INVESTIGATION OF SMOOTH AND NON-SMOOTH PENALTIES FOR REGULARIZED MODEL SELECTION IN REGRESSION

by

Nam Hee Choi

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2009

Doctoral Committee:

Associate Professor Kerby A. Shedden, Co-Chair
Associate Professor Ji Zhu, Co-Chair
Professor Jeremy M. G. Taylor
Assistant Professor Stilian A. Stoev

ACKNOWLEDGEMENTS

I would like to thank all of the wonderful people who made this thesis possible. I am particularly grateful to my advisors, Dr. Kerby Shedden and Dr. Ji Zhu for their patience, guidance and support. I would like to show my gratitude to my committee members, Dr. Jeremy Taylor and Dr. Stilian Stoev, for their invaluable advices. I would also like to thank my fellow students who have made the last five years so enjoyable. Finally, I would like to thank my family: my husband and son, Jaeil and Benjamin Ahn, my parents, Seonok Choi and Misun Lee, and my parents-in-law, Hongmoon Ahn and Yekyung Park. This thesis would not have been possible without their unconditional love and support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

This thesis is focused on identifying the model structure using penalized regression methods and its application to designed experiments or bioinformatics. The primary topics include variable selection with heredity constraints and ranking variables according to their strength of association with a response. Below, I will briefly describe the major components of my thesis.

1.1 Variable Selection with Heredity Constraints

The variable selection problem plays an important role as the presence of a large number of candidate predictors occur in a wide variety of scientific fields. For example, in microarray analysis, the number of predictors to be analyzed is far higher than the sample size.

By selecting a subset of important variables, one wants to achieve accurate predictions and interpretable models. Traditional variable selection methods include forward/backward stepwise regression and all-subsets regression. A drawback of these traditional methods is that they are discrete and unstable processes; a small change in the data could lead to a completely different conclusion selecting a different subset of variables as addressed by [8].

The Lasso [56], on the other hand, by penalizing the L_1 norm of the coefficients,

shrinks some of the coefficient estimates to zeros by penalizing the L_1 norm of the coefficients, so it can estimate the coefficients and select variables at the same time. Due to the L_1 regularization, it introduces bias in the estimates but can reduce the variance of the estimates, so that the mean squared error can be reduced due to the bias-variance tradeoff. The Lasso is a more stable procedure than the traditional methods above.

In this thesis, a regression model that includes main terms and their interaction terms is considered. There can be some model structures in that setting; some sets of predictors may be assumed to be grouped (group structures), while some terms are supposed to be included for other terms to be included (order restrictions). We focus on order restrictions here.

One may want to include a higher order term in a model only when the corresponding lower order terms are also in the model. This is called marginality in linear models [43] and heredity principle in designed experiments [29]. Justifications of this heredity principle are presented in Chapter II. The Lasso and other traditional variable selection methods do not consider this type of order restriction; they treat all variables "flatly". A variable selection that incorporates the order restriction is proposed in Chapter II. It is also shown that the proposed method has theoretically "oracle" properties.

1.2 Penalized Regression Methods and Ranking Variables by Their Strength of Association with a Response

There has been extensive research on variable selection and prediction in regression. Variable selection focuses on differentiating non-zero effects from zero effects, while prediction focuses on accurately predicting a response in the new data. In this thesis, a different approach is considered, which focuses on ranking predictors according to

the size of their effects on a response. This approach may have implications in genetic association studies and other analyses involving regression methods with weak effects and collinear regressors. Especially, in the genetic mapping application, it would be useful to focus on ranking predictors because one would be interested in prioritizing the genetic variants that have the highest association with the trait of interest for further investigation.

One could use a univariate analysis to rank the genetic variants as a pre-screening tool, but the univariate approach does not consider the combined effects of more than one variant. To take into account the presence of other variants, multiple linear regression can be used. However, in genetic data, it is common to have highly correlated predictors, and multiple linear regression is known to perform unsatisfactorily in that situation.

Regularization is often used to improve the multiple linear regression when strong collinearity is present. With the perspective of ranking accuracy, three types of regularized regression methods were considered in this thesis: ridge regression, the Lasso, and the elastic net. They have been studied in various ways, but have not been rigourously studied for the purpose of ranking the effects. The ranking behavior of the regularization methods are analyzed in detail for two- or three-predictor cases. The three methods are applied to simulated data that mimic the correlation structures in SNP genotypes and then compared in terms of ranking performance.

CHAPTER II

Variable Selection with Strong Heredity Constraint and Its Oracle Property

In this chapter, a variable selection method based on the L_1 regularization that simultaneously fits a regression model and identifies important interaction terms is proposed. Unlike most of the existing variable selection methods, the proposed method automatically enforces the heredity constraint, i.e., an interaction term can be included in the model only if the corresponding main terms are also included in the model. Furthermore, we extend our method to generalized linear models, and show that it performs as well as if the true model were given in advance, i.e., the *oracle* property [21, 22]. Numerical results on both simulation data and real data indicate that the proposed method tends to select relevant variables and remove irrelevant variables more effectively and provide better prediction performance than previous work [56, 67, 69].

2.1 Introduction

Consider the usual regression situation: we have training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_i, y_i), \ldots, (\boldsymbol{x}_n, y_n)$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})$ are the predictors and y_i is the response. To model the response y in terms of the predictors $x_1, \ldots, x_p$, one may

consider the linear model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where ϵ is the error term. In many important practical problems, however, the main terms $x_1, \ldots, x_p$ alone may not be enough to capture the relationship between the response and the predictors, and higher order interactions are often of interest to scientific researchers. For example, many complex diseases, such as cancer, involve multiple genetic and environmental risk factors, and scientists are particularly interested in assessing gene-gene and gene-environment interactions.

In this chapter, a regression model with main terms and all possible two-way interaction terms is considered, i.e.,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \alpha_{12}(x_1 x_2) + \alpha_{13}(x_1 x_3) + \cdots + \alpha_{p-1,p}(x_{p-1} x_p) + \epsilon. \quad (2.1)$$

The goal here is to find out which terms, especially which interaction terms, have an important effect on the response. For example, $x_1, \ldots, x_p$ may represent different genetic factors, y may represent a certain phenotype, and we are interested in deciphering how these genetic factors "work together" to determine the phenotype. Later, we extend the setting to generalized linear models and develop an asymptotic theory there.

There are two important challenges in this problem: prediction accuracy and interpretation. We would like our model to accurately predict the future data. Prediction accuracy can often be improved by shrinking the regression coefficients. Shrinkage sacrifices unbiasedness to reduce the variance of the predicted value and hence may improve the overall prediction accuracy. Interpretability is often realized via variable selection. With a large number of variables (including both the main terms and the

interaction terms), possibly larger than the number of observations, we often would like to determine a smaller subset that exhibits the strongest effects.

Variable selection has been studied extensively in the literature, for example, see [7], [56], [21]. In particular, the Lasso [56] has gained much attention in recent years. The Lasso criterion penalizes the L_1-norm of the regression coefficients to achieve a sparse model:

$$\min_{\beta_j,\alpha_{jj'}} \sum_{i=1}^{n}\Big(y_i - \beta_0 - \sum_j \beta_j x_{ij} - \sum_{j<j'} \alpha_{jj'}(x_{ij}x_{ij'})\Big)^2 + \lambda\Big(\sum_j |\beta_j| + \sum_{j<j'} |\alpha_{jj'}|\Big). \tag{2.2}$$

The L_1-norm penalty can shrink some of the fitted coefficients to be exactly zero when making the tuning parameter sufficiently large. However, the Lasso and other methods mentioned above are for the case when the candidate variables can be treated individually or "flatly." When interaction terms exist, there is a natural hierarchy among the variables, i.e., an interaction term can be included in the model only if both of the corresponding main terms are also included in the model. This is referred to as the marginality in generalized linear models [43, 45] or the strong heredity in the analysis of designed experiments [29]. Justifications of effect heredity can be found in [11, 34]. Although it is possible that the true model contains only an interaction term but not the corresponding main terms, it is a relatively rare case. Moreover, a linear transformation of predictors would result in getting the main terms in the model. A generic variable selection method does not enforce the heredity constraint, that is, it may select an interaction term but not the corresponding main terms, and such models are difficult to interpret in practice.

In this chapter, we extend the Lasso method so that it simultaneously fits the regression model and identifies interaction terms obeying the strong heredity constraint. Furthermore, we show that when the regularization parameters are appropriately chosen, our new method has the oracle property [21] and [22], i.e., it performs

as well as if the correct underlying model were given in advance. Such theoretical property has not been previously studied for variable selection with heredity constraints.

[67] and [69] also address the variable selection problem with heredity/marginality constraints. [67] extends the LARS algorithm [18] and enforces the constraint that if an interaction term is to be selected, its "parents," i.e., the corresponding main terms, are either already in the model, or selected together with the interaction term. The criterion for selecting a variable or a set of variables (in the case of selecting an interaction term and its parents) is an "averaged correlation" between the residual vector and the variable or the linear space spanned by the set of variables. [69] suggests a so-called Composite Absolute Penalty (CAP) to enforce the heredity/marginality constraint. In particular, they propose:

$$\begin{aligned}\min_{\beta_j,\alpha_{jj'}} \quad & \sum_{i=1}^{n}\Big(y_i-\beta_0-\sum_j \beta_j x_{ij}-\sum_{j<j'}\alpha_{jj'}(x_{ij}x_{ij'})\Big)^2 \\ & +\lambda\cdot\Big(\max(|\beta_1|,|\alpha_{12}|,\ldots,|\alpha_{1p}|) \\ & +\max(|\beta_2|,|\alpha_{12}|,|\alpha_{23}|,\ldots,|\alpha_{2p}|) \\ & +\cdots \\ & +\max(|\beta_p|,|\alpha_{1p}|,\ldots,|\alpha_{p-1,p}|) \\ & +\sum_{j<j'}|\alpha_{jj'}|\Big).\end{aligned}$$

Note that each vector in the "max(·)" contains a main term and all its "descendants"; if the coefficient for an interaction term is nonzero, there is no increase in the penalty for letting the coefficients of the corresponding main terms move away from zero. Hence if an interaction term is selected, the corresponding main terms are also automatically selected.

However, there are some possible drawbacks with these two methods. For exam-

ple, [69] found in their simulation study that CAP tends to perform worse than the Lasso in terms of prediction accuracy when the interaction effects are relatively large compared to the main effects. The same problem may also occur for [67], because they select a set of variables based on an "average" criterion. If an interaction effect is large while the corresponding main effects are relatively small, the "average" criterion may fail to select the set (of the interaction term and the corresponding main terms) into the model. As we will see in the next sections, our new method does not suffer from this drawback. It regulates the main effects and the interaction effects separately, while still maintaining the heredity/marginality constraint. Numerical results indicate that our method performs well on a wide range of relative sizes for the main effects and the interaction effects.

The rest of this chapter is organized as follows. In Section 2.2, we introduce our new model and an algorithm to fit the model. Asymptotic properties are studied in Section 2.3, and numerical results are in Section 2.4 and 2.5. We conclude this chapter with Section 2.6.

2.2 Strong Heredity Interaction Model

In this section, we extend the Lasso method for selecting interaction terms while at the same time keeping the strong heredity constraint. We call our model the strong heredity interaction model (SHIM). After introducing the model in Section 2.2.1, we develop an algorithm to compute the SHIM estimate in Section 2.2.2. We then extend SHIM to generalized linear models in Section 2.2.3.

2.2.1 Model

We re-parameterize the coefficients for the interaction terms $\alpha_{jj'}$, $j < j'$, $j, j' = 1, \ldots, p$, as $\alpha_{jj'} = \gamma_{jj'}\beta_j\beta_{j'}$, and consider the following model:

(2.3)

$$g(\boldsymbol{x}) = \beta_0+\beta_1 x_1+\cdots+\beta_p x_p+\gamma_{12}\beta_1\beta_2(x_1x_2)+\gamma_{13}\beta_1\beta_3(x_1x_3)+\cdots+\gamma_{p-1,p}\beta_{p-1}\beta_p(x_{p-1}x_p).$$

Notice the difference in the coefficients of the interaction terms between (2.1) and (2.3). In (2.3), the coefficient for the interaction term $(x_jx_{j'})$ is expressed as the product of $\gamma_{jj'}$, β_j and $\beta_{j'}$, instead of a single parameter $\alpha_{jj'}$. By writing the coefficient as a product, the model itself enforces the heredity constraint. That is, whenever the coefficient for either x_j or $x_{j'}$, i.e., β_j or $\beta_{j'}$, is equal to zero, the coefficient for the interaction term $(x_jx_{j'})$ is automatically set to zero; vice versa, if the coefficient for $(x_jx_{j'})$ is not equal to zero, it implies that both β_j and $\beta_{j'}$ are not equal to zero.

For the purpose of variable selection, we consider the following penalized least squares criterion:

$$\min_{\beta_j,\gamma_{jj'}} \sum_{i=1}^{n}(y_i - g(\boldsymbol{x}_i))^2 + \lambda_\beta\Big(|\beta_1| + \cdots + |\beta_p|\Big) + \lambda_\gamma\Big(|\gamma_{12}| + \cdots + |\gamma_{p-1,p}|\Big), \tag{2.4}$$

where $g(\boldsymbol{x})$ is from (2.3), and the penalty is the L_1-norm of the parameters, as in the Lasso (2.2). There are two tuning parameters, λ_β and λ_γ. The first tuning parameter λ_β controls the estimates at the main effect level: if β_j is shrunken to zero, variable x_j and all its "descendants," i.e., the corresponding interaction terms that involve x_j are removed from the model. The second tuning parameter λ_γ controls the estimates at the interaction effect level: if β_j and $\beta_{j'}$ are not equal to zero but the corresponding interaction effect is not strong, $\gamma_{jj'}$ still has the possibility of being zero; so it has the flexibility of selecting only the main terms.

To further improve the criterion (2.4), we apply the adaptive idea which has been used extensively in the literature, including [7], [70], [60], [68], i.e., to penalize

different parameters differently. We consider

(2.5)
$$\min_{\beta_j,\gamma_{jj'}} \sum_{i=1}^{n}(y_i - g(\boldsymbol{x}_i))^2 + \lambda_\beta\Big(w_1^\beta|\beta_1| + \cdots + w_p^\beta|\beta_p|\Big) + \lambda_\gamma\Big(w_{12}^\gamma|\gamma_{12}| + \cdots + w_{p-1,p}^\gamma|\gamma_{p-1,p}|\Big),$$

where w_j^β and $w_{jj'}^\gamma$ are pre-specified weights. The intuition is that if the effect of a variable is strong, we would like the corresponding weight to be small, hence the corresponding parameter is lightly penalized. If the effect of a variable is not strong, we would like the corresponding weight to be large, hence the corresponding parameter is heavily penalized. How to pre-specify the weights w_j^β and $w_{jj'}^\gamma$ from the data is discussed below.

Computing Adaptive Weights

Regarding the adaptive weights w_j^β and $w_{jj'}^\gamma$ for the regression parameters in (2.5), we consider three possibilities:

1. Set all the weights equal to 1. We denote this as "plain."

2. Following [7] and [70], we can compute the weights using the ordinary least squares (OLS) estimates from the training observations:
$$w_j^\beta = \left|\frac{1}{\hat\beta_j^{OLS}}\right|,\ w_{jj'}^\gamma = \left|\frac{\hat\beta_j^{OLS}\cdot\hat\beta_{j'}^{OLS}}{\hat\alpha_{jj'}^{OLS}}\right|$$
where $\hat\beta_j^{OLS}$ and $\hat\alpha_{jj'}^{OLS}$ are the corresponding OLS estimates. We denote this as "Adaptive(OLS)."

3. When $n < p$, the OLS estimates are not available, we can compute the weights using the ridge regression estimates, i.e., replacing all the above OLS estimates with the ridge regression estimates, and we denote this as "Adaptive(Ridge)."

2.2.2 Algorithm

To estimate the parameters β_j and $\gamma_{jj'}$, we can use an iterative approach, i.e., we first fix β_j and estimate $\gamma_{jj'}$, then we fix $\gamma_{jj'}$ and estimate β_j, and we iterate between these steps until the SHIM criterion converges based on the relative difference of criterion values for the two estimates from two consecutive iterations. Since at each step, the value of the objective function (2.5) decreases, the solution is guaranteed to converge.

When β_j, $j = 1, \ldots, p$, are fixed, (2.5) becomes a Lasso problem, hence we can use either the Lars/Lasso algorithm [18] or a quadratic programming package to efficiently solve for $\gamma_{jj'}$, $j < j'$. When $\gamma_{jj'}$, $j < j'$, are fixed, we can sequentially solve for β_j: for each $j = 1, \ldots, p$, we fix $\gamma_{jj'}$, $j < j'$, and $\boldsymbol{\beta}_{[-j]} = (\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$, then (2.5) becomes a simple Lasso problem with only one parameter β_j, and we can solve it with a closed form formula. We note that the sequential strategy of fixing $(p-1)$ β_j's and solving for the other β_j is similar to the shooting algorithm in [26, 25].

In summary, the algorithm proceeds as follows:

1. (Standardization) Center $\boldsymbol{y}$. Center and normalize each term $\boldsymbol{x}_j$, $\boldsymbol{x}_j\boldsymbol{x}_{j'}$, $j < j'$, $j, j' = 1, \ldots, p$.

2. (Initialization) Initialize $\hat{\beta}_j^{(0)}$ and $\hat{\gamma}_{jj'}^{(0)}$, $j < j'$, $j, j' = 1, \ldots, p$, with some plausible values. For example, we can use the least square estimates or the simple regression estimates by regressing the response $\boldsymbol{y}$ on each of the terms. Let $m = 1$.

3. (Update $\hat{\gamma}_{jj'}$) Let

$$\begin{aligned} \tilde{y}_i &= y_i - \hat{\beta}_1^{(m-1)} x_{i1} - \cdots - \hat{\beta}_p^{(m-1)} x_{ip},\ i = 1, \ldots, n \\ \tilde{x}_{i,jj'} &= \hat{\beta}_j^{(m-1)} \hat{\beta}_{j'}^{(m-1)} (x_{ij} x_{ij'}),\ i = 1, \ldots, n; j < j', j, j' = 1, \ldots, p \end{aligned}$$

then

$$\hat{\gamma}_{jj'}^{(m)} = \arg\min_{\gamma_{jj'}} \sum_{i=1}^{n} \Big(\tilde{y}_i - \sum_{j<j'} \gamma_{jj'} \tilde{x}_{i,jj'} \Big)^2 + \lambda_\gamma \sum_{j<j'} w_{jj'}^{\gamma} |\gamma_{jj'}|.$$

4. (Update $\hat{\beta}_j$)

   - Let $\hat{\beta}_j^{(m)} = \hat{\beta}_j^{(m-1)}$, $j = 1, \ldots, p$.
   - For each j in $1, \ldots, p$, let

$$\begin{aligned} \tilde{y}_i &= y_i - \sum_{j' \neq j} \hat{\beta}_{j'}^{(m)} x_{ij'} - \sum_{j'<j'',\ j',j'' \neq j} \hat{\beta}_{j'}^{(m)} \hat{\beta}_{j''}^{(m)} (x_{ij'} x_{ij''}),\ i = 1, \ldots, n \\ \tilde{x}_i &= x_{ij} + \sum_{j'<j} \hat{\gamma}_{j'j}^{(m)} \hat{\beta}_{j'}^{(m)} (x_{ij'} x_{ij}) + \sum_{j'>j} \hat{\gamma}_{jj'}^{(m)} \hat{\beta}_{j'}^{(m)} (x_{ij} x_{ij'}),\ i = 1, \ldots, n \end{aligned}$$

   then

$$\hat{\beta}_j^{(m)} = \arg\min_{\beta_j} \sum_{i=1}^{n} (\tilde{y}_i - \beta_j \tilde{x}_i)^2 + \lambda_\beta w_j^{\beta} |\beta_j|.$$

5. Compute the relative difference between $Q_n\big(\hat{\boldsymbol{\theta}}^{(m-1)}\big)$ and $Q_n\big(\hat{\boldsymbol{\theta}}^{(m)}\big)$:

$$\Delta^{(m)} = \frac{\Big|\ Q_n\big(\hat{\boldsymbol{\theta}}^{(m-1)}\big) - Q_n\big(\hat{\boldsymbol{\theta}}^{(m)}\big)\ \Big|}{\Big|\ Q_n\big(\hat{\boldsymbol{\theta}}^{(m-1)}\big)\ \Big|},$$

   where

$$Q_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - g(\boldsymbol{x}_i))^2 + \lambda_\beta \big(w_1^{\beta} |\beta_1| + \cdots + w_p^{\beta} |\beta_p| \big) + \lambda_\gamma \big(w_{12}^{\gamma} |\gamma_{12}| + \cdots + w_{p-1,p}^{\gamma} |\gamma_{p-1,p}| \big),$$

   for $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \gamma_{12}, \ldots, \gamma_{p-1,p})$.

6. Stop the algorithm if $\Delta^{(m)}$ is small enough. Otherwise, let $m = m + 1$ and go back to step 2.

2.2.3 Extension to Generalized Linear Models

The SHIM method can be naturally extended to likelihood based generalized linear models. Assume that the data $\boldsymbol{V}_i = \{(\boldsymbol{x}_i, y_i)\}, i = 1, \ldots, n$ are collected

independently. Conditioning on $\boldsymbol{x}_i$, suppose Y_i has a density $f(g(\boldsymbol{x}_i),\ y_i)$, where g is a known link function with main terms and all possible interaction terms:

$$\begin{aligned} g(\boldsymbol{x}) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \alpha_{12}(x_1 x_2) + \alpha_{13}(x_1 x_3) + \cdots + \alpha_{p-1,p}(x_{p-1} x_p) \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \gamma_{12}\beta_1\beta_2(x_1 x_2) + \cdots + \gamma_{p-1,p}\beta_{p-1}\beta_p(x_{p-1} x_p). \end{aligned} \tag{2.6}$$

As before, for the purpose of variable selection, we consider the following penalized negative log-likelihood criterion:

$$\min_{\beta_j,\gamma_{jj'}} -\sum_{i=1}^{n} \ell(g(\boldsymbol{x}_i),\ y_i) + \lambda_\beta(w_1^\beta|\beta_1| + \cdots + w_p^\beta|\beta_p|) + \lambda_\gamma(w_{12}^\gamma|\gamma_{12}| + \cdots + w_{p-1,p}^\gamma|\gamma_{p-1,p}|), \tag{2.7}$$

where $\ell(\cdot,\cdot) = \log f(\cdot,\cdot)$ is the conditional log-likelihood of Y. Similar to what we suggested in Section 2.2.1, one can specify the weights w_j^β and $w_{jj'}^\gamma$ using un-penalized maximum likelihood estimates or L_2-penalized maximum likelihood estimates. Later in Section 2.3, we show that under certain regularity conditions, using those un-penalized maximum likelihood estimates for specifying the weights guarantees that SHIM possesses the asymptotic oracle property.

2.3 Asymptotic Oracle Property

In this section, we study the asymptotic behavior of SHIM based on the generalized linear model setting introduced in Section 2.2.3. In Section 2.3.1, we consider the asymptotic properties of SHIM estimates when the sample size n approaches to infinity. Furthermore, in Section 2.3.2, we consider the asymptotic properties of SHIM estimates when the number of covariates p_n also increases as the sample size n increases.

2.3.1 Asymptotic Oracle Property When $n \to \infty$

We show that when the number of predictors is fixed and the sample size approaches to infinity, SHIM possesses the oracle property under certain regularity conditions, that is, it performs as well as if the true model were known in advance [21].

Problem Setup

Let β_j^* and $\alpha_{jj'}^*$ denote the underlying true parameters. We further assume that the true model obeys the strong heredity constraint: $\alpha_{jj'}^* = 0$ if $\beta_j^* = 0$ or $\beta_{j'}^* = 0$. Let $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*\mathsf{T}}, \boldsymbol{\gamma}^{*\mathsf{T}})^{\mathsf{T}}$ where

$$\gamma_{jj'}^* = \begin{cases} \frac{\alpha_{jj'}^*}{\beta_j^* \beta_{j'}^*} & \text{if } \beta_j^* \neq 0 \text{ and } \beta_{j'}^* \neq 0 \\ 0 & \text{otherwise} \end{cases}.$$

We consider the SHIM estimates $\hat{\boldsymbol{\theta}}_n$:

(2.8)

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} - \sum_{i=1}^{n} \ell(g(\boldsymbol{x}_i),\ y_i) + n \sum_{j=1}^{p} \lambda_j^{\beta} |\beta_j| + n \sum_{k<k'} \lambda_{kk'}^{\gamma} |\gamma_{kk'}|,$$

where g is defined in (2.6). Note that $Q_n(\boldsymbol{\theta})$ in (2.8) is equivalent to the criterion in (2.7) by letting $\lambda_j^{\beta} = \frac{1}{n}\lambda_{\beta} w_j^{\beta}$ and $\lambda_{kk'}^{\gamma} = \frac{1}{n}\lambda_{\gamma} w_{kk'}^{\gamma}$. Furthermore, we define

$$\mathcal{A}_1 = \{j :\ \beta_j^* \neq 0\},\ \mathcal{A}_2 = \{(k,k') :\ \gamma_{kk'}^* \neq 0\},\ \mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2,$$

i.e., $\mathcal{A}_1$ contains the indices for main terms whose true coefficients are nonzero, and $\mathcal{A}_2$ contains the indices for interaction terms whose true coefficients are nonzero. Let

$$\begin{aligned} a_n &= \max\{\lambda_j^{\beta},\ \lambda_{kk'}^{\gamma} :\ j \in \mathcal{A}_1,\ (k,k') \in \mathcal{A}_2\} \\ b_n &= \min\{\lambda_j^{\beta},\ \lambda_{kk'}^{\gamma} :\ j \in \mathcal{A}_1^c,\ (k,k') \in \mathcal{A}_2^c,\ k,k' \in \mathcal{A}_1\} \end{aligned}$$

Notice that to compute b_n, we do not consider every case of $\gamma^*_{kk'} = 0$, i.e., $(k, k') \in \mathcal{A}_2^c$. Instead, we only consider the cases where $\gamma^*_{kk'}$ is zero and the two corresponding β^*_k and $\beta^*_{k'}$ are nonzero, i.e., $(k, k') \in \mathcal{A}_2^c$ and $k, k' \in \mathcal{A}_1$.

