ldots \leq \beta_q. \tag{4.9}$$

We do not assume here the interactions are ordered whenever the main effects are ordered. In many cases it might be safe to assume this, and the interactions will satisfy a partial ordering constraint, i.e. $0 \leq \nu_{j,1} \leq ... \leq \nu_{j,q}$ and $0 \leq \nu_{1,k} \leq ... \leq \nu_{p,k}$, $j = 1, ..., p$, $k = 1, ..., q$. The penalty then can be further simplified given these constraints.

Table 4.1: Summary of variables in the advanced colorectal neoplasia data set

Variable	Categories	Male (n = 2160)	Female (n = 2304)
Age group	0 = younger than 65	1910	2019
	1 = older than 65	250	285
Sigmoidoscopy / colonoscopy and polyp history	0 = Unknown screen or polyps	301	314
	1 = Screened and NO polyps	24	14
	2 = No screening	1793	1938
	3 = Screened and polyps	39	38
Number of Relatives with CRC	0 = 0 relatives w/ CRC	1554	1430
	1 = 1 relative w/ CRC	432	575
	2 = 2 or more relatives w/ CRC	174	299
Cigarette smoking, pack-years	0 = 0 pack-year	1172	1537
	1 = greater then 0 and < 20	460	437
	2 = 20 or more pack-years	528	330
Leisure-time vigorous activity	0 = greater than 4 hrs/week	1341	1127
	1 = 2 - 4 hrs /week	161	203
	2 = 0 - 2 hrs/week	114	134
	3 = 0 hrs/week	544	840
Vegetable consumption	0 = 5 or more servings/day	73	141
	1 = less than 5 servings/day	2087	2163
Body mass index (BMI)	0 = less than or equal to 24.9	410	1581
	1 = greater than 24.9 and ≤ 29.9	974	
	2 = greater than 29.9	776	723
NSAID use	0 = Regular user of Aspirin/NSAID	1148	1089
	1 = Nonuser of Aspirin/NSAID	1012	1215
Estrogen use (female)	0 = estrogen use in the past 2 yrs	-	953
	1 = no estrogen use in the past 2 yrs	-	1351

4.4 Data Analysis and Results

Study subjects were aged 50 to 80 years and underwent first-time screening colonoscopy between 12/2004 and 9/2011. Advanced neoplasia, the outcome of interest, is defined as a tubular adenoma greater than 1cm, a polyp with villous histology or high-grade dysplasia, or CRC. Among 4,526 subjects (mean age 57.30 ± 6.78 years; 51.8% women), the prevalence of advanced neoplasia was 7.96%. Among the 4,464 (98.6%) with complete data (mean age 57.25 ± 6.70 years; 51.6% women), the prevalence of advanced neoplasia was 8.36%, including 46 subjects with CRC.

4.4.1 Fitted Models

Data from men and women are analyzed separately. Table 4.1 presents a summary of the variables included in the analysis. There are eight risk factors for men and nine for women. Among the nine variables, eight are *a priori* ordered with greater index associated with higher risk, and one (screening and polyp history) is partially ordered - patients in category 3 are expected to have higher risk than those in category 1. BMI is divided into three categories for men and two categories for women. Some categories have very few cases in them (e.g. categories 1 and 3 of screening and polyp history), which might be problematic under a naive logistic regression.

Table 4.2: Estimated coefficients for men. LR = logistic regression; PLR-1 = penalized logistic regression with only main effects; PLR-2 = penalized logistic regression with main effects and their two-way interactions.

Variable	Categories		LR	PLR-1	PLR-2
Age group	0 = younger than 65				
	1 = older than 65	Age1	0.996	0.798	0.611
Sigmoidoscopy / colonoscopy and polyp history	0 = Unknown screen or polyps				
	1 = Screened and NO polyps	SigCol1	0.610	0.167	0
	2 = No screening	SigCol2	0.739	0.383	0
	3 = Screened and polyps	SigCol3	2.006	0.734	0.317
Number of Relatives with CRC	0 = 0 relatives w/ CRC				
	1 = 1 relative w/ CRC	Rel1	0.196	0.090	0.023
	2 = 2 or more relatives w/ CRC	Rel2	0.346		
Cigarette smoking, pack-years	0 = 0 pack-year				
	1 = greater then 0 and < 20	Packyear1	0.855	0.766	0.197
	2 = 20 or more pack-years	Packyear2	0.971	0.826	0.302
Leisure-time vigorous activity	0 = greater than 4 hrs/week				
	1 = 2 - 4 hrs /week	Act1	-0.276	0	0
	2 = 0 - 2 hrs/week	Act2	0.113	0.022	0
	3 = 0 hrs/week	Act3	0.469	0.412	0.379
Vegetable consumption	0 = 5 or more servings/day				
	1 = less than 5 servings/day	Veg1	-0.047	0	0
Body mass index (BMI)	0 = less than or equal to 24.9				
	1 = greater than 24.9 and ≤ 29.9	BMI1	-0.229	0	0
	2 = greater than 29.9	BMI2	-0.121	0	0
NSAID use	0 = Regular user of Aspirin/NSAID				
	1 = Nonuser of Aspirin/NSAID	NSAID1	0.217	0.167	0
Interactions		Packyear1&2 : SigCol2&3	-	-	0.245
		Packyear1&2 : NSAID1	-	-	0.271
		Packyear1&2 : Rel1&2	-	-	0.159
		Age1 : SigCol2&3	-	-	0.168
		Age1 : NSAID1	-	-	0.067
		SigCol2&3 : NSAID1	-	-	0.142

Table 4.3: Estimated coefficients for women. LR = logistic regression; PLR-1 = penalized logistic regression with only main effects; PLR-2 = penalized logistic regression with main effects and their two-way interactions.

