Fall 2014

Nonconvex selection in nonparametric additive models

Xiangmin Zhang
University of Iowa

NONCONVEX SELECTION IN NONPARAMETRIC ADDITIVE MODELS

by

Xiangmin Zhang

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Statistics
in the Graduate College of
The University of Iowa

December 2014

Thesis Supervisor: Professor Jian Huang

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Xiangmin Zhang

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Statistics at the December 2014 graduation.

Thesis Committee: _____
Jian Huang, Thesis Supervisor

_____
Patrick Breheny

_____
Kung-Sik Chan

_____
Aixin Tan

_____
Dale Zimmerman

ACKNOWLEDGEMENTS

I am deeply indebted to many people for the completion of this work. I am very grateful that my adviser Professor Jian Huang provides me with kind guidance and support and that he is very patient with me all the time. His dedication to creative and high-quality research always inspires me to work harder. I am also very thankful to my dissertation committee members, Professors Patrick Breheny, Kung-Sik Chan, Aixin Tan and Dale Zimmerman, whose suggestions and comments lead to many improvements of the work. In addition, I would like to thank all the professors at the department of Statistics and Actuarial Science for their dedicated service to the department and students. They taught me the foundations and applications of Statistics and prepared me well for any current and future research.

Last, but not least, I would like to express my gratitude to my father Yuzhong Zhang, my mother Shuying Shen and my friends Ying Hu and Dr. Yeh-Fong Chen for being the best family and friends that I can ever hope for. I would not be able to go this far had it not been their long-time selfless support, encouragement and love.

ABSTRACT

High-dimensional data offers researchers increased ability to find useful factors in predicting a response. However, determination of the most important factors requires careful selection of the explanatory variables. In order to tackle this challenge, much work has been done on single or grouped variable selection under the penalized regression framework. Although the topic of variable selection has been extensively studied under the parametric framework, its extensions to more flexible nonparametric models are yet to be explored.

In order to implement the variable selection in nonparametric additive models, I introduce and study two nonconvex selection methods under the penalized regression framework, namely the group MCP and the adaptive group LASSO, aiming at improvements on the selection performances of the more widely known group LASSO method in such models. One major part of the dissertation focuses on the theoretical properties of the group MCP and the adaptive group LASSO. I derive their selection and estimation properties. The application of the presently proposed methods to nonparametric additive models are further examined using simulation. Their applications to areas such as the economics and genomics are presented as well. Under both the simulation studies and data applications, the group MCP and the adaptive group LASSO have shown their advantages over the more traditionally used group LASSO method.

For the proposed adaptive group LASSO that uses the newly proposed weights,

whose recursive application is therefore never studied before, I also derive its theoretical properties under a very general framework. Simulation studies under linear regression are included.

In addition to the theoretical and empirical investigations, throughout the dissertation, several other important issues have been briefly discussed, including the computing algorithms and different ways of selecting tuning parameters.

PUBLIC ABSTRACT

Nowadays in areas such as genetics, behavioral sciences and banking and finance, high dimensional data (i.e. data that have a much greater number of variables than the sample size) are more and more frequently available. On one hand, high-dimensional data offer researchers increased ability to find useful factors in building statistical models and predicting a response variable. On the other hand, determination of the most important factors requires careful selection of variables. In order to tackle this challenge, much work has been done on variable selection using statistical methods. Although the topic of variable selection has been studied under some modeling framework, its extensions to many other models are yet to be explored.

In order to implement the variable selection in the so-called nonparametric additive models that are very useful in many scientific areas, I introduce and study two variable selection methods in my dissertation, aiming at improvements on the variable selection performances of the more traditional variable selection method in such models. The dissertation studies the theoretical properties of the proposed methods and examines these methods using extensive simulation studies. The dissertation also presents their applications to economics and genomics. Under both the simulation studies and data applications, the proposed methods have shown their advantages over the more traditional method.

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

Variable selection is a fundamental topic in statistical modeling. When constructing regression models, people want to select a parsimonious set of variables that adequately explain and predict a response variable without adding too much noise. Traditional variable selection approaches include subset selection and stepwise procedures, such as forward selection and backward elimination. Even when the number of explanatory variables is moderately large, these approaches can become very computationally expensive. Subset selection also suffers from selection instability (Breiman, 1996). More critically, traditional approaches are only applicable when the number of explanatory variables p is smaller than the sample size n. However, nowadays in areas such as bio-informatics, behavioral sciences and finance, high dimensional data are more frequently available. While they offer researchers increased ability to find useful factors in predicting a response, determination of the most important factors requires careful selection of explanatory variables. Therefore, from the perspective of computational geneticists, marketing economists and other researchers in areas where high-dimensional data are frequently encountered, the variable selection aspect of the data analysis presents an enormous challenge. For example, in recent research on the rheumatoid arthritis drug Methotrexate (MTX) (Aslibekyan et al., 2013), researchers considered including genetic variants in the modeling of MTX response. They would like to test the associations between 863 known pharmacogenetic variants and the MTX response in 471 trial participants who completed the trial. In order to address

the large p small n issues such as the one that arose in the above example, the penalized regression method has been developed over recent years to implement variable selection.

Penalized regression is formulated through the objective function Q as the sum of a loss function L and a penalty function P:

$$Q(\boldsymbol{\beta}; \lambda, \gamma) \quad = \quad L(\boldsymbol{\beta}) + P(\boldsymbol{\beta}; \lambda, \gamma), \tag{1.1}$$

where λ and γ are penalty parameters that control how much and how the model coefficient vector $\boldsymbol{\beta}$ is penalized. Consider the simple case of a linear regression model, the centered response vector

$$\boldsymbol{Y} = \sum_{j=1}^{p} \boldsymbol{X}_j \beta_j + \boldsymbol{\epsilon}, \tag{1.2}$$

where for $j \in 1, \ldots, p$, $\boldsymbol{X}_j$ is the $n \times 1$ predictor vector, β_j is the regression coefficient and $\boldsymbol{\epsilon}$ is the random error vector. Denote $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$. In what follows, $\|.\|_2$ denotes the Euclidean norm. For linear regression, the loss function is usually quadratic and the penalty function can be of various forms. Under this general framework, the estimator of $\boldsymbol{\beta}$ is the minimizer to

$$Q(\boldsymbol{\beta}; \lambda, \gamma) \quad = \quad \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} \rho(\beta_j; \lambda, \gamma). \tag{1.3}$$

With a properly chosen penalty function, some β_j's end up with estimated values of

zero, whose corresponding explanatory variables are considered to be excluded from the final model, thus resulting in a reduction of dimensionality.

An early version of the penalized regression method is the ridge regression proposed by Hoerl and Kennard (1970). They studied the l_2 penalty function $\rho(t;\lambda) = \lambda t^2/2$, which alleviates the ill-conditioning of the design matrix. The resulting estimates are biased and none are constrained to zero. No variable selection is performed under ridge regression. Frank and Friedman (1993) generalized the above l_2 penalty to a broad family of l_γ penalty, or the bridge penalty: $\rho(t;\lambda,\gamma) = \lambda|t|^\gamma$, $\gamma > 0$. Bridge regression includes ridge regression and subset selection as special cases when $\gamma = 2$ and $\gamma = 0$, respectively. The bridge penalty facilitates variable selection when $\gamma \leq 1$. Its variable selection properties when $\gamma = 1$ are further studied (Tibshirani, 1996; Chen, Donoho and Saunders, 1998) and designated the name Least Absolute Shrinkage and Selection Operator (LASSO) by Tibshirani (1996). Since then, much research has been done on the LASSO. The LASSO has been shown to have selection consistency under certain regularization conditions by Meinshausen and Bühlmann (2006), Zhao and Yu (2006) and Zhang (2009), among others, which loosely speaking indicates that the method identifies and includes the true "important" explanatory variables in the final model with very high probability. The LASSO does not have the oracle property as if the "important" as well as the "unimportant" explanatory variables have been known before model fitting. But sufficient conditions for oracle inequalities and upper bounds for the estimation error have been derived by Bunea, Tsybakov and Wegkamp (2009), Zhang and Huang (2008), Bickel, Ritov and Tsybakov (2009)

and Meinshausen and Yu (2009), among others. Its prediction performance has also been examined (Greenshtein and Ritov, 2004).

Building on the variable selection strengths of the LASSO, concave penalties have been proposed. They impose milder regularity conditions and remedy the over-selection of the LASSO method when the number of explanatory variables increases. They can reduce the effect of the bias on estimation induced by the LASSO penalty function. Well-known concave penalties include the Smoothly Clipped Absolute Deviation (SCAD) proposed by Fan and Li (2001), with the penalty function

$$\rho_{SCAD}(t;\lambda,\gamma) = \lambda \int_0^{|t|} \left[\mathbf{1}_{\{x\leq\lambda\}} + \frac{(\gamma\lambda - x)_+}{(\gamma-1)\lambda} \mathbf{1}_{\{x>\lambda\}} \right] dx; \tag{1.4}$$

and the Minimax Concave Penalty (MCP) by Zhang (2010) with the penalty function

$$\rho_{MCP}(t;\lambda,\gamma) = \lambda \int_0^{|t|} \left(1 - \frac{x}{\gamma\lambda} \right)_+ dx. \tag{1.5}$$

The LASSO, MCP and SCAD are closely related. The penalty functions of the SCAD and MCP are derived from modifications of the first derivative, or the penalty rate, of the LASSO. Fan and Li (2001) and Zhang (2010) discussed the connection of the SCAD and LASSO, and the MCP and LASSO respectively in their papers. They also proved the oracle property of their proposed methods.

Another improvement on the LASSO that has received much attention is the adaptive LASSO proposed by Zou (2006). It imposes data-dependent weights on the

l_1 norms in the penalty function, which leads to the following objective function

$$Q_{adaptive}(\boldsymbol{\beta};\lambda) = \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} \lambda w_j |\beta_j|. \tag{1.6}$$

In contrast to the LASSO, the adaptive LASSO with weights proposed in Zou (2006) is proved to have oracle property even when p diverges (Huang, Ma and Zhang, 2008). The adaptive LASSO requires initial values of the coefficients, which we denote as $\tilde{\beta}_j$, $j = 1, \ldots, p$, in the following. Originally, Zou proposed $w_j = 1/\tilde{\beta}_j$. More recently, Huang and Zhang (2012) extended the weights to more general forms and obtained results for a number of important models, including the logistic and log-linear regressions. The adaptive LASSO also has applications to various other models, for example Cox's proportional hazard model (Zhang and Lu, 2007) and the graphical model (Fan, Feng and Wu, 2009).

Some recent studies shift attention not only to the use of the concave penalties or adaptive penalties for their superiority in selection properties, but also to extensions of the LASSO, such as the group LASSO (Yuan and Lin, 2006), for their flexibility in dealing with other types of selection problems. Group structure arises naturally in many scientific settings. In genetics studies, take a nucleic acid sequence for example, each nucleotide within the DNA can be expressed by one of the four nucleobases: adenine (A), cytosine (C), guanine (G) and thymine (T). To deal with each categorical nucleotide variable, we can expand it with three (i.e. number of levels - 1) dummy variables, which are considered as forming one group. Another example is when we

use nonparametric additive models, we can express each additive components by basis functions. In both examples, when we try to determine the most important variables, we are no longer directly selecting a single nucleotide or a single component, but a group of dummy variables corresponding to a nucleotide letter expression or a group of basis functions corresponding to a component. Accordingly the individual variable selection is converted to grouped variable selection. We refer to such selection as group selection for the rest of the dissertation. Assume that we are given fixed and pre-determined J groups of variables. We denote the coefficient vector for each group as $\boldsymbol{\beta}_j$, $j = 1, \ldots, J$. More specifically, $\boldsymbol{\beta}_j$ is a vector that contains all coefficients corresponding to all variables of the jth group. $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$. The objective function of the group LASSO for linear regression is

$$Q_{group}(\boldsymbol{\beta}; \lambda) = \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{J} \lambda \|\boldsymbol{\beta}_j\|_2. \tag{1.7}$$

As a result of minimizing (1.7) with respect to $\boldsymbol{\beta}$, the coefficients of the same group are estimated to be all zero's or all non-zero's. Therefore, the group LASSO is particularly suitable for selecting variables when they have inherent group structures. After Yuan and Lin (2006) proposed the group LASSO for linear regression, Kim, Kim and Kim (2006) extended the group LASSO to generalized linear models. Meier, van de Geer and Bühlmann (2008) studied the group LASSO for logistic regression. Zhao, Rocha and Yu (2009) generalized the group LASSO to a family of composite absolute penalties.

So far, much work has been done by others on variable selection under the parametric framework. Nevertheless, underlying parametric assumptions may be too restrictive under many circumstances. In areas such as macroeconomics and genomics, sometimes nonparametric modeling may be more appropriate at the frequent presence of non-linearity in empirical data. Sometimes little is known or justified about the assumptions required by a parametric model, which entitles the use of a more flexible nonparametric model. Due to the importance of nonparametric models, the cases when the number of nonparametric components is smaller than n have been studied in a huge body of literature. Specifically for nonparametric additive models, Stone (1985, 1986) showed that the estimates based on the spline approximations achieve the optimal rate of convergence under a general fixed number of components that is smaller than n. Hastie and Tibshirani (1990) provided a comprehensive presentation of many modeling issues and properties of additive models. Some recent literature has already taken a step forward by studying variable selection in nonparametric models. For example, Bach (2008) and Huang, Horowitz and Wei (2010) applied the group LASSO to nonparametric additive models. Bach (2008) established selection properties assuming a fixed number of covariates while Huang, Horowitz and Wei (2010) proved selection properties under the more general $p > n$ case. Huang, Horowitz and Wei (2010) also proposed an adaptive group LASSO under the same modeling framework. Meier, van de Geer and Bühlmann (2009) proposed a penalized least-square estimator other than the group LASSO for nonparametric additive models, without establishing the selection consistency properties of their procedure. Variable selection

in other nonparametric models, for example varying coefficient models, has also been studied by Wang and Xia (2007), and Wei, Huang and Li (2011), among others.

In this work, I implement new variable selection methods in nonparametric additive models where basis functions are present. More specifically, I investigate two group selection methods, namely the group MCP and the adaptive group LASSO, aiming at improvements on the selection performances of the group LASSO in such models. In Chapter 2, I detail the general settings of the nonparametric additive models in use and study the application of the group MCP in such context at the theoretical level. I briefly demonstrate the proposed method with a data example at the end of this chapter. In Chapter 3, I study a new adaptive group LASSO selection method that uses different data dependent weights from those used in the more traditional adaptive group LASSO method. I lay the theoretical foundation of the proposed new method for general models. I further study the method with simulation studies on linear regression. In Chapter 4, I extend the method introduced in Chapter 3 to nonparametric additive models. I present the theoretical results and include more extensive numerical studies, including simulation studies and data examples, at the end of this chapter. In Chapter 5, I summarize the work and discuss several issues with existing research in this area.

CHAPTER 2
THE GROUP MCP IN NONPARAMETRIC ADDITIVE MODELS

Chapter 2 has two main goals. As briefly discussed in Chapter 1, we aim at studying the variable selection problem in nonparametric additive models. So one main goal of Chapter 2 is to lay out the nonparametric model settings that we use for the rest of the dissertation, as well as the group selection framework under such a model. We also cover various details such as knot selection for spline expansion and general form of the objective function.

Another goal of Chapter 2 is to present the theoretical results of the group MCP under nonparametric additive models. Both the selection performance and estimation performance are examined.

In Section 2.3, we further describe a fast and stable computing algorithm that is used in subsequent numerical studies. In Section 2.4, we present a short data example for demonstration of the proposed method.

2.1 Variable selection in nonparametric additive models

Let $\boldsymbol{X}_j = (X_{1j}, \ldots, X_{nj})', j = 1, \ldots, p$, be the covariate vector and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ be the n-dimensional response vector. Consider the nonparametric additive model, the centered response

$$Y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i, i = 1, \ldots, n, \tag{2.1}$$

where f_j's are unknown functions and ϵ_i is the random error with mean zero and finite variance σ^2. Let $q \in [0, p]$ be a fixed integer. Suppose that the first q f_j's are nonzero and the rest $p - q$ f_j's are zero. The nonzero additive components can be selected and approximated using a group penalized method, such as the group LASSO, group MCP or group SCAD. Assume the range of each component of $\boldsymbol{X}_j$ be $[a, b]$ where $a < b$ are finite numbers. We make the following 3 standard assumptions for nonparametric additive models:

(A1) Define $\|f\|_2 = [\int_a^b f^2(x)dx]^{1/2}$ for a function f, if the integral exists. There exists a constant $c_f > 0$ such that $\min_{1 \le j \le q} \|f_j\|_2 \ge c_f$.

(A2) Let e^* be a non-negative integer and $d = e^* + \alpha > 0.5$ for some $\alpha \in (0, 1]$. Define $\mathcal{F}$ as the class of functions f on $[a, b]$, whose e^*th derivative of f exists and satisfies the Lipschitz condition of order α: $|f^{(e^*)}(s) - f^{(e^*)}(t)| \le C|s - t|^\alpha$ for $s, t \in [a, b]$ and some constant C. $Ef_j(X_j) = 0$ for $f_j \in \mathcal{F}$.

(A3) The density function g_j of $\boldsymbol{X}_j$ satisfies $0 < C_1 \le g_j(x) \le C_2 < \infty$ on $[a, b]$ for some constants C_1 and C_2.

Now partition $[a, b]$ into K sub-intervals. $a = \xi_0 < \xi_1 < \ldots < \xi_K < \xi_{K+1} = b$, where $K \equiv K_n = n^\nu$ with a positive integer $\nu \in (0, 0.5)$ such that $\max_{1 \le k \le K+1} |\xi_k - \xi_{k-1}| = O(n^{-\nu})$. Using the definition phrased after Stone (1985), $\mathcal{S}_n$ is the polynomial spline space of degree $l \ge 1$. Any function $s \in \mathcal{S}_n$ satisfies:

(i) the restriction of s to each sub-interval is a polynomial of degree l;

(ii) for $l \ge 2$ and $0 \le l' \le l - 2$, s is l' times continuously differentiable on $[a, b]$.

The unknown function f_j's in (2.1) can be approximated by functions in $\mathcal{S}_n$. There exists normalized B-spline basis $\{\phi_k, 1 \le k \le m_n\}$ for $\mathcal{S}_n$, where $m_n \equiv K_n + l$. $m_n = O(n^\nu) = O(n^{1/(2d+1)})$ based on the definitions of K_n and l, and assumption (A1). For any $f_{nj} \in \mathcal{S}_n, j = 1, \ldots, p$,

$$f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk}\phi_k(x).$$

According to such an expansion, the numbers of coefficients as well as the basis functions depend on the sample size n, but for simplicity, for the rest of this dissertation, we omit such dependency in the notations.

A general class of group selection method for nonparametric additive model is based on the following penalized least squares criterion. We study the minimizer to

$$Q(\boldsymbol{\beta}; \lambda, \gamma) = \frac{1}{2n}\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{p}\sum_{k=1}^{m_n}\beta_{jk}\phi_k(X_{ij})\right]^2 + \sum_{j=1}^{p}\rho(\left\|\boldsymbol{\beta}_j\right\|_2; w_j\lambda, \gamma), \qquad (2.2)$$

with the constraints

$$\sum_{i=1}^{n}\phi_k(X_{ij}) = 0, \quad j = 1, \ldots, p, \quad k = 1, \ldots, m_n, \qquad (2.3)$$

where w_j's are some given weights, λ and γ are the penalty parameters and $\gamma > 0$. We let $w_j = 1$, $j = 1, \ldots, p$, for the rest of this chapter. Define

$$\bar{\phi}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\phi_k(\boldsymbol{X}_{ij}), \quad \psi_{jk}(\boldsymbol{X}_{ij}) = \phi_k(\boldsymbol{X}_{ij}) - \bar{\phi}_{jk}.$$

Denote $\boldsymbol{Z}_{ij} = (\psi_{j1}(\boldsymbol{X}_{ij}), \ldots, \psi_{jm_n}(\boldsymbol{X}_{ij}))^{'}$, $\boldsymbol{Z}_j = (\boldsymbol{Z}_{1j}, \ldots, \boldsymbol{Z}_{nj})^{'}$, $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_p)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{'}, \ldots, \boldsymbol{\beta}_p^{'})^{'}$. We further assume that $\boldsymbol{Z}_j$ satisfies $\frac{1}{n}\boldsymbol{Z}_j^{'}\boldsymbol{Z}_j = \boldsymbol{I}_{m_n}$. Note that the condition $\boldsymbol{Z}_j$ and $\boldsymbol{Z}_k$ are orthogonal for $j \neq k$ is not required. (2.2) and (2.3) can be simplified to

$$Q(\boldsymbol{\beta}; \lambda, \gamma) = \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} \rho(\left\|\boldsymbol{\beta}_j\right\|_2; \lambda, \gamma). \tag{2.4}$$

2.2 The group MCP in nonparametric additive models

Under the group MCP, the objective function in nonparametric additive models with B-spline expansions is of the following form

$$Q(\boldsymbol{\beta}; \lambda, \gamma) = \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \int_0^{\|\boldsymbol{\beta}_j\|_2} \left(1 - \frac{t}{\gamma\lambda}\right)_+ dt. \tag{2.5}$$

Now Let $\boldsymbol{\Sigma} = \boldsymbol{Z}^{'}\boldsymbol{Z}/n$. Denote its smallest eigenvalue by $\lambda_{min}(\boldsymbol{\Sigma})$. Let $A_1 = \{1, \ldots, q\}$ be the index set for the nonzero nonparametric additive components and $A_0 = \{(q+1), \ldots, p\}$ the index set for the zero nonparametric additive components. $\boldsymbol{\Sigma}_{A_1} = \boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1}/n$, where $\boldsymbol{Z}_{A_1} = (\boldsymbol{Z}_j, j \in A_1)$.

Denote the true regression coefficients by $\boldsymbol{\beta}^o$. Define $\boldsymbol{\beta}_*^o = \min\{\left\|\boldsymbol{\beta}_j\right\|_2 : j \in A_1\}$; $\boldsymbol{\beta}_*^o = \infty$ if A_1 is empty. Define $\hat{\boldsymbol{\beta}}$ to be the minimizer to equation (2.5) and the oracle set $\hat{\boldsymbol{\beta}}^o = \underset{\boldsymbol{b}}{\operatorname{argmin}} \{\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b}\|_2 : \boldsymbol{b}_j = \boldsymbol{0}, \forall j \in A_0\}$.

Theorem 2.1

Suppose $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed as $N(0, \sigma^2)$. If

(λ, γ) satisfies $\gamma > 1/\lambda_{min}(\mathbf{\Sigma})$, $\beta_*^o > \lambda\gamma$ and $n\lambda^2 > \sigma^2$, then

$$P(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^o) = O(n^{-(4d+1)/(2d+1)}) + O(n^{-2d/(2d+1)}) + O_p(n^{-1}).$$

Proof of Theorem 2.1 See Appendix A.1.

Remark An immediate result of Theorem 2.1 is that $P(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^o) \to 0$ when $n \to \infty$. So under certain conditions and when n is large, the solution set to the objective function with the group MCP is equivalent to the oracle set with very high probability. The oracle property further implies the selection consistency of the solution set under large n and fixed p.

Next we consider the case where the number of groups (i.e. the number of candidate components) $p > n$, although the number of true zero components $q < n$. We introduce the sparse Riesz condition (Zhang and Huang, 2008) before proving a similar result.

Definition 2.a (Sparse Riesz Condition)

We say a matrix $\boldsymbol{X}$ satisfies SRC with rank d^* and spectrum bounds c_* and c^*, or $\boldsymbol{X}$ satisfies SRC(d^*, c_*, c^*), if

$$0 < c_* \leq \frac{\|\boldsymbol{X}_A u\|_2^2}{n} \leq c^* < \infty, \quad |A| \leq d^*, \|u\|_2 = 1.$$

Denote $\boldsymbol{P}_A = \boldsymbol{X}_A(\boldsymbol{X}_A'\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A'$, $\boldsymbol{P}_{A_1} = \boldsymbol{X}_{A_1}(\boldsymbol{X}_{A_1}'\boldsymbol{X}_{A_1})^{-1}\boldsymbol{X}_{A_1}'$ and $\boldsymbol{\delta} = \boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta} - \boldsymbol{\epsilon}$.

Lemma 2.2

Suppose $\lambda^2 mn/(8c^* m_n^2) > \|\boldsymbol{\delta}\|_2^2 + \sigma^2$. For $1 \le m \le 8c^* m_n$, we have

$$P\left(2\sqrt{c^* m_n} \max_{\substack{|A|-|A_1|=m \\ A \supset A_1}} \frac{\|(\boldsymbol{P}_A - \boldsymbol{P}_{A_1})\boldsymbol{Y}\|_2}{\sqrt{mn}} > \lambda\right) \to 0 \text{ as } n \to \infty.$$

Proof of Lemma 2.2 See Appendix A.2.