Oracle Property of SHIM

The asymptotic properties of SHIM when the sample size increases are described in the following lemma and theorems. The regularity conditions (C1)-(C3) and the proofs are given in Appendix A.

Lemma II.1. *Assume that $a_n = o(1)$ as $n \to \infty$. Then under the regularity conditions (C1)-(C3), there exists a local minimizer $\hat{\boldsymbol{\theta}}_n$ of $Q_n(\boldsymbol{\theta})$ such that $||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*|| = O_p(n^{-1/2} + a_n)$.*

Lemma II.1 implies that if the tuning parameters λ_j^β and $\lambda_{kk'}^\gamma$ associated with the nonzero coefficients converge to 0 at a rate faster than $n^{-1/2}$, then there exists a local minimizer of $Q_n(\boldsymbol{\theta})$, which is $\sqrt{n}$-consistent.

Theorem II.2. *(Sparsity) Assume that $\sqrt{n}b_n \to \infty$ and the local minimizer $\hat{\boldsymbol{\theta}}_n$ given in Lemma II.1 satisfies $||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*|| = O_p(n^{-1/2})$. Then under the regularity conditions (C1)-(C3),*

$$P(\hat{\boldsymbol{\beta}}_{\mathcal{A}_1^c} = 0) \to 1 \quad \textit{and} \quad P(\hat{\boldsymbol{\gamma}}_{\mathcal{A}_2^c} = 0) \to 1.$$

Theorem II.2 shows that SHIM can consistently remove the noise terms with probability tending to 1. Specifically, when the tuning parameters for the nonzero coefficients converge to 0 faster than $n^{-1/2}$ and those for zero coefficients are big enough so that $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$, then Lemma II.1 and Theorem II.2 imply that the $\sqrt{n}$-consistent estimator $\hat{\boldsymbol{\theta}}_n$ satisfies $P(\hat{\boldsymbol{\theta}}_{\mathcal{A}^c} = 0) \to 1$.

Theorem II.3. *(Asymptotic normality) Assume that $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$.*

Then under the regularity conditions (C1)-(C3), the component $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$ *of the local minimizer* $\hat{\boldsymbol{\theta}}_n$ *given in Lemma II.1 satisfies*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) \to_d N(0, \boldsymbol{I}^{-1}(\boldsymbol{\theta}^*_{\mathcal{A}})),$$

where $\boldsymbol{I}(\boldsymbol{\theta}^*_{\mathcal{A}})$ *is the Fisher information matrix of* $\boldsymbol{\theta}_{\mathcal{A}}$ *at* $\boldsymbol{\theta}_{\mathcal{A}} = \boldsymbol{\theta}^*_{\mathcal{A}}$ *assuming that* $\boldsymbol{\theta}^*_{\mathcal{A}^c} = 0$ *is known in advance.*

In Theorem II.3, we find that the SHIM estimates for nonzero coefficients in the true model have the same asymptotic distribution as they would have if the zero coefficients were known in advance. Therefore, based on Theorem II.2 and Theorem II.3, we can conclude that asymptotically SHIM performs as well as if the true underlying model were given in advance, i.e., it has the oracle property (Fan and Li, 2001), when the tuning parameters satisfy the conditions $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$.

Now the remaining question is how we specify the adaptive weights so that the conditions $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$ are satisfied. It turns out that the Adaptive(MLE) weights introduced in Section 2.2.1 satisfy those conditions. Following the idea in [61], let

$$\begin{aligned}
\lambda^{\beta}_j &= \frac{\log(n)}{n}\lambda_\beta w^{\beta}_j = \frac{\log(n)}{n}\lambda_\beta \left|\frac{1}{\hat{\beta}^{MLE}_j}\right|, \\
\lambda^{\gamma}_{kk'} &= \frac{\log(n)}{n}\lambda_\gamma w^{\gamma}_{kk'} = \frac{\log(n)}{n}\lambda_\gamma \left|\frac{\hat{\beta}^{MLE}_k \cdot \hat{\beta}^{MLE}_{k'}}{\hat{\alpha}^{MLE}_{kk'}}\right|.
\end{aligned}$$

Using the fact that $\hat{\boldsymbol{\beta}}^{MLE}$ and $\hat{\boldsymbol{\alpha}}^{MLE}$ are $\sqrt{n}$-consistent estimates of $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$, it can be easily shown that the tuning parameters λ^{β}_j and $\lambda^{\gamma}_{kk'}$ defined above satisfy the conditions for the oracle property. Therefore, we can conclude that by tuning the two regularization parameters λ_β and λ_γ and using the pre-specified weights Adaptive(MLE), SHIM asymptotically possesses the oracle property.

2.3.2 Asymptotic Oracle Property When $p_n \to \infty$ as $n \to \infty$

In this section, we consider the asymptotic behavior of SHIM when the number of predictors p_n is allowed to approach infinity as well as the sample size n. Similar as in [22], we show that under certain regularity conditions, SHIM still possesses the oracle property.

We first re-define some notations because now the number of predictors p_n changes with the sample size n. We denote the total number of parameters $q_n = (p_n+1)p_n/2$. We add a subscript n to $\boldsymbol{V}$, $f(\cdot,\cdot)$ and $\boldsymbol{\theta}$ to denote that these quantities now change with n. Similarly for $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}$ which are defined in Section 2.3.1, and we let $s_n = |\mathcal{A}_n|$.

Oracle Property of SHIM

The asymptotic properties of SHIM when the number of predictors increases as well as the sample size are described in the following lemma and theorems. The regularity conditions (C4-C6) and the proofs are given in Appendix A.

Lemma II.4. *Assume that the density $f_n(\boldsymbol{V}_n, \boldsymbol{\theta}_n^*)$ satisfies the regularity conditions (C4-C6). If $\sqrt{n}a_n \to 0$ and $q_n^5/n \to 0$ as $n \to \infty$, then there exists a local minimizer $\hat{\boldsymbol{\theta}}_n$ of $Q_n(\boldsymbol{\theta}_n)$ such that*

$$\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*\| = O_p(\sqrt{q_n}(n^{-1/2} + a_n)).$$

Theorem II.5. *Suppose that the density $f_n(\boldsymbol{V}_n, \boldsymbol{\theta}_n^*)$ satisfies the regularity conditions (C4-C6). If $\sqrt{nq_n}a_n \to 0$, $\sqrt{n/q_n}b_n \to \infty$, and $q_n^5/n \to 0$ as $n \to \infty$, then with probability tending to 1, the $\sqrt{n/q_n}$-consistent local minimizer $\hat{\boldsymbol{\theta}}_n$ in Lemma II.4 satisfies the following:*

(a) (Sparsity) $\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n^c} = \mathbf{0}$;

(b) (Asymptotic normality)

$$\sqrt{n}\boldsymbol{A}_n\boldsymbol{I}_n^{1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \to_d N(\mathbf{0}, \boldsymbol{G}),$$

where $\boldsymbol{A}_n$ is an arbitrary $m \times s_n$ matrix with a finite m such that $\boldsymbol{A}_n\boldsymbol{A}_n^{\intercal} \to \boldsymbol{G}$ and $\boldsymbol{G}$ is a $m \times m$ nonnegative symmetric matrix and $\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})$ is the Fisher information matrix of $\boldsymbol{\theta}_{n\mathcal{A}_n}$ at $\boldsymbol{\theta}_{n\mathcal{A}_n} = \boldsymbol{\theta}^*_{n\mathcal{A}_n}$.

Note that because the dimension of $\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}$ approaches to infinity as the sample size n grows, for asymptotic normality of SHIM estimates, we consider an arbitrary linear combination $\boldsymbol{A}_n\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}$, where $\boldsymbol{A}_n$ is an arbitrary $m \times s_n$ matrix with a finite m.

Similar as in Section 2.3.1, now the remaining question is whether the Adaptive(MLE) weights introduced in Section 2.2.1 satisfy the conditions for the oracle property. Let

$$\begin{aligned}
\lambda^{\beta}_{nj} &= \frac{\log(n)\ q_n}{n}\lambda_\beta w^{\beta}_j = \frac{\log(n)\ q_n}{n}\lambda_\beta \left|\frac{1}{\hat{\beta}^{MLE}_j}\right|, \\
\lambda^{\gamma}_{n,kk'} &= \frac{\log(n)\ q_n}{n}\lambda_\gamma w^{\gamma}_{kk'} = \frac{\log(n)\ q_n}{n}\lambda_\gamma \left|\frac{\hat{\beta}^{MLE}_k \cdot \hat{\beta}^{MLE}_{k'}}{\hat{\alpha}^{MLE}_{kk'}}\right|.
\end{aligned}$$

Using the fact that $\hat{\boldsymbol{\beta}}^{MLE}$ and $\hat{\boldsymbol{\alpha}}^{MLE}$ are $\sqrt{n/q_n}$-consistent estimates of $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$ and assuming $q_n^4/n \to 0$, it can be easily shown that the tuning parameters λ^{β}_{nj} and $\lambda^{\gamma}_{n,kk'}$ defined above satisfy the conditions for the oracle property: $\sqrt{nq_n}a_n \to 0$ and $\sqrt{n/q_n}b_n \to \infty$. Therefore, we can conclude that by tuning the two regularization parameters λ_β and λ_γ and using the pre-specified weights Adaptive(MLE), SHIM asymptotically possesses the oracle property.

2.4 Simulation Study

2.4.1 Regression models with random normal covariates

In this section, we use simulation data to demonstrate the efficacy of SHIM, and compare the results with those of the Lasso, a method that does not guarantee the

heredity constraint. Furthermore, we compare the performance of SHIM with two other methods, [69] and [67], which also address the variable selection problem with heredity constraint.

Table 2.1: Simulation study: coefficients of the true models

	x_1	x_2	x_3	x_4	x_1x_2	x_1x_3	x_1x_4	x_2x_3	x_2x_4	x_3x_4
Case 1	7	2	1	1	0	0	0	0	0	0
Case 2	7	2	1	1	1.0	0	0	0.5	0.4	0.1
Case 3	7	2	1	1	7	7	7	2	2	1
Case 4	7	2	1	1	14	14	14	4	4	2
Case 5	0	0	0	0	7	7	7	2	2	1

We mimicked and extended the simulations in [69]. There are $p = 10$ predictors with only the first 4 affecting the response. The total number of candidate terms (including all possible two-way interaction terms) is $p+p(p-1)/2 = 55$. Each of the 10 predictors is normally distributed with mean zero and variance one. The 10 predictors are generated either independently or with correlation $\mathrm{Corr}(X_j, X_{j'}) = 0.5^{|j-j'|}$. With each of independent predictors and correlated predictors, we considered five different cases with coefficients shown in Table 2.1. The signal to noise ratio (SNR) was set to 4.0 in every case.

Case 1 is a model with no interaction effect; Case 2 is a model with interaction effects of moderate size; Case 3 represents a model with interaction effects of large size; and Case 4 is a model where the size of interaction effects is larger than that of the main effects. Case 5 is a model that does not even obey the heredity constraint.

We generated $n = 200$ training observations from each of the above models and $10,000$ test observations. To select the tuning parameters λ's for SHIM, we considered three criteria: GCV, BIC, and the validation error on a validation set with

$m = 200$ observations.

$$\begin{aligned} \text{GCV}(\lambda_1, \lambda_2) &= \frac{\hat{\sigma}^2_{(\lambda_1,\lambda_2)}}{(1 - df_{(\lambda_1,\lambda_2)}/n)^2} \\ \text{BIC}(\lambda_1, \lambda_2) &= \log \hat{\sigma}^2_{(\lambda_1,\lambda_2)} + df_{(\lambda_1,\lambda_2)} \log n/n \\ \text{Validation Error}(\lambda_1, \lambda_2) &= \frac{1}{m} \sum_{i=1}^{m} (y_i^{\text{val}} - \hat{f}_{(\lambda_1,\lambda_2)}(\boldsymbol{x}_i^{\text{val}}))^2 \end{aligned}$$

where $\hat{\sigma}^2_{(\lambda_1,\lambda_2)} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}_{(\lambda_1,\lambda_2)}(\boldsymbol{x}_i))^2$ and $df_{(\lambda_1,\lambda_2)}$ = the degree of freedom for (λ_1, λ_2). We simulated 100 replicates and in each replicate, we considered the three pairs of the (λ_1, λ_2)'s that minimize GCV, BIC, and a validation error respectively.

In the following sections, we compare our method SHIM with other methods in terms of the prediction accuracy and the variable selection performance.

Prediction Performance

We first compare the prediction accuracy of SHIM with those of other methods: oracle, OLS, Lasso, CAP, and CARDS. "oracle" refers to the OLS applied only to the relevant terms, which serves as an optimal bench mark. CAP and CARDS refer to the two previous works, [69] and [67] respectively, which address the variable selection problem with heredity constraint. The latter extends the LARS algorithm [18], and the former suggests a Composite Absolute Penalty (CAP) in order to enforce the heredity constraint.

We compute the mean squared error (MSE) with a test set with $10,000$ observations for measuring the prediction accuracy. Figure 2.1 and Figure 2.2 show the boxplots of the 100 MSEs from 100 replicates when the tuning parameters are chosen based on validation error and GCV, in independent cases and correlated cases, respectively. For the Lasso, SHIM, CAP and CARDS, we choose to plot the results based on validation error because we find the prediction accuracy is the best when validation error is used to select tuning parameters among the three criteria. We

also plot the results based on GCV for a comparison because it performs second-best and a validation set is not always available in real data sets. In Figure 2.1, ALASSO-1 and ASHIM-1 refer to the adaptive Lasso and the adaptive SHIM with the weights based on OLS estimates; in Figure 2.2, ALASSO-2 and ASHIM-2 refer to the adaptive Lasso and the adaptive SHIM with the weights based on ridge regression estimates. The results based on OLS estimates are not shown in Figure 2.2, because they are often not good estimates in correlated cases.

Both figures show both Lasso and SHIM perform much better than OLS; this illustrates that some regularization or shrinkage is crucial for prediction accuracy. Furthermore, SHIM seems to perform consistently better than the Lasso. We can also see that the adaptive weights often help us improve the prediction accuracy for both SHIM and the Lasso, when the validation error is used.

Comparing SHIM (non-adaptive version) with the two other previous works, CAP [69] and CARDS [67], we can see that the prediction accuracy of SHIM is consistently better than CARDS and CAP in both independent and correlated cases, especially when the effect of interaction terms increases.

Variable Selection Performance

We also compare the variable selection performance of SHIM with those of the other methods.

We define "underfitted", "correctly-fitted", and "overfitted" models following [62]. Suppose that we have q candidate terms and there are only $q_0 \leq q$ number of relevant terms in a true model. And let $I_F = \{1, 2, \ldots, q\}$ denote the index set of the full model; $I_T = \{j_1, j_2, \ldots, j_{q_0}\}$ denote the index set of the true model; I denote the index set of the selected model based on any method. Then we define a model as a underfitted model when $I_T \nsubseteq I$, an overfitted model when $I_T \subsetneq I$, and a correctly-

fitted model when $I = I_T$.

In Table 2.2, we present those results for SHIM and the other methods, Lasso, CAP [69] and CARDS [67]. These results are based on the tuning parameters selected by minimizing BIC. We choose to show the results based on BIC because they have the best variable selection performance among the three criteria in our simulation. [62], [59] and [68] show similar results in their papers.

Table 2.2 shows that SHIM and adaptive SHIM tend to select the correct model more often than other methods, since they have the highest number of correctly-fitted models among all methods in most cases. When this is not the case: None of the methods can find the exactly correct models (Case 2), because they would easily miss some of the weak interaction effects, or the methods that enforce the heredity constraint are not supposed to perform well because the true model does not satisfy the heredity constraint (Case 5). In addition, all methods perform similarly, when there is no interaction effect in the true model (Case 1).

We can confirm our conclusion in Figure 2.3 and 2.4. In the two figures, we plot (1-specificity, sensitivity) of the selected models based on BIC. In each replicate, the sensitivity is defined as the proportion of the number of selected relevant terms to the number of relevant terms and specificity is defined as the proportion of the number of excluded irrelevant terms to the number of irrelevant terms.

Each dot in the figures corresponds to each pair of (1-specificity, sensitivity) from one replicate so we should have 100 dots in each plot. If the selected models contain relevant terms and remove irrelevant terms effectively, we would expect the dots to be located at the upper left corner of the plots, as it would mean both sensitivity and specificity are close to 1 simultaneously. We can see that four methods work similarly in Case 1 and 2 where the effects of interaction terms are small. For Case

Table 2.2: Simulation results: variable selection based on BIC. "Underfitted", "Correctly-fitted" and "Overfitted" represent the numbers of replicates that are underfitted, correctly-fitted and overfitted among the 100 replicates. "ALASSO" and "ASHIM" refer to the adaptive Lasso and the adaptive SHIM with the OLS weights for independent cases and the Ridge weights for correlated cases; "CAP" refers to [69] and "CARDS" refers to [67].

		LASSO	ALASSO	SHIM	ASHIM	CAP	CARDS
Independent Cases							
Case 1	Underfitted	21	30	3	14	23	14
	Correctly-fitted	23	27	17	49	42	44
	Overfitted	56	43	80	37	35	42
Case 2	Underfitted	100	100	98	98	97	99
	Correctly-fitted	0	0	0	0	1	0
	Overfitted	0	0	2	2	2	1
Case 3	Underfitted	93	97	17	20	41	59
	Correctly-fitted	0	0	78	78	17	13
	Overfitted	7	3	5	2	42	28
Case 4	Underfitted	99	100	10	19	29	44
	Correctly-fitted	0	0	87	76	12	17
	Overfitted	1	0	3	5	59	39
Case 5	Underfitted	50	64	1	6	29	42
	Correctly-fitted	2	10	0	0	0	0
	Overfitted	48	26	99	94	71	58
Correlated Cases							
Case 1	Underfitted	18	48	22	65	19	17
	Correctly-fitted	38	26	53	31	54	61
	Overfitted	44	26	25	4	27	22
Case 2	Underfitted	95	100	88	93	85	93
	Correctly-fitted	1	0	0	1	1	3
	Overfitted	4	0	12	6	14	4
Case 3	Underfitted	91	99	9	27	22	44
	Correctly-fitted	1	0	88	68	29	36
	Overfitted	8	1	3	5	49	20
Case 4	Underfitted	98	100	3	22	16	33
	Correctly-fitted	0	0	97	72	31	38
	Overfitted	2	0	0	6	53	29
Case 5	Underfitted	59	83	1	14	16	33
	Correctly-fitted	15	6	0	0	0	0
	Overfitted	26	11	99	86	84	67

3, 4 and 5, however, SHIM selects better models more often than the Lasso, CAP and CARDS, as we can see the points for SHIM are more concentrated at the upper left corner of the plots.

2.4.2 Analyzing Designed Experiments Using SHIM

In designed experiments, economic considerations may compel the investigator to use few experiments (runs). Many efficient experimental designs have been proposed in the literature. Among them fractional factorial designs are thoroughly studied and widely used. While the design of experiments literature is replete with research on the construction of the efficient designs, the methodologies of analysis have not received the same amount of attention. Traditional analysis methods (e.g., stepwise, all subset) continue to be a dominating choice for researchers in the DOE area. Wu and Hamada [65] stated three principles in the analysis of the designed experiment: effect sparsity (i.e., only a few of all candidate factors are active), effect hierarchy (e.g., main effects are more likely to be significant than two-factor interactions), and effect heredity (e.g., two-factor interaction x_1x_2 should be in the model only if the main effects x_1 and x_2 are also in the model).

The proposed method appears to be particularly suitable for analyzing the designed experiments, as SHIM encourages effect sparsity and requires effect heredity in the model. In this section we explore the use of SHIM in analyzing designed experiments. We consider a simulation study, in which a minimum-aberration 2^{6-2}_{IV} design was used to generate simulated data. Six two-level factors are studied in a 16-run design, which is defined by $x_5 = x_1x_2x_3$ and $x_6 = x_1x_2x_4$. Similar to those in Table 2.1, four cases of model are considered and shown in Table 2.3.

To study whether or not SHIM can effectively select the correct model, we generated 1,000 simulations and recorded (1-specificity, sensitivity) as in Section 2.4.1.

Table 2.3: DOE example setting: coefficients of the true models.

	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3
Case 1	7	2	1	0	0	0
Case 2	7	2	1	1	0	0
Case 3	7	2	1	7	7	7
Case 4	7	2	1	14	14	14

In each simulation, the data are generated by using the true models of Table 2.3, plus a random error of $N(0, 1)$. We then compare SHIM with three other competing methods: the Lasso, CARDS, and CAP. The results based on BIC-selected models are shown in Figure 2.5. It can be seen that SHIM performs consistently better than other methods in terms of removing irrelevant effects, especially when the heredity property is stronger in the model (i.e., Case 3 and Case 4).

2.5 Real Data Analysis

In this section, we apply our method SHIM to a real dataset. This dataset was from [33] for a case-control study of bladder cancer. It consists of the genotypes on 14 loci and the status of smoking behavior for 201 bladder cancer patients and 214 controls. Four of the genotypes are two-level factors, nine are three-level factors and one is a five-level factor. We represent all genotypes with dummy variables, hence a total of $4 + 2 \times 9 + 4 = 26$ dummy variables. The status of smoking behavior is represented with two predictors: one is a three-level factor (non-smoker, light-smoker, and heavy-smoker), and the other is a continuous variable, measuring the number of packs consumed per year. Since the response variable is binary (case/control), we used the negative binomial log-likelihood as the loss function rather than the squared error.

We randomly split the data into training ($n = 315$) and testing ($N = 100$). Tuning parameters were chosen via five-fold cross-validation based on the training

data. Fitted models were evaluated on the testing data, with the classification rule given by $\text{sgn}(\hat{g}(\boldsymbol{x}))$.

We considered three cases. In the first case, we used only the genetic information, i.e., the 14 loci genetic factors. There are a total of 336 candidate terms, including the main terms and all possible two-way interaction terms (between two different loci). In the second case, we considered the 14 genetic factors and the categorical smoke-status. There are a total of 390 candidate terms, including all possible two-way interaction terms among the genetic factors and the interaction terms between genetic factors and the categorical smoke-status. In the third case, we replaced the categorical smoke-status with the continuous smoke-status, where we considered the interactions between genetic factors and the continuous smoke-status. For comparison, we fitted both the Lasso and SHIM in each case. We used Adaptive(Ridge) as the pre-specified weights because the number of terms is larger than the number of observations in the first two cases. Misclassification errors, sensitivities and specificities (all on the test data) of these models are summarized in Table 2.4. As we can see, the models that use the genetic factors and the continuous smoke-status perform slightly better than other models in terms of the error rate. This may be heuristically understood as that the continuous smoke-status contains more information than the categorical smoke-status.

We then focused on the third case. Terms selected by the adaptive Lasso and the adaptive SHIM are shown in the upper part of Table 2.5. Notice that both methods selected the smoke-status $PackYear$, $GSTM1$ and MPO. The Lasso also selected an interaction term, $NQO1 \times PackYear$, but it does not obey the heredity constraint; on the other hand, SHIM selected the main term $NQO1$, but not the interaction term.

To further assess the terms that were selected, we applied a bootstrap analysis. The lower part of Table 2.5 summarizes the terms that were selected with selection frequency higher than 30% based on $B = 100$ bootstrap samples. As we can see, the five terms selected by SHIM using the training data are the only five terms that had the selection frequency higher than 30% in bootstrap samples. So SHIM is fairly stable in terms of selecting terms. We can also see that the smoke-status was always selected, followed immediately by MPO. The interaction term $NQO1 \times PackYear$ was selected half of the time by the Lasso, but never by SHIM; instead, SHIM selected the main term $NQO1$ half of the time.

These results seem to be consistent with the findings in [33]. The five terms selected by SHIM are among the ones that were shown to have a significant effect on increasing the risk of bladder cancer in [33].