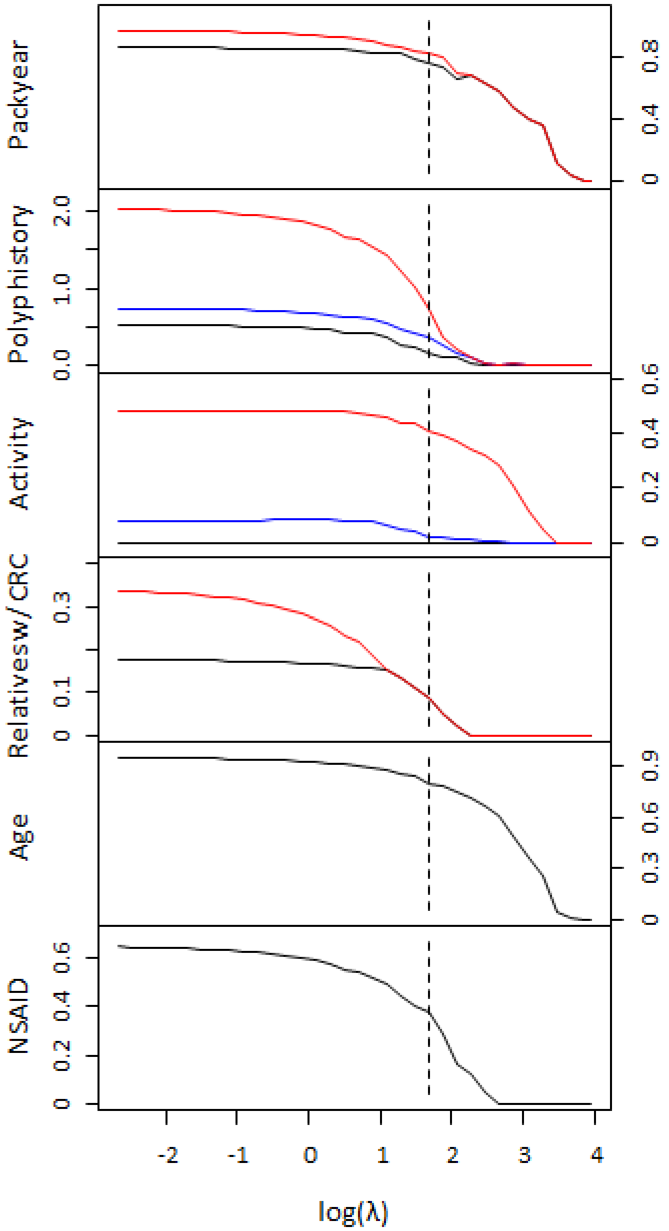
Variable	Categories		LR	PLR-1	PLR-2
Age group	0 = younger than 65				
	1 = older than 65	Age1	0.815	0.692	0.530
Sigmoidoscopy / colonoscopy and polyp history	0 = Unknown screen or polyps				
	1 = Screened and NO polyps	SigCol1	1.152	0.364	0
	2 = No screening	SigCol2	0.937	0.364	0
	3 = Screened and polyps	SigCol3	2.827	1.624	0.617
Number of Relatives with CRC	0 = 0 relatives w/ CRC				
	1 = 1 relative w/ CRC	Rel1	0.330	0.247	0.062
	2 = 2 or more relatives w/ CRC	Rel2	0.567	0.318	
Cigarette smoking, pack-years	0 = 0 pack-year				
	1 = greater then 0 and < 20	Packyear1	0.680	0.562	0.063
	2 = 20 or more pack-years	Packyear2	1.150	0.972	0.482
Leisure-time vigorous activity	0 = greater than 4 hrs/week				
	1 = 2 - 4 hrs /week	Act1	-0.188	0	0
	2 = 0 - 2 hrs/week	Act2	-0.205	0	0
	3 = 0 hrs/week	Act3	0.308	0.265	0.041
Vegetable consumption	0 = 5 or more servings/day				
	1 = less than 5 servings/day	Veg1	-0.314	0	0
Body mass index (BMI)	0 = less than or equal to 29.9				
	1 = greater than 29.9	BMI1	0.521	0.389	0.130
NSAID use	0 = Regular user of Aspirin/NSAID				
	1 = Nonuser of Aspirin/NSAID	NSAID1	0.140	0	0
Estrogen use	0 = estrogen use in the past 2 yrs				
	1 = no estrogen use in the past 2 yrs	Estrogen1	0.708	0.558	0
Interactions		Packyear1&2 : SigCol3	-	-	0.073
		Packyear1&2 : BMI1	-	-	0.087
		Packyear1&2 : Estrogen1	-	-	0.437
		Age1 : SigCol2&3	-	-	0.123
		SigCol2&3 : Estrogen1	-	-	0.093
		Act3 : Estrogen1	-	-	0.321
		BMI1 : Rel1&2	-	-	0.088
		BMI1 : Estrogen1	-	-	0.151
		Rel1&2 : Estrogen1	-	-	0.248

We fit a naive logistic regression, a penalized logistic regression with only main effects (PLR-1), and a penalized logistic regression with main effects and their two-way interactions (PLR-2). Table 4.2 presents the model estimates for men. Because of the natural ordering, all coefficients are expected to be positive. The penalized regression models are able to preserve these orders by dropping unimportant variables and by merging the categories that violate the ordering constraints. This is not guaranteed by the naive logistic regression, where the coefficients for vegetable consumption and BMI are negative, contradictory to common knowledge. These variables are found to be not significant for predicting advanced neoplasia under all models.

Six and five variables are selected, respectively, by the main effect penalized model and the penalized model with interactions. Vegetable consumption and BMI are deemed unimportant variables under both models. The estimated coefficients are shrunk to reach a more stable model. The coefficients for polyp history are shrunk the most since this risk factor is most likely to be correlated with other risk factors. Close categories are grouped simultaneously under both models. For instance, the four-level variable of leisure-time activity can be simplified into two groups, non-active and active, under the penalized model with interactions.

In addition to five main effects, the interaction model selects six interaction terms (of possibly grouped categories). At the same time, some main effect coefficients become much smaller in the interaction model, especially cigarette smoking and polyp

Figure 4.1: Coefficient estimates for men for the six selected variables under the main effect penalized logistic regression model as a function of log(λ). The dotted line represents the value of log(λ) that minimized the BIC.



history. It appears that these variables, as well as NSAID / aspirin use, which is dropped under the interaction model, exhibit risk that is modified by other factors.

For example, cigarette smoking does more harm when other risk factors (i.e. polyp positive, non-user of NSAID / aspirin, and relatives with CRC) are presented. Hence this model sheds more light on how the variables interact and better explains the risk of advanced neoplasia than the main effect model.

Figure 4.1 shows the estimated coefficients for men for the six selected variables under the main effect penalized model as a function of the log tuning parameter $\log(\lambda)$. The dotted line is where the BIC is minimized. From this figure, we gain a glimpse of the relative importance of the risk factors. For instance, cigarette smoking, non-activity, and older age all retain large coefficients for most values of λ, reflecting the significance of their effects on the risk for advanced neoplasia.

Table 4.3 displays the model estimates for female subjects. For women, seven of nine variables are selected by the main effect penalized model. The findings and interpretations are similar to those for men. One thing worth mentioning is that the main effect coefficient of estrogen use is zero under the interaction model because of its significant interactions with many other risk factors. Again, the penalized regression is considered superior and provides more information than the simple logistics regression.

In summary, the penalized logistic regression simultaneously selects important risk factors and provides models with fewer categories. The penalized model with interactions is more desirable since it offers more detailed risk stratification. As the penalized interaction models have only 12 and 16 distinct estimated coefficients for men and women, respectively (compared to a full interaction model which would have 83), these models can be conveniently developed into risk stratification rules for guiding treatment strategy.

Table 4.4: Areas under ROC curves for the risk prediction models. The p-values are for comparing the penalized models to the naive logistic regression.

	Male			Female		
	AUC	95% C.I.	p-value	AUC	95% C.I.	p-value
LR	0.567	(0.531, 0.604)	-	0.573	(0.535, 0.611)	-
PLR-1	0.586	(0.549, 0.623)	0.322	0.589	(0.551, 0.629)	0.339
PLR-2	0.615	(0.578, 0.651)	0.026	0.618	(0.580, 0.657)	0.009

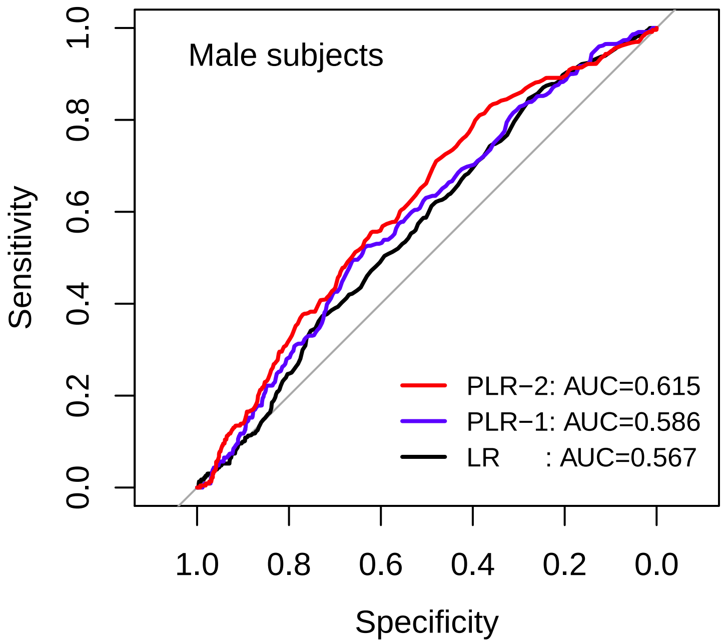
4.4.2 Model Validation

We validate and compare the discriminatory performances of the logistic regression models using receiver operating characteristic (ROC) curves. The area under an ROC curve (AUC) indicates how well a prediction model discriminates between healthy patients and patients with disease. ROC curves are generated by means of 10-fold cross-validation for the three models. The increase in the AUC was evaluated and tested for significance using the test proposed by DeLong et al. [50].