Define T as any set that satisfies

$$A_1 \cup \{j : \|\hat{\boldsymbol{\beta}}_j\|_2 \neq 0\} \subseteq T \subseteq A_1 \cup \{j : \boldsymbol{Z}_j^{'}(\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}_j)/n = \rho^{'}(\|\hat{\boldsymbol{\beta}}_j\|_2; \lambda, \gamma)\hat{\boldsymbol{\beta}}_j/\|\hat{\boldsymbol{\beta}}_j\|_2\}.$$

Theorem 2.3

Suppose $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed as $N(0, \sigma^2)$ and $\boldsymbol{X}$ satisfies SRC(d^*, c_*, c^*) with $d^* \ge (c^*/c_* + 1/2)qm_n$. If (λ, γ) satisfies $\gamma > c_*^{-1}\sqrt{c^*/c_* + 4}$, $\beta_*^o > \lambda\gamma$ and $\lambda^2 mn/(8c^* m_n) > \|\boldsymbol{\delta}\|_2^2 + \sigma^2, 1 \le m \le 8c^* m_n$, then

$$P(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^o) \to 0 \text{ as } n \to \infty.$$

Proof of Theorem 2.3

According to Lemma 1 in Zhang (2010), suppose $\boldsymbol{X}$ satisfies SRC(d^*, c_*, c^*) with $d^* \ge (c^*/c_* + 1/2)qm_n$ and $\gamma > c_*^{-1}\sqrt{c^*/c_* + 4}$, if

$$2\sqrt{c^* m_n} \max_{\substack{|A|-|A_1|=m \\ A \supset A_1}} \frac{\|(\boldsymbol{P}_A - \boldsymbol{P}_{A_1})\boldsymbol{Y}\|_2}{\sqrt{mn}} \le \lambda,$$

then $|T| \leq (c^*/c_* + 1/2)|A_1| = (c^*/c_* + 1/2)q$. So the number of groups (i.e. the number of candidate components) is decreased from p to at most $(c^*/c_* + 1/2)q$. Furthermore, the conditions in Theorem 2.1 are satisfied if the conditions in Theorem 2.3 are satisfied. Combining Lemma 2.2 and Theorem 2.1, the proof is complete. □

Remark Even when the number of groups is divergent, the solution set converges to the oracle set in probability when some conditions are met. The solution set is thus selection consistent when $p > n$.

Theorem 2.4

Suppose $\frac{m_n log(pm_n)}{n} \to 0$ and $\frac{4m_n\lambda^2}{n^2} \to 0$, then

$$\sum_{j=1}^{q} \|\hat{f}_j - f_j\|_2^2 \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Proof of Theorem 2.4 See Appendix A.3.

2.3 Group coordinate descent algorithm

With the methodological development of the penalized regression for variable selection, the algorithms for finding the solutions to the penalized objective function have also received much attention. In principle, the minimization of the objective function can be carried out by non-linear programming algorithms. For the LASSO under the linear regression, well-known efficient algorithms include the Lease Angle Regression (LARS) algorithm proposed by Efron et al. (2004) and the coordinate descent (CD) algorithm brought to attention and discussed by Friedman et al. (2007, 2010), Wu and Lange (2008) and Breheny and Huang (2011). Both LARS and CD

algorithms have extensions to the group variable selection (Yuan and Lin, 2006; Breheny and Huang, 2013). Because the LARS or group LARS cannot handle concave penalties, throughout the rest of the dissertation we focus on the group coordinate descent (GCD) algorithm proposed by Breheny and Huang (2013).

The GCD algorithm is an iterative algorithm that cycle through each group of coefficients. It is especially suitable for optimization problems that have closed form solutions such as the group selection problems we discuss here. Define the soft threshold operator for grouped variable vector $\boldsymbol{x}$ as

$$S(\boldsymbol{x}; t) = \left(1 - \frac{t}{\|\boldsymbol{x}\|_2}\right)_+ \boldsymbol{x}. \tag{2.6}$$

Using the above operator, given any fixed (λ, γ) pair, the solution of each group of predictors to the group LASSO can be expressed as,

$$\hat{\boldsymbol{\beta}}_{gLASSO}(\boldsymbol{x}; \lambda) = S(\boldsymbol{x}; \lambda). \tag{2.7}$$

For the group MCP with $\gamma > 1$, the solution of each group is

$$\hat{\boldsymbol{\beta}}_{gMCP}(\boldsymbol{x}; \lambda, \gamma) = \begin{cases} \frac{\gamma}{\gamma - 1} S(\boldsymbol{x}; \lambda), & \|\boldsymbol{x}\|_2 \leq \lambda\gamma; \\ \boldsymbol{x}, & \|\boldsymbol{x}\|_2 > \lambda\gamma. \end{cases} \tag{2.8}$$

For the group SCAD with $\gamma > 2$, the solution of each group is

$$\hat{\boldsymbol{\beta}}_{gSCAD}(\boldsymbol{x};\lambda,\gamma) = \begin{cases} S(\boldsymbol{x};\lambda), & \|\boldsymbol{x}\|_2 \le 2\lambda; \\ \frac{\gamma-1}{\gamma-2} S(\boldsymbol{x};\frac{\lambda\gamma}{\gamma-1}), & 2\lambda < \|\boldsymbol{x}\|_2 \le \lambda\gamma; \\ \boldsymbol{x}, & \|\boldsymbol{x}\|_2 > \lambda\gamma. \end{cases} \tag{2.9}$$

We now express (2.7), (2.8) and (2.9) in the general form $\hat{\boldsymbol{\beta}}_g(\boldsymbol{x};\lambda,\gamma)$ and the initial values as $\tilde{\boldsymbol{\beta}}^{(0)} = (\tilde{\boldsymbol{\beta}}_1^{(0)},\ldots,\tilde{\boldsymbol{\beta}}_p^{(0)})$. First we initiate the residual $\boldsymbol{r} = \boldsymbol{Y} - \sum_{j=1}^{p} X_j\boldsymbol{\beta}_j^{(0)}$, where $\boldsymbol{Y}$ is the observed centered response vector. Based on the above building blocks, for a given (λ,γ) pair, we enter the cyclic steps of the GCD algorithm: at iteration s, $s = 0, 1, 2, \ldots$,

Step 1 For each group j, $j = 1,\ldots,p$,

1. $\tilde{\boldsymbol{z}}_j^{(s)} = \frac{1}{n} X_j'\boldsymbol{r} + \tilde{\boldsymbol{\beta}}_j^{(s)}$;
2. $\tilde{\boldsymbol{\beta}}_j^{(s+1)} = \hat{\boldsymbol{\beta}}_g(\tilde{\boldsymbol{z}}_j^{(s)};\lambda,\gamma)$;
3. $\boldsymbol{r} = \boldsymbol{r} - X_j(\tilde{\boldsymbol{\beta}}_j^{(s+1)} - \tilde{\boldsymbol{\beta}}_j^{(s)})$.

Step 2 Enter iteration $s+1$.

Step 3 Repeat **Step 1** and **Step 2** until convergence.

The GCD algorithm is fast and stable. The algorithm is proven to converge to a global minimum (Tseng, 2001) for the group LASSO; Local minima are achievable for the group MCP or group SCAD.

2.3.1 Penalty parameter selection

Using the GCD algorithm, we can recover the full solution surface of (λ, γ). Breheny and Huang (2011) discussed a hybrid approach of selecting (λ, γ), which can also be applied to our context. Here for simplicity, assume that in subsequent numerical studies the penalty parameter γ is fixed, or that an optimal or nearly optimal value of γ is known and used. Once we fix the γ, the problem becomes searching for the optimal λ on the solution path to the minimization of (2.4). Let k^* be a positive integer and $\epsilon^* > 0$. Take $\lambda_{max} = \max_j \|X_j \boldsymbol{Y}/n\|_2$, with which all the entries in the solution $\hat{\boldsymbol{\beta}}$ are zero. We calculate the solutions for a pre-determined sequence of k^* λ's equally spaced on a log scale, proceeding from λ_{max} to $\lambda_{min} = \epsilon^* \lambda_{max}$. To pick the optimal λ, and as a result the optimal solution $\hat{\boldsymbol{\beta}}$, a number of methods can be considered. We can use the data-driven K-fold cross validation (CV). It is widely known that CV works fairly well under different selection methods and it seems that it is becoming a standard criterion. Nevertheless, a lack of sufficient theoretical support from a frequentist point of view of the method leads us to also consider the criterion based procedures such as the Akaike information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) (Schwarz, 1978) or extended Bayesian information criterion (EBIC) (Chen and Chen, 2008). These criterion based procedures may or may not perform satisfactory because how to obtain accurate and precise estimation of error under the penalized regression framework as well as how to define the degrees of freedom under various cases remain big challenges of this area.

2.4 Data example: Boston housing data

To illustrate the use of aforementioned group selection methods for nonparametric additive models, in this section we use the Boston housing data for example. Boston housing data are widely studied and publicly available under the R package *MASS*. An error corrected version of the data is accessible in R under the package *spdep*. 506 median values of homes were collected in the Boston Standard Metropolitan Statistical Area in 1970. Also measured were 13 attributes that potentially contribute to the housing prices. The data were analyzed in a large body of literature, including Harrison and Rubinfeld (1978), Breiman and Friedman (1985), Opsomer and Ruppert (1998) and Fan and Jiang (2005). For the purpose of demonstrating the use of the group selection methods, especially the group MCP, in nonparametric additive models under a high-dimensional setting, in our modeling, we add 200 independent $N(0, 1)$ variables to the following 4 variables that were originally collected:

1. RM: average number of rooms per dwelling
2. TAX: full property tax rate per $10,000 per town
3. PTRATIO: pupil/teacher ratio per town
4. LSTAT: percentage of lower status population

As shown in Figure 2.1, the empirical distribution of the median housing values is skewed to the right. As shown in Figure 2.2, the empirical distributions of the 4 important contributing factors are also skewed.

In Breiman and Friedman (1985), Opsomer and Ruppert (1998) and Fan and Jiang (2005), different nonparametric additive models were investigated and the above

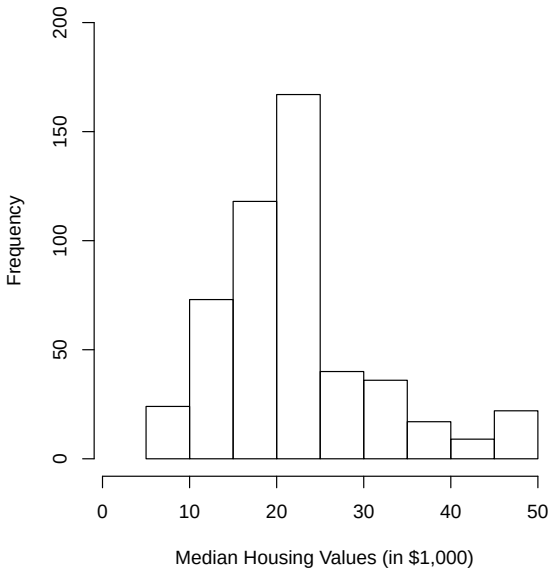

Figure 2.1: Histogram of median housing values

list of 4 variables were chosen as the main factors that affect the response variable MEDV, the median value of owner-occupied home (in $1,000). The selection path and group-wise norm plots under the group MCP, group LASSO and group SCAD are presented in Figure 2.3. Using the 10-fold cross validation (CV), the numbers of correctly selected components are all 4 for the group MCP, group LASSO and group SCAD; the numbers of incorrectly selected components (over-selection) are 2, 5, 4 for the group MCP, group LASSO and group SCAD respectively.

Note that the results are for demonstration purpose only. In another attempt with randomly generated independent noises, we might end up with slightly different results in terms of the number of components selection and the specific components selected. The solution path plots and the group-wised norm plots should look similar. Because the results are not based on simulation studies, we do not make any general

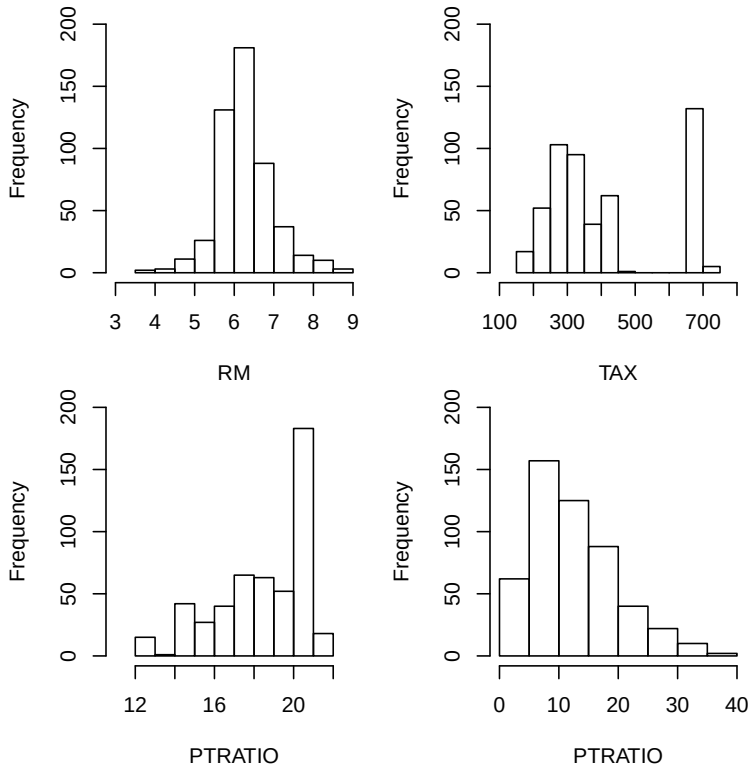

Figure 2.2: Histograms of 4 variables from Boston housing data

conclusion at this moment. We defer more systematic numerical studies on several group selection methods to Chapter 4. We also return to this example at the end of Chapter 4.

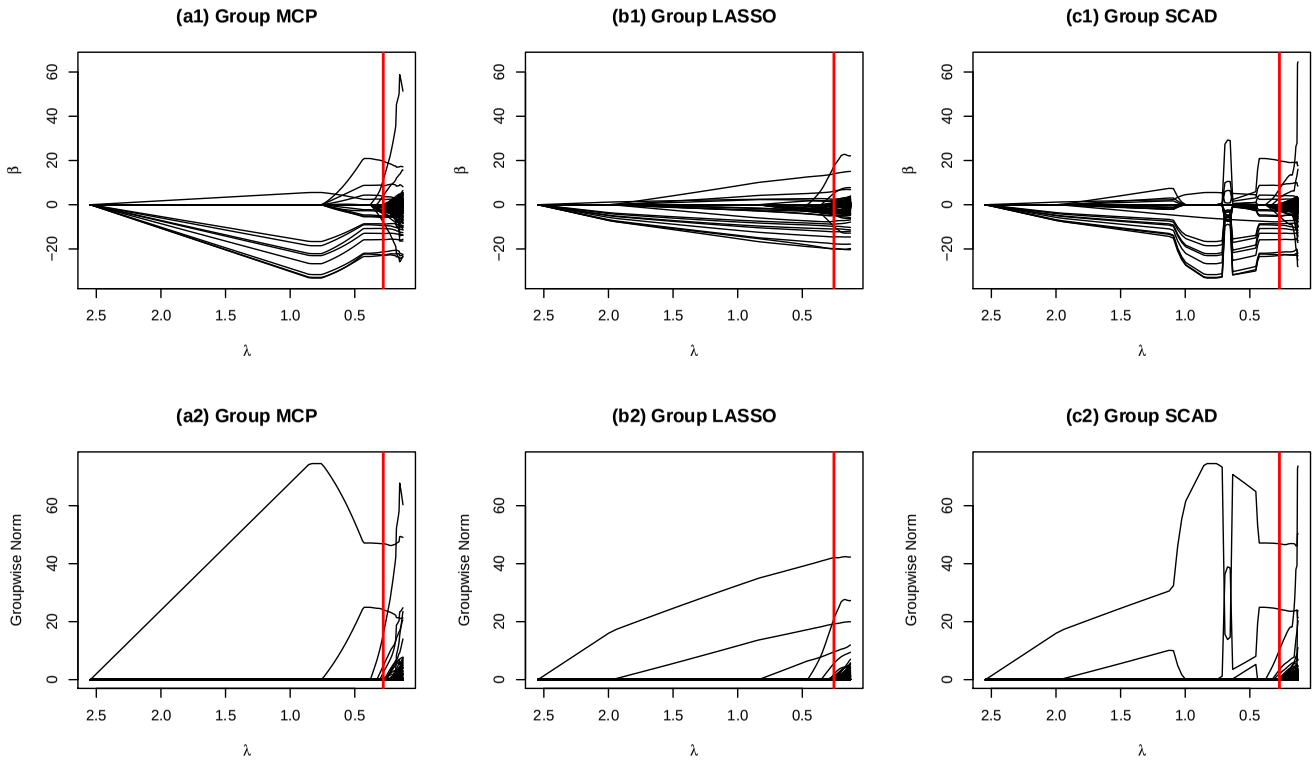

Figure 2.3: Boston housing data results: *(a1) (b1) and (c1) display the solution paths under different selection methods, the vertical line in each plot marks the λ selected using the 10-fold CV; (a2), (b2) and (c2) display the group-wise norms under different selection methods.*

CHAPTER 3
THE ADAPTIVE GROUP LASSO AND ITS RECURSIVE APPLICATION

In this chapter, we study the adaptive group LASSO under general models. In Section 3.1, we present the general framework of adaptive group LASSO. We use the symmetrized version of the Bregman divergence (Bregman, 1967; Nielson and Nock, 2007) to entitle the measure of the difference between proposed adaptive group LASSO estimator and the target estimator under more general models. As a result, in Section 3.2, the theoretical justifications for the adaptive group LASSO lay out avenues for future research on a number of other models. In the same section, we introduce a new way of constructing weights (Huang and Zhang, 2012) and extend and study its utility in the group selection problem. We also examine a recursive application of adaptive group LASSO and show its advantage. In Section 3.3, we carry out simulation studies under linear regression models, with different data generating mechanisms and various correlation structures.

3.1 A general framework for the adaptive group LASSO

Consider a general loss function of the form $l(\boldsymbol{\beta}) = \psi(\boldsymbol{\beta}) - \langle \boldsymbol{\beta}, \mathbf{z} \rangle$, where $\psi(\boldsymbol{\beta})$ is a known convex function, $\mathbf{z}$ is observed and $\boldsymbol{\beta}$ is unknown. Denote the negative gradient of the loss function at $\hat{\boldsymbol{\beta}}$ as $\mathbf{g}$. So $\mathbf{g} = -\dot{l}(\hat{\boldsymbol{\beta}}) = \mathbf{z} - \dot{\psi}(\hat{\boldsymbol{\beta}})$. $\{\boldsymbol{\beta}, \mathbf{z}\} \subset \mathbf{R}^p$ and $\langle \boldsymbol{\beta}, \mathbf{z} \rangle = \boldsymbol{\beta}^{'}\mathbf{z}$. The smooth assumptions on the function ψ in terms of differentiability are invoked by the use of its derivatives.

We consider the group settings where the $\boldsymbol{\beta}$ vector is divided into J groups. Without loss of generality, we assume that the β_i's, $i = 1, \ldots, p$ are ordered by group, which can be easily obtained by rearranging the order of components of $\mathbf{z}$ accordingly. With group index $j = j(i)$ (j as a function of i) and group size d_j with $\sum_{j=1}^{J} d_j = p$, $j = 1, \ldots, J$, $i = 1, \ldots, p$, the unknown coefficient vector can also be expanded as

$$\begin{aligned} \boldsymbol{\beta} &= (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J) \\ &= (\beta_{1(1)}, \ldots, \beta_{1(d_1)}, \beta_{2(1)}, \ldots, \beta_{2(d_2)}, \ldots\ldots, \beta_{J(1)}, \ldots, \beta_{J(d_J)}). \end{aligned}$$

The adaptive group penalized estimator is

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{J} \hat{w}_j \|\boldsymbol{\beta}_j\|_2 \right\}, \tag{3.1}$$

where possibly estimated weights $\hat{w}_j$'s are positive. $\hat{\boldsymbol{\beta}}$ from (3.1) is a global minimizer if and only if it satisfies the Karush-Kuhn-Tucker condition,

$$\begin{cases} \boldsymbol{g}_i = \hat{w}_{j(i)} \lambda \frac{\beta_i}{\|\boldsymbol{\beta}_{j(i)}\|_2}, & \beta_i \neq 0; \\ \boldsymbol{g}_i \in \hat{w}_{j(i)} \lambda [-1, 1], & \beta_i = 0. \end{cases}$$

In order to estimate the performance of the adaptive absolute group penalized estimator $\hat{\boldsymbol{\beta}}$, we introduce the Bregman divergence $D(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}^*) - \langle \dot{l}(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle$ (Bregman, 1967) as the building block for studying $\hat{\boldsymbol{\beta}}$'s proximity to a target vector $\boldsymbol{\beta}^*$ of $\boldsymbol{\beta}$. A symmetrized version of the Bregman divergence (Nielsen

and Nock, 2007) is used, where

$$\Delta(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = D(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + D(\boldsymbol{\beta}^*, \boldsymbol{\beta}) = \left\langle \boldsymbol{\beta} - \boldsymbol{\beta}^*, \dot{\psi}(\boldsymbol{\beta}) - \dot{\psi}(\boldsymbol{\beta}^*) \right\rangle,$$

and it is non-negative in our context due to the convexity of ψ. The symmetrized Bregman divergence can be interpreted as the regret in prediction error, the symmetrized Kullback -Leibler divergence, the spectrum loss for the linear regression, the generalized linear models, and the graphical LASSO, respectively. More detailed interpretations of the symmetrized Bregman divergence in different models can be found in Huang and Zhang (2012)

We assume that $\boldsymbol{\beta}^*$ is sparse in that even though the dimension of $\boldsymbol{\beta}^*$ may be very high, i.e. $p > n$, the true non-zero entries in $\boldsymbol{\beta}^*$ is not. The group settings we are interested in assume that the β_i^*'s in the same group are either all zero or all non-zero. Without loss of generality, let the groups with true zero β_i^*'s be the first q groups, the number of true zero β_i^*'s is then $d^{(q)} = \sum_{j=1}^{q} \sum_{i=1}^{d_j} \beta_{j(i)}^*$. Note that our requirement on the sparsity of $\boldsymbol{\beta}^*$ is on the number of zero β_i^*'s, not on the number of groups with zero β_i^*'s. In other words, we require $d^{(q)} < n$. We denote the index set for non-zero β_i^* as $S_1 = \{i : \beta_i^* \neq 0\}$. S_2 is any set that satisfies $S_2 \supseteq S_1$. To denote the index set for groups with zero β_i^*'s, we use $S_3 = \{j : \|\boldsymbol{\beta}_j^*\|_2 = 0\}$. S_4 is any set that satisfies $S_4 \subseteq S_3$. We also introduce $\mathbf{W} = diag(w_1, w_2, \ldots, w_J)$ with possibly unknown weights $w_j > 0, j = 1, \ldots, J$. $\hat{\mathbf{W}}$ can be viewed as the estimator for $\mathbf{W}$.

Based on aforementioned elements, we define

$$\begin{aligned} z_0^* &= \max_j \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*)_j\|_2; \\ z_1^* &= \max_j \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*)_j/\hat{w}_j\|_2; \\ \Omega_0 &= \{\hat{w}_j \le w_j, \forall j \in S_2\} \cap \{\hat{w}_j \ge w_j, \forall j \in S_2^c\}. \end{aligned}$$

3.2 Theoretical results

Lemma 3.1

(i) In the event $\Omega_0 \cap \{\|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*)_j\|_2 \le \hat{w}_j\lambda, \forall j\}$,

$$\Delta(\boldsymbol{\beta}, \boldsymbol{\beta}^*) \le 2\lambda \sum_{j=1}^{J} \hat{w_j} \|\boldsymbol{\beta}_j^*\|_2 \le 2\lambda \sum_{j=1}^{J} w_j \|\boldsymbol{\beta}_j^*\|_2. \tag{3.2}$$

(ii) In the event $\Omega_0 \cap \{\max_{i \in \text{ group } j} |g_i| \le \hat{w}_j\lambda, \forall j\}$, the error $\mathbf{h} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ satisfies

$$\Delta(\boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\beta}^*) + (\lambda - z_1^*) \sum_{j \in S_2^c} w_j \|\hat{\boldsymbol{\beta}}_j\|_2 \le (\lambda \max_j w_j + z_0^*) \sum_{j \in S_2} \|\mathbf{h}_j\|_2. \tag{3.3}$$

Consequently, in the event $\Omega_0 \cap \{(\lambda \max_j w_j + z_0^*)/(\lambda - z_1^*) \le \xi\}$, $\mathbf{h} \ne \mathbf{0}$ belongs to the sign-restricted cone

$$\mathscr{C}_{\text{-}}(\xi, S) = \left\{\mathbf{b} \in \mathscr{C}(\xi, S) : b_j(\dot{\psi}(\boldsymbol{\beta} + \mathbf{b}) - \dot{\psi}(\boldsymbol{\beta}))_j \le 0, \forall j \in S^c\right\}, \tag{3.4}$$

where

$$\mathscr{C}(\xi, S) = \left\{\mathbf{b} \in \mathbb{R}^p : \sum_{j \in S_2^c} w_j \|\mathbf{b}_j\|_2 \le \xi \sum_{j \in S_2} \|\mathbf{b}_j\|_2 \ne 0\right\}. \tag{3.5}$$

Proof of Lemma 3.1 See Appendix B.1.