Table 2.4: Real data analysis results: misclassification error, sensitivity and specificity on the test data

		Misclassification Error	Sensitivity	Specificity
SHIM using the genetic factors				
LASSO	Plain	0.44	0.48	0.63
	Adaptive	0.41	0.52	0.65
SHIM	Plain	0.36	0.54	0.73
	Adaptive	0.38	0.46	0.77
SHIM using the genetic factors and the categorical smoke-status variable				
LASSO	Plain	0.35	0.58	0.71
	Adaptive	0.37	0.56	0.69
SHIM	Plain	0.35	0.65	0.65
	Adaptive	0.34	0.65	0.67
SHIM using the genetic factors and the continuous smoke-status variable				
LASSO	Plain	0.34	0.60	0.71
	Adaptive	0.32	0.67	0.69
SHIM	Plain	0.33	0.67	0.67
	Adaptive	0.32	0.65	0.71

Table 2.5: Real data analysis results: the upper part lists the terms that were selected using the training data, and the lower part lists the terms that were selected (with selection frequency higher than 30%) based on 100 bootstrap samples. The numbers in the parentheses are the corresponding selection frequencies out of $B = 100$ bootstrap samples. The Lasso and SHIM were used with the genetic factors and the continuous smoke-status variable.

Adaptive LASSO		Adaptive SHIM	
Selected terms using the training data			
$PackYear$		$PackYear$	
$GSTM1$		$GSTM1$	
MPO		MPO	
$(NQO1) \times (PackYear)$		$NQO1$	
—		$MnSOD$	
Selected terms using 100 bootstrap samples			
$PackYear$	(100%)	$PackYear$	(100%)
MPO	(78%)	MPO	(82%)
$(NQO1) \times (PackYear)$	(49%)	$GSTM1$	(57%)
$GSTM1$	(43%)	$NQO1$	(46%)
$NQO1$	(37%)	$MnSOD$	(40%)
$MnSOD$	(36%)	—	—
$(COMT) \times (PackYear)$	(35%)	—	—
$(MPO) \times (PackYear)$	(32%)	—	—
$(XRCC1) \times (PackYear)$	(30%)	—	—

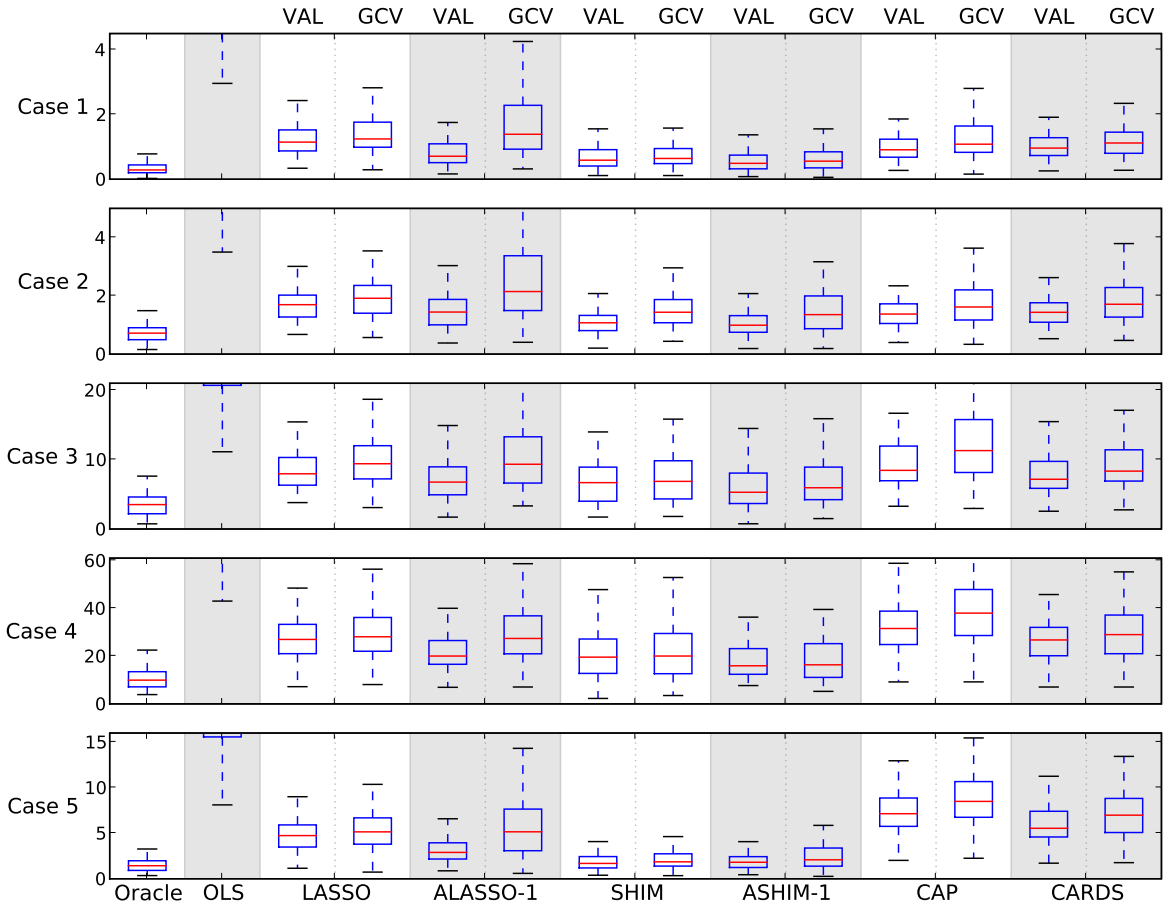
2.6 Discussion

In this chapter, we have extended the Lasso method for simultaneously fitting a regression model and identifying interaction terms. The proposed method automatically enforces the heredity constraint. In addition, it enjoys the "oracle" property under mild regularity conditions. We demonstrate that our new method tends to remove irrelevant variables more effectively and provide better prediction performance than the classical Lasso method, as well as two other more recent work.

The heredity that we have considered in this chapter is the so-called strong heredity, i.e., an interaction term can be included in the model only if both of the corresponding main terms are also included in the model. There is another type of heredity, weak heredity [29], in which only one of the main terms is required to be present when an interaction term is included in the model. Extending our SHIM framework

to enforce the weak heredity is straightforward: instead of re-parameterizing the coefficient for $x_j x_{j'}$ as the product $\gamma_{jj'}\beta_j\beta_{j'}$, we write it as $\gamma_{jj'}(|\beta_j| + |\beta_{j'}|)$. So if the coefficient for $x_j x_{j'}$ is not equal to zero, it implies that at least one of β_j and $\beta_{j'}$ is not equal to zero.

Figure 2.1: Simulation results: the boxplots of MSE values in independent cases. "VAL" refers to when we select the tuning parameters based on validation errors and "GCV" refers to when we use the GCV. "oracle" refers to the OLS applied only to the relevant terms; "ALASSO-1" and "ASHIM-1" refer to the adaptive Lasso and the adaptive SHIM with the weights based on the OLS estimates; "CAP" refers to [69] and "CARDS" refers to [67].



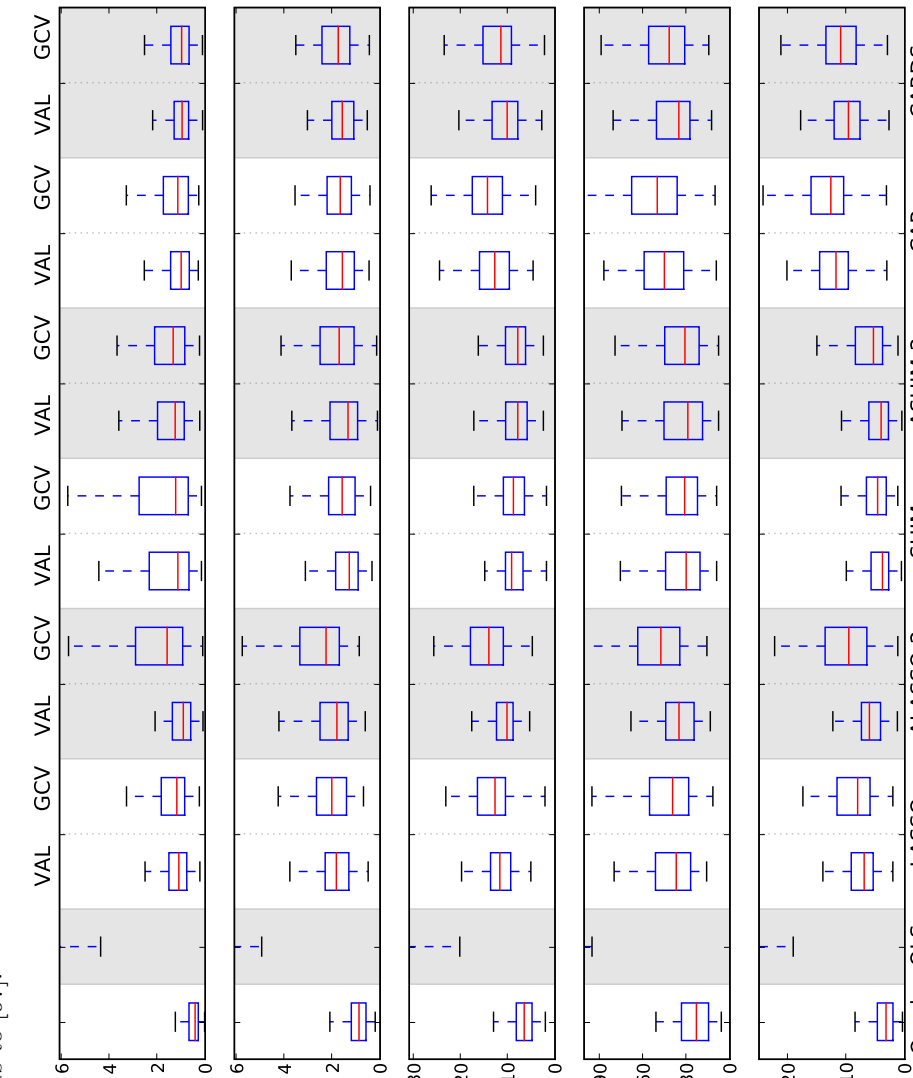


Figure 2.2: Simulation results: the boxplots of MSE values in correlated cases. "VAL" refers to when we select the tuning parameters based on validation errors and "GCV" refers to when we use the GCV. "oracle" refers to the OLS applied only to the relevant terms; "ALASSO-1" and "ASHIM-1" refer to the adaptive Lasso and the adaptive SHIM with the weights based on the OLS estimates; "CAP" refers to [69] and "CARDS" refers to [67].

Figure 2.3: Simulation results: sensitivity and 1 - specificity of the selected models based on BIC in independent cases. Each dot corresponds to each replicate among 100 replicates. "CAP" refers to [69] and "CARDS" refers to [67].

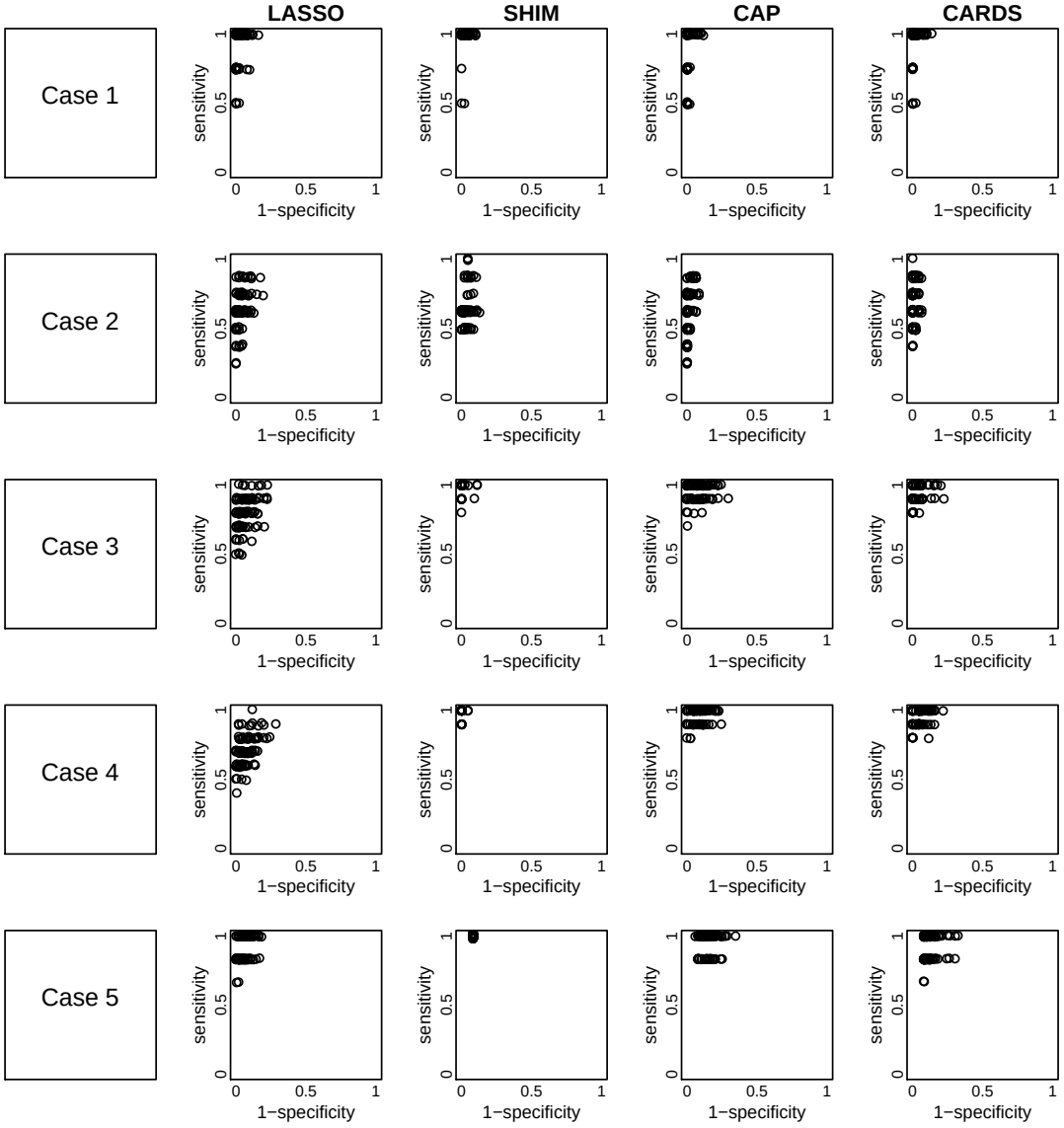


Figure 2.4: Simulation results: sensitivity and 1 - specificity of the selected models based on BIC in correlated cases. Each dot corresponds to each replicate among 100 replicates. "CAP" refers to [69] and "CARDS" refers to [67].

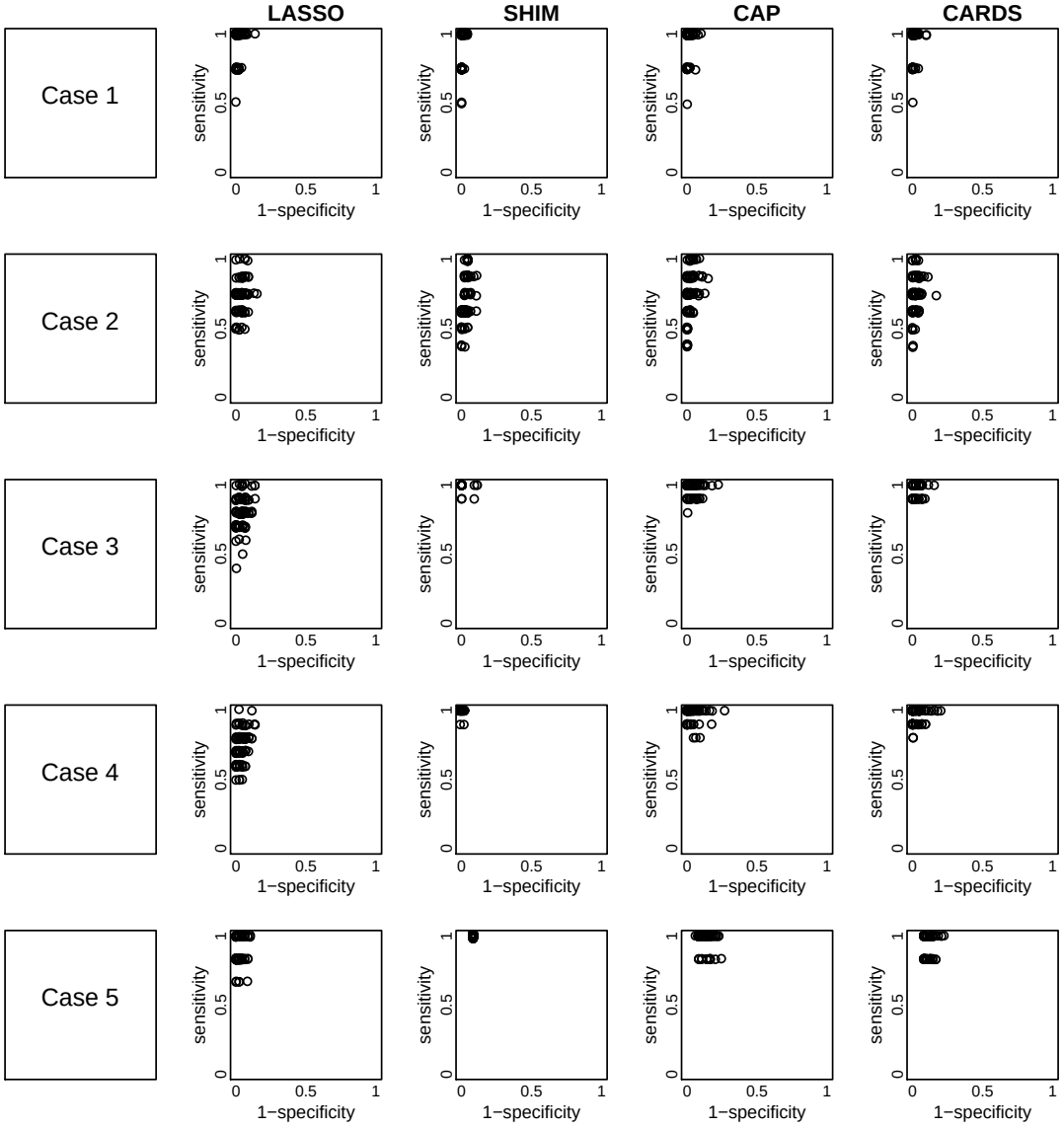
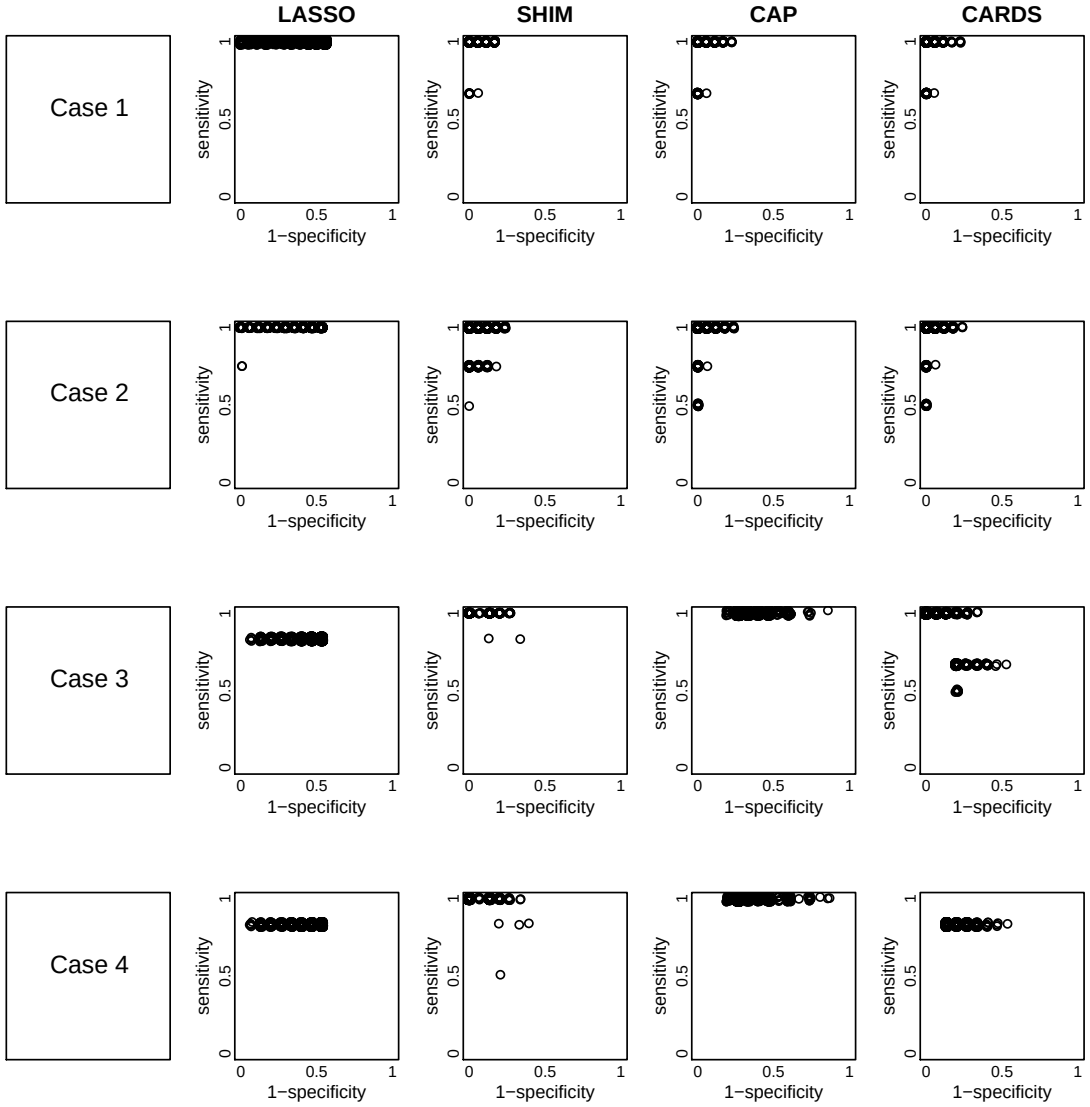


Figure 2.5: DOE example results: sensitivity and 1 - specificity of the selected models based on BIC. Each dot corresponds to each replicate among 1000 replicates. "CAP" refers to [69] and "CARDS" refers to Yuan et al. [67].



2.7 Appendix A

Regularity Conditions for Section 2.3.1

(C1) The observations $\{\boldsymbol{V}_i : i = 1, \ldots, n\}$ are independent and identically distributed with a probability density $f(\boldsymbol{V}, \boldsymbol{\theta})$, which has a common support. We assume the density f satisfies the following equations:

$$E_{\boldsymbol{\theta}}\left[\frac{\partial \log f(\boldsymbol{V}, \boldsymbol{\theta})}{\partial \theta_j}\right] = 0 \text{ for } j = 1, \ldots, \frac{p(p+1)}{2},$$

and

$$\begin{aligned} \boldsymbol{I}_{jk}(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial \theta_j} \log f(\boldsymbol{V}, \boldsymbol{\theta}) \frac{\partial}{\partial \theta_k} \log f(\boldsymbol{V}, \boldsymbol{\theta})\right] \\ &= E_{\boldsymbol{\theta}}\left[-\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f(\boldsymbol{V}, \boldsymbol{\theta})\right]. \end{aligned}$$

(C2) The Fisher information matrix

$$\boldsymbol{I}(\boldsymbol{\theta}) = E\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{V}, \boldsymbol{\theta})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{V}, \boldsymbol{\theta})\right)^{\intercal}\right]$$

is finite and positive definite at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

(C3) There exists an open set ω of Ω that contains the true parameter point $\boldsymbol{\theta}^*$ such that for almost all $\boldsymbol{V}$ the density $f(\boldsymbol{V}, \boldsymbol{\theta})$ admits all third derivatives $(\partial^3 f(\boldsymbol{V}, \boldsymbol{\theta}))/(\partial \theta_j \partial \theta_k \partial \theta_l)$ for all $\boldsymbol{\theta} \in \omega$ and any j, k, $l = 1, \ldots, p(p+1)/2$. Furthermore, there exist functions M_{jkl} such that

$$\left|\frac{\partial^3}{\partial \theta_j \partial \theta_k \partial \theta_l} \log f(\boldsymbol{V}, \boldsymbol{\theta})\right| \leq M_{jkl}(\boldsymbol{V}) \text{ for all } \boldsymbol{\theta} \in \omega,$$

where $m_{jkl} = E_{\boldsymbol{\theta}^*}[M_{jkl}(\boldsymbol{V})] < \infty$.

Regularity Conditions for Section 2.3.2

(C4) The observations $\{\boldsymbol{V}_{ni} : i = 1, \ldots, n\}$ are independent and identically distributed with a probability density $f_n(\boldsymbol{V}_n, \boldsymbol{\theta}_n)$, which has a common support.