The ROC curves of the penalized regression models dominate that of naive logistic regression at most cutoff thresholds for men (Figure 4.2). The naive logistic regression achieves an AUC of 0.567 (95% C.I., 0.531 - 0.604). The penalized regression models achieve AUCs of 0.586 (95% C.I., 0.549 - 0.623) and 0.615 (95% C.I., 0.578 - 0.651) without and with interactions, respectively. The penalized model with interactions performs significantly better (p-value = 0.026) than the naive logistic regression, while the difference between the main effect penalized model and the naive logistic regression is not significant (p-value = 0.322). No statistically significant difference is found between the AUCs of the two penalized models (p-value = 0.394). These findings suggest that the proposed penalized logistic regression models, in particular the model with interactions, have a favorable performance compared to naive logistic regression. The modest discriminatory power suggests the need to find additional strong risk predictors.

Validation is also performed for women and similar improvement in performance is observed (Table 4.4). The ROC curves are shown in Figure 4.3. Again, the penalized model with interactions performs significantly better (p-value = 0.009) than the naive logistic regression. No statistically significant difference is found between the AUCs of the two penalized models (p-value = 0.155).

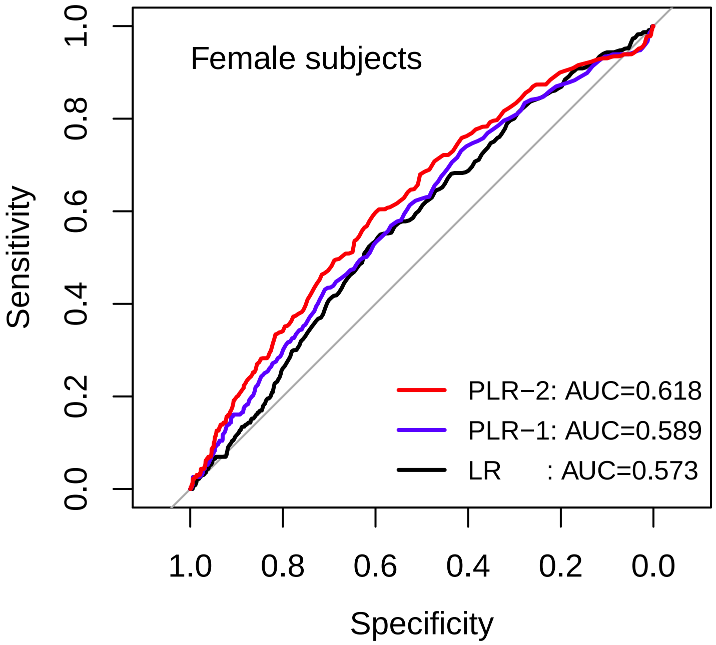
FIGURE 4.2: Receiver-operating characteristic (ROC) curves for the risk prediction models: male subjects.



4.5 Discussion and Conclusions

In this chapter, we have extended the lasso tree strategy and proposed a penalized logistic regression method that automatically selects variables, groups categories, and estimates their coefficients. The model penalizes the $L1$-norm of both the coefficients and their differences. Thus it encourages sparsity in the categories via grouping of the categories, and also sparsity in the variables via variable selection. The method can investigate many variables including their interactions in logistic regression where

Figure 4.3: Receiver-operating characteristic (ROC) curves for the risk prediction models: female subjects.



traditional maximum likelihood based method can break down due to the high number of parameters and insufficient outcome data for certain categories. The order and partial order constraints we put on risk factors in the model incorporates existing scientific findings so that the probability of disease does not decrease at a higher level of risk. The penalty we put on odds ratio coefficients for adjacent categories encourage grouping and lead to parsimonious models. We have applied our method to a recently completed colon cancer screening data. Advantage of our method is seen in terms of both the ROC curves and fitted coefficients for risk factors over the naive logistic regression. The capability for investigating various interactions among numerous risk factors should make our method a powerful tool for cancer risk modeling because currently very few, if any, scientific publications systematically consider interaction terms when there are many risk factors.

This example again illustrate the usefulness of the penalized regression methods. The penalized model is flexible enough to accommodate practical variations. In particular, if no convincing knowledge supports order constraint of a variable, such constraint can be easily dropped from our method. The variables in the colon cancer screening example are entirely categorical, but the penalized regression model can be applied to continuous variables with no extra difficulty. In addition to binary and time to event outcomes, our method can generalize to other types of outcomes such as continuous ones.

Theoretically, the method can also incorporate more than two way interactions. However the computation will be much more involved. Meaningful interpretations of multi-way interactions are extraordinarily difficult to provide. Moreover, when the number of risk factors is large, including multi-way interactions will most likely render the model non-identifiable. In fact, as a preliminary analysis, we attempted to fit a saturated model with the CRC data. With 8 risk factors, the number of parameters in the saturated model is an astonishing 2,304, which is close to the number of data points. The computation hence was very difficult and the result was not meaningful.

Chapter 5

Conclusion and Future Work

This dissertation is motivated by the desire to develop cancer staging systems. In the process, we reframe the task of cancer staging into a model selection context and two model selection methods are proposed for the task: a bootstrap selection method and a penalized regression method, i.e. the lasso tree. The utility of both methods are illustrated on the staging of colon cancer, and their properties studied through simulations.

Of the two approaches, the penalized regression method is considered more promising given its many advantages over the bootstrap selection method. It is more computationally efficient, gives more general and more useful results (the tree-like structure that gives a series of optimal staging systems with different numbers of stages), and is consistent such that it gives the "correct" grouping when the sample size tends to infinity. It gives more regularly configured, and therefore simpler and more interpretable, staging systems. It is also generally applicable to many diseases other than cancers that use risk scores based on risk factors. One example is given in Chapter 4 where a risk stratification model for colorectal cancer is developed using a penalized logistic regression. Using this example, we illustrate how the penalized regression method can be modified to meet different modeling requirements and have applications to a wide range of disease areas and scientific questions.

We expect that the penalized regression method will be used in the future for the staging of other cancers and for the risk stratification or risk assessment of other diseases. Therefore, further investigations will be needed in order to fully understand the properties and performance of the method under different scenarios and make adjustments when needed. In this final chapter, we will point to several interesting directions for future work.

5.1 Cancer Staging and The Lasso Tree

The penalized regression method has general appeal to cancers and many other diseases that use aggregate risk scores based on risk factors. In Chapter 3, the lasso tree is based on the Cox model for censored survival data. But our methodology is applicable in principle to binary outcomes as well, as the example in Chapter 4 illustrated.

In some situations a landmark survival time, such as 5-year or 10-year survival, can be more desirable than using the full survival. A logistic regression model is proposed by Jung for landmark survival analysis [51], and an extension of the lasso tree to this model is also quite possible. Future work to illustrate and evaluate the method's performance on these different models will be important to further understand the method and support its application in a much wider field of medical research and practice.

One difficulty with the development of staging systems has been that the current treatment strategies are stage-dependent. Thus the survival outcomes for patients could be confounded by the actual staging system that was used in their care. A possible remedy is to control for the treatment assignment, which can be done by including the treatment covariate in the penalized proportional hazards model. This would be an interesting and important topic for future work. Otherwise we must acknowledge that our and others' results may combine two or more categories whose outcomes have been rendered similar by varying treatment regiments.