Remark Lemma 3.1 gives the monotonicity of the weight inequality. It shows an analytic properties of the error $\mathbf{h}$ in the sign-restricted cone.

Before deriving the oracle inequality, we introduce two definitions.

Definition 3.a (Quasi Star-shaped Function)

A function $\phi(\mathbf{b})$ is a quasi star-shaped function if $\phi(t\mathbf{b})$ is continuous and non-decreasing in $t \in [0, \infty)$ for all $\mathbf{b} \in \mathbb{R}^p$ and $\lim_{\mathbf{b}\to\mathbf{0}} \phi(\mathbf{b}) = 0$.

Definition 3.b (General Invertibility Factor)

For $0 \leq \eta^* \leq 1$ and any pair of quasi star-shaped functions $\phi_0(\mathbf{b})$ and $\phi(\mathbf{b})$, a general invertibility factor (GIF) over $\mathscr{C}(\xi, S)$ is defined as

$$F(\xi, S; \phi_0, \phi) = \inf \left\{ \frac{\Delta(\boldsymbol{\beta}^* + \mathbf{b}, \boldsymbol{\beta}^*) e^{\phi_0(\mathbf{b})}}{\phi(\mathbf{b}) \sum_{j \in S_2} \|\mathbf{b}_j\|_2} : \mathbf{b} \in \mathscr{C}(\xi, S) \ , \phi_0(\mathbf{b}) \leq \eta^* \right\}.$$

As pointed out in Huang and Zhang (2012), all seminorms are quasi star-shaped functions. So for example in GIF, the functions ϕ_0 and ϕ can simply be the l_1 or the l_2 norm.

Theorem 3.2

Let $0 \leq \eta \leq \eta^* \leq 1$ and $(\phi_0(\mathbf{b}), \phi(\mathbf{b}))$ be a pair of quasi star-shaped functions, then

$$\phi_0(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \eta \ , \ \phi(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \frac{e^{\eta}(\lambda \max_j w_j + z_0^*)}{F(\xi, S; \phi_0, \phi)} \tag{3.6}$$

in the event

$$\Omega_1 = \Omega_0 \cap \left\{ \frac{(\lambda \max_j w_j + z_0^*)}{\lambda - z_1^*} \leq \xi \ , \ \frac{(\lambda \max_j w_j + z_0^*)}{F(\xi, S; \phi_0, \phi)} \leq \eta e^{-\eta} \right\}.$$

Proof of Theorem 3.2 See Appendix B.2.

Now we introduce a new way of constructing weights for the adaptive group LASSO. The new method was first proposed and studied in Huang and Zhang (2012) for variable selection rather than group selection problems. Let $\rho_\lambda(t)$ be a concave penalty with $\dot{\rho}_\lambda(0+) = \lambda$. We use the first derivative of the concave penalty function ρ_λ to construct the weights. Larger initial estimates are assigned smaller weights due to the concavity of the function ρ_λ.

Denote $\tilde{\boldsymbol{\beta}}$ as the vector of initial estimates of the coefficients. The adaptive group penalized estimator using the presently-proposed weight is

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ l(\boldsymbol{\beta}) + \sum_{j=1}^{J} \dot{\rho}_\lambda(\|\tilde{\boldsymbol{\beta}}_j\|_2) \|\boldsymbol{\beta}_j\|_2 \right\}. \tag{3.7}$$

In addition, we can recursively apply the following objective function

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ l(\boldsymbol{\beta}) + \sum_{j=1}^{J} \dot{\rho}_\lambda(\|\hat{\boldsymbol{\beta}}_j^{(k)}\|_2) \|\boldsymbol{\beta}_j\|_2 \right\}, \ k = 0, 1, \ldots. \tag{3.8}$$

We show in Theorem 3.4 that this multistage algorithm is beneficial.

Define the l_2 version of GIF as

$$F_2(\xi, S; \phi_0) = \inf \left\{ \frac{\Delta(\boldsymbol{\beta}^* + \mathbf{b}, \boldsymbol{\beta}^*) e^{\phi_0(\mathbf{b})}}{\sum_{j=1}^{J} \|\mathbf{b}_j\|_2 \sum_{j \in S_3} \|\mathbf{b}_j\|_2} : \ \mathbf{b} \in \mathscr{C}(\xi, S) \ , \phi_0(\mathbf{b}) \leq \eta^* \right\}$$

and the maximum concavity of the group penalty as

$$\kappa = \sup \frac{|\dot{\rho}_\lambda(\|\mathbf{x}_1\|_2) - \dot{\rho}_\lambda(\|\mathbf{x}_2\|_2)|}{\|\mathbf{x}_1 - \mathbf{x}_2\|_2},$$

where $\mathbf{x}_1$, $\mathbf{x}_2 \in \mathbb{R}^d$, d is a positive integer.

Theorem 3.3

Let $\lambda_0 > 0$, $0 < \eta < 1$, $0 < \gamma_0 < 1/\kappa$, $A > 1$ and $\xi \geq (A + 1 - \kappa\gamma_0)/(A - 1)$. Suppose

$$\lambda_0 \left[1 + \frac{A}{1 - \kappa\gamma_0} \right] \leq F(\xi, S, \phi_0, \phi_0) \eta e^{-\eta} \text{ and } F_* \leq F_2(\xi, S; \phi_0) \tag{3.9}$$

for all $S \supseteq S_3$ with $|S \backslash S_3| \leq l^*$. Let $\tilde{\boldsymbol{\beta}}$ be an initial estimator of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ be the minimizer to the adaptive group LASSO objective function (3.7) with $\hat{w}_j = \dot{\rho}_\lambda(\|\tilde{\boldsymbol{\beta}}_j\|_2)/\lambda$ and penalty level $\lambda = A\lambda/(1 - \kappa\gamma_0)$. Then

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{e^\eta}{F_*} \left[\max_j |\dot{\rho}_\lambda(\|\boldsymbol{\beta}_j\|_2)| + \sum_{j \in S} \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2 + \left(\kappa + \frac{1}{\lambda_0 A} - \frac{\kappa}{A} \right) \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \right] \tag{3.10}$$

in the event $\Omega_3 = \left\{ \max_j \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2 \leq \lambda_0 \right\} \cap \left\{ \sum_{j \in S_3} \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2^2 \geq \gamma_0^2 \lambda^2 l^* \right\}$.

Proof of Theorem 3.3 See Appendix B.3.

Theorem 3.4

Use the same $\{\kappa, \S_0, \lambda_0, \eta, \gamma_0, A, \xi, l^*, \lambda, F, F_*\}$ and conditions as those in Theorem 3.3. Let

$$\xi_0 = (\lambda + \lambda_0)/(\lambda - \lambda_0) \text{ and } r_0 = \frac{e^\eta}{F_*}\left[\kappa + \frac{1}{\gamma_0 A} - \frac{\kappa}{A}\right] < 1.$$

Furthermore, suppose

$$\frac{e^\eta[1 + (1 - \kappa\gamma_0)/A]}{F(\xi, S, \phi_0, \phi)} \leq \gamma_0\sqrt{l^*} \tag{3.11}$$

holds, then for $k = 0, 1, \ldots$

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*\|_2 \leq & \frac{(1 - r_0^k)e^\eta\left[\max_j |\dot{\rho}_\lambda(\|\boldsymbol{\beta}_j\|_2)| + \sum_{j\in S}\|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2\right]}{(1 - r_0)F_*} + \\ & \frac{r_0^k e^\eta \lambda\,[1 + (1 - \kappa\gamma_0)/A]}{F(\xi, S, \phi_0, \phi)}\end{aligned} \tag{3.12}$$

in the event

$$\begin{aligned}\Omega_4 = & \left\{\max_j \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2 \leq \lambda_0\right\} \cap \\ & \left\{\frac{e^\eta\left[\max_j |\dot{\rho}_\lambda(\|\boldsymbol{\beta}_j\|_2)| + \sum_{j\in S}\|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2\right]}{(1 - r_0)F_*} \leq \gamma_0\lambda\sqrt{l^*}\right\}.\end{aligned} \tag{3.13}$$

Proof of Theorem 3.4 See Appendix B.4.

3.3 Simulations

3.3.1 Models and methods

Model 1

Consider the model of the general form

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i, \, , \, i = 1, \ldots, n,$$

where ϵ_i's are i.i.d. $n(0,1)$. Assume the covariates $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})' \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Based on an assumption of inherent group structure of group size 5, the correlation matrix is

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_2 & \ldots & \boldsymbol{\Sigma}_2 \\ \boldsymbol{\Sigma}_2 & \boldsymbol{\Sigma}_1 & \ldots & \boldsymbol{\Sigma}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_2 & \boldsymbol{\Sigma}_2 & \ldots & \boldsymbol{\Sigma}_1 \end{bmatrix},$$

where

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & \rho_1 & \rho_1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_1 & \rho_1 & \rho_1 \\ \rho_1 & \rho_1 & 1 & \rho_1 & \rho_1 \\ \rho_1 & \rho_1 & \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_1 & \rho_1 & \rho_1 & 1 \end{bmatrix} = (1-\rho_1)\mathbf{I}_5 + \rho_1 \mathbf{1}_5' \mathbf{1}_5$$

and $\boldsymbol{\Sigma}_2 = \rho_2 \mathbf{1}_5' \mathbf{1}_5$. In other words, within a group, the covariance $cov(X_{ij}, X_{ij'}) = \rho_1$, $j \neq j'$, $j, j' = 1, \ldots, 5$; across groups, the covariance $cov(X_{ij}, X_{i'j'}) = \rho_2$, $i \neq i'$, $j, j' = 1, \ldots, 5$. In the simulation, our choices of ρ_1 and ρ_2 satisfy $0 \leq \rho_1 \leq \rho_2 \leq 1$. We assume that 3 groups of covariates, i.e. 15 covariates, are nonzero. The underlying

model coefficient vector is

$$\boldsymbol{\beta} = (1, 1, -1, -1, 1, \mathbf{0}_5, 1, -1, 1, -1, 1, \mathbf{0}_{10}, -1, -1, 1, -1, -1, \mathbf{0}_{p-30}),$$

where $\mathbf{0}_5$, $\mathbf{0}_{10}$ and $\mathbf{0}_{p-30}$ are zero vector of length 5, 10 and $p - 30$, respectively.

A summary of all the combinations of different parameter values of the model and their estimated signal-to-noise ratio (SNR) based on 500 simulation replications is reported in Table 3.1. The SNR is defined as $Var(\boldsymbol{X}\boldsymbol{\beta})/Var(\boldsymbol{\epsilon})$.

n	p	ρ_1	ρ_2	SNR
100	200	0.0	0.0	15.081
100	500	0.0	0.0	15.216
100	2000	0.0	0.0	15.278
100	200	0.5	0.3	10.237
100	500	0.5	0.3	10.211
100	2000	0.5	0.3	10.195

Table 3.1: Empirical signal-to-noise ratio for Model 1

Model 2

In Yuan and Lin (2006), the authors included an additive model with categorical factors in their simulation studies. Here Model 2 settings are similar, but with different number of factors and more levels.

We use J latent variables $Z_1, \ldots, Z_J \sim N_J(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$ and

$\sigma_{ij} = cov(Z_i, Z_j) = 0.3^{|i-j|}$, $i, j = 1, \ldots, J$. Denote the quantile function of the standard normal distribution as Φ^{-1}. Z_i is $0, 1, 2$ or 3 if it is smaller than $\Phi^{-1}(0.25)$, between $\Phi^{-1}(0.25)$ and $\Phi^{-1}(0.50)$, between $\Phi^{-1}(0.50)$ and $\Phi^{-1}(0.75)$, and larger than $\Phi^{-1}(0.75)$, respectively. The response variable

$$Y_i = 4\mathbf{I}(Z_1 = 0) + 2\mathbf{I}(Z_1 = 1) - 3\mathbf{I}(Z_1 = 2) - 4\mathbf{I}(Z_2 = 0) + \mathbf{I}(Z_2 = 1) + 3\mathbf{I}(Z_2 = 2) -$$
$$\mathbf{I}(Z_3 = 0) - 2\mathbf{I}(Z_3 = 1) - \mathbf{I}(Z_3 = 2) + \epsilon_i,\ i = 1, \ldots, n,$$

where ϵ_i's are i.i.d. $N(0, \sigma^2)$ and $\mathbf{I}$ is the indicator function.

A summary of all the combinations of the model parameters and their estimated SNR based on 500 simulation replications is reported in Table 3.2, where $p = 3J$.

n	p	σ^2	SNR
100	30	1	12.421
100	300	1	12.472
100	900	1	12.504
100	30	2	6.203
100	300	2	6.215
100	900	2	6.256

Table 3.2: Empirical signal-to-noise ratio for Model 2

We use B-splines of degree 3 to approximate each component. The numbers

of knots of the B-splines are decided by the integer closest to $n^{0.3}$. The knots are decided by corresponding percentiles.

Five group penalized methods are applied for comparison. They are (1) the group LASSO, (2) the group MCP, (3) the adaptive group LASSO that uses the inverses of the group LASSO estimates as the weights, (4) the adaptive group LASSO that uses the first derivative of the group MCP penalty on the group LASSO estimates as the weights, and (5) the adaptive group LASSO that is a recursive application of (4). The five methods are given the names "gLASSO", "gMCP" , "agLASSO", "agLASSO1" and "agLASSO2" respectively in the simulation results section. For anywhere the group MCP is applied, the penalty parameter γ is fixed at 2.7.

We run 500 simulation replications on all combinations of different values of p and group selection methods. For both Model 1 and Model 2, the sample size is fixed at $n = 100$. The group coordinate descent algorithm is used for calculating the solution paths. The 10-fold CV and EBIC are used for identifying the optimal penalty parameter λ from a sequence of k^* λ's chosen according to the method detailed in Section 2.3, where $k^* = 100$ and $\epsilon^* = 0.001$. The definition of EBIC (Chen and Chen, 2008) is

$$EBIC(\lambda) = log(RSS) + df\frac{\log n}{n} + \nu df\frac{\log p}{n}, \tag{3.14}$$

where RSS is the residual sum of squares, the degrees of freedom df is chosen as the model size, and ν is set to 0.5.

In order to evaluate the selection properties of each method, we refer to the number of correctly selected variables (CS), the number of incorrectly selected vari-

ables (i.e. over-selection of variables (OS)) or the model size (MS)(i.e. the number of correctly selected variables plus that of incorrect/over-selected variables in the final model) as well as the false discovery rate (FDR) (Benjamini and Hochberg, 1995). The definition of the FDR is

$$\text{FDR} = \frac{\text{OS}}{\text{MS}} = \frac{\text{number of true zero variables in the final model}}{\text{number of variables in the final model}}.$$

To evaluate the estimation performance, we use the following empirical root model error (RME)

$$RME = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}.$$

To help compare the prediction performance of different methods, we use the empirical root prediction error (RPE)

$$RPE = \sqrt{\frac{1}{n^*}\sum_{i=1}^{n^*}(\hat{Y}_i - Y_i)^2},$$

where n^* is the size of the prediction set. In our studies, $n^* = n$.

3.3.2 Simulation results

For Model 1, we report the CS, OS, MS, FDR in two tables: Table 3.3 includes the results under 10-fold CV while Table 3.4 includes the results under EBIC. We present the RME and RPE under different penalty parameter selection methods (10-fold CV and EBIC) in Figure 3.1 and Figure 3.2, respectively.

For Model 2, we report selection, estimation and prediction performances in a

same way: we display the CS, OS, MS, FDR in two tables (Table 3.5 and Table 3.6) and present the RME and RPE under under different penalty parameter selection methods (10-fold CV and EBIC) in two figures (Figure 3.3 and Figure 3.4).

About Model 1 results, using either 10-fold CV or EBIC, all 5 methods select the true non-zero variables. The gMCP and agLASSO seem to work better in terms of FDR. But in terms of RPE, gMCP outperforms agLASSO. The impact of correction is the smallest to gMCP, regardless of penalty parameter selection methods, as reflected in the table summaries as well as the plots.

About Model 2 results, the 10-fold CV methods seem to work better than EBIC in choosing the true non-zero variables but the EBIC method works better in excluding the true zero variables in the final model because it has smaller numbers of over-selection and lower FDR's. The discrepancies in RME and RPE between large SNR case and small SNR case are apparent in the plots, regardless of penalty parameter selection methods used.

Generally speaking, using 10-fold CV, the proposed gMCP and agLASSO1, agLASSO2 perform better than the gLASSO methods, especially in OS. Using EBIC, the proposed gMCP and agLASSO1, agLASSO2 have over-selection results with much larger standard error than that under the gLASSO, which implies that EBIC may not be an ideal selection methods for these selection methods.

Model 1	p	Method	CV			
			CS (s.e.)	OS (s.e.)	MS (s.e.)	FDR (s.e.)
Independent	200	gLASSO	15.00(0.000)	54.40(26.134)	69.40(26.134)	0.75(0.115)
		gMCP	15.00(0.000)	3.27(6.756)	18.27(6.756)	0.11(0.195)
		agLASSO	15.00(0.000)	0.00(0.000)	15.00(0.000)	0.00(0.000)
		agLASSO1	15.00(0.000)	35.55(27.842)	50.55(27.842)	0.63(0.165)
		agLASSO2	15.00(0.000)	28.51(29.064)	43.51(29.064)	0.40(0.391)
	500	gLASSO	15.00(0.000)	76.38(36.040)	91.38(36.040)	0.80(0.099)
		gMCP	15.00(0.000)	4.15(7.691)	19.15(7.691)	0.13(0.216)
		agLASSO	15.00(0.000)	0.00(0.000)	15.00(0.000)	0.00(0.000)
		agLASSO1	14.99(0.224)	47.89(41.551)	62.88(41.559)	0.67(0.179)
		agLASSO2	15.00(0.000)	61.44(22.397)	76.44(22.397)	0.74(0.228)
	2000	gLASSO	15.00(0.000)	111.27(53.766)	126.27(53.766)	0.85(0.084)
		gMCP	15.00(0.000)	4.90(8.771)	19.90(8.771)	0.15(0.232)
		agLASSO	15.00(0.000)	0.00(0.000)	15.00(0.000)	0.00(0.000)
		agLASSO1	14.99(0.100)	66.12(60.151)	81.11(60.161)	0.71(0.172)
		agLASSO2	15.00(0.000)	75.87(13.919)	90.87(13.919)	0.83(0.027)
Correlated	200	gLASSO	15.00(0.000)	64.65(25.843)	79.65(25.843)	0.79(0.088)
		gMCP	15.00(0.000)	2.90(5.921)	17.90(5.921)	0.10(0.187)
		agLASSO	15.00(0.000)	0.00(0.000)	15.00(0.000)	0.00(0.000)
		agLASSO1	14.96(0.446)	38.58(28.517)	53.54(28.559)	0.65(0.163)
		agLASSO2	15.00(0.000)	20.63(23.355)	35.63(23.355)	0.34(0.369)
	500	gLASSO	15.00(0.000)	91.86(36.990)	106.86(36.990)	0.84(0.069)
		gMCP	15.00(0.000)	4.11(7.640)	19.11(7.640)	0.13(0.216)
		agLASSO	15.00(0.000)	0.00(0.000)	15.00(0.000)	0.00(0.000)
		agLASSO1	14.97(0.387)	51.61(43.468)	66.58(43.496)	0.69(0.162)
		agLASSO2	15.00(0.000)	45.58(26.436)	60.58(26.436)	0.62(0.329)
	2000	gLASSO	15.00(0.000)	129.66(55.097)	144.66(55.097)	0.87(0.074)
		gMCP	15.00(0.000)	4.31(8.500)	19.31(8.500)	0.13(0.222)
		agLASSO	15.00(0.000)	0.01(0.100)	15.01(0.100)	0.00(0.005)
		agLASSO1	14.99(0.100)	66.78(57.208)	81.77(57.219)	0.73(0.152)
		agLASSO2	15.00(0.000)	67.83(14.475)	82.83(14.475)	0.81(0.054)

Table 3.3: Model 1 simulation results: selection performances using 10-fold CV

Model 1	p	Method	EBIC			
			CS (s.e.)	OS (s.e.)	MS (s.e.)	FDR (s.e.)
Independent	200	gLASSO	15.00(0.000)	0.47(1.528)	15.47(1.528)	0.02(0.074)
		gMCP	15.00(0.000)	0.00(0.000)	15.00(0.000)	0.00(0.000)
		agLASSO	15.00(0.000)	0.00(0.000)	15.00(0.000)	0.00(0.000)
		agLASSO1	15.00(0.000)	12.26(5.024)	27.26(5.024)	0.43(0.103)
		agLASSO2	15.00(0.000)	7.61(24.537)	22.61(24.537)	0.07(0.242)
	500	gLASSO	14.94(0.948)	0.58(1.664)	15.52(1.933)	0.03(0.081)
		gMCP	15.00(0.000)	0.00(0.000)	15.00(0.000)	0.00(0.000)
		agLASSO	14.95(0.805)	0.00(0.000)	14.95(0.805)	0.00(0.000)
		agLASSO1	15.00(0.000)	13.94(5.886)	28.94(5.886)	0.46(0.109)
		agLASSO2	15.00(0.000)	11.68(29.736)	26.68(29.736)	0.11(0.291)
	2000	gLASSO	14.65(2.244)	0.38(1.400)	15.03(2.695)	0.02(0.067)
		gMCP	15.00(0.000)	0.00(0.000)	15.00(0.000)	0.00(0.000)
		agLASSO	14.68(2.075)	0.00(0.000)	14.68(2.075)	0.00(0.000)
		agLASSO1	15.00(0.000)	16.85(6.412)	31.85(6.412)	0.51(0.097)
		agLASSO2	15.00(0.000)	8.97(26.654)	23.97(26.654)	0.08(0.252)
Correlated	200	gLASSO	14.76(1.830)	0.89(2.041)	15.65(2.818)	0.04(0.098)
		gMCP	15.00(0.000)	0.00(0.000)	15.00(0.000)	0.00(0.000)
		agLASSO	14.79(1.553)	0.00(0.000)	14.79(1.553)	0.00(0.000)
		agLASSO1	15.00(0.000)	13.28(5.595)	28.28(5.595)	0.45(0.104)
		agLASSO2	15.00(0.000)	9.76(26.898)	24.76(26.898)	0.11(0.266)
	500	gLASSO	13.50(4.403)	0.71(1.831)	14.21(4.988)	0.03(0.089)
		gMCP	15.00(0.000)	0.16(3.578)	15.16(3.578)	0.00(0.038)
		agLASSO	13.67(3.926)	0.00(0.000)	13.67(3.926)	0.00(0.000)
		agLASSO1	15.00(0.000)	14.99(6.666)	29.99(6.666)	0.48(0.107)
		agLASSO2	15.00(0.000)	7.37(24.170)	22.37(24.170)	0.06(0.215)
	2000	gLASSO	9.50(7.060)	0.29(1.170)	9.79(7.377)	0.01(0.058)
		gMCP	15.00(0.000)	0.00(0.000)	15.00(0.000)	0.00(0.000)
		agLASSO	10.15(6.349)	0.02(0.316)	10.17(6.325)	0.00(0.063)
		agLASSO1	15.00(0.000)	18.60(8.233)	33.60(8.233)	0.53(0.110)
		agLASSO2	15.00(0.000)	7.08(23.808)	22.08(23.808)	0.04(0.179)

Table 3.4: Model 1 simulation results: selection performances using EBIC

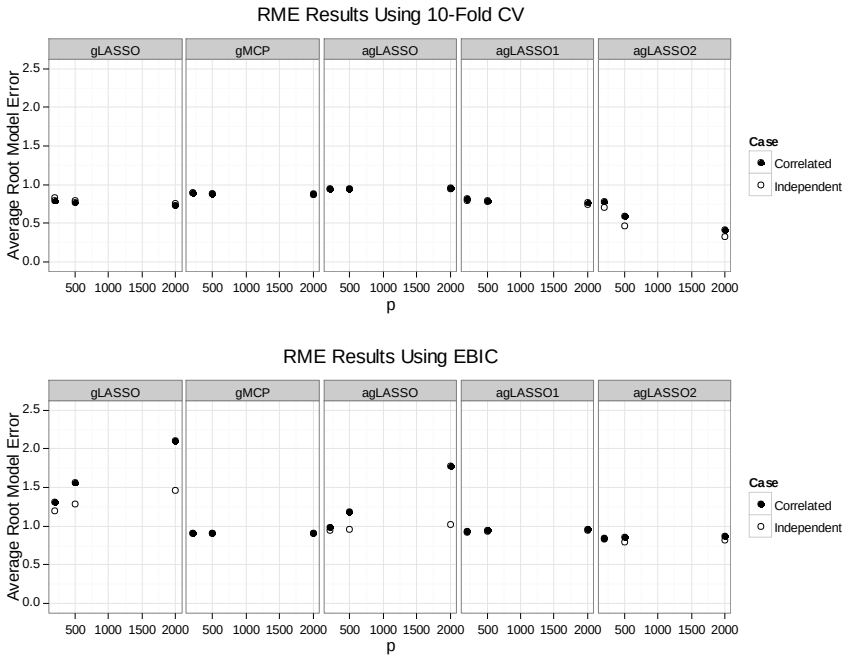

Figure 3.1: Model 1 RME results under different penalty parameter selection methods