We assume the density f_n satisfies the following equations:

$$E_{\boldsymbol{\theta}_n}\left[\frac{\partial \log f_n(\boldsymbol{V}_n, \boldsymbol{\theta}_n)}{\partial \theta_{nj}}\right] = 0 \text{ for } j = 1, \ldots, q_n,$$

and

$$\begin{aligned}\boldsymbol{I}_{jk}(\boldsymbol{\theta}_n) &= E_{\boldsymbol{\theta}_n}\left[\frac{\partial}{\partial \theta_{nj}} \log f_n(\boldsymbol{V}_n, \boldsymbol{\theta}_n) \frac{\partial}{\partial \theta_{nk}} \log f_n(\boldsymbol{V}_n, \boldsymbol{\theta}_n)\right] \\ &= E_{\boldsymbol{\theta}_n}\left[-\frac{\partial^2}{\partial \theta_{nj} \partial \theta_{nk}} \log f_n(\boldsymbol{V}_n, \boldsymbol{\theta}_n)\right].\end{aligned}$$

(C5) $I_n(\boldsymbol{\theta}_n) = E[(\frac{\partial \log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n})(\frac{\partial \log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\theta}_n)}{\partial \boldsymbol{\theta}_n})^{\intercal}]$ satisfies $0 < C_1 < \lambda_{\min}\{I_n(\boldsymbol{\theta}_n)\} \leq \lambda_{\max}\{I_n(\boldsymbol{\theta}_n)\} < C_2 < \infty$ for all n, where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ represent the smallest and the largest eigenvalues of a matrix respectively. Moreover, for any $j, k = 1, 2, \ldots, q_n$,

$$E_{\boldsymbol{\theta}_n}\left\{\frac{\partial \log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\theta}_n)}{\partial \theta_{nj}} \frac{\partial \log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\theta}_n)}{\partial \theta_{nk}}\right\}^2 < C_3 < \infty,$$

and

$$E_{\boldsymbol{\theta}_n}\left\{\frac{\partial^2 \log f_n(\boldsymbol{V}_{n1}, \boldsymbol{\theta}_n)}{\partial \theta_{nj} \partial \theta_{nk}}\right\}^2 < C_4 < \infty.$$

(C6) There exists a large open set $\omega_n \subset \Omega_n \in \mathbb{R}^{q_n}$ which contains the true parameter $\boldsymbol{\theta}_n^*$ such that for almost all $\boldsymbol{V}_{ni}$ the density admits all third derivatives $\partial^3 f_n(\boldsymbol{V}_{ni}, \boldsymbol{\theta}_n)/\partial \theta_{nj} \partial \theta_{nk} \partial \theta_{nl}$ for all $\boldsymbol{\theta}_n \in \omega_n$. Furthermore, there are functions M_{njkl} such that

$$\left|\frac{\partial^3 \log f_n(\boldsymbol{V}_{ni}, \boldsymbol{\theta}_n)}{\partial \theta_{nj} \partial \theta_{nk} \partial \theta_{nl}}\right| \leq M_{njkl}(\boldsymbol{V}_{ni})$$

for all $\boldsymbol{\theta}_n \in \omega_n$ and

$$E_{\boldsymbol{\theta}_n} M_{njkl}^2(\boldsymbol{V}_{ni}) < C_5 < \infty$$

for all q_n, n, and j, k, l.

Proof of Lemma 1

Let $\eta_n = n^{-1/2} + a_n$ and $\{\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta} : ||\boldsymbol{\delta}|| \leq d\}$ be the ball around $\boldsymbol{\theta^*}$, where $\boldsymbol{\delta} = (u_1, \ldots, u_p, v_{12}, \ldots, v_{p-1,p})^\intercal = (\boldsymbol{u}^\intercal, \boldsymbol{v}^\intercal)^\intercal$. Define

$$D_n(\boldsymbol{\delta}) \equiv Q_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) - Q_n(\boldsymbol{\theta}^*).$$

Let $-L_n$ denote the first term of Q_n in (8). For $\boldsymbol{\delta}$ that satisfies $||\boldsymbol{\delta}|| = d$, we have

$$\begin{aligned}
D_n(\boldsymbol{\delta}) &= -L_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) + n\sum_j \lambda_j^\beta\big(|\beta_j^* + \eta_n u_j| - |\beta_j^*|\big) \\
&\quad + n\sum_{k<k'} \lambda_{kk'}^\gamma\big(|\gamma_{kk'}^* + \eta_n v_{kk'}| - |\gamma_{kk'}^*|\big) \\
&\geq -L_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) + n\sum_{j\in\mathcal{A}_1} \lambda_j^\beta\big(|\beta_j^* + \eta_n u_j| - |\beta_j^*|\big) \\
&\quad + n\sum_{(k,k')\in\mathcal{A}_2} \lambda_{kk'}^\gamma\big(|\gamma_{kk'}^* + \eta_n v_{kk'}| - |\gamma_{kk'}^*|\big) \\
&\geq -L_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) - n\eta_n\sum_{j\in\mathcal{A}_1} \lambda_j^\beta|u_j| - n\eta_n\sum_{(k,k')\in\mathcal{A}_2} \lambda_{kk'}^\gamma|v_{kk'}| \\
&\geq -L_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) - n\eta_n^2\Big(\sum_{j\in\mathcal{A}_1}|u_j| + \sum_{(k,k')\in\mathcal{A}_2}|v_{kk'}|\Big) \\
&\geq -L_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) - n\eta_n^2(|\mathcal{A}_1| + |\mathcal{A}_2|)d \\
&= -[\nabla L_n(\boldsymbol{\theta}^*)]^\intercal(\eta_n\boldsymbol{\delta}) - \frac{1}{2}(\eta_n\boldsymbol{\delta})^\intercal[\nabla^2 L_n(\boldsymbol{\theta}^*)](\eta_n\boldsymbol{\delta})(1 + o_p(1)) \\
&\quad -n\eta_n^2(|\mathcal{A}_1| + |\mathcal{A}_2|)d. \qquad (2.9)
\end{aligned}$$

We split (2.9) into three parts:

$$\begin{aligned}
A_1 &= -[\nabla L_n(\boldsymbol{\theta}^*)]^\intercal(\eta_n\boldsymbol{\delta}) \\
A_2 &= -\frac{1}{2}(\eta_n\boldsymbol{\delta})^\intercal[\nabla^2 L_n(\boldsymbol{\theta}^*)](\eta_n\boldsymbol{\delta})(1 + o_p(1)) \\
A_3 &= -n\eta_n^2(|\mathcal{A}_1| + |\mathcal{A}_2|)d
\end{aligned}$$

Then

$$\begin{aligned} A_1 &= -\eta_n[\nabla L_n(\boldsymbol{\theta}^*)]^\intercal \boldsymbol{\delta} \\ &= -\sqrt{n}\eta_n\Big(\frac{1}{\sqrt{n}}\nabla L_n(\boldsymbol{\theta}^*)\Big)^\intercal \boldsymbol{\delta} \\ &= -\sqrt{n}\eta_n\Big(\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\nabla \log f(\boldsymbol{V}_i,\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\Big)^\intercal \boldsymbol{\delta} \\ &= -O_p(\sqrt{n}\eta_n)\boldsymbol{\delta} \\ &= -O_p(n\eta_n^2)\boldsymbol{\delta}, \end{aligned}$$

$$\begin{aligned} A_2 &= \frac{1}{2}n\eta_n^2\Big\{\boldsymbol{\delta}^\intercal\big[-\frac{1}{n}\nabla^2 L_n(\boldsymbol{\theta}^*)\big]\boldsymbol{\delta}\Big\}(1+o_p(1)) \\ &= \frac{1}{2}n\eta_n^2\Big\{\boldsymbol{\delta}^\intercal\big[I(\boldsymbol{\theta}^*)\big]\boldsymbol{\delta}\Big\}(1+o_p(1)) \text{ by the weak law of large numbers.} \end{aligned}$$

Thus,

$$\begin{aligned} D_n(\boldsymbol{\delta}) &\geq A_1 + A_2 + A_3 \\ &= -n\eta_n^2 O_p(1)\boldsymbol{\delta} + \frac{1}{2}n\eta_n^2\Big\{\boldsymbol{\delta}^\intercal\big[I(\boldsymbol{\theta}^*)\big]\boldsymbol{\delta}\Big\}(1+o_p(1)) - n\eta_n^2(|\mathcal{A}_1|+|\mathcal{A}_2|)d. \end{aligned} \tag{2.10}$$

Notice that A_2 dominates the rest terms A_1 and A_3 and is positive since $I(\boldsymbol{\theta})$ is positive definite at $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ from (C2). Therefore, for any given $\epsilon>0$, there exists a large enough constant d such that

$$P\Big\{\inf_{||\delta||=d} Q_n(\boldsymbol{\theta}^*+\eta_n\boldsymbol{\delta}) > Q_n(\boldsymbol{\theta}^*)\Big\} \geq 1-\epsilon.$$

This implies that with probability at least $1-\epsilon$, there exists a local minimizer in the ball $\{\boldsymbol{\theta}^*+\eta_n\boldsymbol{\delta}:\ ||\boldsymbol{\delta}||\leq d\}$. Thus, there exists a local minimizer of $Q_n(\boldsymbol{\theta})$ such that $||\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*|| = O_p(\eta_n)$. $\square$

Proof of Theorem 1

We first consider $P(\hat{\beta}_{\mathcal{A}_1^c} = 0) \to 1$. It is sufficient to show for any $j \in \mathcal{A}_1^c$

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j} < 0 \text{ for } -\epsilon_n < \hat{\beta}_j < 0 \tag{2.11}$$

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j} > 0 \text{ for } 0 < \hat{\beta}_j < \epsilon_n \tag{2.12}$$

with probability tending to 1 where $\epsilon_n = Cn^{-1/2}$ and $C > 0$ is any constant. To show (2.12), notice

$$\begin{aligned}
\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j} &= -\frac{L_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j} + n\lambda_j^{\beta}\mathrm{sgn}(\hat{\beta}_j) \\
&= -\frac{L_n(\boldsymbol{\theta}^*)}{\partial \beta_j} - \sum_{k=1}^{\frac{p(p+1)}{2}} \frac{\partial^2 L_n(\boldsymbol{\theta}^*)}{\partial \beta_j \partial \theta_k}(\hat{\theta}_k - \theta_k^*) \\
&\quad - \sum_{k=1}^{\frac{p(p+1)}{2}} \sum_{l=1}^{\frac{p(p+1)}{2}} \frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}})}{\partial \beta_j \partial \theta_k \partial \theta_l}(\hat{\theta}_k - \theta_k^*)(\hat{\theta}_l - \theta_l^*) + n\lambda_j^{\beta}\mathrm{sgn}(\hat{\beta}_j)
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*$. By (C1)–(C3) and the condition $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\| = O_p(n^{-1/2})$,

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j} = \sqrt{n}\Big\{O_p(1) + \sqrt{n}\lambda_j^{\beta}\mathrm{sgn}(\hat{\beta}_j)\Big\}.$$

As $\sqrt{n}\lambda_j^{\beta} \to \infty$ for $j \in \mathcal{A}_1^c$ from the assumption, the sign of $\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j}$ is dominated by $\mathrm{sgn}(\hat{\beta}_j)$. Therefore,

$$P\Big[\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_j} > 0 \text{ for } 0 < \hat{\beta}_j < \epsilon_n\Big] \to 1 \text{ as } n \to \infty.$$

(2.11) can be shown in the same way.

Next, we prove $P(\hat{\boldsymbol{\gamma}}_{\mathcal{A}_2^c} = 0) \to 1$.

- For (k, k') where $(k, k') \in \mathcal{A}_2^c$ and $k, k' \in \mathcal{A}_1$: we can prove $P(\hat{\boldsymbol{\gamma}}_{kk'} = 0) \to 1$ by a similar reasoning.

- For (k, k') where $(k, k') \in \mathcal{A}_2^c$ and either k or k' is in $\mathcal{A}_1^c$: without loss of generality, assume that $\beta_k^* = 0$. Notice that $\hat{\beta}_k = 0$ implies $\hat{\gamma}_{kk'} = 0$, because if $\hat{\gamma}_{kk'} \neq 0$, then the value of the loss function does not change but the value of the penalty function will increase. Since we already have $P(\hat{\beta}_k = 0) \to 1$, we can conclude $P(\hat{\gamma}_{kk'} = 0) \to 1$ as well.

□

Proof of Theorem 2

Let $Q_n(\boldsymbol{\theta}_{\mathcal{A}})$ denote the objective function Q_n only on the $\mathcal{A}$-component of $\boldsymbol{\theta}$, that is, $Q_n(\boldsymbol{\theta})$ with $\boldsymbol{\theta}_{\mathcal{A}^c}$. Based on Lemma 1 and Theorem 1, we have $P\big(\hat{\boldsymbol{\theta}}_{\mathcal{A}^c} = 0\big) \to 1$. Thus,

$$P\Big[\arg\min_{\boldsymbol{\theta}_{\mathcal{A}}} Q_n(\boldsymbol{\theta}_{\mathcal{A}}) = \big(\mathcal{A}\text{-component of } \arg\min_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta})\big)\Big] \to 1.$$

It means that $\hat{\boldsymbol{\theta}}_{\mathcal{A}}$ should satisfy

$$\left.\frac{\partial Q_n(\boldsymbol{\theta}_{\mathcal{A}})}{\partial \theta_j}\right|_{\boldsymbol{\theta}_{\mathcal{A}} = \hat{\boldsymbol{\theta}}_{\mathcal{A}}} = 0, \ \forall j \in \mathcal{A} \tag{2.13}$$

with probability tending to 1.

Let $L_n(\boldsymbol{\theta}_{\mathcal{A}})$ and $P_{\boldsymbol{\lambda}}(\boldsymbol{\theta}_{\mathcal{A}})$ denote the log-likelihood function of $\boldsymbol{\theta}_{\mathcal{A}}$ and the penalty function of $\boldsymbol{\theta}_{\mathcal{A}}$ respectively so that we have

$$Q_n(\boldsymbol{\theta}_{\mathcal{A}}) = -L_n(\boldsymbol{\theta}_{\mathcal{A}}) + nP_{\boldsymbol{\lambda}}(\boldsymbol{\theta}_{\mathcal{A}}).$$

From (2.13), now we have

$$\nabla_{\mathcal{A}} Q_n(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) = -\nabla_{\mathcal{A}} L_n(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) + n\nabla_{\mathcal{A}} P_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) = \mathbf{0}, \tag{2.14}$$

with probability tending to 1.

- Consider the first term in (2.14). By the Taylor expansion of $-\nabla_{\mathcal{A}} L_n(\boldsymbol{\theta}_{\mathcal{A}})$ at $\boldsymbol{\theta}_{\mathcal{A}} = \boldsymbol{\theta}^*_{\mathcal{A}}$,

$$\begin{aligned} -\nabla_{\mathcal{A}} L_n(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) &= -\nabla_{\mathcal{A}} L_n(\boldsymbol{\theta}^*_{\mathcal{A}}) - \left[\nabla^2_{\mathcal{A}} L_n(\boldsymbol{\theta}^*_{\mathcal{A}}) + o_p(1)\right](\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) \\ &= \sqrt{n}\left[-\frac{1}{\sqrt{n}}\nabla_{\mathcal{A}} L_n(\boldsymbol{\theta}^*_{\mathcal{A}}) + \big(-\frac{1}{n}\nabla^2_{\mathcal{A}} L_n(\boldsymbol{\theta}^*_{\mathcal{A}}) - o_p(1)\big)\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}})\right] \\ &= \sqrt{n}\left[-\frac{1}{\sqrt{n}}\nabla_{\mathcal{A}} L_n(\boldsymbol{\theta}^*_{\mathcal{A}}) + \boldsymbol{I}(\boldsymbol{\theta}^*_{\mathcal{A}})\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) + o_p(1)\right]. \end{aligned}$$

- Consider the second term in (2.14). By the Taylor expansion of $n\nabla_{\mathcal{A}} P_{\boldsymbol{\lambda}}(\boldsymbol{\theta}_{\mathcal{A}})$ at $\boldsymbol{\theta}_{\mathcal{A}} = \boldsymbol{\theta}^*_{\mathcal{A}}$,

$$\begin{aligned} n\nabla_{\mathcal{A}} P_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}) &= n\left\{\left[\begin{array}{c} \lambda^{\beta}_j \mathrm{sgn}(\beta_j) \\ \lambda^{\gamma}_{kk'}\mathrm{sgn}(\gamma_{kk'}) \end{array}\right]_{j\in\mathcal{A}_1,(k,k')\in\mathcal{A}_2} + o_p(1)(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}})\right\} \\ &= \sqrt{n}o_p(1) \end{aligned}$$

  because $\sqrt{n}a_n = o(1)$ and $\|\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}\| = O_p(n^{-1/2})$.

Thus,

$$0 = \sqrt{n}\left[-\frac{1}{\sqrt{n}}\nabla_{\mathcal{A}} L_n(\boldsymbol{\theta}^*_{\mathcal{A}}) + \boldsymbol{I}(\boldsymbol{\theta}^*_{\mathcal{A}})\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) + o_p(1)\right].$$

It follows

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) = \boldsymbol{I}(\boldsymbol{\theta}^*_{\mathcal{A}})^{-1}\ \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\nabla_{\mathcal{A}}\log f(\boldsymbol{V}_i, \boldsymbol{\theta}_{\mathcal{A}}) + o_p(1).$$

Therefore, by central limit theorem,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) \to_d N(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}^*_{\mathcal{A}})).$$

□

Proof of Lemma 2

Let $\eta_n = \sqrt{q_n}(n^{-1/2} + a_n)$ and $\{\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \leq d\}$ be the ball around $\boldsymbol{\theta}_n^*$, where $\boldsymbol{\delta} = (u_1, \ldots, u_{p_n}, v_{12}, \ldots, v_{p_n-1,p_n})^\intercal = (\boldsymbol{u}^\intercal, \boldsymbol{v}^\intercal)^\intercal$. It is sufficient to show that for any $\epsilon > 0$, there is a large constant d such that

$$P\left\{\inf_{\|\delta\|=d} Q_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) > Q_n(\boldsymbol{\theta}_n^*)\right\} \geq 1 - \epsilon,$$

because it implies that with probability at least $1 - \epsilon$, there exists a local minimum in the ball $\{\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \leq d\}$. Define

$$D_n(\boldsymbol{\delta}) \equiv Q_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) - Q_n(\boldsymbol{\theta}_n^*).$$

Let $-L_n$ and nP_n denote the first and the second terms of Q_n in (9). For any $\boldsymbol{\delta}$ satisfying $||\boldsymbol{\delta}|| = d$, we have

$$\begin{aligned}
D_n(\boldsymbol{\delta}) &= -L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) + nP_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) - nP_n(\boldsymbol{\theta}_n^*) \\
&\geq -L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) \\
&\quad +n\Big\{\sum_{j\in\mathcal{A}_{n1}} \lambda_{nj}^{\beta}(|\beta_j + \eta_n u_j| - |\beta_j|) + \sum_{(k,k')\in\mathcal{A}_{n2}} \lambda_{n,kk'}^{\gamma}(|\gamma_{kk'} + \eta_n v_{kk'}| - |\gamma_{kk'}|)\Big\} \\
&\geq -L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) - n\eta_n\Big\{\sum_{j\in\mathcal{A}_{n1}} \lambda_{nj}^{\beta}|u_j| + \sum_{(k,k')\in\mathcal{A}_{n2}} \lambda_{n,kk'}^{\gamma}|v_{kk'}|\Big\} \\
&\geq -L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) - n\eta_n\Big\{\sum_{j\in\mathcal{A}_{n1}} a_n|u_j| + \sum_{(k,k')\in\mathcal{A}_{n2}} a_n|v_{kk'}|\Big\} \\
&\geq -L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) - n\eta_n(\sqrt{s_n}a_n)d \\
&\geq -L_n(\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}_n^*) - n\eta_n^2 d.
\end{aligned}$$

By Taylor expansion,

$$\begin{aligned}
D_n(\boldsymbol{\delta}) &\geq -\nabla^\intercal L_n(\boldsymbol{\theta}_n^*)(\eta_n\boldsymbol{\delta}) - \frac{1}{2}(\eta_n\boldsymbol{\delta})^\intercal\nabla^2 L_n(\boldsymbol{\theta}_n^*)(\eta_n\boldsymbol{\delta}) - \frac{1}{6}\nabla^\intercal\big\{\boldsymbol{\delta}^\intercal\nabla^2 L_n(\tilde{\boldsymbol{\theta}}_n)\boldsymbol{\delta}\big\}\boldsymbol{\delta}\eta_n^3 - n\eta_n^2 d \\
&\equiv A_1 + A_2 + A_3 + A_4,
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_n$ lies between $\boldsymbol{\theta}_n^* + \eta_n\boldsymbol{\delta}$ and $\boldsymbol{\theta}_n^*$. We first consider A_1.

$$\begin{aligned} |A_1| &= |-\nabla^\intercal L_n(\boldsymbol{\theta}_n^*)(\eta_n\boldsymbol{\delta})| \\ &\leq \eta_n\|\nabla^\intercal L_n(\boldsymbol{\theta}_n^*)\|\|\boldsymbol{\delta}\| \\ &= O_p(\eta_n\sqrt{nq_n})d = O_p(n\eta_n^2)d. \end{aligned}$$

Next, since we have

$$\Big\|\frac{1}{n}\nabla^2 L_n(\boldsymbol{\theta}_n^*) + \boldsymbol{I}_n(\boldsymbol{\theta}_n^*)\Big\| = o_p\Big(\frac{1}{q_n}\Big) \tag{2.15}$$

by Chebyshev's inequality and (C5), we can show that

$$\begin{aligned} A_2 &= -\frac{1}{2}\eta_n^2\Big[\boldsymbol{\delta}^\intercal\nabla^2 L_n(\boldsymbol{\theta}_n^*)\boldsymbol{\delta}\Big] \\ &= -\frac{1}{2}\boldsymbol{\delta}^\intercal\left[\frac{1}{n}\Big\{\nabla^2 L_n(\boldsymbol{\theta}_n^*) - E(\nabla^2 L_n(\boldsymbol{\theta}_n^*))\Big\}\right]\boldsymbol{\delta}\cdot n\eta_n^2 - \frac{1}{2}\boldsymbol{\delta}^\intercal\frac{1}{n}E(\nabla^2 L_n(\boldsymbol{\theta}_n^*))\boldsymbol{\delta}\cdot n\eta_n^2 \\ &= \frac{1}{2}n\eta_n^2\boldsymbol{\delta}^\intercal\boldsymbol{I}_n(\boldsymbol{\theta}_n^*)\boldsymbol{\delta} - \frac{1}{2}n\eta_n^2 d^2 o_p(1). \end{aligned}$$

Moreover, by Cauchy-Schwarz inequality, (C6), and the conditions $\sqrt{n}a_n \to 0$ and $q_n^5/n \to 0$,

$$\begin{aligned} |A_3| &= \Big|-\frac{1}{6}\nabla^\intercal\big\{\boldsymbol{\delta}^\intercal\nabla^2 L_n(\tilde{\boldsymbol{\theta}}_n)\boldsymbol{\delta}\big\}\boldsymbol{\delta}\eta_n^3\Big| \\ &= \frac{1}{6}\eta_n^3\Big|\sum_{i=1}^n\sum_{j,k,l=1}^{q_n}\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\theta_{nj}\partial\theta_{nk}\partial\theta_{nl}}\delta_j\delta_k\delta_l\Big| \\ &\leq \eta_n^3\sum_{i=1}^n\Big(\sum_{j,k,l=1}^{q_n}M_{njkl}^2(\boldsymbol{V}_{ni})\Big)^{1/2}\|\boldsymbol{\delta}\|^3 \\ &= n\eta_n^3\ O_p\big(q_n^{3/2}\big)\ \big(q_n O(1)\big)^{1/2}\|\boldsymbol{\delta}\|^2 \\ &= n\eta_n^2\ O_p\big(\eta_n q_n^2\big)d^2 \\ &= n\eta_n^2 o_p(1)d^2. \end{aligned}$$

A_2 dominates the rest terms A_1, A_3 and A_4 for a sufficiently large $\boldsymbol{\delta}$, and is positive because $\boldsymbol{I}_n(\boldsymbol{\theta}_n^*)$ is positive definite by (C5). $\square$

Proof of Theorem 3

Proof of (a)

We first prove $P(\hat{\beta}_{nj} = 0) \to 1$ for $j \in \mathcal{A}_{n1}^c$ as $n \to \infty$. It is enough to show that with probability tending to 1, for any $j \in \mathcal{A}_{n1}^c$,

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_{nj}} < 0 \text{ for } -\epsilon_n < \hat{\beta}_{nj} < 0 \tag{2.16}$$

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_{nj}} > 0 \text{ for } 0 < \hat{\beta}_{nj} < \epsilon_n \tag{2.17}$$

where $\epsilon_n = Cn^{-1/2}$ and $C > 0$ is any constant. To show (2.17), we consider a Taylor expansion of $\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_{nj}}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_n^*$.

$$\begin{aligned}
\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_{nj}} &= -\frac{\partial L_n(\hat{\boldsymbol{\theta}}_n)}{\partial \beta_{nj}} + n\lambda_{nj}^{\beta}\text{sgn}(\hat{\beta}_{nj}) \\
&= -\frac{\partial L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}} - \sum_{k=1}^{q_n} \frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}\partial \theta_{nk}}(\hat{\theta}_{nk} - \theta_{nk}^*) \\
&\quad - \sum_{k=1}^{q_n}\sum_{l=1}^{q_n} \frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \beta_{nj}\partial \theta_{nk}\partial \theta_{nl}}(\hat{\theta}_{nk} - \theta_{nk}^*)(\hat{\theta}_{nl} - \theta_{nl}^*) \\
&\quad + n\lambda_{nj}^{\beta}\text{sgn}(\hat{\beta}_{nj}) \\
&\equiv I_1 + I_2 + I_3 + I_4
\end{aligned} \tag{2.18}$$

where $\tilde{\boldsymbol{\theta}}_n$ lies between $\boldsymbol{\theta}_n^*$ and $\hat{\boldsymbol{\theta}}_n$. By Chebyshev's inequality,

$$I_1 = -\sum_{i=1}^{n} \frac{\partial \log f_n(\boldsymbol{V}_{ni}, \boldsymbol{\theta}_n^*)}{\partial \beta_{nj}} = O_p(\sqrt{n}) = O_p(\sqrt{nq_n}).$$