In many situations, it is important to incorporate "outside information", such as prior beliefs, medical knowledge or experience, practical considerations or constraints, preference, etc., into the process of developing cancer staging systems. For instance, we can incorporate the information from the current AJCC system into the regression model by posing a heavier penalty on the differences between cells that are in the same stage according to AJCC, which leads to a staging system that might look more like the AJCC. The same strategy can be applied to incorporate other prior knowledge besides AJCC. In our study, we have chosen arbitrarily the ratios of heavier penalties to lighter penalties, i.e. ratios between the lasso parameter λ's. A less arbitrary way to choose these penalties would be an interesting topic for future work. For example, these ratios between λ's can also be treated as tuning parameters for the penalized regression, and be estimated by optimizing certain criteria, such as BIC.

Another example of incorporating preference and prior belief into the modeling procedure is to pose heavier penalties on cells with small sample sizes. By doing this, categories of lower prevalence will be forced to aggregate more progressively and therefore it results in a more balanced distribution of stage sample sizes. Other possibilities include incorporating prior information using Bayesian modeling methods. The prior domain knowledge can be quantified in the form of prior distributions. Then the posterior distributions of model parameters can be obtained for the penalized regression model using Bayesian methodology. These are only three examples of how "outside information" can be incorporated into our lasso tree method by using different penalties, i.e. by choosing different sets of tuning parameter λ's, or by using prior distributions. Many possibilities remain to be explored in this area.

One of the advantages of using the bootstrap selection, as we mentioned in Chapter 2, is that it provides a ranges of staging systems given k, the target number of stages, and the means to compare them. It gives inference procedures such as confidence intervals for not only the optimally selected, but all candidate systems, which enables us to evaluate the relative performance of any candidate systems of interest. This cannot be said for the lasso tree method; it directly computes the estimated coefficients which then gives the number of stages. In our experience, output from the lasso tree typically only contains one or two systems for a certain k. This might be a disadvantage of the lasso tree, that is, it would not be able to give, for example, the "best 10" staging systems for a given k.

5.2 CRC and Penalized Regression Methods

In Chapter 4, we applied the penalized logistic regression model to the risk stratification of CRC. Our models estimate the probability of developing advanced neoplasia over a prespecified time interval from data collected from a recently completed large cohort study. There have been a large amount of scientific investigations on the topic of CRC risk stratification and we hope our study would be a further contribution to the growing literature. Some of the "important" risk factors selected by our models confirm the findings in other studies. The capability for investigating various interactions among numerous risk factors should make our results valuable and our method a powerful tool for cancer risk modeling because currently very few, if any, scientific publications systematically consider interaction terms when there are many risk factors.

We acknowledge the preliminary nature of our data and analysis. Our data are reasonably representative of the US population, yet external validations will still be required to support further evaluation of the prognostic models across increasingly diverse settings. Among the fitted models, the ones with interactions are particularly interesting. Validation of these models will need to be further evaluated.

Although the fitted penalized models have shown advantages, they are still far from establishing a recommendation or changing medical practice. Many important practical questions remain: How can the fitted models be translated into clinical rule for deciding screening regimens? The sparsity-enforcing property of the penalty results in fewer variables as well as fewer categories in the final model, leading to a relatively small number of unique predicted values. These predicted values could be further "grouped" and used as decision rules for guiding a management strategy. If this can be done, then, should we recommend less frequent or no screening test for "low risk" patients? And what amount of benefit can be derived from this recommendation? The solution is also an essentially medical one which combines issues of treatment regimen distinctions, diagnostic ease, and clinical practice. Further work is needed from both the statistical community and the medical community to address these questions. Ultimately, we hope this preliminary attempt will be a contribution to the development of risk stratification models for CRC and aid physicians and their patients in deciding on screening regimens.

Applications of the penalized logistic regression to the risk stratification for other diseases would be an important topic for future work. This may include both well-established and emerging areas of disease stratification. For example, the penalized logistic regression model can be used to reexamine the well-known Gail model that established the risk assessment model for breast cancer [52]. Using data from the Breast Cancer Detection Demonstration Project (BCDDP), Gail et. al. developed a model for the absolute risk of breast cancer for women in a given age interval. It takes into account seven key risk factors for breast cancer, including age, age at first period, age at the time of the birth of her first child (or has not given birth), family history of breast cancer (mother, sister or daughter), number of past breast biopsies, number of breast biopsies showing atypical hyperplasia, and race/ethnicity. According to this model, women with a five-year risk of 1.67 percent or higher are classified as “high-risk”. Noticing that all seven risk factors are categorical and most of them ordinal, similar to the CRC risk model, the proposed penalized regression is well suited for modeling such breast cancer risk assessment models. A revisit of the BCDDP data or analyses of other breast cancer datasets using the penalized regression method might be of interest. In particular, the ability of our method to detect important interactions between risk factors might prove useful and add value to the existing risk assessment models for breast cancer.

Our penalized regression method has some limitations. The computation could be demanding when high-dimensional data are involved; the procedure in Section 4.3.2 might not be adequate for computing the estimates. The LAR algorithm of Efron et al. solves efficiently a wide spectrum of lasso problems [53] by exploiting the fact that the solution profiles are piecewise linear functions of the L1-bound. However, an LAR-style algorithm for quickly solving the fused lasso type problem can be much more complex because of the many possible ways that the active sets of constraints can change. This would present interesting challenges for future work.

Similar to the lasso tree method, the selection of the tuning parameter λ needs to be further investigated. A possible improvement of the current penalized regression method would be incorporating the sparse group lasso idea that utilizes different sets of penalties (or λ's) [54]. In the penalized model with interactions, interactions terms between two risk factors can be consider a group. It might be desirable to pose heavier penalties (λ_1) on the groups as a whole, such that when the interaction between two factors is weak, we drop the whole group all together. Only when the interaction as a group is strong enough the penalty will be lifted and another set of penalty (λ_2) will be posed on each individual interaction terms to determine their estimates. The benefit of using this sparse group lasso is that fewer interaction "groups" will remain in the final model, but when the interaction between two certain factors are present we would have more detailed information on each interaction terms.

Appendix A

Proofs of Consistency for the Lasso Tree

This Appendix gives the proofs of the two theorems presented in Chapter 3 which establish the consistency of the lasso tree grouping procedure. Before we present the proofs, we first establish the reparameterizations of the penalized regression model in (3.5). Then we follow steps similar to the proofs in Fan and Li [34] and Zhang and Lu [37] in proving the theorems.

A.1 Reparameterization

There are many different ways to reparameterize the model by rewriting the β's in the T×N table using their neighboring differences. Figure A.1 shows two examples. If we look at the $p \times q$ table by columns, then the β's can be reparameterized into $\theta = (\theta_{1,1}, ..., \theta_{p,q})^T$, where

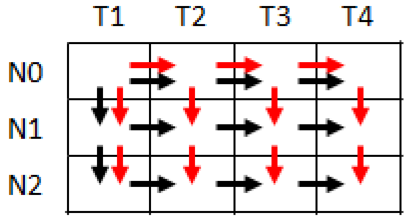
$$\theta_{j,k} = \begin{cases} \beta_{1,1} & \text{if } j = 1 \text{ and } k = 1 \\ \beta_{1,k} - \beta_{1,k-1} & \text{if } j = 1 \text{ and } k \geq 2 \\ \beta_{j,k} - \beta_{j-1,k} & \text{if } j \geq 2 \end{cases},$$

i.e. the inter-column (for the first row) and intra-column differences (red arrows in Figure A.1). Similarly, if we consider the $p \times q$ table by rows, then we can reparameterize the model with $\delta = (\delta_{1,1}, ..., \delta_{p,q})^T$, where

$$\delta_{j,k} = \begin{cases} \beta_{1,1} & \text{if } j = 1 \text{ and } k = 1 \\ \beta_{j,1} - \beta_{j-1,1} & \text{if } j \geq 2 \text{ and } k = 1 \\ \beta_{j,k} - \beta_{j,k-1} & \text{if } k \geq 2 \end{cases},$$

i.e. the inter-row (for the first column) and intra-row differences (black arrows in Figure A.1).