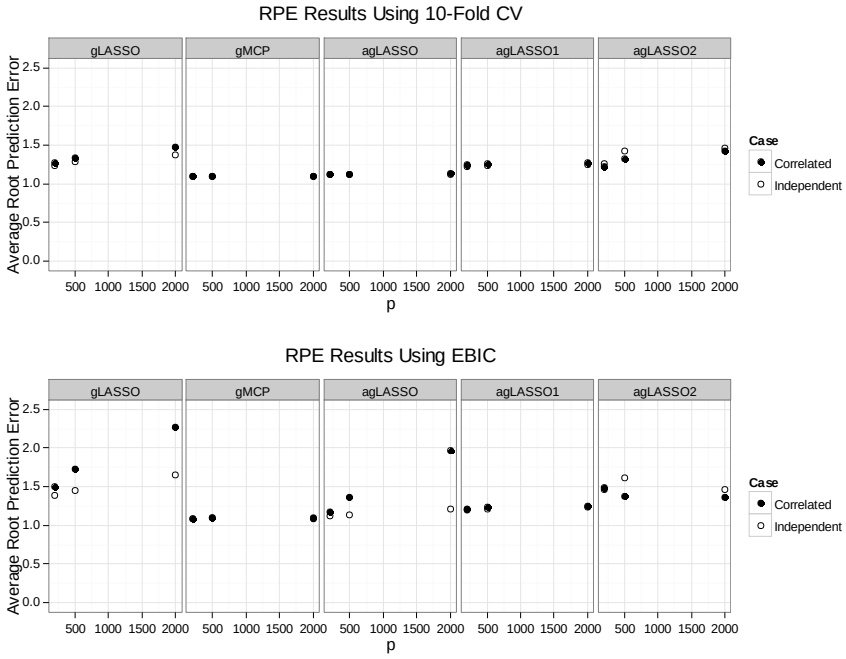

Figure 3.2: Model 1 RPE results under different penalty parameter selection methods

Model 2	p	Method	CV			
			CS (s.e.)	OS (s.e.)	MS (s.e.)	FDR (s.e.)
Large SNR	30	gLASSO	9.00(0.000)	12.40(5.431)	21.40(5.431)	0.55(0.143)
		gMCP	9.00(0.000)	1.82(2.878)	10.82(2.878)	0.12(0.173)
		agLASSO	9.00(0.000)	3.17(3.582)	12.17(3.582)	0.21(0.196)
		agLASSO1	9.00(0.000)	4.22(5.216)	13.22(5.216)	0.23(0.229)
		agLASSO2	9.00(0.000)	2.75(4.445)	11.75(4.445)	0.16(0.212)
	300	gLASSO	9.00(0.000)	44.43(24.154)	53.43(24.154)	0.79(0.115)
		gMCP	8.99(0.190)	5.84(7.072)	14.83(7.079)	0.28(0.261)
		agLASSO	6.68(1.256)	0.00(0.000)	6.68(1.256)	0.00(0.000)
		agLASSO1	8.99(0.134)	13.18(21.720)	22.18(21.722)	0.37(0.293)
		agLASSO2	9.00(0.000)	19.91(12.052)	28.91(12.052)	0.62(0.199)
	900	gLASSO	9.00(0.000)	60.28(33.997)	69.28(33.997)	0.83(0.116)
		gMCP	8.99(0.190)	6.95(8.009)	15.94(8.021)	0.31(0.270)
		agLASSO	6.62(1.215)	0.00(0.000)	6.62(1.215)	0.00(0.000)
		agLASSO1	8.99(0.190)	12.04(23.280)	21.02(23.287)	0.36(0.285)
		agLASSO2	9.00(0.000)	32.56(13.642)	41.56(13.642)	0.75(0.107)
Small SNR	30	gLASSO	9.00(0.000)	12.93(5.374)	21.93(5.374)	0.56(0.133)
		gMCP	8.86(0.642)	2.30(3.272)	11.15(3.427)	0.15(0.186)
		agLASSO	8.98(0.232)	3.26(3.710)	12.24(3.724)	0.21(0.201)
		agLASSO1	8.97(0.299)	4.39(5.469)	13.36(5.498)	0.24(0.232)
		agLASSO2	8.94(0.420)	2.74(3.939)	11.68(3.998)	0.17(0.203)
	300	gLASSO	8.96(0.353)	45.38(27.717)	54.34(27.745)	0.78(0.122)
		gMCP	8.71(0.893)	6.89(7.633)	15.59(7.931)	0.31(0.268)
		agLASSO	6.58(1.187)	0.00(0.000)	6.58(1.187)	0.00(0.000)
		agLASSO1	8.86(0.642)	13.67(22.477)	22.53(22.559)	0.39(0.289)
		agLASSO2	8.98(0.232)	22.07(12.242)	31.05(12.270)	0.65(0.191)
	900	gLASSO	8.87(0.602)	60.18(38.216)	69.05(38.282)	0.82(0.136)
		gMCP	8.34(1.244)	6.90(8.403)	15.24(8.913)	0.30(0.279)
		agLASSO	6.46(1.084)	0.01(0.134)	6.47(1.090)	0.00(0.015)
		agLASSO1	8.65(0.962)	15.23(27.633)	23.88(27.821)	0.38(0.306)
		agLASSO2	8.99(0.134)	32.62(13.266)	41.61(13.276)	0.76(0.101)

Table 3.5: Model 2 simulation results: selection performances using 10-fold CV

Model 2	p	Method	EBIC			
			CS (s.e.)	OS (s.e.)	MS (s.e.)	FDR (s.e.)
Large SNR	30	gLASSO	8.98(0.232)	0.72(1.515)	9.70(1.541)	0.06(0.112)
		gMCP	8.90(0.529)	0.04(0.327)	8.94(0.627)	0.00(0.027)
		agLASSO	8.96(0.353)	0.01(0.190)	8.97(0.402)	0.00(0.016)
		agLASSO1	8.91(0.512)	0.78(1.509)	9.69(1.637)	0.06(0.113)
		agLASSO2	8.96(0.353)	0.01(0.134)	8.96(0.378)	0.00(0.011)
	300	gLASSO	8.67(0.940)	0.55(1.305)	9.22(1.717)	0.04(0.101)
		gMCP	8.71(0.885)	0.07(0.460)	8.78(1.018)	0.01(0.038)
		agLASSO	6.14(0.642)	0.00(0.000)	6.14(0.642)	0.00(0.000)
		agLASSO1	8.74(0.842)	0.79(1.734)	9.53(2.021)	0.06(0.118)
		agLASSO2	8.96(0.353)	1.36(10.744)	10.32(10.755)	0.02(0.114)
	900	gLASSO	8.36(1.232)	0.46(1.225)	8.82(1.852)	0.04(0.096)
		gMCP	8.60(1.023)	0.55(4.661)	9.14(4.818)	0.02(0.093)
		agLASSO	6.07(0.440)	0.00(0.000)	6.07(0.440)	0.00(0.000)
		agLASSO1	8.63(0.983)	1.75(10.222)	10.38(10.328)	0.08(0.152)
		agLASSO2	8.96(0.327)	0.98(8.248)	9.95(8.255)	0.03(0.106)
Small SNR	30	gLASSO	8.56(1.066)	0.69(1.523)	9.25(1.999)	0.05(0.113)
		gMCP	7.51(1.501)	0.02(0.232)	7.53(1.537)	0.00(0.019)
		agLASSO	7.80(1.471)	0.00(0.000)	7.80(1.471)	0.00(0.000)
		agLASSO1	7.73(1.484)	0.51(1.205)	8.24(2.185)	0.04(0.098)
		agLASSO2	7.78(1.476)	0.01(0.190)	7.79(1.486)	0.00(0.019)
	300	gLASSO	7.45(1.501)	0.35(1.041)	7.81(2.018)	0.03(0.088)
		gMCP	7.05(1.432)	0.19(3.368)	7.24(3.753)	0.00(0.049)
		agLASSO	6.11(0.559)	0.00(0.000)	6.11(0.559)	0.00(0.000)
		agLASSO1	7.26(1.482)	0.62(6.264)	7.88(6.561)	0.03(0.099)
		agLASSO2	7.87(1.456)	4.62(19.812)	12.49(20.125)	0.05(0.202)
	900	gLASSO	6.96(1.439)	0.26(0.980)	7.22(1.854)	0.02(0.083)
		gMCP	7.06(1.436)	6.69(17.642)	13.75(18.420)	0.11(0.284)
		agLASSO	6.06(0.461)	0.00(0.000)	6.06(0.461)	0.00(0.000)
		agLASSO1	7.27(1.483)	7.93(23.876)	15.20(24.473)	0.14(0.285)
		agLASSO2	7.88(1.451)	7.88(25.535)	15.76(25.911)	0.10(0.255)

Table 3.6: Model 2 simulation results: selection performances using EBIC

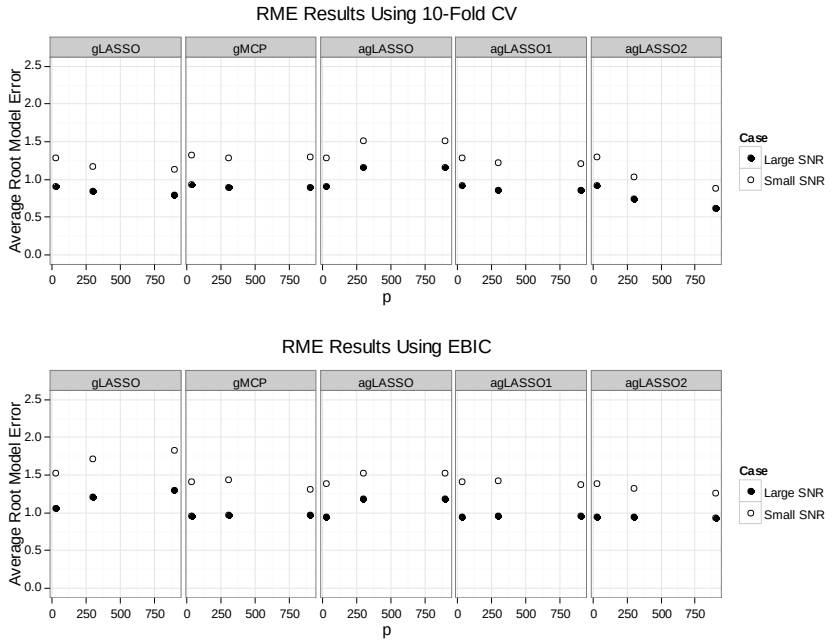

Figure 3.3: Model 2 RME results under different penalty parameter selection methods

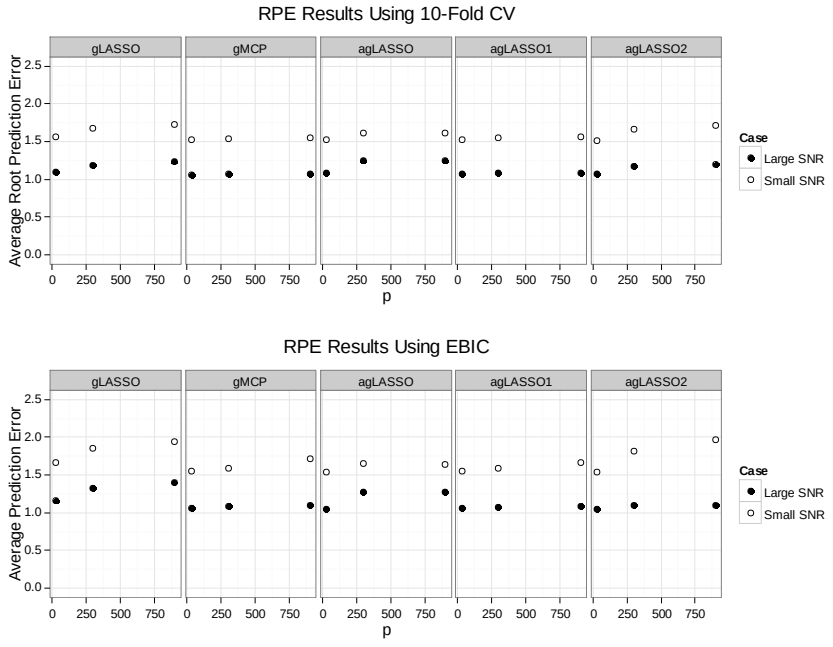

Figure 3.4: Model 2 RPE results under different penalty parameter selection methods

CHAPTER 4
THE ADAPTIVE GROUP LASSO IN NONPARAMETRIC ADDITIVE MODELS

In this chapter, we study the application of the adaptive group LASSO in nonparametric additive models. The specific weights for the adaptive group LASSO are proposed and studied in Chapter 3 for general models. In Section 4.1, theoretical properties of the method are present. In Section 4.2, we carry out extensive simulation studies on the proposed adaptive group LASSO as well the group MCP methods proposed in Chapter 2. In Section 4.3, we apply the two presently proposed group selection methods to data sets in different areas.

4.1 The adaptive group LASSO in nonparametric additive models

We use the general framework settings that are specified for the group selection in nonparametric additive models in section 2.1. We still assume the standard assumptions (A1) - (A3) and use the normalized B-spline basis $\{\phi_k, 1 \leq k \leq m_n\}$ for $\mathcal{S}_n$, whose standardizations are contained in $\mathbf{Z}$. We first obtain the initial estimator $\tilde{\boldsymbol{\beta}}$ as the minimizer to the group LASSO objective function

$$Q(\boldsymbol{\beta}; \lambda^*) = \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 + \lambda^* \sum_{j=1}^{p} \|\boldsymbol{\beta}_j\|_2. \tag{4.1}$$

Then we use the $\tilde{\boldsymbol{\beta}}$ in the following objective function with the group adaptive penalty,

$$Q(\boldsymbol{\beta}; \lambda, \gamma) = \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} \dot{\rho}(\|\tilde{\boldsymbol{\beta}}_j\|_2; \lambda, \gamma) \|\boldsymbol{\beta}_j\|_2, \tag{4.2}$$

where $\dot{\rho}$ is the first derivative of the MCP function. The minimizer $\hat{\boldsymbol{\beta}}$ to (4.2) is the estimator under our proposed adaptive group LASSO in nonparametric additive models.

Define

$$sgn_0(|x|) = \begin{cases} 0, & |x| > 0; \\ 1, & |x| = 0. \end{cases} \tag{4.3}$$

If $sgn_0(\|\hat{\boldsymbol{\beta}}\|_2) = sgn_0(\|\boldsymbol{\beta}\|_2)$, $j = 1, \ldots, p$, then we say $\hat{\boldsymbol{\beta}} =_0 \boldsymbol{\beta}$.

Theorem 4.1

If

$$\frac{m_n}{n} + \frac{\lambda m_n}{n} \to 0 \text{ and } \frac{\sqrt{nlog((p-q)m_n)}}{\lambda} + \frac{n}{\lambda m_n^{d+1/2}} \to 0, \tag{4.4}$$

then

$$(a) P(\hat{\boldsymbol{\beta}} =_0 \boldsymbol{\beta}) \to 1; \tag{4.5}$$

$$(b) \sum_{j=1}^{q} \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_2^2 = O_p\left(\frac{m_n^2}{n}\right) + O_p\left(\frac{m_n}{n}\right) + O\left(\frac{4\lambda^2 m_n}{n^2}\right); \tag{4.6}$$

$$(c) \sum_{j=1}^{q} \|\hat{f}_j - f_j\|_2^2 = O_p\left(\frac{m_n}{n}\right) + O_p\left(\frac{1}{n}\right) + O\left(\frac{4\lambda^2}{n^2}\right). \tag{4.7}$$

Proof of Theorem 4.1

The details of the proof is omitted because the detailed proof of Theorem 3, Lemma 5 and Lemma 6 from Huang, Horowitz and Wei (2010) can be easily applied to obtain the theoretical results in our theorem. The conditions needed for our theorem are slightly different, resulting in different convergence bounds in (4.6) and (4.7).

Another major modification here is that with $\mathbf{v} = \left(w_j\hat{\boldsymbol{\beta}}_j/(2\|\hat{\boldsymbol{\beta}}_j\|_2), j \in A_1\right)$, we have

$$\|\mathbf{v}\|_2 = \sum_{j\in A_1} w_j^2 = \sum_{j\in A_1} \left(\dot{\rho}(\|\tilde{\boldsymbol{\beta}}_j\|_2; \lambda, \gamma)/\lambda\right)^2 \leq \sum_{j\in A_1} 1 = q \tag{4.8}$$

because the proposed weights $w_j = \dot{\rho}(\|\tilde{\boldsymbol{\beta}}_j\|_2; \lambda, \gamma)/\lambda \in [0, 1]$. □

4.2 Simulations

4.2.1 Models and methods

The simulation models borrow heavily from Yuan and Lin (2006) and Meier, van de Geer and Bühlmann (2009), where only the group LASSO method was investigated. Here cases with different signal-to-noise ratio (SNR) are examined. For nonparametric additive models in the following, the SNR is defined as

$$\text{SNR} = \frac{Var(f(\boldsymbol{X}))}{Var(\boldsymbol{\epsilon})}.$$

Consider the model of the general form

$$Y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i = f_1(X_{i1}) + f_2(X_{i2}) + f_3(X_{i3}) + f_4(X_{i4}) + \epsilon_i, \quad i = 1, \ldots, n, \tag{4.9}$$

where $f_j(X_{ij}) = 0$ for $j = 5, \ldots, p$ and ϵ_i's are i.i.d. $N(0, \sigma^2)$. In the following, Model 3 and Model 4 use model of the same general form but different functions for f_1 to f_4. In addition, independent as well as correlated cases are considered under both models, but the correlation structures used are different for each model.

Model 3

For the model in (4.9),

$$f_1(x) = -sin(2x),$$

$$f_2(x) = x^2 - 25/12,$$

$$f_3(x) = x,$$

$$f_4(x) = e^{-x} - 2/5 \cdot sinh(5/2).$$

$\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip}), i = 1, \ldots, n$, and they are assumed to come from a multivariate normal distribution $N_{m_np}(\mathbf{0}, \boldsymbol{\Sigma})$.

Independent Case (SNR ≈ 15):

$\boldsymbol{\Sigma} = \mathbf{I}_{m_np}$, the m_np-dimensional identity matrix. As a result, entries in any $\boldsymbol{X}_i$ are i.i.d. $N(0, 1)$.

Correlated Case (SNR ≈ 6.7):

$\boldsymbol{\Sigma} = \{\sigma_{jj'}\}$, where $\sigma_{jj'} = 0.5^{|j-j'|}$, $j, j' = 1, \ldots p$, is the covariance between any two entries X_j and $X_{j'}$ in $\boldsymbol{X}$.

Model 4

For the model in (4.9),

$$f_1(x) = 5x,$$

$$f_2(x) = 3(2x-1)^2,$$

$$f_3(x) = \frac{4sin(2\pi x)}{2 - sin(2\pi x)},$$

$$f_4(x) = 0.6sin(2\pi x) + 1.2cos(2\pi x) + 1.8sin^2(2\pi x) + 2.4cos^3(2\pi x) + 3.0sin^3(2\pi x),$$

$$\sigma = 1.27.$$

Any X_j in $\boldsymbol{X} = (X_1, \ldots, X_p)$ is generated from

$$\frac{W_j + tU}{1+t}, \quad j = 1, \ldots, p,$$

where W_j and U are i.i.d $Uniform(0,1)$.

Independent Case (SNR ≈ 9):

when $t = 0$, all X_j's are independently and identically uniformly distributed.

Correlated Case (SNR ≈ 7.9):

when $t = 1$, all pair-wise X_j and $X_{j'}$ are correlated with a constant correlation 0.5, $j, j' = 1, \ldots, p$.

We use B-splines of degree 3 to approximate each component. The numbers of knots of the B-splines are decided by the integer closest to $n^{0.3}$. The knots are decided

by corresponding percentiles. Under $n = 100$ and 300, different p's (50, 200 or 500) are used.

4 group penalized methods are applied for comparison. They are (1) the group LASSO, (2) the group MCP, (3) the adaptive group LASSO that uses the first derivative of the group MCP penalty as the weights, and (4) the adaptive group LASSO that is a recursive application of (3). The 4 methods are referred to as gLASSO, gMCP, agLASSO1 and agLASSO2 respectively in the simulation results section (Section 4.2.2). For anywhere the group MCP is applied, the penalty parameter γ is fixed at 2.7.

We run 500 simulation replications on all combinations of different values of n, p and group selection methods. The group coordinate descent algorithm is used for calculating the solution paths. The 10-fold CV and EBIC are used for identifying the optimal penalty parameter λ from a sequence of k^* λ's chosen according to the method detailed in Section 2.3, where $k^* = 100$ and $\epsilon^* = 0.001$. The definition and parameter value choices of EBIC are the same as used in Section 3.4.

In order to evaluate the selection properties of each method, we study the number of correctly selected components (CS) and the number of incorrectly selected components (OS) as well as the false discovery rate (FDR) and false negative rate (FNR). The definition of the FDR is

$$\text{FDR} = \frac{\text{OS}}{\text{MS}} = \frac{\text{number of true zero components in the final model}}{\text{number of components in the final model}}.$$

The definition of the FNR is

$$\begin{aligned}\text{FNR} &= 1 - \frac{\text{CS}}{\text{number of true non-zero components in the underlying model}} \\ &= 1 - \frac{\text{number of true non-zero components in the final model}}{\text{number of true non-zero components in the underlying model}}.\end{aligned}$$

To evaluate the estimation performance, we use the empirical root model error (RME). To help compare the prediction performance of different methods, we use the empirical root prediction error (RPE). The definitions of RME and RPE used can be found in Section 3.3.1.

4.2.2 Simulation results

For Model 3, we report the CS and OS in two tables: Table 4.1 includes the results under 10-fold CV while Table 4.2 includes the results under EBIC. We present the FDR, FNR, RME and RPE in Figure 4.1, Figure 4.2, Figure 4.3 and Figure 4.4 respectively. All 4 figures contain results under different penalty parameter selection methods (10-fold CV and EBIC) and sample sizes (n = 100, 300).

For Model 4, we report selection, estimation and prediction performances in the same way: we display the CS and OS in tables (Table 4.3 and Table 4.4) and present the FDR, FNR, RME and RPE in figures (Figure 4.5, Figure 4.6, Figure 4.7 and Figure 4.8, respectively).