Next,

$$\begin{aligned}
I_2 &= -\sum_{k=1}^{q_n} \frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}\partial \theta_{nk}}(\hat{\theta}_{nk} - \theta_{nk}^*) \\
&= -\sum_{k=1}^{q_n} \left[\frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}\partial \theta_{nk}} - E\Big[\frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}\partial \theta_{nk}}\Big]\right](\hat{\theta}_{nk} - \theta_{nk}^*) - \sum_{k=1}^{q_n} E\Big[\frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial \beta_{nj}\partial \theta_{nk}}\Big](\hat{\theta}_{nk} - \theta_{nk}^*) \\
&\equiv K_1 + K_2.
\end{aligned}$$

By Cauchy-Schwarz inequality and (C5),

$$\begin{aligned} |K_1| &\leq \left[\sum_{k=1}^{q_n}\Big\{\frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial\beta_{nj}\partial\theta_{nk}} - E\Big[\frac{\partial^2 L_n(\boldsymbol{\theta}_n^*)}{\partial\beta_{nj}\partial\theta_{nk}}\Big]\Big\}^2\right]^{1/2} \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*\| \\ &= O_p(\sqrt{nq_n})O_p(\sqrt{q_n/n}) \\ &= O_p(\sqrt{nq_n})o_p(1) = o_p(\sqrt{nq_n}). \end{aligned}$$

Again, by Cauchy-Schwarz inequality and (C5),

$$\begin{aligned} |K_2| &= n\Big|\sum_{k=1}^{q_n} \boldsymbol{I}_n(\boldsymbol{\theta}_n^*)_{(j,k)}(\hat{\theta}_{nk} - \theta_{nk}^*)\Big| \\ &\leq n\Big[\sum_{k=1}^{q_n} \boldsymbol{I}_n(\boldsymbol{\theta}_n^*)_{(j,k)}^2\Big]^{1/2}\Big[\sum_{k=1}^{q_n}(\hat{\theta}_{nk} - \theta_{nk}^*)^2\Big]^{1/2} \\ &= n\ O(1)O_p(\sqrt{q_n/n}) = O_p(\sqrt{nq_n}). \end{aligned}$$

Therefore, $I_2 = O_p(\sqrt{nq_n})$.

$$\begin{aligned} I_3 &= -\sum_{k=1}^{q_n}\sum_{l=1}^{q_n}\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}}(\hat{\theta}_{nk} - \theta_{nk}^*)(\hat{\theta}_{nl} - \theta_{nl}^*) \\ &= -\sum_{k=1}^{q_n}\sum_{l=1}^{q_n}\left[\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}} - E\Big[\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}}\Big]\right](\hat{\theta}_{nk} - \theta_{nk}^*)(\hat{\theta}_{nl} - \theta_{nl}^*) \\ &\quad -\sum_{k=1}^{q_n}\sum_{l=1}^{q_n} E\Big[\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}}\Big](\hat{\theta}_{nk} - \theta_{nk}^*)(\hat{\theta}_{nl} - \theta_{nl}^*) \\ &\equiv K_3 + K_4. \end{aligned}$$

By Cauchy-Schwarz inequality and (C6),

$$\begin{aligned} |K_4| &\leq \left[\sum_{k=1}^{q_n}\sum_{l=1}^{q_n} n^2\Big\{E\Big[\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}}\Big]\Big\}^2\right]^{1/2} \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*\|^2 \\ &\leq \Big[q_n^2 n^2 C_5\Big]^{1/2} O_p(q_n/n) \\ &= O_p(q_n^2) = O_p(\sqrt{nq_n})O_p(\sqrt{q_n^3/n}) = O_p(\sqrt{nq_n})o_p(1) \\ &= o_p(\sqrt{nq_n}). \end{aligned}$$

By Cauchy-Schwarz inequality and (C6),

$$\begin{aligned} |K_3| &\leq \left[\sum_{k=1}^{q_n}\sum_{l=1}^{q_n}\Big\{\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}} - E\Big[\frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}\partial\theta_{nk}\partial\theta_{nl}}\Big]\Big\}^2\right]^{1/2} \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^*\|^2 \\ &= \Big[nq_n^2 O_p(1)\Big]^{1/2} O_p(q_n/n) \\ &= o_p(\sqrt{nq_n}). \end{aligned}$$

Thus, $I_1 + I_2 + I_3 = O_p(\sqrt{nq_n})$. Therefore, returning to (2.18),

$$\begin{aligned} \frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}} &= O_p(\sqrt{nq_n}) + n\lambda_{nj}^{\beta}\mathrm{sgn}(\hat{\beta}_{nj}) \\ &= \sqrt{nq_n}\Big\{O_p(1) + \sqrt{\frac{n}{q_n}}\lambda_{nj}^{\beta}\mathrm{sgn}(\hat{\beta}_{nj})\Big\}. \end{aligned}$$

Since $\sqrt{n/q_n}b_n \to \infty$, $\mathrm{sgn}(\hat{\beta}_{nj})$ dominates the sign of $\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}}$ when n is large. Therefore, for $0 < \hat{\beta}_{nj} < \epsilon_n$, $\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\beta_{nj}} > 0$ with probability tending to 1 as $n \to \infty$. (2.16) can be shown in the same way.

Next, we prove $P(\hat{\boldsymbol{\gamma}}_{n\mathcal{A}_{n2}^c} = 0) \to 1$.

- For (k, k') where $(k, k') \in \mathcal{A}_{n2}^c$ and $k, k' \in \mathcal{A}_{n1}$: we can prove $P(\hat{\boldsymbol{\gamma}}_{n,kk'} = 0) \to 1$ by a similar reasoning.

- For (k, k') where $(k, k') \in \mathcal{A}_{n2}^c$ and either k or k' is in $\mathcal{A}_{n1}^c$: without loss of generality, assume that $\beta_{nk}^* = 0$. Notice that $\hat{\beta}_{nk} = 0$ implies $\hat{\gamma}_{n,kk'} = 0$, because if $\hat{\gamma}_{n,kk'} \neq 0$, then the value of the loss function does not change but the value of the penalty function will increase. Since we already have $P(\hat{\beta}_{nk} = 0) \to 1$, we can conclude $P(\hat{\gamma}_{n,kk'} = 0) \to 1$ as well.

Proof of (b)

We want to show that with probability tending to 1,

$$\begin{aligned}
\sqrt{n}\boldsymbol{A}_n\boldsymbol{I}_n^{1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}-\boldsymbol{\theta}^*_{n\mathcal{A}_n}) &= \sqrt{n}\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}-\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
(2.19) \qquad &= \sqrt{n}\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big\{\frac{1}{n}\nabla L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})+o_p(n^{-1/2})\Big\} \\
&= \frac{1}{\sqrt{n}}\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\sum_{i=1}^{n}\Big[\nabla L_{ni}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big] \\
&\quad +o_p(\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\mathbf{1}_{(s_n\times 1)}) \\
&= \frac{1}{\sqrt{n}}\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\sum_{i=1}^{n}\Big[\nabla L_{ni}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big]+o_p(1) \\
&\equiv \sum_{i=1}^{n}\boldsymbol{Y}_{ni}+o_p(1) \\
(2.20) \qquad &\to_d N(\mathbf{0},\boldsymbol{G}),
\end{aligned}$$

where $\boldsymbol{Y}_{ni}=\frac{1}{\sqrt{n}}\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big[\nabla L_{ni}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big]$. We will show (2.19) and (2.20) in (I) and (II) respectively.

(I) We want to show $\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}-\boldsymbol{\theta}^*_{n\mathcal{A}_n})=\frac{1}{n}\nabla L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})+o_p(\frac{1}{\sqrt{n}})$. We know that with probability tending to 1,

$$\mathbf{0}=\nabla_{\mathcal{A}_n}Q_n(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n})=-\nabla_{\mathcal{A}_n}L_n(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n})+n\nabla_{\mathcal{A}_n}P_{\lambda_n}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}).$$

By Taylor expansion at $\boldsymbol{\theta}=\boldsymbol{\theta}^*_{n\mathcal{A}_n}$

$$\begin{aligned}
\mathbf{0} &= -\nabla_{\mathcal{A}_n}L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})-\Big[\nabla^2_{\mathcal{A}_n}L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big](\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}-\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad -\frac{1}{2}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}-\boldsymbol{\theta}^*_{n\mathcal{A}_n})^{\intercal}\Big[\nabla^2_{\mathcal{A}_n}\big(\nabla_{\mathcal{A}_n}L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\big)\Big](\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}-\boldsymbol{\theta}^*_{n\mathcal{A}_n})+n\nabla_{\mathcal{A}_n}P_{\lambda_n}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}).
\end{aligned}$$

Thus,

$$\begin{aligned}
\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) &= -\frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad + \Big\{\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) + \frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big\}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&= \frac{1}{n}\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad -\frac{1}{2}\frac{1}{n}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n})^{\intercal}\Big[\nabla^2_{\mathcal{A}_n}\big(\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\big)\Big](\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad -\nabla_{\mathcal{A}_n} P_{\lambda_n}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&\quad + \Big\{\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) + \frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big\}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}).
\end{aligned}$$

Therefore, it is sufficient to show that

$$\begin{aligned}
&-\frac{1}{2}\frac{1}{n}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n})^{\intercal}\Big[\nabla^2_{\mathcal{A}_n}\big(\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\big)\Big](\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) - \nabla_{\mathcal{A}_n} P_{\lambda_n}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
&+ \Big\{\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) + \frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big\}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \\
\equiv\ & B_1 + B_2 + B_3 \\
=\ & o_p(n^{-1/2}).
\end{aligned}$$

First, by Cauchy-Schwarz inequality and (C6),

$$\begin{aligned}
\|B_1\|^2 &\le \frac{1}{n^2}\|\nabla^2_{\mathcal{A}_n}\big(\nabla_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\big)\|^2\|\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}\|^4 \\
&\le \frac{1}{n^2}\sum_{j,k,l\in\mathcal{A}_n}\Big\{\sum_{i=1}^n M_{njkl}(\boldsymbol{V}_{ni})\Big\}^2\|\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}\|^4 \\
&= \frac{1}{n^2}\sum_{j,k,l\in\mathcal{A}_n} n^2 O_p(1) O_p(\frac{q_n^2}{n}) \\
&= O_p(q_n^5/n^2) \\
&= o_p(1/n).
\end{aligned}$$

Second, because $a_n = o(1/\sqrt{nq_n})$ from the condition of the theorem,

$$\begin{aligned}
\|B_2\|^2 &= \left\|(\lambda^{\beta}_{n1}\mathrm{sgn}(\beta^*_{n1}), \ldots, \lambda^{\gamma}_{n,(p_n-1,p_n)}\mathrm{sgn}(\gamma^*_{n,(p_n-1,p_n)}))^{\intercal}\right\|^2 \\
&\leq s_n\Big[\max\{\lambda^{\beta}_{nj}, \lambda^{\gamma}_{n,kk'} : j \in \mathcal{A}_{n1}, (k,k') \in \mathcal{A}_{n2}\}\Big]^2 \\
&= s_n a_n^2 = s_n o(1/nq_n) \\
&= o(1/n).
\end{aligned}$$

Third, based on (2.15), it can be shown that

$$\begin{aligned}
\|B_3\|^2 &\leq \|\boldsymbol{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) + \frac{1}{n}\nabla^2_{\mathcal{A}_n} L_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\|^2 \|\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}\|^2 \\
&= o_p(1/q_n^2)O_p(q_n/n) = o_p(1/nq_n) \\
&= o_p(1/n).
\end{aligned}$$

Therefore,

$$B_1 + B_2 + B_3 = o_p(n^{-1/2}).$$

(II) Now we show $\sum_{i=1}^n \boldsymbol{Y}_{ni} + o_p(1) \to_d N(\boldsymbol{0}, \boldsymbol{G})$ where $\boldsymbol{Y}_{ni} = \frac{1}{\sqrt{n}}\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})$ $\Big[\nabla_{\mathcal{A}_n} L_{ni}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\Big]$. It is enough to show that $\boldsymbol{Y}_{ni}$, $i = 1, \ldots, n$ satisfies the conditions for Lindeberg-Feller central limit theorem [57]. For any given $\epsilon > 0$, by Cauchy-Schwarz inequality,

$$\begin{aligned}
\sum_{i=1}^n E\Big[\|\boldsymbol{Y}_{ni}\|^2 I\{\|\boldsymbol{Y}_{ni}\| > \epsilon\}\Big] &= nE\Big[\|\boldsymbol{Y}_{n1}\|^2 I\{\|\boldsymbol{Y}_{n1}\| > \epsilon\}\Big] \\
&\leq n\Big[E\|\boldsymbol{Y}_{n1}\|^4\Big]^{1/2}\Big[E(1\{\|\boldsymbol{Y}_{n1}\| > \epsilon\})\Big]^{1/2} \\
&= n\ B_4^{1/2}\ B_5^{1/2}.
\end{aligned}$$

$$\begin{aligned}
B_4 &= \frac{1}{n^2}E\|\boldsymbol{A}_n\boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\nabla_{\mathcal{A}_n} L_{ni}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\|^4 \\
&\leq \frac{1}{n^2}\|\boldsymbol{A}_n^{\intercal}\boldsymbol{A}_n\|^2\|\boldsymbol{I}_n^{-1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\|^2 E[\nabla^{\intercal}_{\mathcal{A}_n} L_{n1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})\nabla_{\mathcal{A}_n} L_{n1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})]^2 \\
&= \frac{1}{n^2}\lambda^2_{\max}(\boldsymbol{A}_n^{\intercal}\boldsymbol{A}_n)\lambda^2_{\max}(\boldsymbol{I}_n^{-1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}))O(s_n^2) \\
&= O(q_n^2/n^2).
\end{aligned}$$

By Markov inequality,

$$\begin{aligned} B_5 &= P(\|\boldsymbol{Y}_{n1}\| > \epsilon) \\ &\leq \frac{E\|\boldsymbol{Y}_{n1}\|^2}{\epsilon^2} \\ &= O(q_n/n). \end{aligned}$$

Therefore,

$$\sum_{i=1}^{n} E\Big[\|\boldsymbol{Y}_{ni}\|^2 1\{\|\boldsymbol{Y}_{ni}\| > \epsilon\}\Big] = nO(q_n/n)O(\sqrt{q_n/n}) = o(1).$$

Moreover,

$$\begin{aligned} \sum_{i=1}^{n} \mathrm{Cov}(\boldsymbol{Y}_{ni}) &= n\mathrm{Cov}(\boldsymbol{Y}_{n1}) \\ &= \boldsymbol{A}_n \boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) E[\nabla_{\mathcal{A}_n} L_{n1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \nabla^{\intercal}_{\mathcal{A}_n} L_{n1}(\boldsymbol{\theta}^*_{n\mathcal{A}_n})] \boldsymbol{I}_n^{-1/2}(\boldsymbol{\theta}^*_{n\mathcal{A}_n}) \boldsymbol{A}_n^{\intercal} \\ &= \boldsymbol{A}_n \boldsymbol{A}_n^{\intercal} \to G. \end{aligned}$$

Since $\boldsymbol{Y}_{ni}$, $i = 1, \ldots, n$ satisfies the conditions for Lindeberg-Feller central limit theorem, we conclude $\sum_{i=1}^{n} \boldsymbol{Y}_{ni} + o_p(1) \to_d N(\boldsymbol{0}, \boldsymbol{G})$. $\square$

CHAPTER III

Penalized Regression Methods and Ranking Variables by Their Strength of Association with a Response

Recently regularization using various penalties has been proposed to improve the performance of prediction and variable selection. In this chapter, a different perspective on the performance for regularized regression methods is considered - ranking predictors according to their strength of association with a response. This perspective can be useful in highlighting the predictor variables that have the largest effect on a response. It can be practically useful in genetic mapping applications in that one might want to prioritize genetic variants based on their association with the trait of interest with taking account of the effects of other variants. Specifically, three regularization methods, ridge regression, the Lasso and the elastic net, are considered for ranking variables by effect size. First, by analyzing two- or three-predictor cases, the explicit situations are determined where L_1 or L_2 regularization improves, decreases, or has no effect on ranking performance. Then in the simulation studies, the ranking performance of the three methods were compared in 38 population models based on various tuning methods. Ridge regression based on L_2 regularization outperformed the two methods that involve the L_1 penalty especially when R^2 is low. We note that there are other literatures [44, 40] that also consider estimating ranks. They consider ranking the subjects such as teachers, schools and so forth based on their

performance using regression approach. However, our approach is different from their approach in that we consider ranking predictor variables based on their effects on the response in a regression.

3.1 Introduction

In a genome-wide association study (GWAS), univariate testing is a typical way to preliminarily find genetic variants that are potentially associated with a trait of interest. For example, log odds ratios can be used for qualitative traits and Pearson correlation coefficients for quantitative traits for univariate testing. After Z-scores or the p-values are obtained from the test, multiple testing adjustments are usually made to highlight important variants.

Once we find a set of genetic variants that are highly associated with the trait, we would want to continue by considering how those genetic variants are related to the trait in a multivariate way. In other words, we would want to know what is the "unique" effect of each variant on the trait when other variants are taken account of. Considering the effects of other variants, a variant that had a strong univariate association with the trait might turn out to be redundant. On the other hand, a variant that showed a weak univariate association could turn out to have a significant effect on the trait in a multivariate sense. So we set our goal to prioritize genetic variants according to their "unique" effects on the trait so that they could be used further investigation. Rephrasing our goal in a regression setting, it is to rank predictor variables according to their unique association with a response. Fitting a multiple linear regression could provide one possible answer for this.

However, there are some difficulties that reside in multiple linear regression: high correlations that exist between the variants. It is known that OLS performs poorly

when predictor variables are highly correlated. With highly correlated predictors, the OLS estimates tend to have high variance resulting in unstable estimates. Due to the challenges in applying multiple regression in this setting, alternative procedures are often used. For example, the genetic data can be reduced to a count of the number of high-risk genetic variants per subject, followed by a simple linear regression or correlation analysis with the trait [63, 36, 2]. While it has some utility for prediction, it has the disadvantage of failing to provide any insight into the potential complementary roles of the different genetic variants in terms of their influence on the trait.

The collinearity problem has been extensively studied in regression, especially in the context of prediction performance. Specifically, ridge regression moderates the collinearity problem. By controlling the squared L_2 norm of the regression coefficients while minimizing the squared error loss, ridge regression introduces the bias in estimating coefficient estimates but reduces the variance of the estimates. So the mean squared error (MSE) of the coefficient estimates can be substantially reduced when the predictor variables are moderately or strongly correlated.

A more recent development has been the introduction of new types of penalties that in some situations perform better than ridge regression. Two notable approaches are the Lasso [56], which uses an L_1 penalty in place of ridge regression's squared L_2 penalty, and the elastic net [72], which uses both L_1 and squared L_2 penalties. When using penalties involving the L_1 norm, some coefficient estimates can be exactly zero, allowing variable selection to be carried out as part of coefficient estimation. In addition, the introduction of exact zeros into the coefficient estimates leads to better predictive performance when the true regression coefficient vector is "sparse," meaning that it contains a substantial fraction of zero or negligible coefficients.

Penalized regression methods have most commonly been used when the primary goal is prediction. In genetic mapping studies, prediction can be an important goal, but the genetic contribution to a trait may be too low for prediction to be of practical use. A related but distinct goal is to understand which variants contribute unique information about the trait variation. While regression modeling must be used cautiously in this way (e.g. [5, 24]), it can nevertheless provide additional insight into the relationships between genetic variants and traits compared to looking exclusively at univariate relationships [41, 66, 55].

Our goal is to assess the performance of penalized regression methods for ranking variables according to their unique effects on the response, with a focus on situations where the R^2 is low and substantial collinearity is present. After setting up the problem and notation in Section 3.2, we consider the 2- and 3-dimensional cases in Section 3.3 to investigate in what settings L_2 regularization improves over OLS and in what settings L_1 regularization improves over L_2 regularization or vice versa. In Section 3.4, simulation results based on various sets of models are shown and the performance of regularization in effect ranking and prediction is compared for ridge regression, the Lasso and the elastic net. Section 3.5 discusses possible implications of our findings for data analysis.

3.2 Model Estimation and Variable Ranking

In this section, the methods that we used to rank variables are elaborated. Algorithms for estimating ridge regression, the Lasso and the elastic net are explained in Section 3.2.1, the criteria for choosing the regularization parameters are introduced in Section 3.2.2, and in Section 3.2.3, ranking based on those regression-based methods are discussed. Finally in Section 3.2.4, the criterion we used to evaluate the

ranking performance is introduced.

3.2.1 Regression Model Fitting

Let $Y = (Y_1, \ldots, Y_n)'$ denote the response vector with sample size n, and let $X \in \mathcal{R}^{n \times p}$ denote the design matrix containing the data on p predictor variables. For simplicity, and following convention in ridge regression, we will center Y and all columns of X, and fit all regression models without an intercept. Ridge regression, the Lasso, and the elastic net estimate β by minimizing the following loss functions (3.1)-(3.3), respectively

$$\|Y - X\beta\|_2^2 + \lambda_2 \|\beta\|_2^2 \tag{3.1}$$

$$\|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 \tag{3.2}$$

$$\|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2. \tag{3.3}$$

The L_1 and squared L_2 norms are defined in the usual way: $\|\beta\|_2^2 = \sum_j \beta_j^2$ and $\|\beta\|_1 = \sum_j |\beta_j|$. Since the loss functions are convex, the solutions are unique. For fixed tuning parameter values, coefficient estimates for the elastic net (3.3) were obtained using a cyclical coordinate descent method introduced in Friedman et al. [25]. The Lasso can be solved using the same method since it is a special case of the elastic net when $\lambda_2 = 0$. Cyclical coordinate descent starts with initial values of $\hat{\beta}_j$. Then for each $j = 1, \ldots, p$, $\hat{\beta}_j$ is updated by minimizing (3.3) while fixing the values of all other coefficients. This update has a simple closed form. After all

coefficients are updated, the algorithm repeats the cyclical update until it converges. For speedup and stability, we start from very large λ_1 and λ_2 values so that all $\hat{\beta}_j$ are zero, and for decreasing λ values, we use the solutions for the next largest tuning parameter as an initial value when calculating the solution for the current tuning parameters (called pathwise coordinate descent by Friedman et al. [25]). For further speedup, we restricted the updates to nonzero coefficient estimates between complete cyclic updates, following [25].

3.2.2 Tuning

To set the value of the tuning parameter λ_2 in ridge regression, generalized cross validation (GCV) [28] is recognized as performing well, and is the only approach considered here. For the Lasso and elastic net, there is no clearly favored approach for setting λ_1 and/or λ_2. we considered AIC, BIC, and a tuning set approach. For AIC and BIC, a Gaussian likelihood with constant error variance and independent errors was used. The degrees of freedom was the effective degrees of freedom for ridge regression $\mathrm{tr}[(X'X + \lambda I)^{-1}X'X]$, where X contains only the columns corresponding to non-zero coefficient estimates. For the tuning set approach, two independent data sets of the same size were generated, and the tuning parameter was set to the value that optimized the prediction MSE on the second data set when estimating coefficients on the first data set. We expect the tuning set approach to give results that are somewhat similar to cross-validation, which we did not use due to the high computational cost of doing cross-validation in an extensive simulation study. To actually carry out the tuning, we calculated the criterion (e.g. GCV) at each point in fixed, finite sets of values. These sets were $\Lambda_1 = \{0, 10^{-4}, 10^{-3}, \ldots, 10^3, 10^4, 10^5\}$ for λ_1, and $\Lambda_2 = \{0, 10^{-4}, 10^{-3}, \ldots, 10^3, 10^4, 10^5\}$ for λ_2. For the elastic net, the Cartesian product $\Lambda_1 \times \Lambda_2$ (i.e. every pair of values) was considered.

3.2.3 Ranking Regression Effects

Our goal is to rank the variables according to population regression effects β_j (for "signed analysis") or $|\beta_j|$ (for "magnitude analysis"). For estimating the ranking based on data, coefficient estimates $\hat{\beta}_j$ or $|\hat{\beta}_j|$ can be used but also one could consider using some standardized quantity. This will be discussed further below.

Although it is well known that prediction performance benefits from regularization, it is not obvious how regularization affects the ranking performance. Under regularization, the variance of the difference between two coefficient estimates is reduced but at the same time, the size of the expected value of the estimate difference is also reduced. Therefore, the benefit of regularization on the ranking performance depends on the rate at which those two values shrink.

We assume that predictor variables are standardized so that they have zero mean and unit variance. Thus, $\beta_j/\mathrm{SD}(X_j)$ can be interpreted as the expected change in the response for each unit change in the original predictor.

We can use coefficient estimates $\hat{\beta}_j$ to rank regression effects, but also we can consider Z-scores $\hat{\beta}_j/\mathrm{SD}(\hat{\beta}_j)$. The motivation for considering Z-scores is that in some situations using Z-scores might serve to control the estimation variance of coefficient estimates, although the rank based on the expected Z-scores may be different from the rank based on the population coefficients β_j. Below, we find it is possible that ranking by Z-scores can be more precise than ranking by coefficient estimates.