Figure A.1: Schematic showing two examples of reparameterizing the β's. Red arrows are reparameterizion with the inter-column and intra-column differences; and black arrows are reparameterizion with the inter-row and intra-row differences.



Under such parameterizations, θ and δ have the following relationship

$$
\begin{aligned}
\theta_{j,1} &= \delta_{j,1}, \quad j = 1, ..., p \ ; \\
\theta_{1,k} &= \delta_{1,k}, \quad k = 1, ..., q \ ; \\
\theta_{j,k} &= \sum_{l=1}^{k} \delta_{j,l} - \sum_{l=2}^{k} \delta_{j-1,l}, \quad j \geq 2 \text{ and } k \geq 2 \ ; \\
\delta_{j,k} &= \sum_{l=1}^{j} \theta_{l,k} - \sum_{l=2}^{j} \theta_{l,k-1}, \quad j \geq 2 \text{ and } k \geq 2 \ .
\end{aligned}
$$

Both parameterizations are equivalent mathematically to the original model. The adaptive lasso tree can then be reformulated under the two parameterizations as

$$
\text{argmin} \left\{ -\ell_z(\theta) + n\lambda_n \sum_{(j,k)\neq(1,1)} \frac{|\theta_{j,k}|}{|\tilde{\theta}_{j,k}|} + n\lambda_n \sum_{j=2}^{p} \sum_{k=2}^{q} \frac{|\sum_{l=1}^{j} \theta_{l,k} - \sum_{l=2}^{j} \theta_{l,k-1}|}{|\sum_{l=1}^{j} \tilde{\theta}_{l,k} - \sum_{l=2}^{j} \tilde{\theta}_{l,k-1}|} \right\} \tag{A.1}
$$

and

$$\operatorname{argmin}\left\{-\ell_{z'}(\delta)+n\lambda_n\sum_{(j,k)\neq(1,1)}\frac{|\delta_{j,k}|}{|\tilde{\delta}_{j,k}|}+n\lambda_n\sum_{j=2}^{p}\sum_{k=2}^{q}\frac{|\sum_{l=1}^{k}\delta_{j,l}-\sum_{l=2}^{k}\delta_{j-1,l}|}{|\sum_{l=1}^{k}\tilde{\delta}_{j,l}-\sum_{l=2}^{k}\tilde{\delta}_{j-1,l}|}\right\}, \tag{A.2}$$

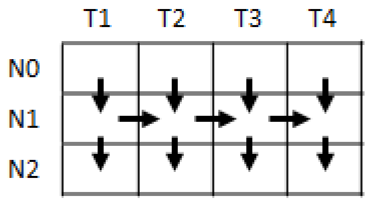
respectively. The third term in each formulation represents the differences between neighboring cells that are not directly represented by the θ's (or δ's), i.e. the "complementary arrows" in the T×N table. In general, these "complementary arrows" can always be written as some linear combinations of the parameters, e.g. $\nu(\theta)$, under any parameterization.

The grouping of β implies that some components of θ and δ are exactly 0. Let $\Theta = \{(j,k) : \theta^0_{j,k} \neq 0\}$ and $\Delta = \{(j,k) : \delta^0_{j,k} \neq 0\}$. Using the relationship between θ and δ, we can write

$$\Theta' = \{(j,k) : \sum_{l=1}^{j}\theta^0_{l,k}-\sum_{l=2}^{j}\theta^0_{l,k-1}\neq 0,\ j\geq 2, k\geq 2\} = \{(j,k) : (j,k)\in\Delta,\ j\geq 2, k\geq 2\},$$
$$\Delta' = \{(j,k) : \sum_{l=1}^{k}\delta^0_{j,l}-\sum_{l=2}^{k}\delta^0_{j-1,l}\neq 0,\ j\geq 2, k\geq 2\} = \{(j,k) : (j,k)\in\Theta,\ j\geq 2, k\geq 2\}\ .$$

Note that these are only two examples of reparameterization. By moving the arrows around we can achieve numerous other reparameterizations. A third example is shown in Figure A.2. These reparameterizations will be used in the following sections to prove the consistency of the lasso tree grouping procedure.

Figure A.2: Schematic showing a third example of reparameterizing the β's.



A.2 Proof of Theorem 1

Proving Theorem 1 is equivalent to proving the root-n consistency for the estimator under any reparameterization. Here we show the proof for the parameterization using θ. That is, when $\sqrt{n}\lambda_n = O_p(1)$, the penalized likelihood estimator $\hat{\theta}$ from (A.1) satisfies $\|\hat{\theta} - \theta^0\| = O_p(n^{-1/2})$, where θ^0 is the true value of θ.

Following the notation in Andersen and Gill [55], define the counting and at-risk processes $N_i(t) = I\{T_i \leq t, T_i \leq C_i\}$ and $Y_i(t) = I\{T_i \geq t, C_i \geq t\}$, respectively. The covariate z is allowed to be time-dependent, denote by $z(t)$. Without loss of generality, assume that $t \in [0,1]$. The Fisher information matrix

$$I(\theta^0) = \int_0^1 \nu(\theta^0, t) s^{(0)}(\theta^0, t) h_0(t) dt$$

is finite positive definite, where

$$\nu(\theta,t) = \frac{s^{(2)}(\theta,t)}{s^{(0)}(\theta,t)} - \Big(\frac{s^{(1)}(\theta,t)}{s^{(0)}(\theta,t)}\Big)\Big(\frac{s^{(1)}(\theta,t)}{s^{(0)}(\theta,t)}\Big)^T ,$$

and $s^{(k)}(\theta,t) = E[z(t)^{\bigotimes k}Y(t)\text{exp}\{\theta^T z(t)\}]$, $k = 0, 1, 2$. The regularity conditions (A) - (D) used in Andersen and Gill [55] are assumed throughout the section.

The log-partial likelihood $\ell(\theta)$ can be written as

$$\ell(\theta) = \sum_{i=1}^{n}\int_0^1 \theta^T z_i(s)dN_i(s) - \int_0^1 \log\Big\{\sum_{i=1}^{n} Y_i(s)\text{exp}(\theta^T z_i(s))\Big\}d\bar{N}(s), \qquad \text{(A.3)}$$

where $\bar{N} = \sum_{i=1}^n N_i$. Using Theorem 4.1 and Lemma 3.1 of Andersen and Gill [55], it follows that for each θ in a neighborhood of θ_0,

$$\frac{1}{n}\{\ell(\theta) - \ell(\theta^0)\} = \int_0^1 \Big[(\theta-\theta^0)^T s^{(1)}(\theta^0,t) - \log\Big\{\frac{s^{(0)}(\theta,t)}{s^{(0)}(\theta^0,t)}\Big\}s^{(0)}(\theta^0,t)\Big]\lambda_0(t)dt$$
$$+O_p(\frac{\|\theta-\theta^0\|}{\sqrt{n}}).$$

Define $s_n(\theta) = \partial\ell(\theta)/\partial\theta$ and $\nabla s_n(\theta) = \partial s_n(\theta)/\partial\theta^T$. We have $s_n(\theta^0)/\sqrt{n} = O_p(1)$ and $\nabla s_n(\theta^0)/n = I(\theta^0) + o_p(1)$.