Model 3	p	Method	n = 100		n = 300	
			CS (s.e.)	OS (s.e.)	CS (s.e.)	OS (s.e.)
Independent	50	gLASSO	3.69(0.712)	4.07(4.176)	3.99(0.109)	8.02(6.506)
		gMCP	3.57(0.784)	0.47(1.033)	3.98(0.172)	0.74(2.078)
		agLASSO1	3.95(0.313)	5.84(6.774)	4.00(0.000)	7.50(9.574)
		agLASSO2	4.00(0.063)	1.62(2.877)	4.00(0.000)	1.34(4.326)
	200	gLASSO	3.39(0.923)	4.89(5.445)	3.93(0.337)	10.97(9.983)
		gMCP	3.47(0.880)	0.68(1.343)	3.94(0.301)	1.25(3.166)
		agLASSO1	3.89(0.507)	7.38(7.931)	4.00(0.000)	12.02(17.617)
		agLASSO2	3.99(0.126)	2.32(3.626)	4.00(0.000)	1.69(4.521)
	500	gLASSO	3.25(0.894)	5.68(6.722)	4.00(0.045)	15.40(13.314)
		gMCP	3.44(0.867)	0.92(1.468)	3.99(0.118)	1.32(3.443)
		agLASSO1	3.90(0.470)	9.37(9.994)	4.00(0.000)	12.37(20.007)
		agLASSO2	3.99(0.148)	2.80(3.947)	4.00(0.000)	1.82(5.514)
Correlated	50	gLASSO	3.25(0.892)	3.73(4.495)	3.98(0.189)	10.45(6.995)
		gMCP	3.30(0.906)	0.56(1.060)	3.94(0.330)	0.70(1.922)
		agLASSO1	3.85(0.411)	4.96(5.990)	3.99(0.100)	6.59(8.690)
		agLASSO2	3.98(0.154)	0.92(1.389)	4.00(0.000)	0.59(1.465)
	200	gLASSO	2.83(0.928)	4.32(5.433)	3.86(0.473)	13.60(11.951)
		gMCP	3.07(0.997)	0.79(1.274)	3.86(0.496)	1.05(2.778)
		agLASSO1	3.72(0.620)	6.95(8.942)	3.98(0.188)	10.78(16.189)
		agLASSO2	3.97(0.202)	1.35(1.746)	4.00(0.045)	0.74(1.555)
	500	gLASSO	2.60(0.847)	4.63(5.900)	3.97(0.227)	18.00(14.753)
		gMCP	2.87(1.013)	0.87(1.414)	3.94(0.308)	1.11(2.872)
		agLASSO1	3.64(0.647)	8.29(10.052)	4.00(0.000)	11.88(20.241)
		agLASSO2	3.97(0.187)	2.04(2.871)	4.00(0.000)	0.84(1.700)

Table 4.1: Model 3 simulation results: selection performances using 10-fold CV

Model 3	p	Method	n = 100		n = 300	
			CS (s.e.)	OS (s.e.)	CS (s.e.)	OS (s.e.)
Independent	50	gLASSO	0.91(1.167)	0.01(0.077)	4.00(0.045)	0.02(0.133)
		gMCP	3.51(0.855)	1.27(4.067)	4.00(0.045)	0.11(1.396)
		agLASSO1	3.88(0.545)	6.42(11.889)	4.00(0.000)	6.19(13.887)
		agLASSO2	3.90(0.351)	5.79(4.555)	4.00(0.000)	6.16(5.006)
	200	gLASSO	0.51(0.740)	0.00(0.045)	4.00(0.063)	0.02(0.125)
		gMCP	3.04(1.243)	1.46(4.037)	4.00(0.000)	0.08(0.859)
		agLASSO1	3.54(1.089)	21.61(54.582)	4.00(0.000)	9.14(23.651)
		agLASSO2	3.87(0.449)	5.87(4.295)	4.00(0.000)	6.93(4.848)
	500	gLASSO	0.42(0.627)	0.01(0.100)	3.98(0.150)	0.03(0.157)
		gMCP	2.59(1.346)	1.15(3.736)	4.00(0.000)	0.11(1.013)
		agLASSO1	3.23(1.318)	40.93(126.648)	4.00(0.000)	9.18(25.844)
		agLASSO2	3.86(0.453)	6.34(4.463)	4.00(0.000)	6.55(4.817)
Correlated	50	gLASSO	0.76(0.902)	0.00(0.000)	3.83(0.522)	0.06(0.242)
		gMCP	2.94(1.125)	2.60(6.188)	4.00(0.045)	0.38(2.569)
		agLASSO1	3.60(0.804)	8.32(15.373)	4.00(0.000)	5.58(12.853)
		agLASSO2	3.89(0.377)	6.55(5.852)	4.00(0.000)	5.48(4.544)
	200	gLASSO	0.51(0.731)	0.01(0.089)	3.44(0.845)	0.00(0.063)
		gMCP	2.37(1.260)	2.36(5.972)	4.00(0.000)	0.40(2.853)
		agLASSO1	3.09(1.185)	29.01(65.866)	4.00(0.000)	10.77(25.427)
		agLASSO2	3.85(0.456)	6.68(5.642)	4.00(0.000)	6.09(4.835)
	500	gLASSO	0.38(0.625)	0.01(0.089)	3.09(0.946)	0.01(0.092)
		gMCP	1.90(1.233)	1.08(3.908)	3.99(0.067)	0.29(2.248)
		agLASSO1	2.69(1.344)	38.88(127.628)	4.00(0.000)	8.58(24.187)
		agLASSO2	3.86(0.430)	6.23(4.864)	4.00(0.045)	5.36(4.696)

Table 4.2: Model 3 simulation results: selection performances using EBIC

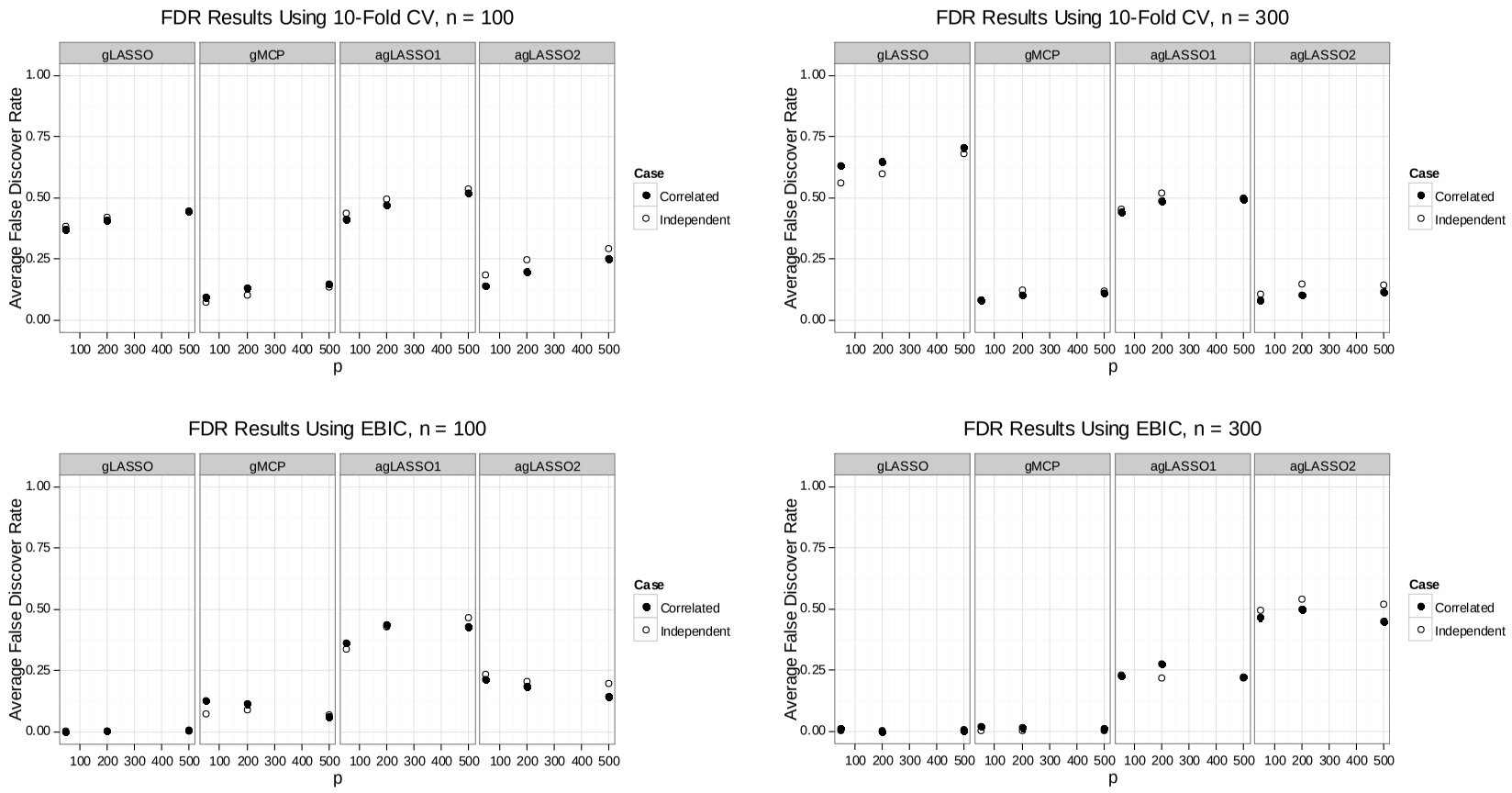

Figure 4.1: Model 3 FDR results under different penalty parameter selection methods and sample sizes

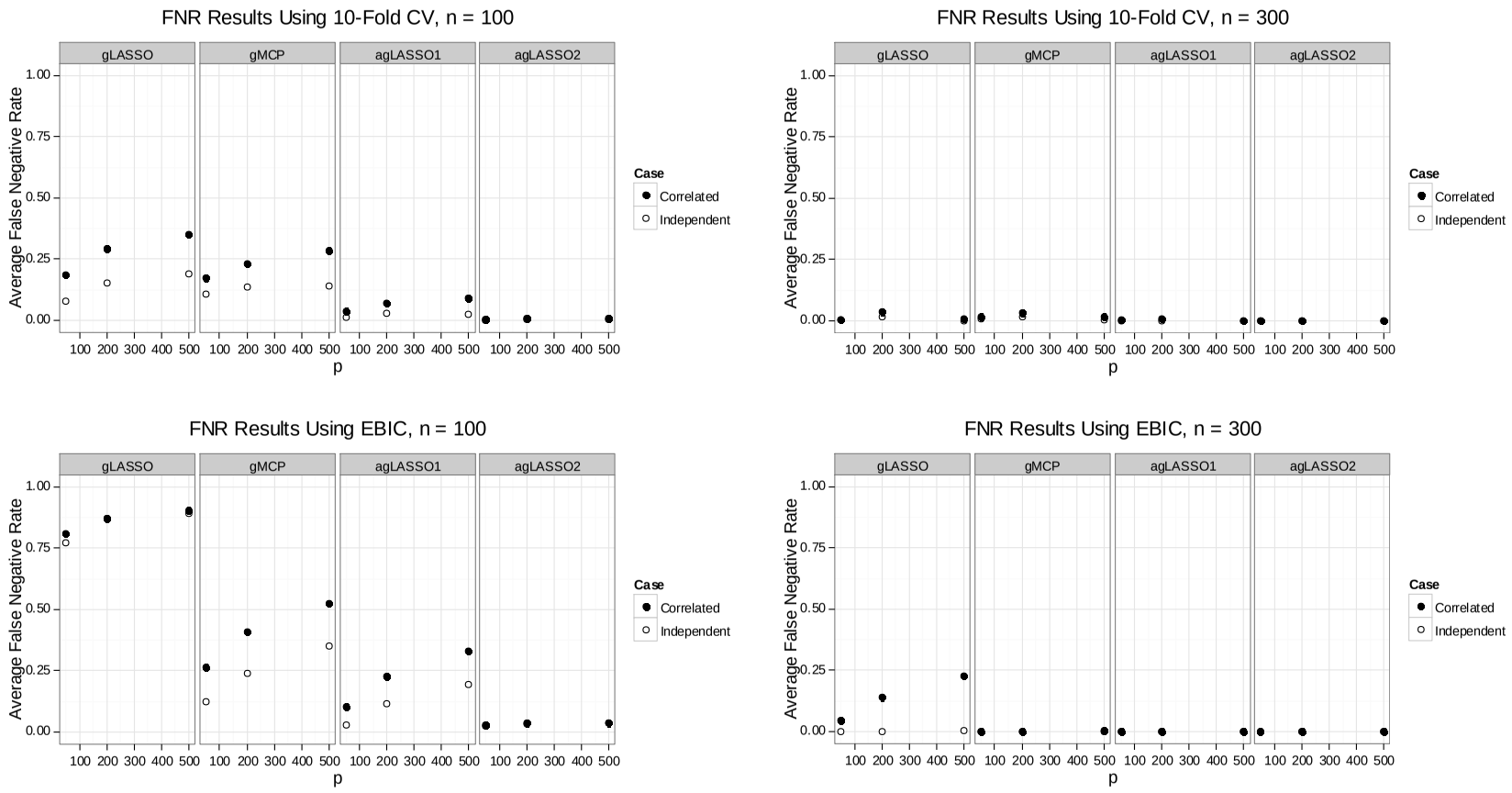

Figure 4.2: Model 3 FNR results under different penalty parameter selection methods and sample sizes

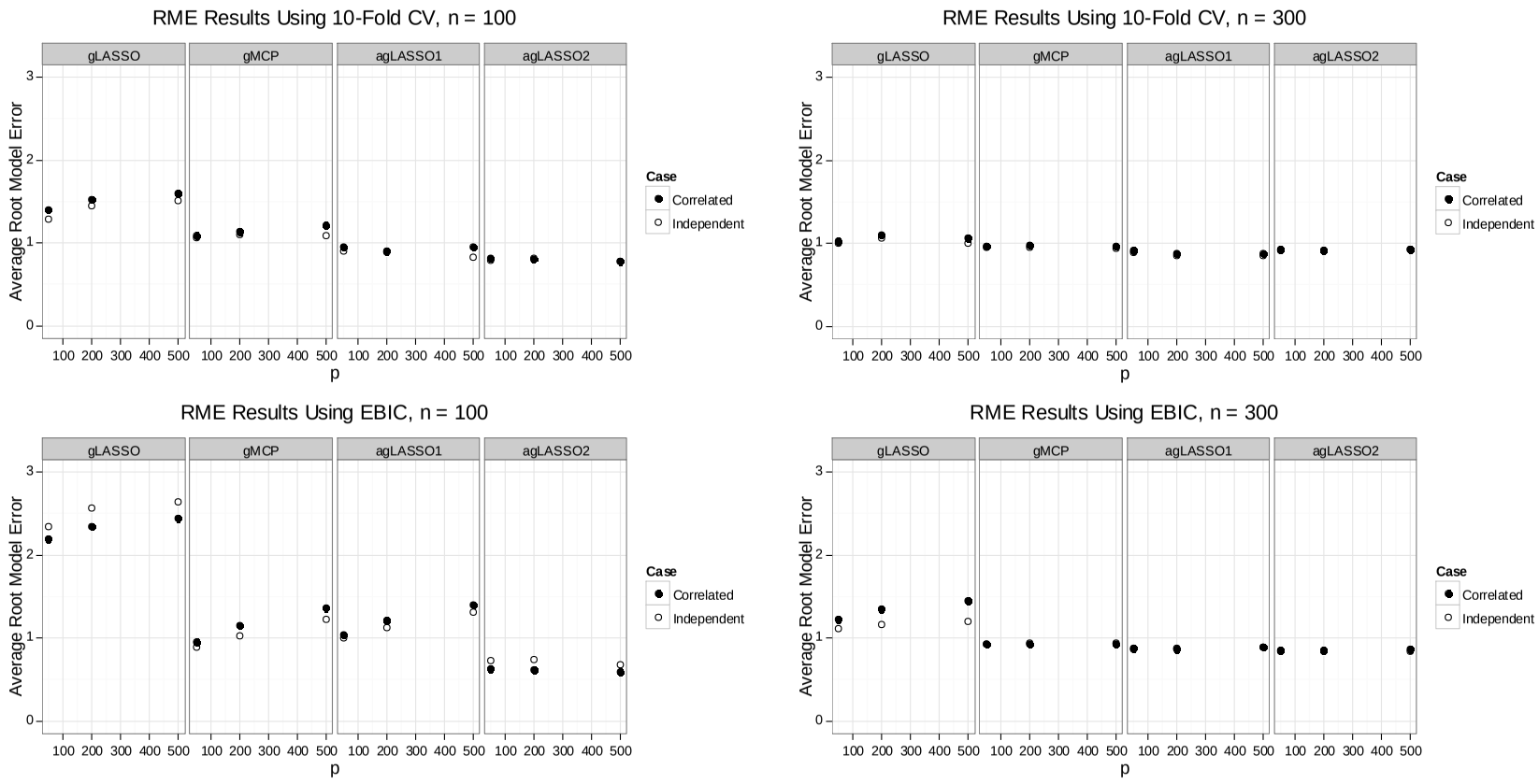

Figure 4.3: Model 3 RME results under different penalty parameter selection methods and sample sizes

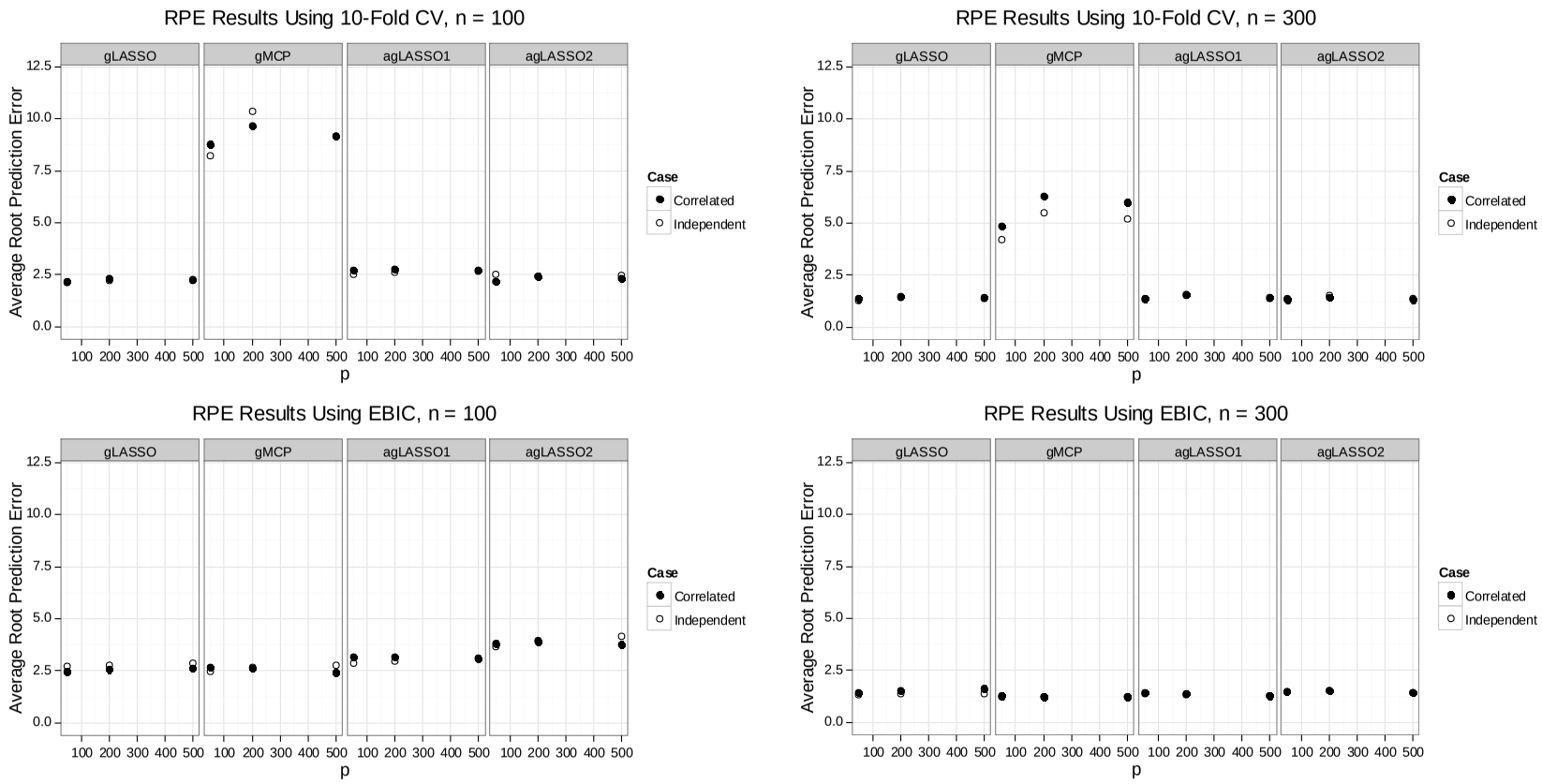

Figure 4.4: Model 3 RPE results under different penalty parameter selection methods and sample sizes

Model 4	p	Method	n = 100		n = 300	
			CS (s.e.)	OS (s.e.)	CS (s.e.)	OS (s.e.)
Independent	50	gLASSO	3.99(0.100)	13.72(5.308)	4.00(0.000)	15.81(6.465)
		gMCP	3.93(0.252)	0.93(1.082)	4.00(0.000)	0.92(1.889)
		agLASSO1	4.00(0.000)	4.21(5.410)	4.00(0.000)	3.16(4.990)
		agLASSO2	4.00(0.000)	0.55(0.856)	4.00(0.000)	0.30(0.731)
	200	gLASSO	3.93(0.249)	20.91(8.457)	4.00(0.000)	27.36(13.134)
		gMCP	3.90(0.298)	1.25(1.277)	4.00(0.000)	1.71(2.998)
		agLASSO1	4.00(0.000)	5.57(7.417)	4.00(0.000)	5.72(10.602)
		agLASSO2	4.00(0.000)	0.85(0.979)	4.00(0.000)	0.91(1.460)
	500	gLASSO	3.83(0.383)	23.81(10.166)	4.00(0.000)	33.96(17.004)
		gMCP	3.86(0.349)	1.36(1.193)	4.00(0.000)	1.62(3.045)
		agLASSO1	4.00(0.000)	7.28(8.570)	4.00(0.000)	5.79(11.884)
		agLASSO2	4.00(0.000)	1.11(1.058)	4.00(0.000)	1.25(1.702)
Correlated	50	gLASSO	3.26(0.740)	7.68(5.973)	4.00(0.000)	17.89(5.319)
		gMCP	3.10(0.775)	0.80(1.089)	4.00(0.045)	1.15(1.797)
		agLASSO1	3.99(0.077)	4.93(5.990)	4.00(0.000)	3.20(4.456)
		agLASSO2	4.00(0.045)	0.38(0.837)	4.00(0.000)	0.25(0.585)
	200	gLASSO	2.76(0.733)	10.24(8.425)	4.00(0.045)	30.29(11.099)
		gMCP	2.73(0.738)	1.13(1.395)	3.99(0.100)	2.00(2.626)
		agLASSO1	3.99(0.089)	8.12(9.222)	4.00(0.000)	5.13(8.977)
		agLASSO2	4.00(0.000)	0.48(0.874)	4.00(0.000)	0.57(1.058)
	500	gLASSO	2.46(0.601)	11.41(10.003)	3.99(0.109)	39.17(15.933)
		gMCP	2.53(0.697)	1.23(1.487)	3.99(0.077)	2.93(3.640)
		agLASSO1	3.98(0.140)	9.59(10.899)	4.00(0.000)	5.27(8.928)
		agLASSO2	4.00(0.063)	0.59(0.927)	4.00(0.000)	0.87(1.469)

Table 4.3: Model 4 simulation results: selection performances using 10-fold CV

Model 4	p	Method	n = 100		n = 300	
			CS (s.e.)	OS (s.e.)	CS (s.e.)	OS (s.e.)
Independent	50	gLASSO	1.36(1.334)	0.00(0.045)	4.00(0.000)	0.02(0.140)
		gMCP	3.56(0.600)	0.21(1.508)	4.00(0.045)	0.00(0.000)
		agLASSO1	3.98(0.172)	1.91(5.831)	4.00(0.000)	0.54(2.968)
		agLASSO2	3.98(0.194)	8.28(3.290)	4.00(0.000)	21.39(6.328)
	200	gLASSO	0.62(0.837)	0.00(0.000)	3.99(0.089)	0.01(0.118)
		gMCP	3.24(0.844)	0.26(1.568)	4.00(0.045)	0.04(0.894)
		agLASSO1	3.90(0.419)	3.50(17.831)	4.00(0.000)	1.21(8.898)
		agLASSO2	3.96(0.202)	6.34(4.475)	4.00(0.000)	15.25(10.703)
	500	gLASSO	0.44(0.622)	0.00(0.000)	3.99(0.109)	0.01(0.089)
		gMCP	2.96(0.966)	0.34(1.785)	4.00(0.063)	0.09(1.376)
		agLASSO1	3.82(0.460)	8.30(54.028)	4.00(0.000)	1.49(11.187)
		agLASSO2	3.91(0.297)	4.73(4.618)	4.00(0.000)	7.00(10.059)
Correlated	50	gLASSO	1.91(0.379)	0.01(0.077)	2.50(0.774)	0.00(0.045)
		gMCP	2.08(0.686)	0.02(0.363)	3.55(0.646)	0.00(0.000)
		agLASSO1	3.84(0.369)	0.97(2.304)	4.00(0.000)	1.74(7.556)
		agLASSO2	3.96(0.246)	6.61(2.409)	4.00(0.045)	13.26(3.887)
	200	gLASSO	1.71(0.491)	0.00(0.000)	2.18(0.413)	0.00(0.000)
		gMCP	1.89(0.640)	0.00(0.045)	3.25(0.811)	0.00(0.000)
		agLASSO1	3.75(0.443)	1.22(2.405)	4.00(0.000)	2.91(14.251)
		agLASSO2	3.93(0.316)	6.33(2.869)	4.00(0.063)	13.22(4.228)
	500	gLASSO	1.59(0.571)	0.00(0.045)	2.14(0.361)	0.00(0.000)
		gMCP	1.78(0.609)	0.00(0.045)	3.04(0.847)	0.04(0.850)
		agLASSO1	3.68(0.482)	1.39(3.202)	4.00(0.000)	1.51(9.733)
		agLASSO2	3.94(0.298)	5.92(3.062)	4.00(0.045)	13.56(4.013)