To compare the accuracy of ranking by Z-scores and coefficient estimates, we consider a simple case with two predictors where β_1 and β_2 are the true coefficients and $\hat{\beta}_1$ and $\hat{\beta}_2$ are their estimates. By translation and scaling, it is sufficient to consider the case where $\beta_2 = 1$, $|\beta_1| < 1$, $\mathrm{var}(\hat{\beta}_1) = 1$, and $\mathrm{var}(\hat{\beta}_2) = \tau^2$. Our question is whether

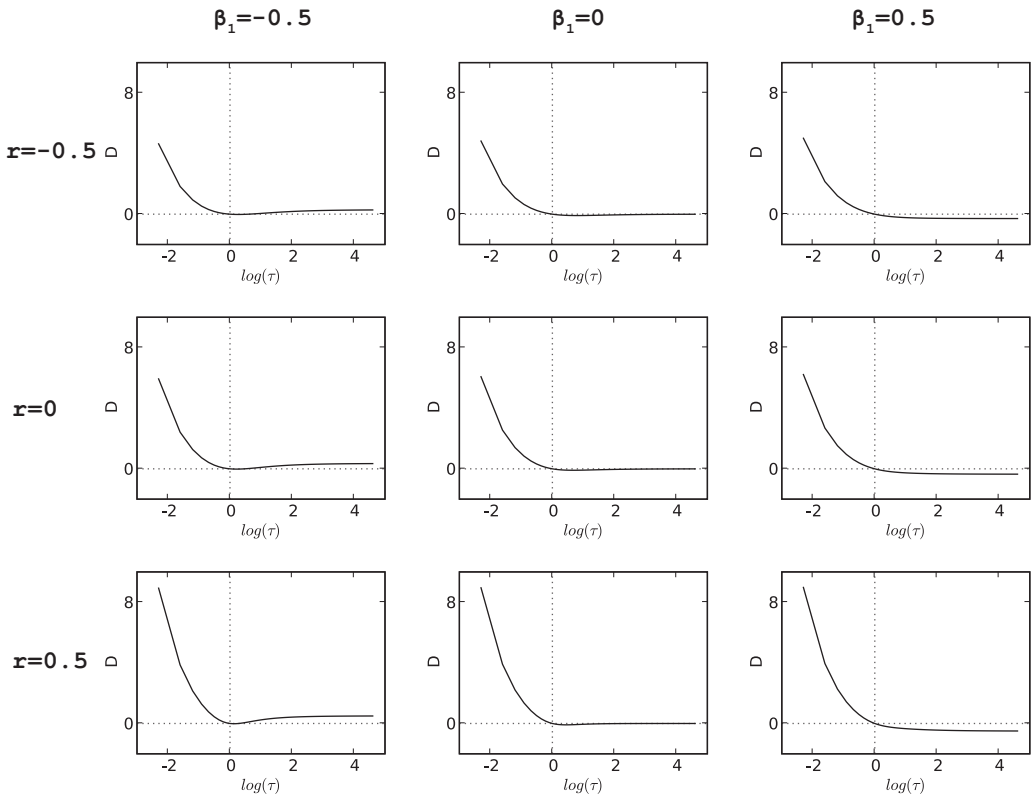


Figure 3.1: Comparison of using Z-scores versus using coefficient estimates for ranking. The figure shows the difference D in (3.5) or (3.6) versus $\log(\tau)$. The three rows correspond to when $r = -0.5$, $r = 0$ and $r = 0.5$, and the three columns correspond to when $\beta_1 = -0.5$, $\beta_1 = 0$ and $\beta_1 = 0.5$

$$P(\hat{\beta}_2 > \hat{\beta}_1) < P(\hat{\beta}_2/\tau > \hat{\beta}_1), \tag{3.4}$$

implying that Z-scores have a higher probability of getting a correct ranking than coefficient estimates. Assuming that $\hat{\beta}_1$ and $\hat{\beta}_2$ are OLS estimates, they are unbiased and approximately normally distributed. We first assume that the two estimates are uncorrelated. Treating the probabilities in (3.4) as normal, we can reformulate (3.4) as (3.5).

$$D \equiv \frac{\tau^{-1} - \beta_1}{\sqrt{2}} - \frac{1 - \beta_1}{\sqrt{1 + \tau^2}} > 0. \tag{3.5}$$

It can be easily shown that (3.5) holds whenever $\tau < 1$, and in some circumstances (depending on the value of β_1), (3.5) also holds when $\tau > 1$. The second row of Figure 3.1 illustrates this result. They show the difference D in (3.5) versus $\log(\tau)$ in uncorrelated cases. We can see that D is positive when $\log(\tau) < 0$ in all three plots. And when β_1 is negative as in the left plot, D is non-negative regardless of τ.

Second, we consider what happens when the two estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are correlated. When the correlation between the two estimates is r, it is sufficient to consider whether

$$D \equiv \frac{\tau^{-1} - \beta_1}{\sqrt{2 - 2r}} - \frac{1 - \beta_1}{\sqrt{1 + \tau^2 - 2r\tau}} > 0 \tag{3.6}$$

to show Z-scores should be used to estimate the ranking. By numerical experiments, we can obtain similar results for correlated cases. The first row and the third row of Figure 3.1 illustrate those results. They show the difference D in (3.6) versus $\log(\tau)$. We can see that D is positive when $\log(\tau) < 0$ in all three plots. And when

β_1 is negative as in the left plot, D is non-negative regardless of τ. It suggests that ranking by Z-scores is more accurate than ranking by coefficient estimates in wider variety of situations.

The analysis above is based on signed analysis. Since it is hard to study magnitude analysis analytically due to the effect of the absolute value signs, we used simulation to compare the performance of z-scores and coefficient estimates. Similar to signed analysis, we found that Z-scores perform better than coefficient estimates when $\tau < 1$, that is, the variance of the larger effect is smaller than that of the smaller effect. It is different from signed analysis that coefficient estimates always seems to perform better than Z-scores when $\tau > 1$. So for magnitude analysis, the situation is more balanced and there is no clear advantage or disadvantage to either of the two approaches. For further analysis below, we consider both ranking by Z-scores and ranking by coefficient estimates. We emphasize that even when Z-scores are used to estimate ranking, we evaluate the performance based on accurate ranking of the actual effects β_j or $|\beta_j|$, not on ranking of the expected Z-scores or the expected magnitudes of Z-scores.

As a technical point, to calculate the Z-scores we need to be able to calculate the standard errors of the coefficient estimates. For the elastic net, the standard error of nonzero coefficient estimates are calculated using the sandwich formula following Fan and Li [21].

$$\widehat{\mathrm{cov}}(\hat{\beta}) = (X'X + \lambda_2 I)^{-1} X'X (X'X + \lambda_2 I)^{-1} \hat{\sigma}^2, \tag{3.7}$$

where X contains only the columns corresponding to non-zero coefficient estimates. The residual variance σ^2 is estimated as the sum of squared residuals divided by $n - \mathrm{df}$, where df is the degrees of freedom as defined above. The standard errors

for ridge regression and the Lasso can be calculated from (3.7) as special cases. For the Lasso and elastic net, a tolerance threshold of 10^{-6} was set such that coefficient estimates smaller in magnitude than the threshold were deemed to be exact zeros, and were not standardized when calculating Z-scores.

3.2.4 Performance Evaluation

The main criterion for ranking performance evaluation was following the concordance score (CS), which is closely related to the Mann-Whitney formulation of the the area under the Receiver Operating Characteristics (ROC) curve. For signed analysis with the approach based on Z-scores, the CS is defined as

$$\frac{\sum_{i \neq j} \Big[\mathcal{I}(z_i > z_j) \cdot \mathcal{I}(\beta_i > \beta_j) + \mathcal{I}(z_i = z_j) \cdot \mathcal{I}(\beta_i > \beta_j) \cdot 0.5 \Big]}{\sum_{i \neq j} \mathcal{I}(\beta_i > \beta_j)}, \tag{3.8}$$

where z_j is the Z-score (the coefficient estimate divided by its standard error) for the effect of each predictor variable. For the approach based on coefficient estimates, z_j are replaced with coefficient estimates $\hat{\beta}_j$. For magnitude analysis, Z-scores z_j, coefficients β_j and coefficient estimates $\hat{\beta}_j$ are replaced with their magnitudes.

Note that a CS of 1 corresponds to perfect ranking whereas a CS of $1/2$ is expected from random guessing. Also, note that even when the Lasso estimates for a pair of coefficients with different values are exactly zeros, the CS still gets 0.5 as it would get in the case of random guessing. In the sense that the CS does not decrease due to the ties resulting from zero estimates, the CS is fair to L_1 and L_2 regularization. We considered ranking performance under both "oracle tuning," in which the CS was maximized over the set of tuning parameters, and ranking performance using data-adaptive tuning criteria such as AIC, BIC, and GCV. The distribution of CS values for independent data sets was approximated using simulation. Pairs of methods were

compared (e.g. the Lasso compared to ridge regression) based on the distribution of differences in CS values for two methods using the same underlying data set.

For comparison, we also considered predictive performance, based on the mean squared prediction error on a large ($n = 10,000$) independent validation set. In this case, we also considered oracle tuning, which optimized the prediction error on the validation set over the tuning parameters, and tuning using the data-driven criteria. We compared methods based on a relative MSE score

$$\text{rMSE} = \frac{\text{MSE}_1 - \text{MSE}_2}{\text{MSE}_{\text{ENO}}}, \tag{3.9}$$

where MSE_1 and MSE_2 are the MSE values for the two methods, and MSE_{ENO} is the MSE for the elastic net using oracle tuning. Within the class of methods considered here, MSE_{ENO} is the smallest MSE that can be achieved, but it is still larger than the residual variance $\text{var}(Y|X)$.

3.3 Analytic and Numerical Results for Two, Three and Higher Dimensions

In this section, we consider the cases where only two or three predictors exist in the model to better understand how the regularization of ridge regression and the Lasso affect the ranking performance. Also, we generalize the results to higher dimensions when it is possible. In Section 3.3.1, we examine how ridge regression influences the accuracy of ranking estimation compared to OLS and in Section 3.3.2, ridge regression is compared with the Lasso.

3.3.1 Comparison of Ridge Regression and OLS

The simulation studies in Section 3.4 show that the ranking performance of ridge regression is generally good when using data-adaptive tuning. While we note it is

possible that ridging can help us to accurately estimate the ranking, one might still wonder in what circumstances the regularization by ridge regression improves the ranking performance. One might expect that ridging would help us to accurately rank coefficients when predictors are highly correlated because OLS is expected to perform poorly in that situation. In the analytical assessment below, however, we found that multicollinearity is not a sufficient condition for ridging to improve the ranking performance. The specific situations where ridging improves OLS are not clearly known. In this section, we consider broad ranges of situations with various correlation structures of predictors and relative effect sizes and investigate how the benefit of ridging changes with different settings. We focus on signed analysis based on coefficient estimates in this section, in which OLS and ridge regression can be studied analytically.

Below, we show that the effect of ridge regression is different for two-predictor case and three-predictor case: ridge regression has no effect on ranking performance in the two-predictor case, while in the three-predictor case, the effect of ridge regression is complicated and depends on how the third variable relates to the first two variables. Then, the results are generalized to higher dimensions. Note that these results are true only when we assume all predictors are standardized so they have zero mean and unit variance [9].

For the analysis below, assume a linear model in which the error term is independent and identically distributed with normal distribution of mean zero and variance σ^2. Let $\hat{\beta}_{j\lambda}$ be the estimate of the coefficient β_j by ridge regression with a tuning parameter λ. Assuming $\beta_j > \beta_k$, let $P_{jk} = P(\hat{\beta}_{j\lambda} > \hat{\beta}_{k\lambda})$ can be calculated using the bivariate normal distribution of $(\hat{\beta}_{j\lambda}, \hat{\beta}_{k\lambda})$, providing the probability that a correct ranking occurs. Note that the joint distribution of coefficient estimates does not nec-

essarily have to be bivariate normal, but we just need the assumption that all linear combinations of coefficient estimates follow a common location/scale family. Also, note that the CS is the average of these probabilities over all pairs where $\beta_j > \beta_k$. Analytic calculations in this section were done using a computer algebra software.

For the two-predictor case, assume that the predictor cross-product matrix has the form

$$X'X = n \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix},$$

where n is the sample size and r is the correlation coefficient between the two predictors. Without loss of generality, we assume $\beta_1 > \beta_2$. This situation is depicted in Figure 3.2a.

The probability of obtaining a correct ranking, $P_{12} = P(\hat{\beta}_{1\lambda} > \hat{\beta}_{2\lambda})$, can be analytically calculated based on the bivariate normal sampling distribution of $(\hat{\beta}_{1\lambda}, \hat{\beta}_{2\lambda})$. It can be easily shown that P_{12} is $\Phi(T_{12})$, where Φ is the standard normal cumulative distribution function and

$$T_{12} = E(\hat{\beta}_{1\lambda} - \hat{\beta}_{2\lambda})/SD(\hat{\beta}_{1\lambda} - \hat{\beta}_{2\lambda}) \tag{3.10}$$

$$= (\beta_1 - \beta_2)\frac{\sqrt{n(1-r)}}{\sqrt{2}\sigma}. \tag{3.11}$$

Note that T_{12} does not depend on λ, which means ridging does not affect the probability of obtaining a correct ranking. In the case of two standardized predictors, the standard error for $\hat{\beta}_{1\lambda}$ and $\hat{\beta}_{2\lambda}$ are the same. Therefore, using either coefficient estimates or Z-scores to estimate the ranks gives the same result for the probability of getting a correct ranking.

As mentioned earlier in Section 3.2.3, both the expected value and the standard deviation of $\hat{\beta}_{1\lambda} - \hat{\beta}_{2\lambda}$ in (3.10) decreases in magnitude as λ increases. Since the numerator and the denominator shrink at the same time, whether regularization

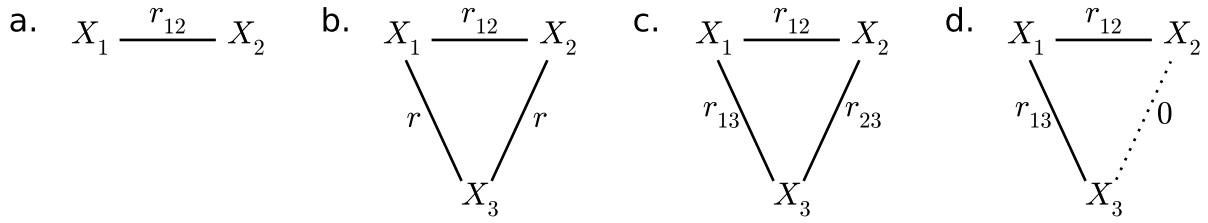


Figure 3.2: Schematic depiction of covariate relationships that influence how ridging affects CS. In **a** and **b**, ridging has no effect. In **c**, ridging can improve, decrease, or have no effect on ranking performance depending on model parameters. In **d**, ridging can improve the performance if coefficient estimates are compared, but has no effect if Z-scores are compared.

improve ranking accuracy depends on the speeds at which the two values decrease. In the case of two predictors above, it is not only that they decrease at the same rate but also the terms related to λ in the numerator and the denominator are canceled.

Next we consider the case of three predictors. First, a special situation is considered where the third variable is equally correlated to each of the first two variables. In that case, the predictor cross-product matrix has the form

$$X'X = n \begin{pmatrix} 1 & r & s \\ r & 1 & s \\ s & s & 1 \end{pmatrix}.$$

This situation is depicted in Figure 3.2b. In this case the probability of obtaining a correct rank for the pair of β_1 and β_2 is again $\Phi(T_{12})$ with T_{12} as in (3.11). Thus, ridging does not have any effect on ranking the two variables as in two-predictor case when the third variable has the same relationships with each of the first two variables. Note that the standard errors for $\hat{\beta}_{1\lambda}$ and $\hat{\beta}_{2\lambda}$ are again same, so ranking by Z-scores and ranking by coefficient estimates give identical results.

It can be analytically shown that this result can be generalized to high dimensional cases. When $r_{1k} = r_{2k}$ for $k \neq 1, 2$, the probability of correctly ranking β_1 and β_2

is $\Phi(T_{12})$ in which T_{12} has the same form as (3.11). More generally, if $r_{ik} = r_{jk}$ for $k \neq i, j$, then the probability of correctly ranking β_i and β_j does not depend on L_2 regularization because $T_{ij} = (\beta_i - \beta_j)\sqrt{n(1 - r_{ij})}/\sqrt{2}\sigma$ does not depend on λ.

To analytically show that T_{12} does not depend on λ when $r_{1k} = r_{2k}$ for $k \neq 1, 2$ in high dimensional cases, we begin by blocking $X'X$ as follows

$$X'X = \begin{pmatrix} K & A' \\ A & J \end{pmatrix},$$

where K is a 2×2 matrix,

$$K = \begin{pmatrix} n & nr_{12} \\ nr_{12} & n \end{pmatrix},$$

A is a $(p-2) \times 2$ matrix with two identical columns, and J is a $(p-2) \times (p-2)$ strictly positive definite matrix. Also, we can rewrite T_{12} as

$$T_{12}(\lambda) = \left[\frac{\beta' X'X(X'X + \lambda I)^{-1} DD'(X'X + \lambda I)^{-1} X'X\beta}{D'(X'X + \lambda I)^{-1} X'X(X'X + \lambda I)^{-1} D} \right]^{1/2},$$

where $D = (1, -1, 0, \ldots, 0)'$. We claim that $T_{12}(\lambda)$ is a constant function of λ when $X'X$ has the structure given above.

First, we will use a change of variables from X to $Z = XQ$ to simplify the problem. Note that if Q is any orthogonal matrix such that $QD = D$, then the denominator of T_{12} is unchanged. This follows by direct calculation. The condition $QD = D$ will be satisfied if Q has the form

$$Q = \begin{pmatrix} I_2 & 0 \\ 0 & Q_2 \end{pmatrix}$$

where Q_2 is a $(p-2) \times (p-2)$ orthogonal matrix. Note also that if we apply such a transform, then $Z'Z$ has the same structure as $X'X$ with regard to the matrix A

having two identical columns.

$$Z'Z = \begin{pmatrix} K & AQ_2 \\ Q_2'A' & JQ_2 \end{pmatrix}$$

Assume that we factor $X_2 = USV'$ using the singular value decomposition where X_2 is a $n \times (p-2)$ matrix that contains all variables other than the first two variables. If we take $Q_2 = V$, then the block representation of $Z'Z$, as above, has the property that JQ_2 is diagonal. Thus, if we show that T_{12} is constant in λ when JQ_2 is diagonal, it will follow that T_{12} is constant as a function of λ for all J. Using the inversion of block matrices and algebraic calculations, we can show that

$$T_{12}(\lambda) = \left[\frac{\beta' Q Z' Z (Z'Z + \lambda I)^{-1} D D' (Z'Z + \lambda I)^{-1} Z' Z Q' \beta}{D'(Z'Z + \lambda I)^{-1} Z' Z (Z'Z + \lambda I)^{-1} D} \right]^{1/2},$$

is a constant function of λ given that $Z'Z$ has the structure as above. Therefore, we now proved that ridging does not affect the probability of correctly ranking β_1 and β_2 when $r_{1k} = r_{2k}$ for $k \neq 1, 2$. Similarly, we can prove that ridging does not affect the probability of correctly ranking β_i and β_j when $r_{ik} = r_{jk}$ for $k \neq i, j$. In a special case where all r_{ij} are the same for $i \neq j$, all $T_{ij}(\lambda)$ will be constant functions of λ, so riding would not affect the overall ranking performance.

Returning to the 3-predictor case, as depicted in Figure 3.2c, now we allow a cross-product matrix to be general as in the form

$$X'X = n \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{pmatrix}.$$

In this general setting, the benefit of ridging becomes more clear. Now the expression for T_{12} is complicated and depends on all six parameters including three parameters in β and three parameters in $X'X$ as well as the regularization parameter λ. The

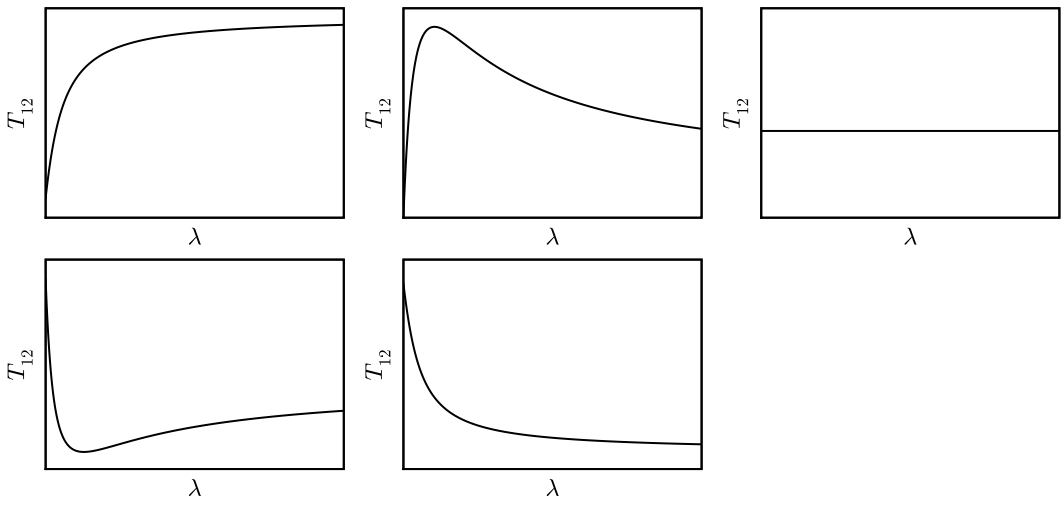


Figure 3.3: Shapes of $T_{12}(\lambda)$ as a function of λ. $T_{12}(\lambda)$ can have various shapes depending on the model parameters including correlations between predictors and true effects.

standard errors of $\hat{\beta}_{1\lambda}$ and $\hat{\beta}_{2\lambda}$ are no longer identical in the general case. In more detailed discussion below, we focus on ranking by coefficient estimates rather than Z-scores for simplicity.

By using a computer algebra software, we calculated $T_{12}(\lambda)$ in general 3-predictor case. Omitting the lower order terms in λ, $T_{12}(\lambda)$ has the form

$$T_{12}(\lambda) = \frac{\sqrt{n}\Big(\Big\{(\beta_1-\beta_2)(1-r_{12})+\beta_3(r_{13}-r_{23})\Big\}\lambda^2+\ldots\Big)\Big/\Big(\lambda^3+\ldots\Big)}{\sigma\sqrt{\Big(2(1-r_{12})\lambda^4+\ldots\Big)\Big/\Big(\lambda^3+\ldots\Big)^2}}. \tag{3.12}$$

Note that as λ goes to infinity, $T_{12}(\lambda)$ converges to $\frac{\sqrt{n}}{\sigma}\Big((\beta_1-\beta_2)(1-r_{12})+\beta_3(r_{13}-r_{23})\Big)\Big/\sqrt{2(1-r_{12})}$. Also, note that as the sample size n goes to infinity, $T_{12}(\lambda)$ converges to infinity, so the probability of getting a correct ranking becomes 1, which is naturally expected.

Depending on the model parameters in $T_{12}(\lambda)$, the value of $T_{12}(\lambda)$ has various shapes as a function of λ. Figure 3.3 illustrates the shapes of $T_{12}(\lambda)$. The value of

$T_{12}(\lambda)$ can increase, increase then decrease, stay the same, decrease then increase, or decrease as λ grows. In any case, it converges to a certain value as shown above.

In order to figure out when ridging can improve the performance, we first look at the sign of $T_{12}(\infty) - T_{12}(0)$. If the sign of $T_{12}(\infty) - T_{12}(0)$ is positive, it would mean that at least, the extreme case of ridge regression (this is equivalent to univariate analysis) has better ranking performance than OLS. We note that the approach of looking at the sign of $T_{12}(\infty) - T_{12}(0)$ can easily miss the cases where some amount of ridging is helpful but the univariate analysis is worse than OLS. But it could at least serve as a lower bound of the cases where ridging improves the ranking accuracy.