Denote the penalized log partial likelihood function by

$$Q(\theta) = \ell_z(\theta) - n\lambda_n \sum_{(j,k)\neq(1,1)} \frac{|\theta_{j,k}|}{|\tilde{\theta}_{j,k}|} - n\lambda_n \sum_{j=2}^{p}\sum_{k=2}^{q} \frac{|\sum_{l=1}^{j}\theta_{l,k} - \sum_{l=2}^{j}\theta_{l,k-1}|}{|\sum_{l=1}^{j}\tilde{\theta}_{l,k} - \sum_{l=2}^{j}\tilde{\theta}_{l,k-1}|} .$$

It is sufficient to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{ \sup_{\|u\|=C} Q(\theta^0 + n^{-1/2}u) < Q(\theta^0) \right\} \geq 1 - \varepsilon. \tag{A.4}$$

This implies that with probability at least $1-\varepsilon$ that there exists a local maximum in the ball $\{\theta^0 + n^{-1/2}u :\ \|u\| \leq C\}$. Hence, there exists a local maximizer $\hat{\theta}$ such that $\|\hat{\theta} - \theta^0\| = O_p(n^{-1/2})$.

By Taylor's expansion, we have

$$\begin{aligned}
\frac{1}{n}\Big\{\ell(\theta^0 + n^{-1/2}u) - \ell(\theta^0)\Big\} &= \frac{1}{n}\Big(s_n^T(\theta^0)/\sqrt{n}\Big)u - \frac{1}{2n}u^T\Big(\nabla s_n(\theta^0)/n\Big)u + \frac{1}{n}u^T o_p(1)u \\
&= -\frac{1}{2n}u^T\Big\{I(\theta^0) + o_p(1)\Big\}u + \frac{1}{n}O_p(1)\sum_{j=1}^{p}\sum_{k=1}^{q}|u_{j,k}| \\
&= -C^2 n^{-1}O_p(1) + Cn^{-1}O_p(1)
\end{aligned}$$

where $u = (u_{1,1}, ..., u_{p,q})^T$. Then we have

$$\begin{aligned}
D_n(u) \equiv\ & \frac{1}{n}\Big\{Q(\theta^0 + n^{-1/2}u) - Q(\theta^0)\Big\} \\
=\ & \frac{1}{n}\Big\{\ell(\theta^0 + n^{-1/2}u) - \ell(\theta^0)\Big\} - \lambda_n \sum_{(j,k)\neq(1,1)} \left(\frac{|\theta^0_{j,k} + n^{-1/2}u_{j,k}|}{|\tilde{\theta}_{j,k}|} - \frac{|\theta^0_{j,k}|}{|\tilde{\theta}_{j,k}|}\right) \\
& - \lambda_n \sum_{j=2}^{p}\sum_{k=2}^{q}\left(\frac{|\sum_{l=1}^{j}\theta^0_{l,k} - \sum_{l=2}^{j}\theta^0_{l,k-1} + n^{-1/2}(\sum_{l=1}^{j}u_{l,k} - \sum_{l=2}^{j}u_{l,k-1})|}{|\sum_{l=1}^{j}\tilde{\theta}_{l,k} - \sum_{l=2}^{j}\tilde{\theta}_{l,k-1}|}\right. \\
& \left. - \frac{|\sum_{l=1}^{j}\theta^0_{l,k} - \sum_{l=2}^{j}\theta^0_{l,k-1}|}{|\sum_{l=1}^{j}\tilde{\theta}_{l,k} - \sum_{l=2}^{j}\tilde{\theta}_{l,k-1}|}\right)
\end{aligned}$$

$$\begin{aligned}&\leq -C^2 n^{-1} O_p(1) + C n^{-1} O_p(1) + \frac{1}{\sqrt{n}} \lambda_n \sum_{(j,k)\in\Theta} \frac{|u_{j,k}|}{|\tilde{\theta}_{j,k}|} \\ &+ \frac{1}{\sqrt{n}} \lambda_n \sum_{(j,k)\in\Theta'} \frac{|\sum_{l=1}^{j} u_{l,k} - \sum_{l=2}^{j} u_{l,k-1}|}{|\sum_{l=1}^{j} \tilde{\theta}_{l,k} - \sum_{l=2}^{j} \tilde{\theta}_{l,k-1}|} .\end{aligned} \tag{A.5}$$

Since the maximum likelihood estimator $\tilde{\theta}$ satisfies $\|\tilde{\theta}-\theta^0\| = O_p(n^{-1/2})$, we have, for $(j,k) \in \Theta$,

$$\begin{aligned}\frac{1}{|\tilde{\theta}_{j,k}|} &= \frac{1}{|\theta^0_{j,k}|} - \frac{\text{sign}(\theta^0_{j,k})}{(\theta^0_{j,k})^2}(\tilde{\theta}_{j,k} - \theta^0_{j,k}) + o_p(|\tilde{\theta}_{j,k} - \theta^0_{j,k}|) \\ &= \frac{1}{|\theta^0_{j,k}|} + \frac{O_p(1)}{\sqrt{n}},\end{aligned}$$

and for $(j,k) \in \Theta'$,

$$\begin{aligned}\frac{1}{|\sum_{l=1}^{j} \tilde{\theta}_{l,k} - \sum_{l=2}^{j} \tilde{\theta}_{l,k-1}|} &= \frac{1}{|\sum_{l=1}^{j} \theta^0_{l,k} - \sum_{l=2}^{j} \theta^0_{l,k-1}|} - \frac{\text{sign}(\sum_{l=1}^{j} \theta^0_{l,k} - \sum_{l=2}^{j} \theta^0_{l,k-1})}{(\sum_{l=1}^{j} \theta^0_{l,k} - \sum_{l=2}^{j} \theta^0_{l,k-1})^2} \times \\ &\quad \left\{ \sum_{l=1}^{j} (\tilde{\theta}_{l,k} - \theta^0_{l,k}) - \sum_{l=2}^{j} (\tilde{\theta}_{l,k-1} - \theta^0_{l,k-1}) \right\} + o_p(|\tilde{\theta} - \theta^0|) \\ &= \frac{1}{|\sum_{l=1}^{j} \theta^0_{l,k} - \sum_{l=2}^{j} \theta^0_{l,k-1}|} + \frac{O_p(1)}{\sqrt{n}}.\end{aligned}$$

In addition, since $\sqrt{n}\lambda_n = O_p(1)$, we have the third term in (A.5)

$$\begin{aligned}\frac{1}{\sqrt{n}} \lambda_n \sum_{(j,k)\in\Theta} \frac{|u_{j,k}|}{|\tilde{\theta}_{j,k}|} &= \frac{1}{\sqrt{n}} \lambda_n \sum_{(j,k)\in\Theta} \Big(\frac{|u_{j,k}|}{|\theta^0_{j,k}|} + \frac{|u_{j,k}|}{\sqrt{n}} O_p(1) \Big) \\ &\leq \frac{C}{n} \sqrt{n} \lambda_n O_p(1) = C n^{-1} O_p(1),\end{aligned}$$

and similarly, the fourth term in (A.5)

$$\frac{1}{\sqrt{n}}\lambda_n \sum_{(j,k)\in\Theta'} \frac{|\sum_{l=1}^{j} u_{l,k} - \sum_{l=2}^{j} u_{l,k-1}|}{|\sum_{l=1}^{j} \tilde{\theta}_{l,k} - \sum_{l=2}^{j} \tilde{\theta}_{l,k-1}|} \leq Cn^{-1}O_p(1).$$

Therefore in (A.5), by choosing a sufficiently large C, the first term is of the order C^2n^{-1}. The rest of the terms are of the order Cn^{-1}, which are dominated by the first term. Hence (A.4) holds and this completes the proof of Theorem 1. When the β's are naturally ordered as in the cancer staging example, i.e. $\beta_{j,1} \leq \ldots \leq \beta_{j,q}$ and $\beta_{1,k} \leq \ldots \leq \beta_{p,k},\ \ j = 1, \ldots, p,\ k = 1, \ldots, q$, the penalized likelihood estimator (3.13) will be obtained under the ordering constraint and the absolute values in the penalties can be dropped. This does not impact the proofs and the same result can be reached.