Table 4.4: Model 4 simulation results: selection performances using EBIC

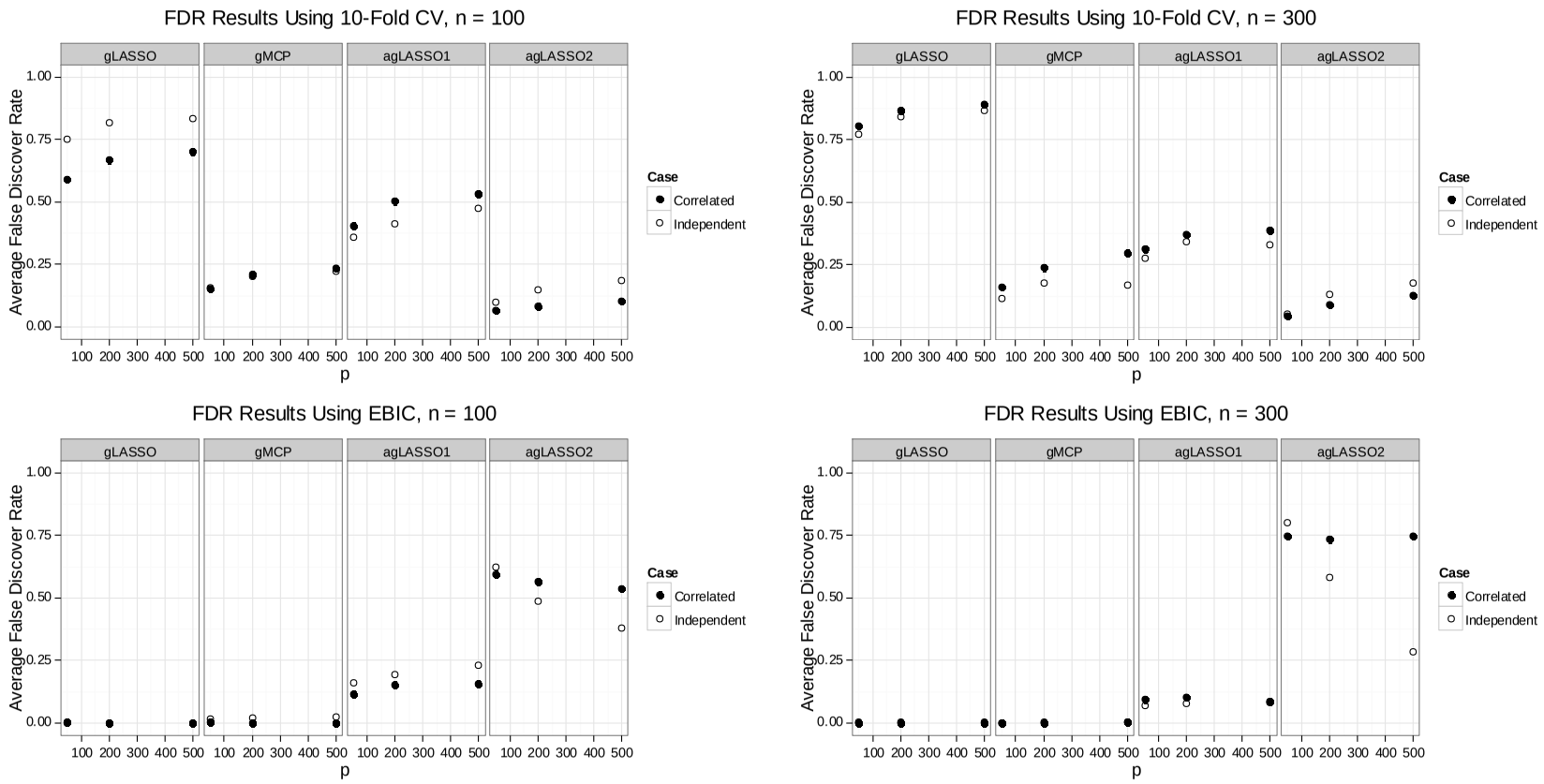

Figure 4.5: Model 4 FDR results under different penalty parameter selection methods and sample sizes

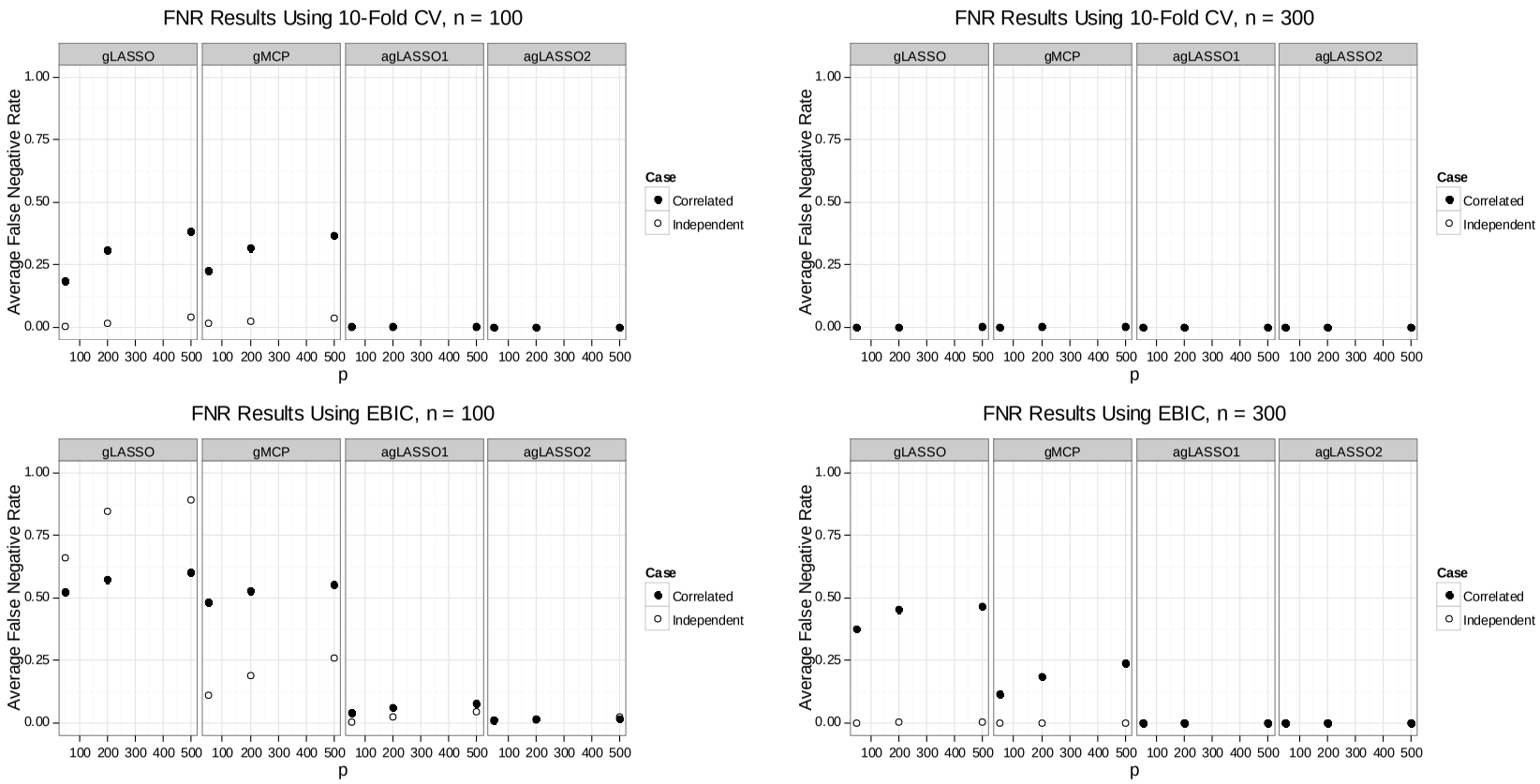

Figure 4.6: Model 4 FNR results under different penalty parameter selection methods and sample sizes

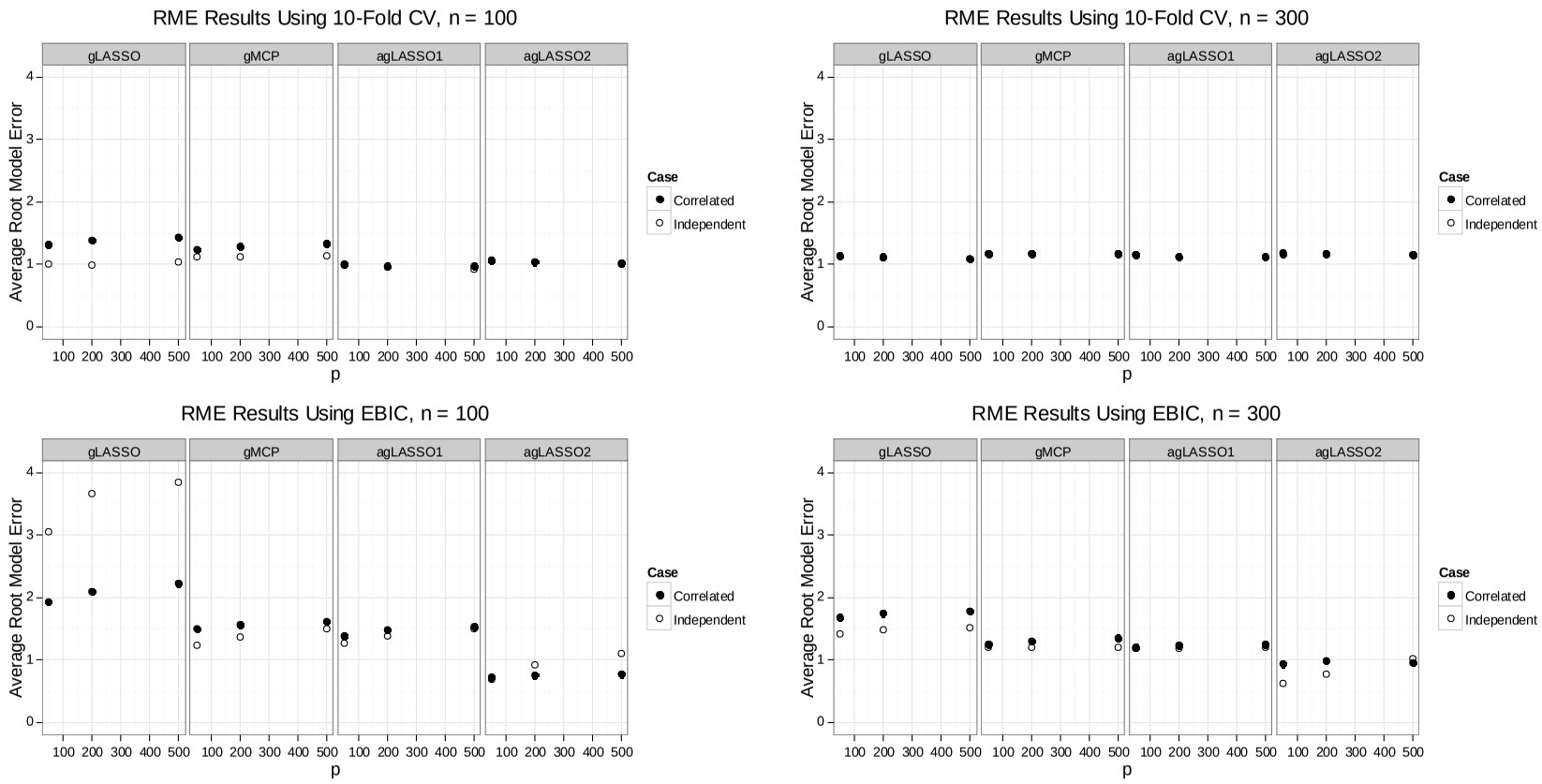

Figure 4.7: Model 4 RME results under different penalty parameter selection methods and sample sizes

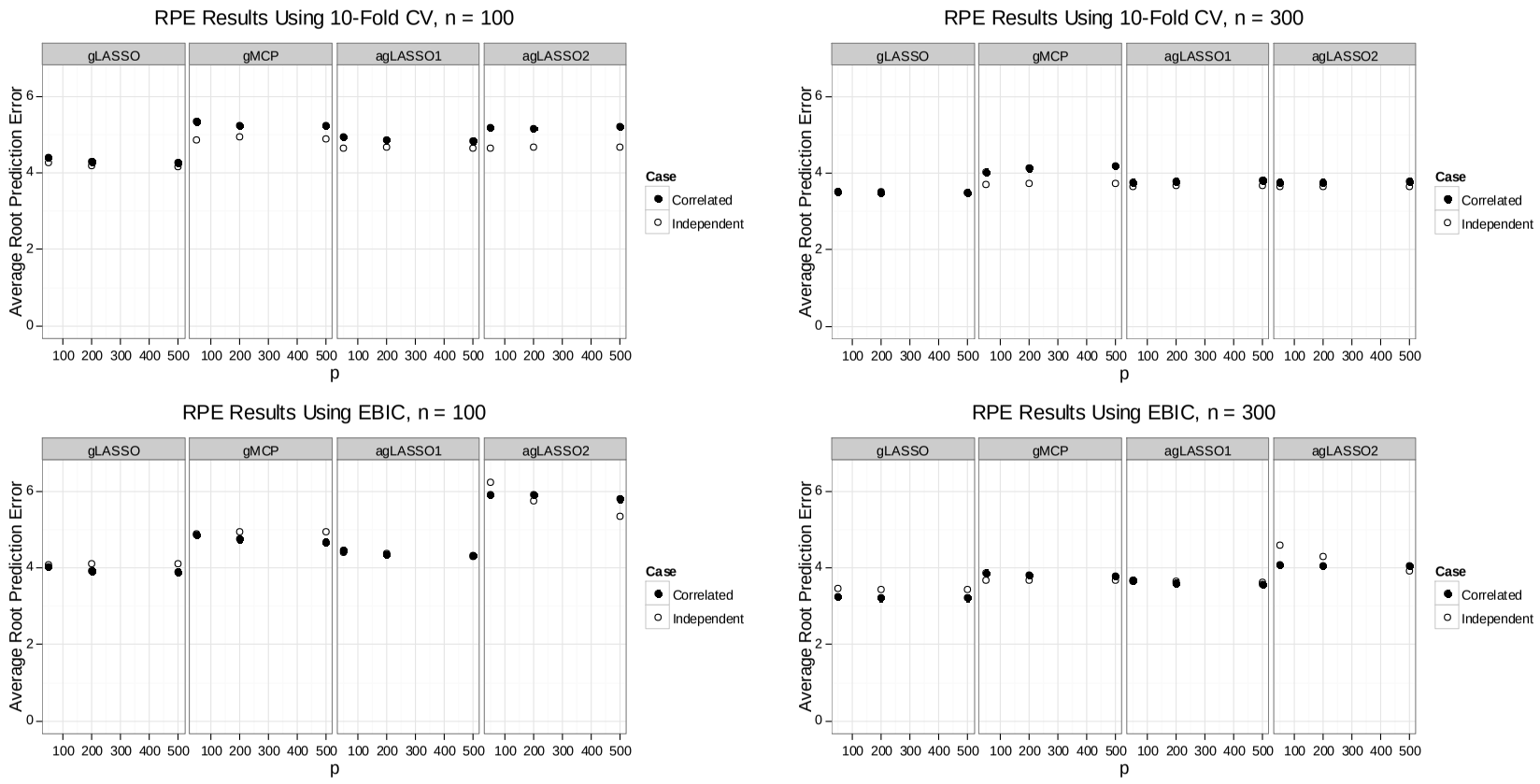

Figure 4.8: Model 4 RPE results under different penalty parameter selection methods and sample sizes

About Model 3 results, under 10-fold CV, agLASSO1 and agLASSO2 seem to perform better than gLASSO in terms of CS; gMCP and agLSSO2 seem to perform better than gLASSO in terms of OS. Several results under EBIC, for example the CS of gLASSO and OS of agLASSO1, are far from satisfactory, which implies that EBIC may not be a good choice of penalty parameter selection method under such cases.

About Model 4 results, under 10-fold CV, we have similar findings as those for Model 3 about CS and OS. One noticeable impact of the correlation is that the CS and FDR's of gMCP and gLASSO under the correlated cases are much smaller than those under the independent cases. Once again, some results under EBIC leave the method a questionable choice for gLASSO or gLASSO2.

4.3 Data examples

4.3.1 Breast cancer data

To demonstrate the use of the two proposed group selection methods, the group MCP and the adaptive group LASSO, in nonparametric additive models, we study their applications in breast cancer data collected under The Cancer Genome Atlas (TCGA) project. The data are available under the website `http://cancergenome.nih.gov/`. We extract a subset of the gene expression that used Agilent mRNA expression microarrays. Our data set contains 17814 genes and 253 patients. Our goal is to pick the genes that are closely related to the gene BRCA1, a tumor suppressor gene whose mutation is considered as breast cancer oncogene.

We first screen the genes using the method described in Huang et al. (2013) in

order to exclude genes whose expression data that do not vary much. More specifically, we select genes whose expression data have

(a) an effect size (mean/variance) greater than 1;

(b) a standard deviation greater than 0.6;

(c) a correlation with BRCA1 greater than 1.

The screening results in a final analysis set with 1587 genes, BRCA1 gene included.

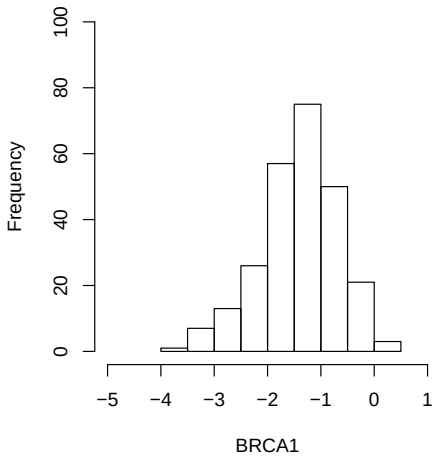

Figure 4.9: Histogram of BRAC1 expression data

A preliminary look at the final analysis set shows that the response variable BRCA1 is skewed to the left (Figure 4.9). We also arbitrarily select the 300th, 600th, 900th, 1200th genes from the alphabetically ordered 1587 genes and present their histograms in Figure 4.10. We observe skewness from all 4 genes.

The gene association study was previously carried out by Huang et al. (2013)

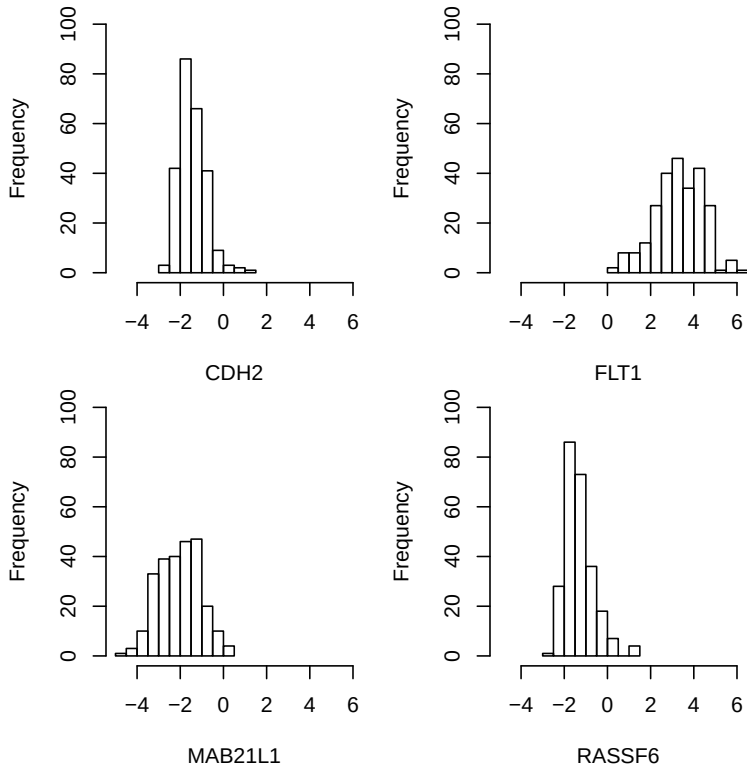

Figure 4.10: Histograms of 4 arbitrarily selected genes

using LASSO when they study a newly proposed semi-parametric variable selection method. The skewness in data leads us to consider nonparametric models such as the nonparametric additive models used below. We expand each covariates using B-spline basis functions. The detailed method is the same as the one used in simulation studies in this dissertation. In addition to the LASSO method, we perform 4 group selection methods: the group LASSO, the group MCP, the adaptive group LASSO described in Section 3.2 as well as its recursive application studied in Section 3.2. We denote these 5 variable selection methods as LASSO, gLASSO, gMCP, agLASSO1, agLASSO2, respectively. Their selection results using 10-fold CV are summarized in Table 4.5.

	LASSO	gLASSO	gMCP	agLASSO1	agLASSO2
Genes Selected	139	28	3	15	30
Root Model Error	0.210	0.466	0.623	0.484	0.416

Table 4.5: Breast cancer data selection results

Several observations are made by scrutinizing the lists of genes selected by different variable selection methods. For this specific data set, agLASSO2 method includes all the genes selected by agLASSO1 method and more. A number of genes are selected by the majority of the methods. In Table 4.6 we report genes that are selected by at least 3 out of the 5 selection methods.

	LASSO	gLASSO	gMCP	agLASSO1	agLASSO2
AASS	✓			✓	✓
C16orf59		✓		✓	✓
CCDC40	✓	✓		✓	✓
DSN1	✓	✓	✓	✓	✓
FOXG1		✓	✓	✓	✓
GBX2		✓	✓	✓	✓
ITGBIBP2	✓	✓			✓
KLHDC8A	✓	✓			✓
RCN3	✓	✓			✓
RSU1	✓			✓	✓
SGOL2	✓	✓		✓	✓

Table 4.6: Breast cancer data selected genes

4.3.2 Boston housing data revisit

At the end of Chapter 2, we used Boston housing data to demonstrate the application of the group MCP method in nonparametric additive models. Here we once again use the data set and add 200 $N(0,1)$ random variables to the candidate component list. Out of the 506 housing data records, we use 300 randomly selected records for estimation and the rest 206 for prediction. We repeat such a partition 20 times on the same data set that has 204 total variables (4 "important" variables and 200 artificial noise variables) under the 10-fold CV. In contrast to the Chapter 2 data example, we now perform 5 variable selection methods, namely the LASSO, the group LASSO, the group MCP, the adaptive group LASSO described in Section 3.2 as well as its recursive application studied in Section 3.2. We denote them as LASSO, gLASSO, gMCP, agLASSO1, agLASSO2 respectively when we present the results in Table 4.7.

	LASSO (s.e.)	gLASSO (s.e.)	gMCP (s.e.)	agLASSO1 (s.e.)	agLASSO2 (s.e.)
Model Size	11.05 (8.624)	7.50 (3.693)	3.30 (2.203)	6.70 (5.630)	4.10 (0.553)
FDR	0.57 (0.205)	0.39 (0.245)	0.26 (0.262)	0.37 (0.185)	0.04 (0.082)
FNR	0.13 (0.128)	0.06 (0.111)	0.48 (0.112)	0.15 (0.150)	0.03 (0.077)
RME	5.20 (0.293)	4.15 (0.387)	4.19 (0.317)	3.97 (0.431)	3.85 (0.211)
RPE	5.31 (0.344)	4.76 (0.617)	4.68 (0.427)	4.72 (0.606)	4.50 (0.500)

Table 4.7: Boston housing data selection results

CHAPTER 5
SUMMARY AND DISCUSSION

In this dissertation, I study the theoretical properties of the group MCP and the adaptive group LASSO in high-dimensional nonparametric additive models. I also present theoretical results on the the adaptive group LASSO and its recursive application in general models. The proposed methods have advantages in terms of selection and estimation properties. Several simulation examples are shown. They provide empirical evidence of the advantages and performance of the proposed methods. The simulation examples also consider various issues that are discussed briefly in the following and that require more explorations in future studies.

Despite the improved performances of the group MCP and the proposed adaptive group LASSO in nonparametric additive models, compared with the group LASSO, their applications in other nonparametric models are yet to be explored. In addition, the implementation of the adaptive group LASSO and its recursive application in generalized linear models is another open problem.

One interesting and at the same time difficult task for all single or grouped variable selection under the penalized regression framework is the selection of the penalty parameter(s). Even though the theoretical investigation and interpretation of the cross validation method is lacking, the information based criterion as alternatives may not be reliable, according to the simulation results in this dissertation. Two terms remain unclear under the penalized regression, especially under the penalized regression with nonconvex penalty functions: the definition of the degrees of freedom

and the accurate estimation of the model error. There have been some attempts at defining degrees of freedom under the penalized regression (Zou, Hastie and Tibshirani, 2007; Tibshirani and Taylor, 2012, among others). Intuitively, one may guess the degrees of freedom under the penalized regression framework to be the number of explanatory variables (or the number of components for nonparametric models) in the final model. This is the definition we used in our simulation studies and data applications and the one used in many simulation studies in literature where information based criterion is considered. For estimating the model error, there is still significant need for studies to show robust results that are applicable under more general contexts, which would eventually entitle statistical inference under the penalized regression framework.

Another issue that has been covered in the numerical studies in this dissertation but not on a theoretical level is the impact of correlation into the model when using penalized regression. Not surprisingly, the simulation studies have show effects of correlation to various extent.

APPENDIX A
PROOFS FOR CHAPTER 2

A.1 Proof of Theorem 2.1

The proof essentially follows the proof of Theorem 4.1 of Huang, Breheny and Ma (2012). Additional changes considering the approximation error of splines are made. The Karush-Kuhn-Tucker (KKT) condition states that it is necessary and sufficient for $\hat{\boldsymbol{\beta}}$ to satisfy the following equations,

$$\begin{cases} 2\boldsymbol{Z}_j^{'}(\boldsymbol{Y}-\boldsymbol{Z}\hat{\boldsymbol{\beta}})/n = \rho^{'}(\|\hat{\boldsymbol{\beta}}_j\|_2;\lambda,\gamma), & j\in A_1; \\ \|\boldsymbol{Z}_j^{'}(\boldsymbol{Y}-\boldsymbol{Z}\hat{\boldsymbol{\beta}})\|_2/n \le \lambda, & j\in A_0. \end{cases}$$

When $\|\hat{\boldsymbol{\beta}}_j^{o}\|_2 \ge \lambda\gamma$, $\rho^{'}(\|\hat{\boldsymbol{\beta}}_j^{o}\|_2;\lambda,\gamma)=0$ for all $j\in A_1$. Since $\lambda_{min}(\boldsymbol{\Sigma})>1/\gamma$, the objective function (2.4) becomes strictly convex. According to the KKT condition, $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}^{o}$ holds in the intersection of the event $\Omega_1(\lambda)=\{\|\boldsymbol{Z}_j^{'}(\boldsymbol{Y}-\boldsymbol{Z}\hat{\boldsymbol{\beta}}_j^{o})\|_2/n\le\lambda,\forall j\in A_0\}$ and the event $\Omega_2(\lambda)=\{\|\hat{\boldsymbol{\beta}}_j^{o}\|_2\ge\lambda\gamma,\forall j\in A_1\}$. So $P(\hat{\boldsymbol{\beta}}\neq\hat{\boldsymbol{\beta}}^{o})=P\left(\overline{\Omega_1(\lambda)\bigcup\Omega_2(\lambda)}\right)=P\left(\overline{\Omega_1(\lambda)}\bigcap\overline{\Omega_2(\lambda)}\right)\le P\left(\overline{\Omega_1(\lambda)}\right)+P\left(\overline{\Omega_2(\lambda)}\right)=(1-P(\Omega_1(\lambda)))+(1-P(\Omega_2(\lambda)))$.