Although it is straightforward to compute $T_{12}(\infty) - T_{12}(0)$ based on (3.12), we re-expressed $T_{12}(\lambda)$ by representing β in spherical coordinate to better understand and determine the situations where L_2 regularization improves the ranking performance. As β is reformulated as

$$\begin{aligned} \beta_1 &= b\cos\phi\sin\theta, \\ \beta_2 &= b\sin\phi\sin\theta, \qquad \text{where b} = \|\beta\| \\ \beta_3 &= b\cos\theta, \end{aligned}$$

$T_{12}(\lambda)$ can be expressed as a function of the two key quantities, $D = r_{23} - r_{13}$ and $Q = (\beta_1 - \beta_2)/\beta_1$ and several other quantities defined below. The importance of D is related to the special situation described above: ridging has no effect on ranking performance when $r_{13} = r_{23}$, i.e., $D = 0$. Q defines the relative difference between the two effects β_1 and β_2 and is intuitively relevant to ranking performance. With the spherical representation, the quantity Q can be represented as $1 - \tan\phi$. In addition to D and Q, $T_{12}(\lambda)$ depends on r_{12}, $M = (r_{13} + r_{23})/2$, b, θ, n and σ^2. Note that the sign of $T_{12}(\lambda)$ does not depend on $b \geq 0$. Moreover, the sign of $T_{12}(\lambda)$ does

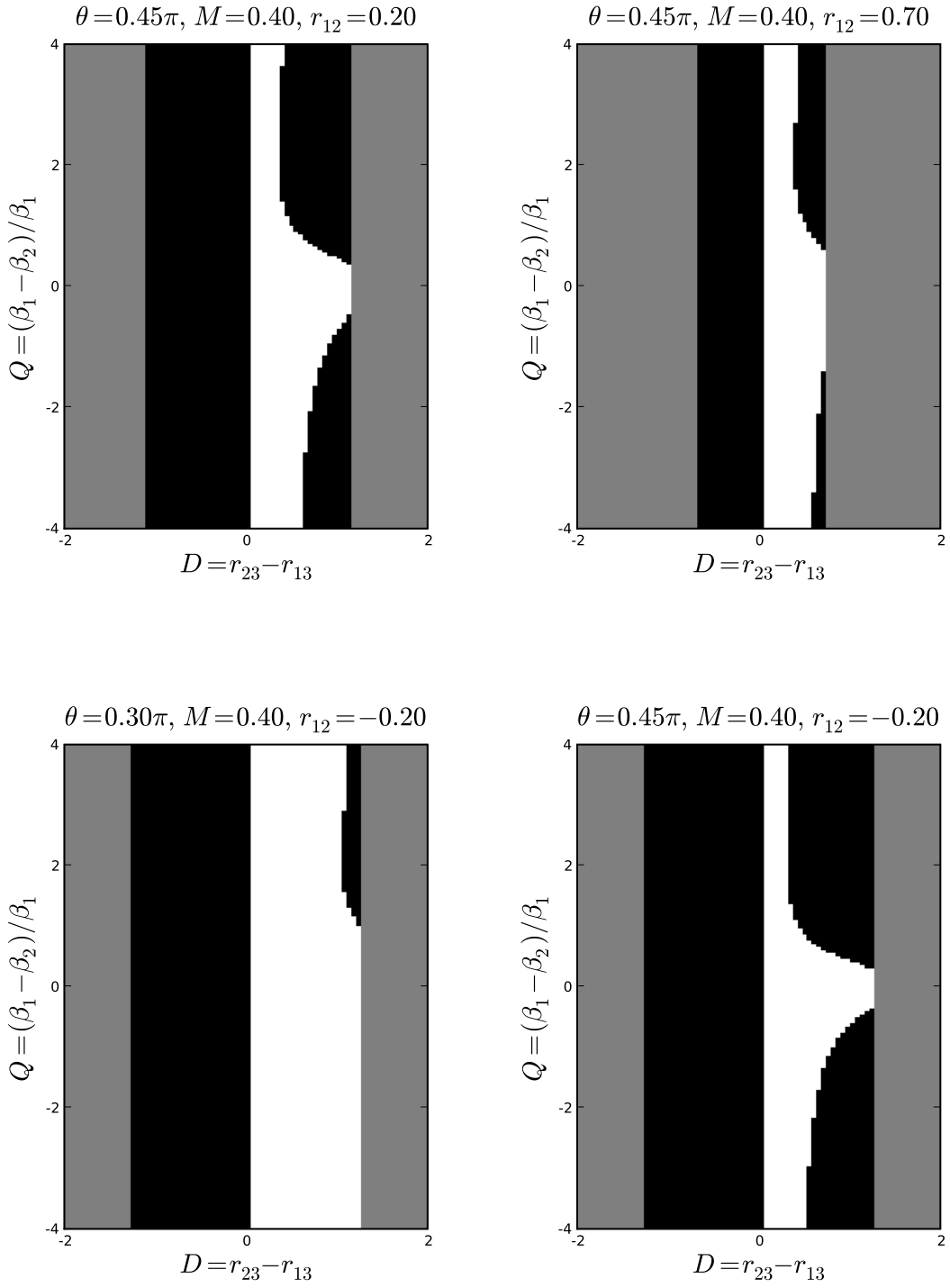


Figure 3.4: Plots of the sign of $T_{12}(\infty) - T_{12}(0)$ on the plane of D versus Q. The black region represents the cases when the sign of $T_{12}(\infty) - T_{12}(0)$ is positive implying ridging improves the ranking accuracy; the white region represents negative cases; the grey region represents the infeasible cases due to the non-positive definiteness of $X'X$.

not depend on n and σ^2 as n and σ^2 only appear as a factor of $\sqrt{n}/\sigma$ in $T_{12}(\lambda)$. Therefore, the sign of $T_{12}(\lambda)$ only depends on D, Q, r_{12}, M and θ.

Besides D and Q, $T_{12}(\lambda)$ also depends on M and θ. M is the average correlation between X_3 and the two variables X_1 and X_2 that we focus on. And θ can be interpreted as a component that regulates the effect size of β_3 relative to the effect sizes of β_1 and β_2. Intuitively, one would expect that ridging would be less beneficial when the relative effect size of β_3 is small compared to the effect sizes of β_1 and β_2, since otherwise ridging could improve the ranking performance when adding a third variable that has no effect at all.

Returning to the question of determining in what situation ridging improves the performance, a few examples of the numerical results on the sign of $T_{12}(\infty) - T_{12}(0)$ are shown in Figure 3.4 based on the parameterizations explained above. It shows the sign of $T_{12}(\infty) - T_{12}(0)$ on the plane of D versus Q. The black region represents the cases when the sign of $T_{12}(\infty) - T_{12}(0)$ is positive, implying that ridging improves the ranking accuracy; the white region represents when it is negative implying ridging harms or doesn't improve the ranking performance; the grey region represents the infeasible cases due to the non-positive definiteness of $X'X$. The first row shows the results for two different r_{12} values for fixed θ and M; the second row shows the results for two different θ values when others are fixed.

By algebraic calculation, it can be shown that the sign of $T_{12}(\infty) - T_{12}(0)$ depends on a quadratic function with respect to D when other values are fixed. We can confirm this in Figure 3.4: there are at most two sign changes along any horizonal lines. Furthermore, we can see that there is always a sign change at $D = 0$. It is related to the fact that ridging has no effect when the two correlations r_{13} and r_{23} are identical as shown in the special case above. It can also be algebraically shown

that $T_{12}(\infty) - T_{12}(0)$ is zero whenever $D = 0$. Looking at the first row of Figure 3.4, we can see that for different r_{12} values, the ranking performance of ridge regression is different as well as the feasible area based on the positive definiteness of $X'X$. The second row of Figure 3.4 shows results for $\theta = 0.30\pi$ and $\theta = 0.45\pi$. When $\theta = 0.30\pi$, the relative effect size of β_3 compared to the effect sizes of β_1 and β_2 is smaller than when $\theta = 0.45\pi$. As discussed above, ridging is expected to be less beneficial when the relative effect size of β_3 is small. We can confirm this in Figure 3.4: the black region is larger when $\theta = 0.45\pi$ compared to $\theta = 0.30\pi$.

As noted above, the approach of looking at the sign of $T_{12}(\infty) - T_{12}(0)$ only focuses on the extreme version of ridge regression when $\lambda = \infty$ and can miss the cases where ridging improves the accuracy with a smaller λ but decreases the accuracy with a larger λ. So now we consider when ridging would improve the ranking performance with relatively small values of λ. The sign of the derivative of $T_{12}(\lambda)$ with respect to λ at $\lambda = 0$, i.e. $\text{sgn}[T'_{12}(0)]$, can provide one answer to that question because it would reveal whether $T_{12}(\lambda)$ increases or decreases at small values of λ. If either $T_{12}(\infty) - T_{12}(0)$ or $T'_{12}(0)$ is positive, ridge regression would be guaranteed to improve the ranking performance with some value of λ, although still we could miss some cases where ridging improve the accuracy. However, a large fraction of the situations where ridging improve the performance for some value of λ have at least one of $T_{12}(\infty) - T_{12}(0)$ or $T'_{12}(0)$ being positive.

Figure 3.5 shows the example results on the sign of $T'_{12}(0)$. It shows the sign of $T'_{12}(0)$ on the plane of D versus Q. As in the results on $T_{12}(\infty) - T_{12}(0)$, the black region represents the cases when $T'_{12}(0)$ is positive implying ridging improve the ranking performance for small λ and the white region represents the cases when $T'_{12}(0)$ is negative implying ridging doesn't improve the ranking performance for small

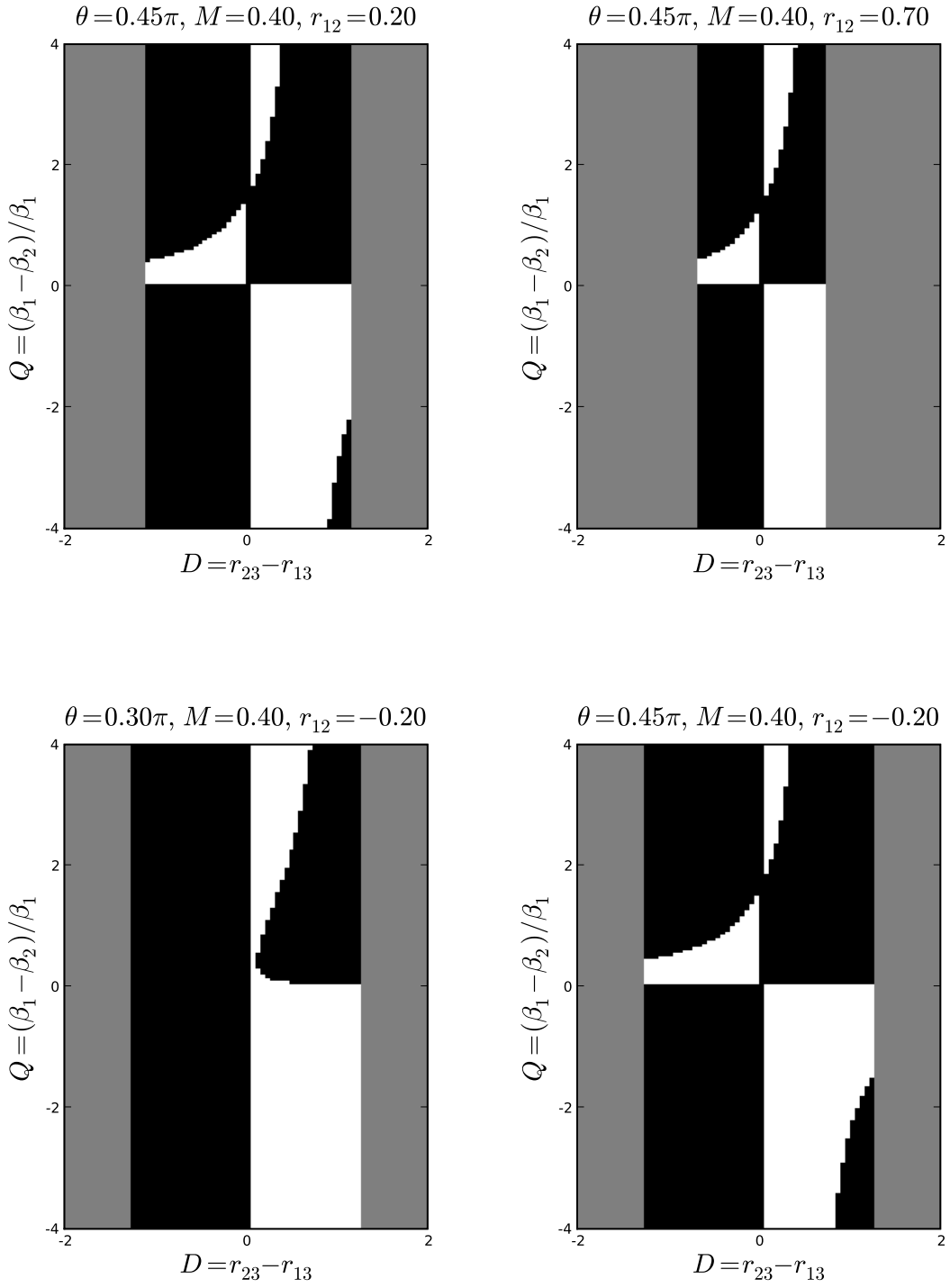


Figure 3.5: Plots of the sign of $T'_{12}(0)$ on the plane of D versus Q. The black region represents the cases when the sign of $T'_{12}(0)$ is positive implying ridging improve the ranking accuracy; the white region represents negative cases; the grey region represents the infeasible cases due to the non-positive definiteness of $X'X$.

λ. The first row shows the results for two different r_{12} values for fixed θ and M; the second row shows the results for two different θ values when others are fixed.

Similar to $T_{12}(\infty) - T_{12}(0)$ case, it can be shown that the numerator of $T'_{12}(0)$ is a quadratic function with respect to D when other values are fixed. The denominator of $T'_{12}(0)$ does not affect the sign of $T'_{12}(0)$ because it is always positive. In Figure 3.5, we can confirm the sign of $T'_{12}(0)$ depends on a quadratic function of D: there are at most two sign changes along any horizonal lines. Furthermore, we can see that there is always a sign change at $D = 0$, which corresponds to the special case where ridging has no effect.

Looking at the first row of Figure 3.5, we can see that for different r_{12} values, the ranking performance of ridge regression is different as well as the feasible area based on the positive definiteness of $X'X$. The second row of Figure 3.5 shows results for $\theta = 0.30\pi$ and $\theta = 0.45\pi$. As discussed above, ridging is expected to be less beneficial when the relative effect size of β_3 is small. It is not necessarily true in the comparison of the second row in Figure 3.5, but it can be true when we consider the union of the black regions of Figure 3.4 and Figure 3.5.

Those figures were chosen to illustrate a few examples when ridging would improve the ranking performance, so they do not show all possible cases. To further summarize the results, the area of the black region where $T'_{12}(0) > 0$ (or $T'_{12}(0) > 0$, respectively) was considered. Then the area was numerically minimized over possible values of θ, M and r_{12}. Based on a fine grid of values for $0 \leq \theta \leq 2\pi$, $-2 \leq M \leq 2$ and $-1 \leq r_{12} \leq 1$, the numerical results suggest that the area is equal to or more than half of the total area, which means ridging "often" improves ranking performance over OLS in some sense. We note that the plots were considered to be in the domain of $-4 \leq Q \leq 4$, but similar results appear to hold for other domain.

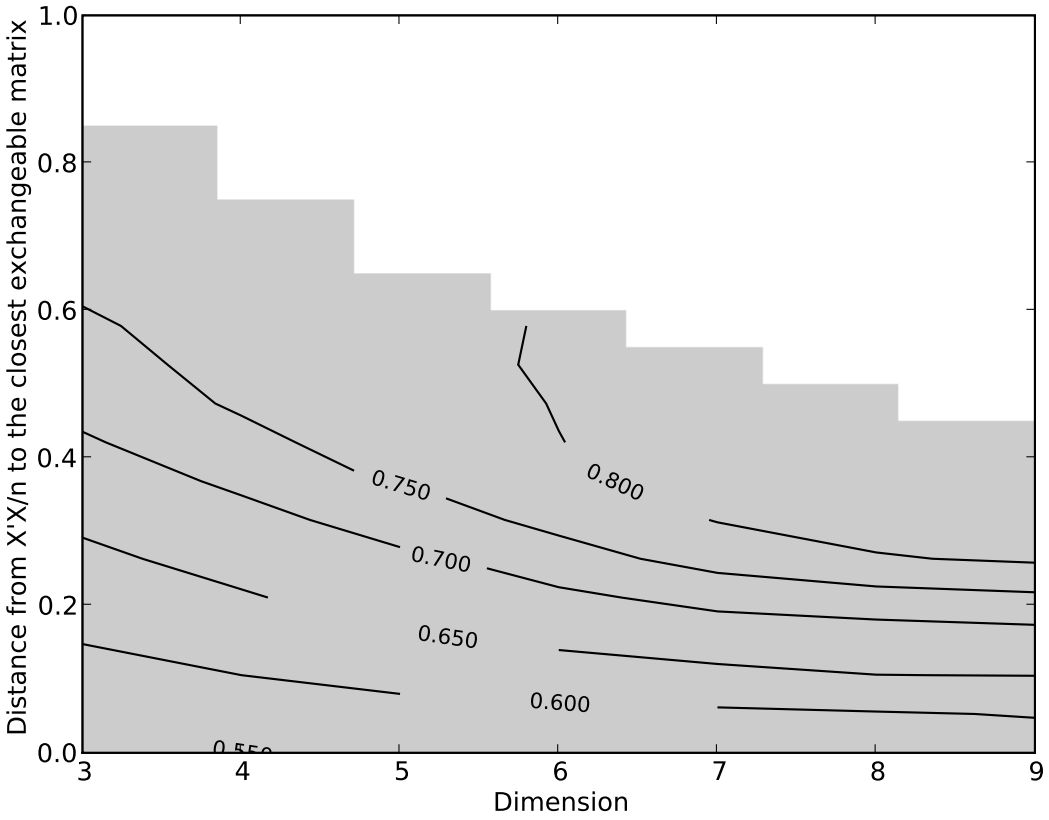


Figure 3.6: The proportion of models with various dimensions (horizontal axis) and with a given degree of non-exchangeability in $X'X/n$ (vertical axis) for which ridging improves ranking performance for small values of λ.

Next, we generalized the results to high dimensional cases by looking at the sign of $T'_{12}(0)$ since it is a good way to identify when ridge regression improves the ranking performance for small values of λ. Recalling that when $r_{ik} = r_{jk}$ for $k \neq i, j$, the probability of correctly ranking β_i and β_j does not depend on λ, the CS criterion does not depend on λ when all correlations between any two variables are same, i.e., the correlation matrix is an exchangeable matrix. Therefore, strong correlations between predictors are not sufficient for ridging to be beneficial. Instead, the heterogeneity in the correlations between covariates plays an important rule on whether ridging improve the ranking accuracy over OLS.

To further explore this issue, we considered the proportion of models in which the slopes of $T_{12}(\lambda)$ at $\lambda = 0$ are positive among all possible models in Figure 3.6. That is, we considered the conditional probabilities

$$E(J(M, \beta)|D(M) \in B) \tag{3.13}$$

where M is uniformly distributed on the set of all $p \times p$ correlation matrices, and β is independent of M, and is uniformly distributed on the unit sphere constrained to $\beta_1 > \beta_2$. The function $J(M, \beta)$ is the indicator that in the population defined by $X'X/n = M$, along with the vector of regression coefficients β, the slope of $T_{12}(\lambda)$ is positive when evaluated at $\lambda = 0$. The scalar-valued function $D(M)$ is the Frobenius norm of the difference between M and the closest exchangeable matrix to M, and B is an interval on the positive real line.

If X is a $(p + 1) \times p$ iid array of standard normal values, then $M = X'X$ has a Wishart distribution with density function

$$\frac{1}{2^{p(p+1)/2}\Gamma_p((p+1)/2)} \exp\left(-\mathrm{tr}M/2\right).$$

We can reparameterize $M = (M_s, M_d)$, where $M_s(i,j) = M_{ij}/\sqrt{M_{ii}M_{jj}}$ and $M_d(i) = M_{ii}$. By applying the change of variables formula, we can show that the joint density of M_s, M_d has the form $\pi_s(M_s) \cdot \pi_d(M_d)$, where

$$\pi_d(M_d) = c_d \exp(-\sum_i M_d(i)/2) \cdot \prod_i M_d(i)^{(p-1)/2}$$

and $\pi_s(M_s) = c_s$, and c_d and c_s are constants. This implies that M_s is uniformly distributed on the set of all correlation matrices. Therefore, we can generate uniformly distributed correlation matrices with density π_s by forming the correlation matrices from X matrices that are $(p+1) \times p$ iid arrays of standard normal values.

Assuming that we sample matrices $M_1, \ldots, M_m$ from a distribution with density $f(\cdot)$ and sample $\beta_1, \ldots, \beta_m$ from their correct marginal distribution. Then the conditional expectation in (3.13) can be estimated as

$$\sum_i J(M_i, \beta) w_i / \sum_i w_i, \tag{3.14}$$

where $w_i = f(M_i)\mathcal{I}(D(M_i) \in B)$.

When we sample matrices from the density π_s like above, there are few or no observations with which to form the average if B is close to zero. Thus we need to consider distributions f with more mass close to exchangeable matrices. Thus we need to consider distributions f with more mass close to exchangeable matrices. One way to do this is to consider exchangeable matrices M where $M_{ii} = 1$ and $M_{ij} = x$ for $i \neq j$, with x sampled from some distribution g. This matrix will always be a correlation matrix if the support of g is $(0, 1)$. More generally, we can consider matrices of the form $\lambda M_1 + (1-\lambda)M_2$, where M_1 is an exchangeable matrix in which the off-diagonals are simulated from g and M_2 is simulated from π_s. To estimate the conditional expectation, we need to evaluate the density from which M was sampled.

This has the form of a deconvolution

$$f(M) = \int_x g(x)\pi_s\left((M - \lambda M_1)/(1-\lambda)\right) dx.$$

If we choose g to be uniform on $(0, 1)$, this reduces to

$$f(M) = c_s \cdot L,$$

where L is the length of the set of x such that $M - \lambda M_1$ is strictly positive definite. By superimposing the resulting estimates for conditional expectations from $\lambda = 0,\ 0.5,\ 0.8$, we could obtain a complete map for the proportions of models in which the slopes of $T_{12}(\lambda)$ at $\lambda = 0$ are positive among all possible models as shown in Figure 3.6

In Figure 3.6, we can see that when $D(M)$ is small, i.e., when the correlation matrix is close to an exchangeable matrix, the proportions of models where ridging improve the ranking performance is close to $1/2$, while the proportions increase up to around $0.8 \sim 0.85$ as $D(M)$ increases. In this respect, ridge regression improves the ranking performance more often than not and is likely to improve the performance more often when the correlation matrix of predictors has more heterogeneous off-diagonal elements.

Finally, we consider a situation depicted in Figure 3.2d, where X_2 and X_3 are uncorrelated. It can be algebraically shown that in this situation ridging can affect the ranking performance when using the approach based on coefficient estimates but ridging has no effect when using the approach based on Z-scores. It suggests that Z-scores are a form of regularization, and in some cases no further improvement results from ridge regression regularization.

We conclude this section with a 3-predictor example that illustrates a situation

where ridging provides a very substantial improvement in the probability of correctly ranking β_1 and β_2 based on $\hat{\beta}_{1\lambda}$ and $\hat{\beta}_{2\lambda}$:

$$X'X = 100 \begin{pmatrix} 1 & 0.4 & 0.4 \\ 0.4 & 1 & -0.4 \\ 0.4 & -0.4 & 1 \end{pmatrix} \qquad \beta = (1, 0.8, 0.9)'. \tag{3.15}$$

This model has $n = 100$ observations. Setting the residual variance to give a population R^2 of 0.4, the value of T_{12} ranges from 0.4 when $\lambda = 0$ to 4.3 when $\lambda = 1000$ (corresponding to correct ranking probabilities ranging from 0.66 to nearly 1).

3.3.2 Comparison of Ridge Regression and the Lasso

Having established that ridge regression can improve the ranking performance more often than not, we next considered how the Lasso and ridge regression perform compared to each other. It is not straightforward to analytically compare ridge regression and the Lasso, so in this section we use numerical experiments to explain the difference between ranking behaviors of the two methods.

A clear difference between the L_1 penalty and the squared L_2 penalty is their limiting behavior. When there is no regularization ($\lambda = 0$), ridge regression and the Lasso are equivalent to OLS. And when there is a small amount of regularization with small values of λ, ridge regression and the Lasso would work similarly. As λ grows, the coefficient estimates of both methods converge to zero. However, the ranking by the two methods are not necessarily the same, because the Lasso estimates are shrunken to exactly zero as λ grows while ridge regression estimates are not. Ranking by ridge regression estimates is equivalent to ranking by univariate analysis when λ approaches infinity. On the other hand, as λ approaches infinity, the Lasso point estimates reach zero, and the CS is 0.5. We conjectured that the difference between

ranking performance of the Lasso and ridge regression would be partly related to how well the Lasso can approximate the ranking based on the univariate analysis.

To explore these matters, we considered three examples below. The first example is the model defined in (3.15) where regularization by ridge regression greatly improves the ranking performance. And the following two models are considered where ridge regression does not monotonically increase the ranking performance:

$$X'X = 100 \begin{pmatrix} 1 & 0.7 & 0.3 \\ 0.7 & 1 & -0.4 \\ 0.3 & -0.4 & 1 \end{pmatrix} \qquad \beta = (1, 0.8, 1.4)'. \tag{3.16}$$

$$X'X = 100 \begin{pmatrix} 1 & 0.5 & 0.4 \\ 0.5 & 1 & -0.3 \\ 0.4 & -0.3 & 1 \end{pmatrix} \qquad \beta = (1, -0.8, 0.7)'. \tag{3.17}$$

Figure 3.7 shows ranking results for the three examples. Note that we used $R^2 = 0.1$ for the three examples. Figure 3.7 plots the expected value of CS against the probability of getting a rank identical to univariate ranking based on ridge regression (dashed lines) and the Lasso (dotted lines) based on 2000 simulations. The three rows of plots show the results for the three models defined in (3.15), (3.16) and (3.17), respectively. The left column shows the results based on ranking by coefficient estimates and the right column shows the results based on ranking by Z-scores.