A.3 Proof of Theorem 2

We now show that the adaptive lasso tree identifies the correct grouping. It is equivalent to showing that, for any difference between two neighboring cells in the T×N table (i.e. arrows in Figure A.1 or A.2), if its true value is 0, i.e. the two cells belong to the same group, then its estimator must be exactly 0 under the lasso tree penalized model. This must be true under any parameterization; the penalized likelihood to be maximized is convex and hence there must exist a global solution. Under the two parameterizations in Section A.1, for example, this means $\hat{\theta}_{j,k} = 0$ for all $(j,k) \in \Theta^C$ and $\hat{\delta}_{j,k} = 0$ for all $(j,k) \in \Delta^C$.

With some abuse of notation, here and below we will use θ to denote the differences between neighboring cells under any parameterization. It is sufficient to show that, for any given $\theta_{j,k}$ such that $\theta^0_{j,k} = 0$, there exists a parameterization under which $\hat{\theta}_{j,k} = 0$.

Since the penalized likelihood is maximized subject to $\beta_{j,1} \leq \ldots \leq \beta_{j,q}$ and $\beta_{1,k} \leq \ldots \leq \beta_{p,k},\ j = 1, \ldots, p,\ k = 1, \ldots, q$, the reparameterization is maximized subject to

$$\theta \geq 0 \ \text{ and } \ \nu(\theta) \geq 0,$$

where $\nu(\theta) = \{\nu_l(\theta),\ l = 1, ..., m\}$ are m linear combinations of θ representing the "complementary arrows" in the T×N table, i.e. the m pairwise differences which are not presented by θ's directly under this particular parameterization. The penalized log partial likelihood function can be written as

$$Q(\theta) = \ell_z(\theta) - n\lambda_n \sum \frac{\theta_{j,k}}{|\tilde{\theta}_{j,k}|} - n\lambda_n \sum \frac{\nu_l(\theta)}{|\nu_l(\tilde{\theta})|}\ . \tag{A.6}$$

Let θ^* denote a sequence of θ satisfying $\|\theta^* - \theta^0\| = O_p(n^{-1/2})$. Proving Theorem 2 is then equivalent to proving that, for any $\theta_{j,k}$ such that $\theta^0_{j,k} = 0$, there exists a parameterization under which the maximizer of $Q(\theta^*)$ is $\hat{\theta}^*$ where $\hat{\theta}^*_{j,k} = 0$. It is sufficient to show that, for any given $\theta_{j,k}$ such that $\theta^0_{j,k} = 0$, with probability tending to 1,

$$\frac{\partial Q(\theta)}{\partial \theta_{j,k}} < 0 \quad \text{for} \quad \theta_{j,k} \in (0, Cn^{-1/2}) \tag{A.7}$$

for any θ satisfying $\|\theta - \theta^0\| = O_p(n^{-1/2})$.

In order to show (A.7), we will examine the partial derivatives of the three components of $Q(\theta)$ respectively. The first term involves the partial derivative of $\ell(\theta)$. For each θ in a neighborhood of θ^0, by (A.3) and Taylor expansion,

$$\ell(\theta) = \ell(\theta^0) + nf(\theta) + O_p(\sqrt{n}\|\theta - \theta^0\|),$$

where $f(\theta) = -\frac{1}{2}(\theta - \theta^0)^T\{I(\theta^0) + o(1)\}(\theta - \theta^0)$. We have

$$\frac{\partial \ell(\theta)}{\partial \theta_{j,k}} = O_p(n^{1/2}). \tag{A.8}$$

The second term of $Q(\theta)$ involves $\theta_{j,k}$ directly. Since $\sqrt{n}(\tilde{\theta}_{j,k} - 0) = O_p(1)$ and $\sqrt{n}\lambda_n = O_p(1)$, we have the partial derivative

$$\begin{aligned} -n\lambda_n \frac{1}{|\tilde{\theta}_{j,k}|} &= -(n\lambda_n)n^{1/2}\frac{1}{|\sqrt{n}\tilde{\theta}_{j,k}|} \\ &= -(n\lambda_n)O_p(n^{1/2}) = -O_p(n), \end{aligned} \tag{A.9}$$

which dominates the first term when n is large. If the third term in (A.6), the linear combinations, does not involve $\theta_{j,k}$, then

$$\frac{\partial Q(\theta)}{\partial \theta_{j,k}} = O_p(n^{1/2}) - O_p(n) \tag{A.10}$$

Hence $\partial Q(\theta)/\partial \theta_{j,k}$ is negative when n is large, which establishes (A.7) and completes the proof.

When $\theta_{j,k}$ is involved in the linear combinations, we can always find a parameterization where $\theta_{j,k}$ has a positive contribution to all the linear combinations in which it is involved. This way, the partial derivative of the third term becomes

$$-n\lambda_n \sum \frac{1}{|\nu_l(\tilde{\theta})|},$$

where the summation is over all $\nu_l(\theta)$ that involve $\theta_{j,k}$. Notice that $\nu(\theta)$ are also differences between neighboring cells, and some of their true values might be 0. Without loss of generality, denote by $v_l(\theta)$, $l = 1, ..., m_1$, the linear combinations that involve $\theta_{j,k}$, among which $v_l(\theta)$, $l = 1, ..., m_0$, $m_0 \leq m_1$, satisfy $\nu_l(\theta^0) = 0$. Then we have $\sqrt{n}\nu_l(\tilde{\theta}) = O_p(1)$, $l = 1, ..., m_0$, and

$$
\begin{aligned}
-n\lambda_n \sum_{l=1}^{m_1} \frac{1}{|\nu_l(\tilde{\theta})|} &= -n\lambda_n n^{1/2} \sum_{l=1}^{m_0} \frac{1}{|\sqrt{n}\nu_l(\tilde{\theta})|} - n\lambda_n \sum_{l=m_0+1}^{m_1} \frac{1}{|\nu_l(\tilde{\theta})|} \\
&= -(n\lambda_n) n^{1/2} O_p(1) - (n\lambda_n) O_p(1) \\
&= -O_p(n) - O_p(n^{1/2}).
\end{aligned} \tag{A.11}
$$

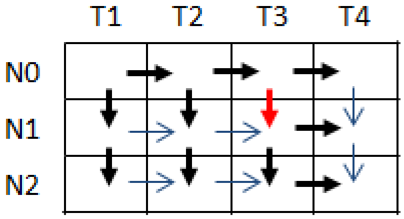
Hence the derivative becomes

$$
\begin{aligned}
\frac{\partial Q(\theta)}{\partial \theta_{j,k}} &= O_p(n^{1/2}) - O_p(n) - O_p(n) - O_p(n^{1/2}) \\
&= O_p(n^{1/2}) - O_p(n),
\end{aligned} \tag{A.12}
$$

which is negative when n is large. This again establishes (A.7) and completes the proof.