First we bound $1-P(\Omega_1(\lambda))$. Define $\boldsymbol{\delta}=\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}-\boldsymbol{\epsilon}$ and $\hat{\boldsymbol{\beta}}_{A_1}^{o}=(\hat{\boldsymbol{\beta}}_j^{o},j\in A_1)$. $(\hat{\boldsymbol{\beta}}_{A_1}^{o})^{'}=(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Y}/n$, so $\boldsymbol{Z}\hat{\boldsymbol{\beta}}^{o}=\boldsymbol{Z}_{A_1}\hat{\boldsymbol{\beta}}_{A_1}^{o}$. Denote $\boldsymbol{H}=\boldsymbol{I}_n-\boldsymbol{Z}_{A_1}(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}$.

$$
\begin{aligned}
1 - P(\Omega_1(\lambda)) &= 1 - P(\|\boldsymbol{Z}_j^{'}(\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}^{o})\|_2/n \leq \lambda, \forall j \in A_0) \\
&= P(\|\boldsymbol{Z}_j^{'}(\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}^{o})\|_2/n > \lambda, \forall j \in A_0) \\
&\leq \sum_{j \in A_0} P(\|\boldsymbol{Z}_j^{'}(\boldsymbol{Y} - \boldsymbol{Z}_{A_1}\hat{\boldsymbol{\beta}}_{A_1}^{o})\|_2/n > \lambda) \\
&= \sum_{j \in A_0} P(\|\boldsymbol{Z}_j^{'}[\boldsymbol{I}_n - \boldsymbol{Z}_{A_1}(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}]\boldsymbol{Y}\|_2/n > \lambda) \\
&= \sum_{j \in A_0} P(\|\boldsymbol{Z}_j^{'}[\boldsymbol{I}_n - \boldsymbol{Z}_{A_1}(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}](\boldsymbol{Z}_{A_1}\boldsymbol{\beta}_{A_1} + \boldsymbol{\epsilon} + \boldsymbol{\delta})\|_2/n > \lambda) \\
&= \sum_{j \in A_0} P(\|\boldsymbol{Z}_j^{'}[\boldsymbol{I}_n - \boldsymbol{Z}_{A_1}(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}](\boldsymbol{\epsilon} + \boldsymbol{\delta})\|_2/n > \lambda) \\
&= \sum_{j \in A_0} P(\|\boldsymbol{Z}_j^{'}\boldsymbol{H}(\boldsymbol{\epsilon} + \boldsymbol{\delta})\|_2/n > \lambda) \\
&= \sum_{j \in A_0} P(\|\boldsymbol{Z}_j^{'}\boldsymbol{H}(\boldsymbol{\epsilon} + \boldsymbol{\delta})\|_2^2/n^2 > \lambda^2) \\
&\leq \sum_{j \in A_0} E(\|\boldsymbol{Z}_j^{'}\boldsymbol{H}(\boldsymbol{\epsilon} + \boldsymbol{\delta})\|_2^2)/(n^2\lambda^2) \\
&\leq \sum_{j \in A_0} \frac{E\|\boldsymbol{Z}_j^{'}\boldsymbol{H}\boldsymbol{\epsilon}\|_2^2}{n^2\lambda^2} + \frac{E\|\boldsymbol{Z}_j^{'}H\boldsymbol{\delta}\|_2^2}{n^2\lambda^2}. \qquad \text{(A.1)}
\end{aligned}
$$

$\|\boldsymbol{Z}_j^{'}\boldsymbol{H}\boldsymbol{\epsilon}\|_2^2/\sigma^2$ follows χ^2 distribution with degrees of freedom m_n, which implies

$$
\sum_{j \in A_0} E\|\boldsymbol{Z}_j^{'}\boldsymbol{H}\boldsymbol{\epsilon}\|_2^2/(n^2\lambda^2) = \sum_{j \in A_0} m_n\sigma^2/n^2\lambda^2 = (p-q)m_n\sigma^2/n^2\lambda^2 = O(n^{-(4d+1)/(2d+1)}). \qquad \text{(A.2)}
$$

According to the results in the proof of Lemma 6 in Huang, Horowitz and Wei (2010),

$\max_{j\in A_0} \|\boldsymbol{Z}_j^{'}\boldsymbol{H}\boldsymbol{\delta}\|_2 \leq O(1)n[O_p(m_n^{-1})q]^{1/2}m_n^{-d}$, so

$$\begin{aligned}\sum_{j\in A_0} E\|\boldsymbol{Z}_j^{'}\boldsymbol{H}\boldsymbol{\delta}\|_2^2/(n^2\lambda^2) &= (p-q)O(n^2)O_p(m_n^{-1})qm_n^{-2d}/(n^2\lambda^2)\\ &= O_p(n^{-1/(2d+1)})O(n^{-2d/(2d+1)})\\ &= O_p(n^{-1}). \qquad \text{(A.3)}\end{aligned}$$

By (A.2) and (A.3), it follows that (A.1) becomes $1-P(\Omega_1(\lambda)) = O(n^{-(4d+1)/(2d+1)}) + O_p(n^{-1})$.

Next we bound $1-P(\Omega_2(\lambda))$. By triangular inequality, $\|\hat{\boldsymbol{\beta}}_j^o\|_2 + \|\hat{\boldsymbol{\beta}}_j^o - \boldsymbol{\beta}_j^o\|_2 \geq \|\boldsymbol{\beta}_j^o\|_2$. So $\|\hat{\boldsymbol{\beta}}_j^o\|_2 \geq \|\boldsymbol{\beta}_j^o\|_2 - \|\hat{\boldsymbol{\beta}}_j^o - \boldsymbol{\beta}_j^o\|_2$ and $P(\|\hat{\boldsymbol{\beta}}_j^o\|_2 \geq \lambda\gamma) \geq P(\|\boldsymbol{\beta}_j^o\|_2 - \|\hat{\boldsymbol{\beta}}_j^o - \boldsymbol{\beta}_j^o\|_2 \geq \lambda\gamma)$. So we get

$$\begin{aligned}1-P(\Omega_2(\lambda)) &= 1-P(\|\hat{\boldsymbol{\beta}}_j^o\|_2 \geq \lambda\gamma, \forall j \in A_1)\\ &\leq 1-P(\|\boldsymbol{\beta}_j^o\|_2 - \|\hat{\boldsymbol{\beta}}_j^o - \boldsymbol{\beta}_j^o\|_2 \geq \lambda\gamma, \forall j \in A_1)\\ &= P(\|\hat{\boldsymbol{\beta}}_j^o - \boldsymbol{\beta}_j^o\|_2 > \|\boldsymbol{\beta}_j^o\|_2 - \lambda\gamma, \forall j \in A_1)\\ &\leq \sum_{j\in A_1} P(\|\hat{\boldsymbol{\beta}}_j^o - \boldsymbol{\beta}_j^o\|_2 > \boldsymbol{\beta}_*^o - \lambda\gamma)\\ &= \sum_{j\in A_1} P(\|\sqrt{n}(\hat{\boldsymbol{\beta}}_j^o - \boldsymbol{\beta}_j^o)\|_2^2 > n(\boldsymbol{\beta}_*^o - \lambda\gamma)^2)\\ &\leq \sum_{j\in A_1} E\|\sqrt{n}(\hat{\boldsymbol{\beta}}_j^o - \boldsymbol{\beta}_j^o)\|_2^2/(n(\boldsymbol{\beta}_*^o - \lambda\gamma)^2)^2. \qquad \text{(A.4)}\end{aligned}$$

Let T_j be a $m_n \times qm_n$ matrix with I_{m_n} in the jth block and 0 elsewhere. $\sqrt{n}(\hat{\boldsymbol{\beta}}_j^o - \boldsymbol{\beta}_j^o) =$

$T_j(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}(\boldsymbol{\epsilon}+\boldsymbol{\delta})/\sqrt{n}$, so

$$\begin{aligned}
\|\sqrt{n}(\hat{\boldsymbol{\beta}}_j^o - \boldsymbol{\beta}_j^o)\|_2^2 &= \|T_j(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}(\boldsymbol{\epsilon}+\boldsymbol{\delta})/\sqrt{n}\|_2^2 \\
&\leq \|T_j(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\epsilon}/\sqrt{n}\|_2^2 + \|T_j(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\delta}/\sqrt{n}\|_2^2 \\
&\leq \|T_j\|_2^2\|(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1/2}\|_2^2\|(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1/2}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\epsilon}/\sqrt{n}\|_2^2 + \\
&\quad \|T_j(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\delta}/\sqrt{n}\|_2^2 \\
&\leq \|(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1/2}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\epsilon}/\sqrt{n}\|_2^2/\lambda^* + \\
&\quad \|T_j(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\delta}/\sqrt{n}\|_2^2, \qquad \text{(A.5)}
\end{aligned}$$

where λ^* denotes the maximal eigenvalue of $(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1/2}$.

Combine (A.4) and (A.5), we get

$$\begin{aligned}
1 - P(\Omega_2(\lambda)) &\leq \sum_{j\in A_1} E\|(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1/2}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\epsilon}/\sqrt{n}\|_2^2/(n\lambda^*(\boldsymbol{\beta}_*^o - \lambda\gamma)^2) + \\
&\quad E\|T_j(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\delta}/\sqrt{n}\|_2^2/(n(\boldsymbol{\beta}_*^o - \lambda\gamma)^2) \\
&= \sum_{j\in A_1} E\|(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1/2}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\epsilon}/\sqrt{n}\|_2^2/(n\lambda^*(\boldsymbol{\beta}_*^o - \lambda\gamma)^2) + \\
&\quad E\|T_j(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\delta}\|_2^2/(\boldsymbol{\beta}_*^o - \lambda\gamma)^2. \qquad \text{(A.6)}
\end{aligned}$$

$\|(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1/2}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\epsilon}/\sqrt{n}\|_2^2/(n\sigma^2)$ follows χ^2 distribution with degrees of freedom m_n,

which implies

$$\sum_{j\in A_1} E\|(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1/2}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\epsilon}/\sqrt{n}\|_2^2/(n\lambda^*(\boldsymbol{\beta}_*^o-\lambda\gamma)^2) = qm_n\sigma^2/(n\lambda^*(\boldsymbol{\beta}_*^o-\lambda\gamma)^2)$$

$$= O(n^{-2d/(2d+1)}). \tag{A.7}$$

In the proof of Lemma 5 in Huang, Horowitz and Wei (2010),

$$\max_{j\in A_1}\|T_j\boldsymbol{Z}_j^{'}H\boldsymbol{\delta}\|_2 \le nO_p(1)O_p(m_n^{-1})[O_p(m_n)q]^{1/2}m_n^{-d},$$

so

$$\begin{aligned} E\|T_j(\boldsymbol{Z}_{A_1}^{'}\boldsymbol{Z}_{A_1})^{-1}\boldsymbol{Z}_{A_1}^{'}\boldsymbol{\delta}\|_2^2/(n(\boldsymbol{\beta}_*^o-\lambda\gamma)^2)^2 &\le O_p(1)O_p(m_n^{-2})O_p(m_n)qm_n^{-2d} \\ &= O_p(m_n^{-1})O(m_n^{-2d}) \\ &= O_p(n^{-1/(2d+1)})O(n^{-2d/(2d+1)}) \\ &= O_p(n^{-1}). \end{aligned} \tag{A.8}$$

As a result of (A.7) and (A.8), $1-P(\Omega_2(\lambda)) = O(n^{-2d/(2d+1)}) + O_p(n^{-(3+2d)/(2d+1)})$. Combining the upper bounds of $1-P(\Omega_2(\lambda))$ and $1-P(\Omega_2(\lambda))$,

$$P(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^o) = O(n^{-(4d+1)/(2d+1)}) + O(n^{-2d/(2d+1)}) + O_p(n^{-1}).$$

□

A.2 Proof of Lemma 2.2

For any A such that $A \supset A_1$ and $|A| - |A_1| = m$, we have

$$\begin{aligned}\|(\boldsymbol{P}_A - \boldsymbol{P}_{A_1})\boldsymbol{\epsilon}\|_2^2 &\le \|\boldsymbol{P}_A\boldsymbol{\epsilon}\|_2^2 + \|\boldsymbol{P}_{A_1}\boldsymbol{\epsilon}\|_2^2 \\ &\le \|\boldsymbol{P}_A\|_2^2\|\boldsymbol{\epsilon}\|_2^2 + \|\boldsymbol{P}_{A_1}\|_2^2\|\boldsymbol{\epsilon}\|_2^2 \\ &= 2\|\boldsymbol{\epsilon}\|_2^2.\end{aligned}$$

Similarly, $\|(\boldsymbol{P}_A - \boldsymbol{P}_{A_1})\boldsymbol{\delta}\|_2^2 \le 2\|\boldsymbol{\delta}\|_2^2$. So

$$\begin{aligned}&P\left(2\sqrt{c^* m_n}\max_{\substack{|A|-|A_1|=m\\ A\supset A_1}}\frac{\|(\boldsymbol{P}_A - \boldsymbol{P}_{A_1})\boldsymbol{Y}\|_2}{\sqrt{mn}} > \lambda\right) \\ &= P\left(2\sqrt{c^* m_n}\max_{\substack{|A|-|A_1|=m\\ A\supset A_1}}\frac{\|(\boldsymbol{P}_A - \boldsymbol{P}_{A_1})(\boldsymbol{Z}_{\boldsymbol{A_1}}\boldsymbol{\beta}_{\boldsymbol{A_1}} + \boldsymbol{\epsilon} + \boldsymbol{\delta})\|_2}{\sqrt{mn}} > \lambda\right) \\ &= P\left(4c^* m_n\max_{\substack{|A|-|A_1|=m\\ A\supset A_1}}\frac{\|(\boldsymbol{P}_A - \boldsymbol{P}_{A_1})(\boldsymbol{\epsilon} + \boldsymbol{\delta})\|_2^2}{mn} > \lambda^2\right) \\ &= P\left(\max_{\substack{|A|-|A_1|=m\\ A\supset A_1}}\|(\boldsymbol{P}_A - \boldsymbol{P}_{A_1})(\boldsymbol{\epsilon} + \boldsymbol{\delta})\|_2^2 > \frac{\lambda^2 mn}{4c^* m_n}\right) \\ &\le P\left(\max_{\substack{|A|-|A_1|=m\\ A\supset A_1}}\|(\boldsymbol{P}_A - \boldsymbol{P}_{A_1})\boldsymbol{\epsilon}\|_2^2 + \max_{\substack{|A|-|A_1|=m\\ A\supset A_1}}\|(\boldsymbol{P}_A - \boldsymbol{P}_{A_1})\boldsymbol{\delta}\|_2^2 > \frac{\lambda^2 mn}{4c^* m_n}\right) \\ &\le P\left(2\|\boldsymbol{\epsilon}\|_2^2 + 2\|\boldsymbol{\delta}\|_2^2 > \frac{\lambda^2 mn}{4c^* m_n}\right) \\ &= P\left(\|\boldsymbol{\epsilon}\|_2^2 + \|\boldsymbol{\delta}\|_2^2 > \frac{\lambda^2 mn}{8c^* m_n}\right) \\ &= P\left(\left\|\frac{\boldsymbol{\epsilon}}{\sigma}\right\|_2^2 > \frac{\lambda^2 mn}{8c^* m_n\sigma^2} - \frac{\|\boldsymbol{\delta}\|_2^2}{\sigma^2}\right). \qquad \text{(A.9)}\end{aligned}$$

Under the assumption that ϵ_i are independent and identically distributed as

$N(0, \sigma^2)$, $\|\boldsymbol{\epsilon}/\sigma\|_2^2$ follows χ^2 distribution with degrees of freedom n. According to Lemma A.1 in Huang, Breheny and Ma (2012), which is a restatement of results in Laurent and Massart (2000), for a random variable X that follows χ^2 distribution with degrees of freedom k, and $t > 1$,

$$P(X \geq kt) \leq exp(-k(\sqrt{2t-1}-1)^2/4). \tag{A.10}$$

Now for (A.10), use

$$t = \frac{\lambda^2 mn}{8c^* m_n \sigma^2} - \frac{\|\boldsymbol{\delta}\|_2^2}{\sigma^2}; \quad k = n.$$

Then the term $(\sqrt{2t-1}-1)^2$ becomes

$$\begin{aligned}
&\left[\sqrt{2\left(\frac{\lambda^2 mn}{8c^* m_n \sigma^2} - \frac{\|\boldsymbol{\delta}\|_2^2}{\sigma^2}\right) - 1} - 1\right]^2 \\
&\leq 2\left(\frac{\lambda^2 mn}{8c^* m_n \sigma^2} - \frac{\|\boldsymbol{\delta}\|_2^2}{\sigma^2}\right) \\
&\leq \frac{\lambda^2 mn}{4c^* m_n \sigma^2} \\
&= O(n^{2d/(2d+1)}).
\end{aligned}$$

So (A.9) becomes

$$P\left(\left\|\frac{\boldsymbol{\epsilon}}{\sigma}\right\|_2^2 > \frac{\lambda^2 mn}{8c^* m_n \sigma^2} - \frac{\|\boldsymbol{\delta}\|_2^2}{\sigma^2}\right) \leq exp(-nO(n^{2d/(2d+1)})).$$

The result follows. □

A.3 Proof of Theorem 2.4

The proof largely follows the proof of Theorem 1 of Huang Horowitz and Wei (2010). Additional changes considering the MCP penalty are made. According to the definition of $\hat{\boldsymbol{\beta}}$,

$$\begin{aligned} & \frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{Z}\hat{\boldsymbol{\beta}}\|_2^2+\lambda\sum_{j=1}^{p}\int_0^{\|\hat{\boldsymbol{\beta}}_j\|_2}\left(1-\frac{t}{\gamma\lambda}\right)_+dt \\ \leq\ & \frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}^o\|_2^2+\lambda\sum_{j=1}^{p}\int_0^{\|\boldsymbol{\beta}_j^o\|_2}\left(1-\frac{t}{\gamma\lambda}\right)_+dt. \end{aligned} \tag{A.11}$$

Define $A_2=\{j:\hat{\boldsymbol{\beta}}_j\neq 0 \text{ or } \boldsymbol{\beta}_j^o\neq 0\}$. Denote $\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}^o$ as $\boldsymbol{\eta}$. We can rewrite (A.11) as

$$\begin{aligned} & \frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{Z}_{A_2}\hat{\boldsymbol{\beta}}_{A_2}\|_2^2+\lambda\sum_{j\in A_2}\int_0^{\|\hat{\boldsymbol{\beta}}_j\|_2}\left(1-\frac{t}{\gamma\lambda}\right)_+dt \\ \leq\ & \frac{1}{2n}\|\boldsymbol{\eta}\|_2^2+\lambda\sum_{j\in A_2}\int_0^{\|\boldsymbol{\beta}_j^o\|_2}\left(1-\frac{t}{\gamma\lambda}\right)_+dt. \end{aligned} \tag{A.12}$$

The first term of the left hand side of (A.12)

$$\begin{aligned} \|\boldsymbol{Y}-\boldsymbol{Z}_{A_2}\hat{\boldsymbol{\beta}}_{A_2}\|_2^2 &= \|(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}^o)-(\boldsymbol{Z}_{A_2}\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{Z}\boldsymbol{\beta}^o)\|_2^2 \\ &= \|(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}^o)-(\boldsymbol{Z}_{A_2}\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{Z}_{A_2}\boldsymbol{\beta}^o_{A_2})\|_2^2 \\ &= \|\boldsymbol{\eta}-(\boldsymbol{Z}_{A_2}\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{Z}_{A_2}\boldsymbol{\beta}^o_{A_2})\|_2^2 \\ &\leq \|\boldsymbol{\eta}\|_2^2-2\boldsymbol{\eta}'\boldsymbol{Z}_{A_2}(\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{\beta}^o_{A_2})+ \\ &\quad \|(\boldsymbol{Z}_{A_2}\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{Z}_{A_2}\boldsymbol{\beta}^o_{A_2})\|_2^2. \end{aligned} \tag{A.13}$$

Plug (A.13) into (A.12) and rearrange (A.12), we get

$$\begin{aligned}
& \frac{1}{2n}\|\boldsymbol{Z}_{A_2}(\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{\beta}^o_{A_2})\|_2^2-\frac{1}{n}\boldsymbol{\eta}'\boldsymbol{Z}^o_{A_2}(\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{\beta}_{A_2}) \\
\leq\ & \lambda\sum_{j\in A_1}\left[\int_0^{\|\boldsymbol{\beta}^o_j\|_2}\left(1-\frac{t}{\gamma\lambda}\right)_+dt-\int_0^{\|\hat{\boldsymbol{\beta}}_j\|_2}\left(1-\frac{t}{\gamma\lambda}\right)_+dt\right] \\
=\ & \lambda\sum_{j\in A_1}\int_{\|\hat{\boldsymbol{\beta}}_j\|_2}^{\|\boldsymbol{\beta}^o_j\|_2}\left(1-\frac{t}{\gamma\lambda}\right)_+dt \\
\leq\ & \lambda\sum_{j\in A_1}\int_{\|\hat{\boldsymbol{\beta}}_j\|_2}^{\|\boldsymbol{\beta}^o_j\|_2}1dt \\
\leq\ & \lambda\left|\sum_{j\in A_1}(\|\hat{\boldsymbol{\beta}}_j\|_2-\|\boldsymbol{\beta}^o_j\|_2)\right| \\
\leq\ & \lambda\sqrt{|A_1|}\|\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{\beta}^o_{A_2}\|_2, \qquad \text{(A.14)}
\end{aligned}$$

where $|A_1|$ denotes the cardinality of A_1. Now the second term on the left hand side of (A.14),

$$\begin{aligned}
& \frac{1}{n}\boldsymbol{\eta}'\boldsymbol{Z}_{A_2}(\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{\beta}^o_{A_2}) \\
\leq\ & \frac{1}{n}|\boldsymbol{\eta}'\boldsymbol{Z}_{A_2}(\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{\beta}^o_{A_2})| \\
\leq\ & \frac{1}{n}\|\boldsymbol{Z}_{A_2}(\boldsymbol{Z}'_{A_2}\boldsymbol{Z}_{A_2})^{-1}\boldsymbol{Z}'_{A_2}\boldsymbol{\eta}\|_2\cdot\|\boldsymbol{Z}_{A_2}(\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{\beta}^o_{A_2})\|_2 \\
\leq\ & \frac{1}{2n}\left[2\|\boldsymbol{Z}_{A_2}(\boldsymbol{Z}'_{A_2}\boldsymbol{Z}_{A_2})^{-1}\boldsymbol{Z}'_{A_2}\boldsymbol{\eta}\|_2^2+\frac{1}{2}\|\boldsymbol{Z}_{A_2}(\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{\beta}^o_{A_2})\|_2^2\right]. \qquad \text{(A.15)}
\end{aligned}$$

The last step of (A.15) is derived by the Cauchy-Schwarz inequality. By (A.14) and (A.15),

$$\frac{1}{4n}\|\boldsymbol{Z}_{A_2}(\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{\beta}^o_{A_2})\|_2^2\leq\lambda\sqrt{|A_1|}\|\hat{\boldsymbol{\beta}}_{A_2}-\boldsymbol{\beta}^o_{A_2}\|_2+\frac{1}{n}\|\boldsymbol{Z}_{A_2}(\boldsymbol{Z}'_{A_2}\boldsymbol{Z}_{A_2})^{-1}\boldsymbol{Z}'_{A_2}\boldsymbol{\eta}\|_2^2. \quad \text{(A.16)}$$

According to Theorem 1 of Huang, Horowitz and Wei (2010), if we denote $\mathbf{\Sigma}_{A_2} = \boldsymbol{Z}'_{A_2}\boldsymbol{Z}_{A_2}/n$ and $\lambda_{min}(\mathbf{\Sigma}_{A_2})$ as its smallest eigenvalue,

$$\|\boldsymbol{Z}_{A_2}(\hat{\boldsymbol{\beta}}_{A_2} - \boldsymbol{\beta}^o_{A_2})\|_2^2 \geq n\lambda_{min}(\mathbf{\Sigma}_{A_2})\|\hat{\boldsymbol{\beta}}_{A_2} - \boldsymbol{\beta}^o_{A_2}\|_2^2. \tag{A.17}$$

So (A.16) becomes,

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_{A_2} - \boldsymbol{\beta}^o_{A_2}\|_2^2 - \frac{4\lambda\sqrt{|A_1|}}{\lambda_{min}(\mathbf{\Sigma}_{A_2})}\|\hat{\boldsymbol{\beta}}_{A_2} - \boldsymbol{\beta}^o_{A_2}\|_2 \\ -\frac{4}{n\lambda_{min}(\mathbf{\Sigma}_{A_2})}\|\boldsymbol{Z}_{A_2}(\boldsymbol{Z}'_{A_2}\boldsymbol{Z}_{A_2})^{-1}\boldsymbol{Z}'_{A_2}\boldsymbol{\eta}\|_2^2 < 0.\end{aligned} \tag{A.18}$$

Solve the quadratic inequality (A.18) with respect to $\|\hat{\boldsymbol{\beta}}_{A_2} - \boldsymbol{\beta}^o_{A_2}\|_2$, we get an upper bound

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_{A_2} - \boldsymbol{\beta}^o_{A_2}\|_2 &\leq \sqrt{\frac{4\lambda^2|A_1|}{\lambda^2_{min}(\mathbf{\Sigma}_{A_2})} + \frac{4\|\boldsymbol{Z}_{A_2}(\boldsymbol{Z}'_{A_2}\boldsymbol{Z}_{A_2})^{-1}\boldsymbol{Z}'_{A_2}\boldsymbol{\eta}\|_2^2}{n\lambda_{min}(\mathbf{\Sigma}_{A_2})}} + \frac{2\lambda\sqrt{|A_1|}}{\lambda_{min}(\mathbf{\Sigma}_{A_2})} \\ &\leq 2\sqrt{\frac{4\lambda^2|A_1|}{\lambda^2_{min}(\mathbf{\Sigma}_{A_2})} + \frac{4\|\boldsymbol{Z}_{A_2}(\boldsymbol{Z}'_{A_2}\boldsymbol{Z}_{A_2})^{-1}\boldsymbol{Z}'_{A_2}\boldsymbol{\eta}\|_2^2}{n\lambda_{min}(\mathbf{\Sigma}_{A_2})}}.\end{aligned}$$

So

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_{A_2} - \boldsymbol{\beta}^o_{A_2}\|_2^2 &\leq \frac{16\lambda^2|A_1|}{\lambda^2_{min}(\mathbf{\Sigma}_{A_2})} + \frac{16\|\boldsymbol{Z}_{A_2}(\boldsymbol{Z}'_{A_2}\boldsymbol{Z}_{A_2})^{-1}\boldsymbol{Z}'_{A_2}\boldsymbol{\eta}\|_2^2}{n\lambda_{min}(\mathbf{\Sigma}_{A_2})} \\ &= O_p\left(\frac{m_n^2 log(pm_n)}{n}\right) + O_p\left(\frac{m_n}{n}\right) + O\left(\frac{1}{mn^{2d-1}}\right) + \left(\frac{4m_n^2\lambda^2}{n^2}\right).\end{aligned}$$

The last step is from the proof of Theorem 1 in Huang, Horowitz and Wei (2010).