To see how we can always find a parameterization where $\theta_{j,k}$ has a positive contribution to all the linear combinations $\nu(\theta)$ in which it is involved, let us look at one example. Figure A.3 shows a parameterization where the thick black (and red) arrows represent the θ's and the thin blue arrows represent the pairwise differences which are not presented by θ's directly. Suppose the one red, vertical arrow is the $\theta_{j,k}$ to be proved. Then under this parameterization, it has a positive contribution to all the linear combinations, i.e. the blue arrows, in which it is involved. In fact, in general $\theta_{j,k}$ would only be involved in the differences / arrows one column to the left and one column to the right of itself. In this example, they are the two horizontal blue arrows between T2 and T3 and the two vertical blue arrows in column T4. Among them, the differences to the left of $\theta_{j,k}$ always involve a positive $\theta_{j,k}$; these arrows are always pointing at $\theta_{j,k}$. The differences to the right, on the other hand, have two different effects. The horizontal arrows to the right, pointing away from $\theta_{j,k}$, always involve a negative $\theta_{j,k}$. While the vertical arrows to the right, parallel to $\theta_{j,k}$, involve a positive $\theta_{j,k}$. In this example, blue arrows to the right are only present in vertical forms, and hence only involve positive $\theta_{j,k}$. Therefore, as long as all $\nu(\theta)$ to the right of $\theta_{j,k}$ (the one we want to prove in (A.7)) are vertical arrows, i.e. differences between rows, we will have positive involvement of $\theta_{j,k}$ in all $\nu(\theta)$, which assures the proof of Theorem 2. This kind of parameterization is in fact not hard to find for any $\theta_{j,k}$, that is, any arrow in the two-way table.

FIGURE A.3: Schematic showing an example of parameterization (in thick arrows) where $\theta_{j,k}$ (the thick red arrow) has a positive contribution to all the linear combinations $\nu(\theta)$ (thin blue arrows) in which it is involved.



Bibliography

[1] American Joint Committee on Cancer. *AJCC Cancer Staging Manual*. Springer, New York, 6th edition, 2002.

[2] American Joint Committee on Cancer. *AJCC Cancer Staging Manual*. Springer, New York, 7th edition, 2010.

[3] National Cancer Institute. National cancer institute fact sheet: Cancer staging, 2004. http://www.cancer.gov/cancertopics/factsheet/detection/staging.

[4] M. Gönen and M. R. Weiser. Whither TNM? *Semin Oncol*, 37:27–30, 2010.

[5] A. B. Benson, D. Schrag, M. R. Somerfield, et al. American Society of Clinical Oncology recommendations on adjuvant chemotherapy for stage II colon cancer. *J Clin Oncol*, 22:3408–3419, 2004.

[6] P. A. Groome, K. M. Schulze, W. J. Mackillop, et al. A comparison of published head and neck stage groupings in carcinomas of the tonsillar region. *Cancer*, 92: 1484–1494, 2001.

[7] A. W. Lee, W. Foo, S. C. Law, et al. Staging of nasopharyngeal carcinoma: From Ho's to the new UICC system. *Int J Cancer*, 842:179–187, 1999.

[8] C. B. Begg, L. D. Cramer, E. S. Venkatraman, and J. Rosai. Comparing tumor staging and grading systems: a case study and a review of the issues, using thymoma as a model. *Statistics in Medicine*, 19:1997–2014, 2005.

[9] T. Hothorn and A. Zeileis. Generalized maximally selected statistics. *Biometrics*, 64:1263–1269, 2008.

[10] L. Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24:2350–2383, 1996.

[11] M. R. Weiser, M. Gönen, J. F. Chou, M. W. Kattan, and D. Schrag. Predicting survival after curative colectomy for cancer: individualizing colon cancer staging. *J Clin Oncol*, 22:4796–4802, 2011.

[12] American Joint Committee on Cancer. Missions and objectives, 2000. http://www.cancerstaging.org.

[13] National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch. Surveillance, Epidemiology, and End Results (SEER) Program Public-Use Data, 2000. http://www.seer.cancer.gov.

[14] W. Sauerbrei and M. Schumacher. A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine*, 11: 2093–2109, 1992.

[15] M. Schemper and J. Stare. Explained variation in survival analysis. *Statistics in Medicine*, 15:1999–2012, 1996.

[16] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18:2529–2545, 1999.

[17] P. W. Kasteleyn. The statistics of dimers on a lattice. *Physica*, 27:1209–1225, 1961.

[18] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[19] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati. Evaluating the yield of medical tests. *Journal of the American Medical Association*, 247: 2543–2546, 1982.

[20] F. E. Harrell, K. L. Lee, and D. B. Mark. Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387, 1996.

[21] D. Faraggi and R. Simon. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Statistics in Medicine*, 15: 2203–2213, 1996.

[22] B. Lausen and M. Schumacher. Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Computational Statistics and Data Analysis*, 21:307–326, 1996.

[23] D. Siegmund. Confidence sets in change-point problems. *International Statistical Review*, 56:31–48, 1988.

[24] Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA*, 285:2486–2497, 2001.

[25] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1983.

[26] R. Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16:385–395, 1997.

[27] R. Tibshirani, M. Saunders, S. Rosset, and J. Zhu. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, 67(1):91–108, 2005.

[28] R. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39:1335–1371, 2011.

[29] G. E. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[30] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[31] P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1978.

[32] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7: 221–264, 1997.

[33] J. Fan and R. Li. Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30:74–79, 2002.

[34] J. Fan and R. Li. Variables selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[35] H. Wang, G. Li, and C. L. Tsai. Regression coefficient and autoregressive order shrinkage and selection via lasso. *Journal of the Royal Statistical Society Series B*, 69:63–78, 2006.

[36] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

[37] H. Zhang and W. Lu. Adaptive-lasso for Cox's proportional hazards model. *Biometrika*, 94:691–703, 2007.

[38] G. Wahba. A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, 13:1378–1402, 1985.

[39] K. C. Li. Asymptotic optimality for cp, cl , cross-validation and generalized cross validation: discrete index set. *The Annals of Statistics*, 15:958–975, 1987.

[40] J. H. Watson, H. C. Sox, R. K. Neff, and L. Goldman. Clinical prediction rules: Application and methodological standards. *N Engl J Med*, 313:793–799, 1985.

[41] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression.* John Wiley & Sons, New York, 1989.

[42] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun. Cancer statistics, 2009. *CA Cancer J Clin*, 59:225–249, 2009.

[43] B. Levin B, D. A. Lieberman, B. McFarland, et al. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology*, 134:1570–1595, 2008.

[44] D. K. Rex, D. A. Johnson, J. C. Anderson, et al. American College of Gastroenterology guidelines for colorectal cancer screening 2008. *Am J Gastroenterol*, 104:739–750, 2009.

[45] U.S. Preventive Services Task Force. Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med*, 149: 627–637, 2008.

[46] T. F. Imperiale, D. R. Wagner, C. Y. Lin, G. N. Larkin, J. D. Rogge, and D. F. Ransohoff. Risk of advanced proximal neoplasms in asymptomatic adults according to the distal colorectal findings. *N Engl J Med*, 343:169–174, 2000.

[47] E. Quintero, A. Castells, L. Bujanda, et al. Colonoscopy versus fecal immunochemical testing in colorectal-cancer screening. *N Engl J Med*, 366:697–706, 2012.

[48] A. N. Freedman, M. L. Slattery, R. Ballard-Barbash, G. Willis, B. J. Cann, D. Pee, M. H. Gail, and R. M. Pfeiffer. Colorectal cancer risk prediction tool for

white men and women without known susceptibility. *J Clin Oncol*, 27:686–693, 2009.

[49] P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society B*, 46:149–192, 1984.

[50] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. *Biometrics*, 44:837–845, 1988.

[51] S. H. Jung. Regression analysis for long-term survival rate. *Biometrika*, 83: 227–232, 1996.

[52] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81:1879–86, 1989.

[53] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32:407–499, 2004.

[54] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *Technical report, Department of Statistics, Stanford University*, 2010.

[55] P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10:1100–1120, 1982.