Now by property of splines in de Boor (2001),

$$\frac{c_1}{m_n}\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^o\|_2^2 \leq \|\hat{f}_j - f_{nj}\|_2^2 \leq \frac{c_2}{m_n}\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^o\|_2^2$$

for some constants c_1 and c_2. Thus $\|\hat{f}_j - f_{nj}\|_2^2 = O_p\left(\frac{m_n log(pm_n)}{n}\right) + O_p\left(\frac{1}{n}\right) + O\left(\frac{1}{mn^{2d}}\right) + O\left(\frac{4m_n\lambda^2}{n^2}\right)$. Combine this bound with the direct result of splines that $\|f_{nj} - f_j\|_2^2 = O(m_n^{-2d})$ and the triangular inequality $\|\hat{f}_j - f_j\|_2^2 \leq \|\hat{f}_j - f_{nj}\|_2^2 + \|f_{nj} - f_j\|_2^2$, we get $\|\hat{f}_j - f_j\|_2^2 = O_p\left(\frac{m_n log(pm_n)}{n}\right) + O_p\left(\frac{1}{n}\right) + O\left(\frac{1}{mn^{2d}}\right) + O\left(\frac{4m_n\lambda^2}{n^2}\right) + O(m_n^{-2d})$. Since $\frac{m_n log(pm_n)}{n} \to 0$, $\frac{4m_n\lambda^2}{n^2} \to 0$, when we use $m_n = O(n^{1/(2d+1)})$, the result follows. □

APPENDIX B
PROOFS FOR CHAPTER 3

B.1 Proof of Lemma 3.1

(i) The left hand side of (3.2)

$$\begin{aligned}\Delta(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) &= \langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*, \dot{\psi}(\boldsymbol{\beta}) - \dot{\psi}(\boldsymbol{\beta}^*) \rangle \\ &= \langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*, \mathbf{z} + \dot{l}(\hat{\boldsymbol{\beta}}) - \dot{\psi}(\boldsymbol{\beta}^*) \rangle \\ &= \langle \hat{\boldsymbol{\beta}}, \mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*) \rangle - \langle \hat{\boldsymbol{\beta}}, -\dot{l}(\hat{\boldsymbol{\beta}}) \rangle - \langle \boldsymbol{\beta}^*, \mathbf{z} + \dot{l}(\hat{\boldsymbol{\beta}}) - \dot{\psi}(\boldsymbol{\beta}^*) \rangle. \end{aligned} \tag{B.1}$$

The first term on the right hand side of (B.1)

$$\langle \hat{\boldsymbol{\beta}}, \mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*) \rangle = \sum_{i=1}^{p} \hat{\beta}_i (\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_i \leq \sum_{j=1}^{J} \|\hat{\boldsymbol{\beta}}_j\|_2 \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2 \leq \lambda \sum_{j=1}^{J} \hat{w}_j \|\hat{\boldsymbol{\beta}}_j\|_2. \tag{B.2}$$

The second term on the right hand side of (B.1)

$$\langle \hat{\boldsymbol{\beta}}, -\dot{l}(\hat{\boldsymbol{\beta}}) \rangle = \sum_{i=1}^{p} \hat{\beta}_i g_i = \sum_{i:\hat{\beta}_i \neq 0} \hat{\beta}_i \lambda \hat{w}_j(i) \frac{\hat{\beta}_i}{\|\hat{\boldsymbol{\beta}}_{j(i)}\|_2} = \lambda \sum_{j=1}^{J} \hat{w}_j \|\hat{\boldsymbol{\beta}}_j\|_2. \tag{B.3}$$

The last term on the right hand side of (B.1)

$$
\begin{aligned}
-\langle \boldsymbol{\beta}^*, \mathbf{z} + \dot{l}(\hat{\boldsymbol{\beta}}) - \dot{\psi}(\boldsymbol{\beta}^*)\rangle &= \langle \boldsymbol{\beta}^*, -(\mathbf{z} + \dot{\psi}(\boldsymbol{\beta}^*))\rangle + \langle \boldsymbol{\beta}^*, -\dot{l}(\hat{\boldsymbol{\beta}})\rangle \\
&\leq \sum_{j=1}^{J} \|\boldsymbol{\beta}_j^*\|_2 \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*)_j\|_2 + \sum_{j=1}^{J} \lambda \hat{w}_j \|\boldsymbol{\beta}_j^*\|_2 \\
&\leq \sum_{j=1}^{J} \lambda \hat{w}_j \|\boldsymbol{\beta}_j^*\|_2 + \sum_{j=1}^{J} \lambda \hat{w}_j \|\boldsymbol{\beta}_j^*\|_2 \\
&= 2\lambda \sum_{j=1}^{J} \lambda \hat{w}_j \|\boldsymbol{\beta}_j^*\|_2. \qquad \text{(B.4)}
\end{aligned}
$$

Combining (B.2), (B.3) and (B.4), we get the first inequality in (3.2). In the event Ω_0, the proof of the second inequality in (3.2) is trivial.

(ii) We partition $\Delta(\boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\beta}^*)$ into $\Delta(\boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\beta}^*)_{S_2^c}$ and $\Delta(\boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\beta}^*)_{S_2}$. So

$$
\begin{aligned}
\Delta(\boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\beta}^*)_{S_2^c} &= \left\langle \mathbf{h}_{S_2^c}, (\dot{\psi}(\boldsymbol{\beta}) - \dot{\psi}(\boldsymbol{\beta}^*))_{S_2^c} \right\rangle \\
&= \left\langle \mathbf{h}_{S_2^c}, (\mathbf{z} + \dot{l}(\hat{\boldsymbol{\beta}}) - \dot{\psi}(\boldsymbol{\beta}^*))_{S_2^c} \right\rangle \\
&= \left\langle \mathbf{h}_{S_2^c}, (\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_{S_2^c} \right\rangle - \left\langle \mathbf{h}_{S_2^c}, -\dot{l}(\hat{\boldsymbol{\beta}})_{S_2^c} \right\rangle \\
&= \sum_{i \in S_4^c} \hat{\beta}_i (\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_i - \sum_{j \in S_2^c} \lambda \hat{w}_j \|\hat{\boldsymbol{\beta}}_j\|_2 \\
&\leq \sum_{j \in S_2^c} \|\hat{\beta}_i\|_2 \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2 - \sum_{j \in S_2^c} \lambda \hat{w}_j \|\hat{\boldsymbol{\beta}}_j\|_2 \\
&= \sum_{j \in S_2^c} \hat{w}_j \|\hat{\beta}_i\|_2 \left\| \frac{(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j}{\hat{w}_j} \right\|_2 - \sum_{j \in S_2^c} \lambda \hat{w}_j \|\hat{\boldsymbol{\beta}}_j\|_2 \\
&= (z_1^* - \lambda) \sum_{j \in S_2^c} \lambda \hat{w}_j \|\hat{\boldsymbol{\beta}}_j\|_2. \qquad \text{(B.5)}
\end{aligned}
$$

The other part

$$\begin{aligned}
\Delta(\boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\beta}^*)_{S_2} &= \left\langle \mathbf{h}_{S_2}, (\dot{\psi}(\boldsymbol{\beta}) - \dot{\psi}(\boldsymbol{\beta}^*))_S \right\rangle \\
&= \left\langle \mathbf{h}_{S_2^c}, (\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*) - \mathbf{g})_{S_2^c} \right\rangle \\
&= \sum_{i \in S_4} h_i (\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*) - \mathbf{g})_i \\
&\leq \sum_{j \in S_2} \|\mathbf{h}_j\|_2 \|\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*) - \mathbf{g}\|_2 \\
&\leq \sum_{j \in S_2} \|\mathbf{h}_j\|_2 \|\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*)\|_2 + \|\mathbf{g}\|_2 \\
&\leq (\lambda \max_j \hat{w}_j + z_0^*) \sum_{j \in S_2} \|\mathbf{h}_j\|_2 \\
&\leq (\lambda \max_j w_j + z_0^*) \sum_{j \in S_2} \|\mathbf{h}_j\|_2. \qquad \text{(B.6)}
\end{aligned}$$

Rearranging the inequality after inserting (B.5) and (B.6) into $\Delta(\boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\beta}^*) = \Delta(\boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\beta}^*)_{S_2^c} + \Delta(\boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\beta}^*)_{S_2}$ completes the proof. □

B.2 Proof of Theorem 3.2

Define $f(t) = \Delta(\boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\beta}^*) = \frac{\partial}{\partial t} \left[\psi(\boldsymbol{\beta}^* + t\mathbf{h}) - t\langle \mathbf{h} \, , \, \dot{\psi}(\boldsymbol{\beta}) \rangle \right]$. $f(t)$ is increasing in t due to the convexity of $\psi(\boldsymbol{\beta})$, so for $0 \leq t \leq 1$,

$$f(t) \leq f(1) = \Delta(\boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\beta}^*) < (\lambda \max_j w_j + z_0^*) \sum_{j \in S_2} \|\mathbf{h}_j\|_2. \qquad \text{(B.7)}$$

For $\phi_0(t\mathbf{h}) \leq \eta^*$,

$$F(\xi, S; \phi_0, \phi) \leq \frac{\Delta(\boldsymbol{\beta}^* + t\mathbf{h}, \boldsymbol{\beta}^*) e^{\phi_0(t\mathbf{h})}}{\phi(t\mathbf{h}) \sum_{j \in S_2} \|t\mathbf{h}_j\|_2} = \frac{f(t) e^{\phi_0(t\mathbf{h})}}{\phi(t\mathbf{h}) \sum_{j \in S_2} \|\mathbf{h}_j\|_2}. \tag{B.8}$$

Rearrange (B.8) and according to (B.7) , in the event Ω_1, we get

$$\phi_0(t\mathbf{h}) e^{-\phi_0(t\mathbf{h})} \leq \frac{f(t)}{\sum_{j \in S_2} \|\mathbf{h}_j\|_2 F(\xi, S; \phi_0, \phi)} < \frac{\lambda \max_j w_j + z_0^*}{F(\xi, S; \phi_0, \phi)} \leq \eta e^{-\eta}. \tag{B.9}$$

Then by contradiction, we can easily prove $\phi_0(\mathbf{h}) \leq \eta \leq \eta^*$. This completes the proof. □

B.3 Proof of Theorem 3.3

Let $\mathbf{h} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, $w_j = \hat{w}_j$ and $S_0 = \{j : \|\hat{\boldsymbol{\beta}}_j\|_2 > \gamma_0 \lambda\} \cup S_3$. According to the definition of κ, for $j \notin S_0$

$$\kappa \geq \frac{\left| \dot{\rho}_\lambda(\|\tilde{\boldsymbol{\beta}}_j\|_2) - \dot{\rho}_\lambda(\|\mathbf{0}+\|_2) \right|}{\|\tilde{\boldsymbol{\beta}}_j\|_2} = \frac{\left| \dot{\rho}_\lambda(\|\tilde{\boldsymbol{\beta}}_j\|_2) - \lambda \right|}{\|\tilde{\boldsymbol{\beta}}_j\|_2}, \tag{B.10}$$

so $\dot{\rho}_\lambda(\|\tilde{\boldsymbol{\beta}}_j\|_2) \geq (1 - \kappa\gamma_0)/\lambda$ and the weight $w_j = \dot{\rho}_\lambda(\|\tilde{\boldsymbol{\beta}}_j\|_2)/\lambda \geq (1 - \kappa\gamma_0)$. As a result, $z_1^* = \max_j \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*)_j / \hat{w}_j\|_2 \leq \gamma_0/(1 - \kappa\gamma_0)$ and

$$\begin{aligned}
\frac{\lambda \max_j w_j + z_0^*}{\lambda - z_1^*} &\leq \frac{\lambda \max_j w_j + \lambda_0}{\lambda - \frac{\lambda_0}{1-\kappa\gamma_0}} \\
&= \frac{\frac{A\gamma_0}{1-\kappa\gamma_0} \max_j w_j + \lambda_0}{\frac{A\gamma_0}{1-\kappa\gamma_0} - \frac{\lambda_0}{1-\kappa\gamma_0}} \\
&= \frac{A \max_j w_j + 1 - \kappa\gamma_0}{A-1} \\
&\leq \frac{A + 1 - \kappa\gamma_0}{A-1} \\
&= \xi.
\end{aligned}$$

We have proved that one of the two conditions needed for applying theorem 3.1, we continue with proving the other condition:

$$\lambda \max_j w_j + z_0^* \leq \lambda + \lambda_0 = \lambda_0 \left(1 + \frac{A}{1-\kappa\gamma_0}\right) \leq F(\xi, S; \phi_0, \phi)\eta e^{-\eta}.$$

Now apply theorem 3.1, we get $\phi_0(\mathbf{h}) \leq \eta$.

By definition of $F_2(\xi, S; \phi_0)$, F_* and by Lemma 3.2, we have

$$e^{-\eta} F_* \sum_{j=1}^{J} \|\mathbf{h}_j\|_2 \sum_{j \in S_3} \|\mathbf{h}_j\|_2 \leq \Delta(\boldsymbol{\beta}^* + \mathbf{h}, \boldsymbol{\beta}^*) \leq (\lambda \max_j w_j + z_0^*) \sum_{j \in S_3} \|\mathbf{h}_j\|_2.$$

So

$$e^{-\eta} F_* \sum_{j=1}^{J} \|\mathbf{h}_j\|_2 \leq \lambda \max_j w_j + z_0^*. \tag{B.11}$$

For the first term on the right hand side of (B.11), because $\lambda \hat{w}_j = \dot{\rho}_\lambda(\|\tilde{\boldsymbol{\beta}}_j\|_2) \leq$

$\dot{\rho}_\lambda(\|\boldsymbol{\beta}_j^*\|_2) + \kappa\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2$, we get

$$\begin{aligned}\lambda \max_j w_j &\leq max_j \left[\left|\dot{\rho}_\lambda(\|\boldsymbol{\beta}_j^*\|_2)\right| + \kappa\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2\right] \\ &\leq max_j \left|\dot{\rho}_\lambda(\|\boldsymbol{\beta}_j^*\|_2)\right| + \kappa\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2. \end{aligned} \tag{B.12}$$

The second term on the right hand side of (B.11) is bounded by

$$\begin{aligned}\sum_{j\in S_3} \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2 &\leq \sum_{j\in S_0} \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2 + \sum_{j\in S_0 \supseteq S_3} \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2 \\ &\leq \sum_{j\in S_0} \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2 + \lambda_0 |S_0 \supseteq S_3|^{1/2} \\ &\leq \sum_{j\in S_0} \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2 + \lambda_0 \sqrt{\frac{\|\tilde{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^*\|_2^2}{\gamma_0^2 \lambda^2 l^*}} \\ &\leq \sum_{j\in S_0} \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2 + \frac{1 - \kappa\gamma_0}{\gamma_0 A}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2. \end{aligned} \tag{B.13}$$

Rearrangements after substituting (B.12) and (B.13) into (B.11) leave us

$$\sum_{j=1}^{J} \|\mathbf{h}_j\|_2 \leq \frac{e^\eta}{F_*}\left[\max_j \left|\dot{\rho}_\lambda(\|\boldsymbol{\beta}_j\|_2)\right| + \sum_{j\in S} \|(\mathbf{z} - \dot{\psi}(\boldsymbol{\beta}^*))_j\|_2 + \left(\kappa + \frac{1}{\lambda_0 A} - \frac{\kappa}{A}\right)\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2\right]. \tag{B.14}$$

Because $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \sum_{j=1}^{J} \|\mathbf{h}_j\|_2$, the proof is complete. □

B.4 Proof of Theorem 3.4

Define

$$R^{(k)} = \begin{cases} \frac{e^{\eta}\lambda[1+(1-\kappa\gamma_0)/A]}{F(\xi,S,\phi_0,\phi)}, & k=0, \\ (1-r_0^k)R^{(\infty)} + r_0^k R^{(0)}, & k=1,2,\ldots, \\ \frac{e^{\eta}\left[\max_j |\dot{\rho}_\lambda(\|\boldsymbol{\beta}_j\|_2)| + \sum_{j\in S}\|(\mathbf{z}-\dot{\psi}(\boldsymbol{\beta}^*))_j\|_2\right]}{(1-r_0)F_*}, & k=\infty. \end{cases}$$

Then to prove (3.12) is to prove

$$\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*\|_2 \le (1-r_0^k)R^{(\infty)} + r_0^k R^{(0)},\ k=0,1,\ldots. \tag{B.15}$$

$R^{(0)} \le \gamma_0\lambda\sqrt{l^*}$ by assumption. In the event $\{R^{(\infty)} \le \gamma_0\lambda\sqrt{l^*}\}$, $R^{(l)} \le \gamma_0\lambda\sqrt{l^*}$. Now we prove (B.15) by induction. When $l=0$, according to Theorem 3.2,

$$\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_2 \le \frac{e^{\eta}(\lambda+\lambda_0)}{F(\xi,S,\phi_0,\phi)}, \tag{B.16}$$

so the inequality (B.15) holds. If (B.15) holds for $l=k-1$, then when $l=k$,

$$
\begin{aligned}
\|\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^*\|_2 &\leq (1 - r_0)R^{(\infty)} + r_0 R^{(k-1)} \\
&\leq (1 - r_0)R^{(\infty)} + r_0 \left[(1 - r_0)R^{(\infty)} + r_0 R^{(k-2)}\right] \\
&\leq \dots \\
&= \left[(1 - r_0) + (1 - r_0)r_0 + (1 - r_0)r_0^2 + \dots\right] R^{(\infty)} + r_0^k R^{(0)} \\
&= (1 - r_0^k)R^{(\infty)} + r_0^k R^{(0)}.
\end{aligned}
$$

□

REFERENCES

[1] Aslibekyan, S., Brown, E. E., Reynolds, R. J., Redden, D. T., Morgan, S., Baggott, J. E., Sha, J., Moreland, L. W., O'Dell, J. R., Curtis, J. R., Mikuls, T. R., Bridges Jr, S. L. & Arnett, D. K. (2013). Genetic variants associated with methotrexate efficacy and toxicity in early rheumatoid arthritis: results from the treatment of early aggressive rheumatoid arthritis trial. *The Pharmacogenomics Journal*, **14**(1), 48-53.

[2] Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, **19**(6), 716-723.

[3] Bach, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, **9**, 1179-1225.

[4] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289-300.

[5] Bickel, P. J., Ritov, Y. A., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, **37**(4), 1705-1732.

[6] Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, **7**(3), 200-217.

[7] Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, **5**(1), 232.

[8] Breheny, P., & Huang, J. (2013). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*.

[9] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, **24**(6), 2350-2383.

[10] Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, **80**(391), 580-598.

[11] Bunea, F., Tsybakov, A., & Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, **1**, 169-194.

[12] Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**(3), 759-771.

[13] Chen, S. S., Donoho, D. L., & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, **20**(1), 33-61.

[14] de Boor, C. (2001). A Practical Guide to Splines. New York: Springer.

[15] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32**(2), 407-499.

[16] Fan, J., Feng, Y., & Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, **3**(2), 521.

[17] Fan, J., & Jiang, J. (2005). Nonparametric inferences for additive models. *Journal of the American Statistical Association*, **100**(471), 890-907.

[18] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**(456), 1348-1360.

[19] Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**(2), 109-135.

[20] Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, **1**(2), 302-332.

[21] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent.*Journal of Statistical Software*, **33**(1), 1-22.

[22] Greenshtein, E., & Ritov, Y. A. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, **10**(6), 971-988.

[23] Harrison Jr, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, **5**(1), 81-102.

[24] Hastie, T., & Tibshirani, R. Generalized Additive Models. 1990. CRC Press.

[25] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55-67.

[26] Huang, J., Breheny, P., & Ma, S. (2012). A selective review of group selection in high-dimensional models. *Journal of Statistical Science*, **27**(4), 481-499.

[27] Huang, J., Horowitz, J. L. & Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, **38**(4), 2282-2313.

[28] Huang, J., Ma, S., & Zhang, C. H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, **18**(4), 1603-1618.

[29] Huang, J., Ma, S., Zhang, C. H., & Zhou, Y. (2013). Semi-Penalized Inference with Direct False Discovery Rate Control in High-Dimensions. arXiv preprint arXiv:1311.7455.

[30] Huang, J., & Zhang, C. H. (2012). Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. *The Journal of Machine Learning Research*, **13**(1), 1839-1864.

[31] Kim, Y., Kim, J., & Kim, Y. (2006). Blockwise sparse regression. *Statistica Sinica*, **16**(2), 375-390.

[32] Laurent, B., & Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, **28**(5), 1302-1338.

[33] Meier, L., van de Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 53-71.

[34] Meier, L., van de Geer, S. & Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics*, **37**(6B), 3779-3821.

[35] Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, **34**(3), 1436-1462.

[36] Meinshausen, N., & Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, **37**(1), 246-270.

[37] Nielsen, F., & Nock, R. (2007). On the centroids of symmetrized bregman divergences. arXiv preprint arXiv:0711.3242.

[38] Opsomer, J. D., & Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association*, **93**(442), 605-619.

[39] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461-464.

[40] Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13**(2) 689-705.

[41] Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, **14**(2), 590-606.

[42] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

[43] Tibshirani, R., & Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, **40**(2), 1198-1232.

[44] Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, **109**(3), 475-494.

[45] Wang, H., & Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, **104**(486), 747-757.

[46] Wei, F., Huang, J., & Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, **21**(4), 1515-1540.

[47] Wu, T. T., & Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, **2**(1), 224-244.

[48] Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(1), 49-67.

[49] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**(2), 894-942.

[50] Zhang, C. H., & Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, **36**(4), 1567-1594.

[51] Zhang, H. H., & Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, **94**(3), 691-703.

[52] Zhang, T. (2009). Some sharp performance bounds for least squares regression with L1 regularization. *The Annals of Statistics*, **37**(5A), 2109-2144.

[53] Zhao, P., Rocha, G., & Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, **37**(6A), 3468-3497.

[54] Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research*, **7**, 2541-2563.

[55] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**(476), 1418-1429.

[56] Zou, H., Hastie, T., & Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, **35**(5), 2173-2192.