First, the the results for the model defined in (3.15) show that both ridge regression and the Lasso can improve the ranking performance over OLS. They start off from a point that corresponds to OLS, stay at the same path while increasing, and then diverge to the two different extreme cases where $\lambda = \infty$. Ridge regression, as λ_2

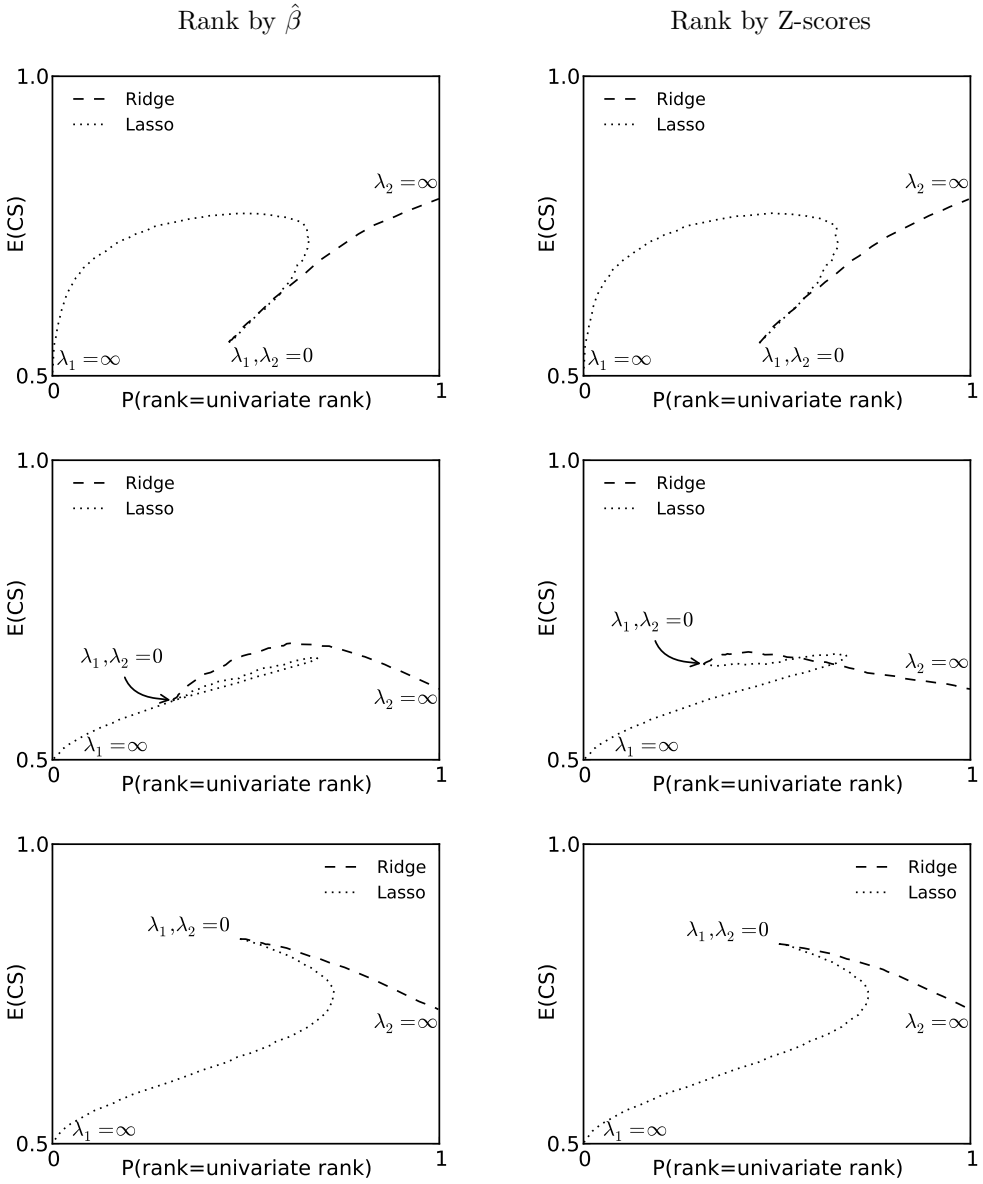


Figure 3.7: The expected value of CS versus the probability of getting a rank identical to univariate ranking. The plots in each row show the results for each model defined in (3.15), (3.16) and (3.17), respectively. The left column shows the results based on ranking by coefficient estimates and the right column by Z-scores.

grows, continues to improve the ranking performance (since the expected value of CS increases) and also the probability of getting a rank identical the univariate analysis converges to 1. On the other hand, for the Lasso, the ranking performance is improved as it becomes more univariate-like, but at some point, it deviates from the path of ridge regression, and its ranking performance deteriorates to the CS of 0.5 as all estimates will be zero when $\lambda_1 = \infty$. Comparing $\hat{\beta}$-based approach and Z-scores-based approach, they don't differ much in the patterns. Overall, this example shows the case where the univariate analysis ($\lambda_2 = \infty$) can rank the variables better than any regularized regression methods.

Second, the results for the model defined in (3.16) illustrates the case when some amount of L_2 regularization improves the ranking performance but excessive regularization decreases the performance. As in the results for the model in (3.15), the ranking performance of the Lasso increases as its rank becomes more univariate-like, but the performance decreases as it starts to become less univariate-like. Again, ranking by coefficient estimates and ranking by Z-scores have similar results.

Third, the results for the model defined in (3.17) show the case where OLS ($\lambda_1, \lambda_2 = 0$) has better ranking performance than any other methods. The ranking performance of ridge regression decreases as it converges to the univariate ranking. Ranking by the Lasso initially becomes more univariate-like, however, different from the previous examples, its ranking performance decreases at the same time. In this example, both ridge regression and the Lasso cannot improve the ranking performance. This situation is a relatively rare case as we showed that ridge regression can improve the ranking performance more often than not in Section 3.3.1. We note that in practice nearly optimal points can be selected by adaptive tuning in all three examples when the sample size is reasonable.

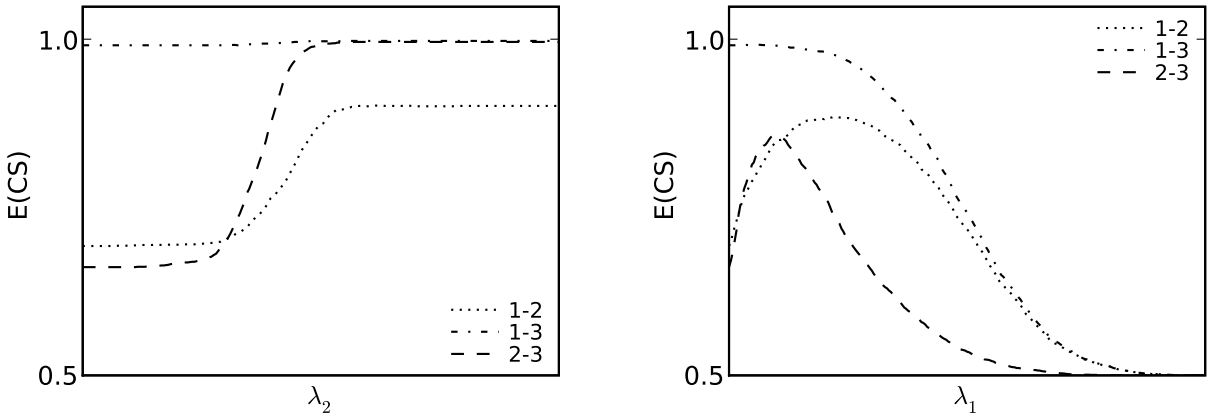


Figure 3.8: The expected value of CS for each pair among the three pairs when $\beta = (1, 0.3, -0.2)$ and $(r_{12}, r_{13}, r_{23}) = (0.6, -0.4, 0.3)$. The left plot shows the results for ridge regression and the right plot shows the results for the Lasso.

Besides, there is another important difference between L_1 regularization and L_2 regularization. The Lasso usually has the best ranking performance at the interior point of λ_1 or at $\lambda_1 = 0$ (it is not possible to achieve the best performance when $\lambda = \infty$ because all coefficient estimates are set to zero). On the other hand, ridge regression can have the best ranking performance at anywhere of $0 \leq \lambda_2 \leq \infty$. In many examples, we could observe that it is possible that the L_2 regularization continues to improve the performance of variable ranking as λ_2 increases.

Considering how the optimal value of the regularization parameter can be chosen in each method, we find it interesting to think about what would be the optimal regularization parameters for comparing each pair among all possible pairs of variables. Figure 3.8 shows an example to explore this issue. As in the previous examples, we assumed a three-predictor model with $\beta = (1, 0.3, -0.2)$, $(r_{12}, r_{13}, r_{23}) = (0.6, -0.4, 0.3)$ and $R^2 = 0.1$. The plots show the expected value of CS for each pair among the three pairs. The left plot shows the results for ridge regression and the right plot shows the results for the Lasso. Ridge regression continues to improve the pairwise ranking

as λ_2 grows, so the optimal λ_2 values for all three pairs occur in a wide interval of λ_2 values. Meanwhile, the Lasso has the optimal pairwise performance at three different values of λ_1, where the peaks are only partially overlapping. Noting that the overall CS is the average of the pairwise CSs over all coefficient pairs where $\beta_j > \beta_k$, the optimal values of λ for the pairwise CSs should overlap in order to let a single value of λ give the optimal overall CS. We observed that in general ridge regression tends to perform well over broader and more overlapping ranges of λ values than the Lasso.

3.4 Simulation Studies

In the previous section, the benefit of L_1 and L_2 regularization for the ranking performance was discussed in two- or three-dimension cases. While interesting aspects of regularization on the ranking performance were considered, it still remains a question how they will work in high dimensional models.

3.4.1 Population Models

To understand what happens with higher dimensional models, as described in Table 3.1, seven families of β vectors were defined to consider plausible conditional mean relationships $EY_i = \sum_j \beta_j X_{ij}$ between a quantitative trait Y and genetic variants in X. Each family of population models contains a set of related five or six β vectors, giving a total of 38 β vectors. We will refer to the k^{th} model in Family q as model $q.k$. For each family, the second column in Table 3.1 defines the value of β in terms of the family parameter. The notation $x\{n\}$ indicates that the value x is repeated n times in sequence (numbers not followed by $\{n\}$ have an implicit $\{1\}$). For example, for Family 1 when $\alpha = 0.2$ we have

$$\beta = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0.2, 1, 0.2, 1, 0, 0, 0, 0, 0, 0, 0, 0)',$$

Family	β	Range	Dim(β)
1	$0\{8\}, \alpha, 1, \alpha, 1, 0\{8\}$	$\alpha = 0, 0.2, \ldots, 1$	$p = 20$
2	$(1, 0\{k\})\{5\}$	$k = 1, 3, \ldots, 9$	$p = 10, 20, \ldots, 50$
3	$(1, 1, 0\{k\})\{5\}$	$k = 1, 3, \ldots, 9$	$p = 15, 25, \ldots, 55$
4	$(1, -1, 0\{k\})\{5\}$	$k = 1, 3, \ldots, 9$	$p = 15, 25, \ldots, 55$
5	$(1, -0.5, 0\{k\}, 1, 0.5, 0\{k\})\{5\}$	$k = 0, 2, \ldots, 10$	$p = 20, 40, \ldots, 120$
6	$0, 1/(k-1), 2/(k-1), \ldots, (k-2)/(k-1), 1$	$k = 2, 4, \ldots, 10$	$p = 2, 4, \ldots, 10$
7	$-0.5, -0.5 + 1/(k-1), \ldots, -0.5 + (k-2)/(k-1), 0.5$	$k = 2, 4, \ldots, 10$	$p = 2, 4, \ldots, 10$

Table 3.1: The population structures used to evaluate the performance of regularized regression methods.

and for Family 2 when $k = 2$ we have

$$\beta = (1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0)'.$$

Based on a particular β vector and genetic data matrix X (generated as described below), phenotype data was considered to follow a linear model with mean $X\beta$ and variance σ^2, where σ^2 was set to provide a given R^2 value, describing the proportion of trait variance explained by the genetic variables. The R^2 values were set to either 0.1 or 0.5, with $R^2 = 0.1$ being more representative of what is expected in genetic analyses of complex traits [23, 58]. The dimension of β in our models ranged up to 120, but most models had less than 50 variants. This is realistic for the number of genetic variants that might arise in a typical genetic mapping study.

The β vectors were selected to give plausible patterns of effects for SNP's in one or a small number of genomic regions. The models range from being very sparse to non-sparse. For example, Family 2 has only a few non-zero effects that are well-separated and weakly dependent, while families such as 6 and 7 that have contributions from nearly all the variants in the model. In addition, some families, for example 3, have "reinforcing" effects in the sense that positively correlated variants have effects in the same direction. Families 4 and 5 have "masking" effects that are positively correlated but have opposite signs.

3.4.2 Predictor Data

For the predictor data X, we considered two types of data: simulated to match SNP data from human subjects and data simulated from a simple parametric model. Our simulated SNP dataset was generated using the GWASimulator program [37] which simulates biallelic SNP genotypes that have mean and local correlation structure similar to that in a given set of phased measured genotypes. As the input set for GWASimulator, we used phased genotypes from the HapMap project [12] for 60 individuals (120 phased chromosomes) in the HapMap CEU sample (Utah residents with ancestry from northern and western Europe). We then selected from the GWASimulator output data for the 22518 SNPs on chromosome 1 that were assayed on the Illumina platform at the Sanger Institute. The data were partitioned into 148 non-overlapping blocks of adjacent SNPs of size 150. SNPs were eliminated if the minor allele frequency was below 0.05. An iterative procedure was applied to remove SNP pairs with correlation greater than 0.9: at each step in the procedure, the pair with the greatest correlation was identified and one SNP in the pair was selected at random and dropped; the procedure continued until no SNP pairs with correlation greater than 0.9 remained. If the final block was shorter than the length of β, it was discarded. Otherwise the initial segment of the block with with length equal to the length of β was used. Finally, each SNP was standardized to have zero mean and unit variance.

Our simple parametric model is a Gaussian AR(1) model with a correlation of 0.8 at lag 1. This is continuous data, whereas most measurements of genetic variation are coded as categories, but since the performance of regression depends on the correlation structure among the predictors, we believe these results are still relevant. All simulations are based on 300 replications of a sample involving 500 independent

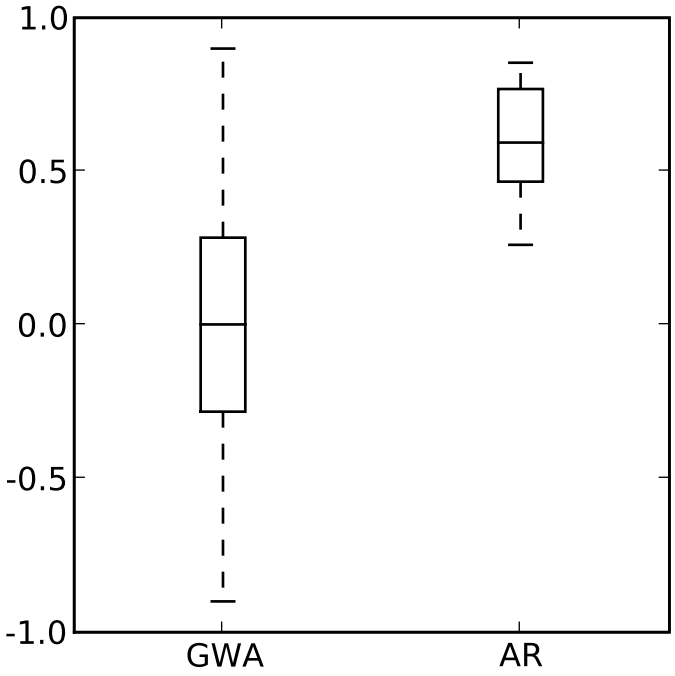


Figure 3.9: Comparison of pairwise correlations in GWASimulator data and AR(1) data.

subjects. The iterative procedure described above was applied to the data for eliminating variables that have pairwise correlations above 0.9. Again, predictor variables were standardized to have zero mean and unit variance.

Although multicollinearity is present in both GWASimulator data and AR(1) data, the two datasets have somewhat different correlation structures. For each data, assuming $p = 55$ as in model 3.5 or model 4.5, the pairwise correlation coefficients between predictors were calculated. Then the correlations between adjacent predictors, r_{ij} for $i \neq j$ and $|i - j| < 5$, were selected for comparison. For GWASimulator data, $(Q_{0.05}, Q_{0.25}, Q_{0.50}, Q_{0.75}, Q_{0.95}) = (-0.685, -0.282, 0.002, 0.284, 0.680)$, while for AR(1) data, $(Q_{0.05}, Q_{0.25}, Q_{0.50}, Q_{0.75}, Q_{0.95}) = (0.379, 0.467, 0.594, 0.768, 0.814)$, where Q_p is $(p \times 100)$th percentile. The boxplots of these correlations are shown in Figure 3.9. The predictor correlations in GWASimulator data range from about

-0.9 to 0.9, so the adjacent predictors can be highly correlated in either positive or negative direction. On the other hand, the adjacent predictors in AR(1) data are positively correlated and the correlations do not vary as much as the correlations in GWASimulator data.

In the performance evaluation using GWASimulator data, a set of X matrices corresponding to a sequence of blocks along chromosome 1 was constructed as described above, and for each X matrix, a single Y vector was generated following each of the 38 population models. Note that the X matrices in this case are not repeated samples from a fixed underlying distribution. For AR(1) data, independent and identically distributed X matrices were obtained, and for each X matrix a single Y vector was generated following each of the 38 population models defined above.

3.4.3 Performance for Variable Ranking

As discussed above, three methods are considered: ridge regression, the Lasso and the elastic net. In this section, the ranking performance of those three methods are compared for both AR(1) data and GWASimulator data. For evaluating the ranking performance, the results based on magnitude analysis with Z-scores are presented, but we note that the results based on other approaches (either signed analysis or coefficient estimates instead of Z-scores) are similar to the results below.

Among the two values (0.1 and 0.5) of R^2, $R^2 = 0.1$, which more represents a situation in genetics study, is first considered. For tuning the regularization parameters, oracle tuning is first discussed and then data-adaptive tuning is considered. Oracle tuning can be thought of as the tuning that chooses the optimal performance that each method can achieve, and data-adaptive tuning would represent the tuning in which each method's practical performance is selected.

Under oracle tuning, the elastic net includes both the Lasso and ridge regression

as special cases, so the CS difference between the elastic net and either of the Lasso or ridge regression must be non-negative. Figure 3.10 shows the CS differences among each pair of methods for AR(1) and GWASimulator data. The results are presented as the boxplots of the CS differences of one method labeled on the upper right margin relative to another method labeled on the upper left margin. For example, the left plot in the first row of Figure 3.10 shows the boxplots of the CS of the elastic net minus the CS of ridge regression. Therefore, a positive difference indicates that the method on the right side performed better in terms of CS. The results shown in Figure 3.10 indicate that if nearly optimal tuning is achieved, the elastic net provides a small benefit relative to ridge regression and a larger benefit relative to the Lasso, and ridge regression performs somewhat better than the Lasso. For the comparison of ridge regression to the Lasso, the gains for ridge regression are larger and somewhat more consistent across models when looking at AR(1) data compared to GWASimulator data.

Using data-adaptive tuning, the situation changes somewhat. Figure 3.11 shows the CS differences between each pair of methods using GCV to tune ridge regression and AIC to tune the Lasso and elastic net procedures. In this situation, ridge regression outperforms the Lasso as it did in the oracle case, but now ridge regression also outperforms the elastic net as well. This is presumably due to data-adaptive tuning being more difficult for the elastic net due to the presence of two tuning parameters.

We also considered BIC and test set tuning for the Lasso and elastic net procedures. These results (not shown) are similar to those shown in Figure 3.11 – ridge regression outperforms the elastic net and the Lasso, and the elastic net outperforms the Lasso. Since the test set tuning procedure approximates an upper bound to optimal data-adaptive tuning, these results suggest that the relatively better perfor-

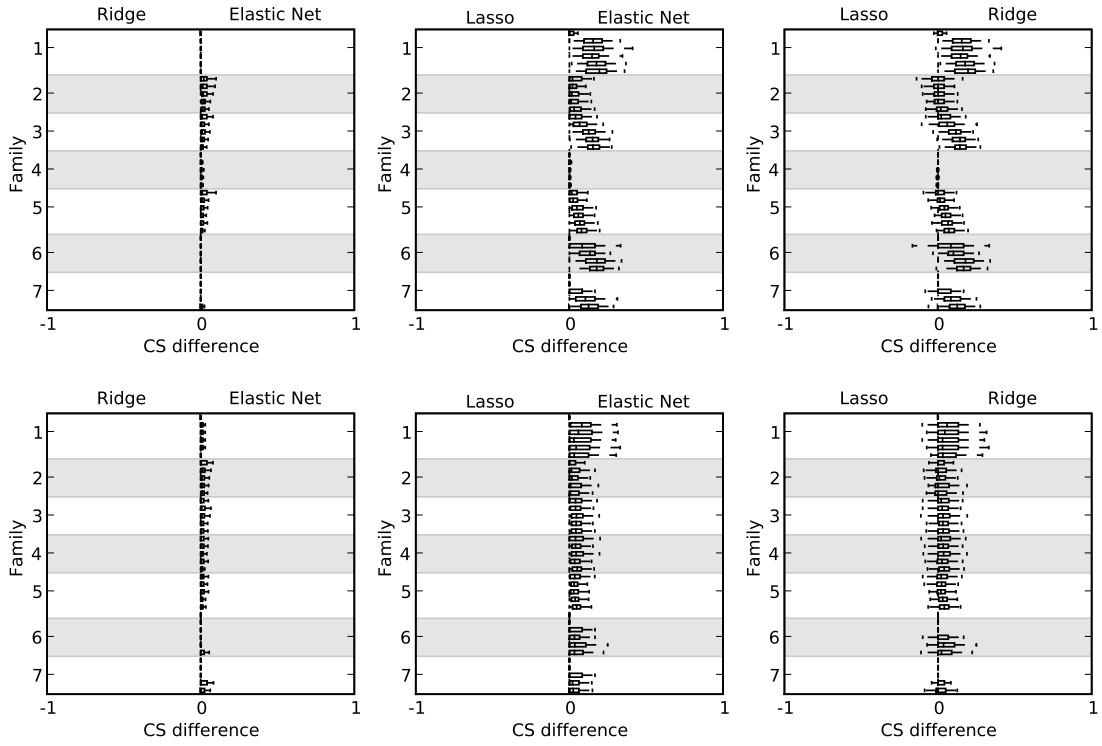


Figure 3.10: Pairwise comparisons of CS among the three methods using oracle tuning for AR(1) data (top) and GWASimulator data (bottom) when $R^2 = 0.1$.

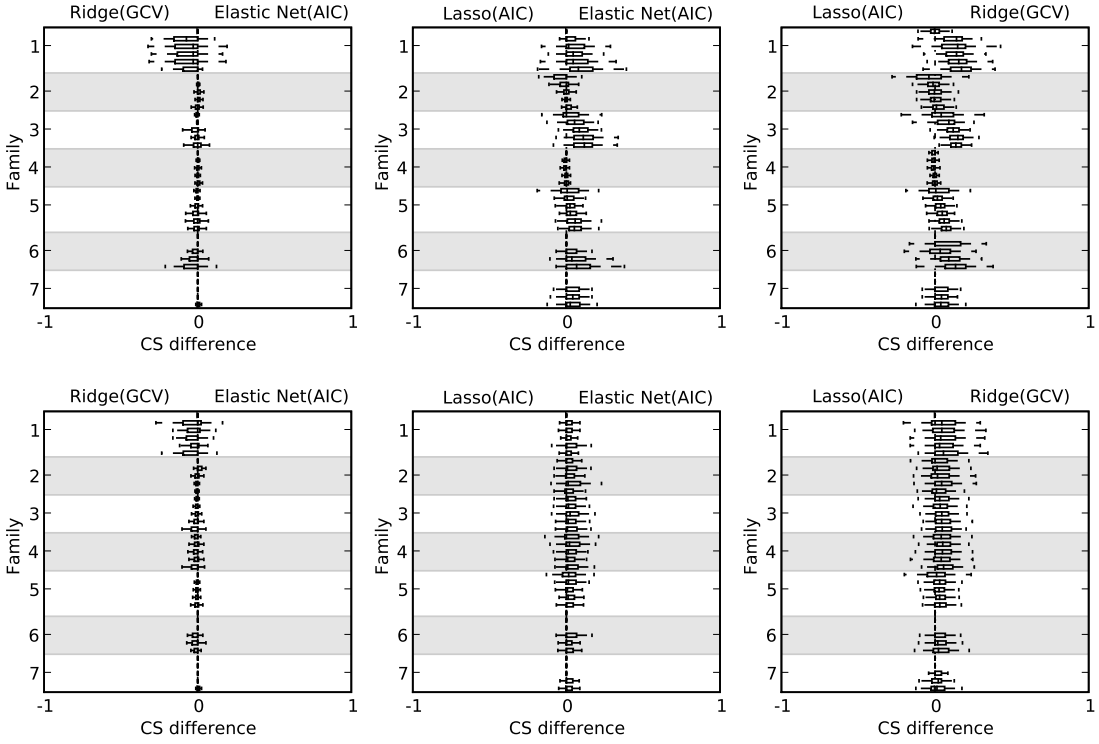


Figure 3.11: Pairwise comparisons of CS among the three methods using data-adaptive tuning for AR(1) data (top) and GWASimulator data (bottom) when $R^2 = 0.1$.

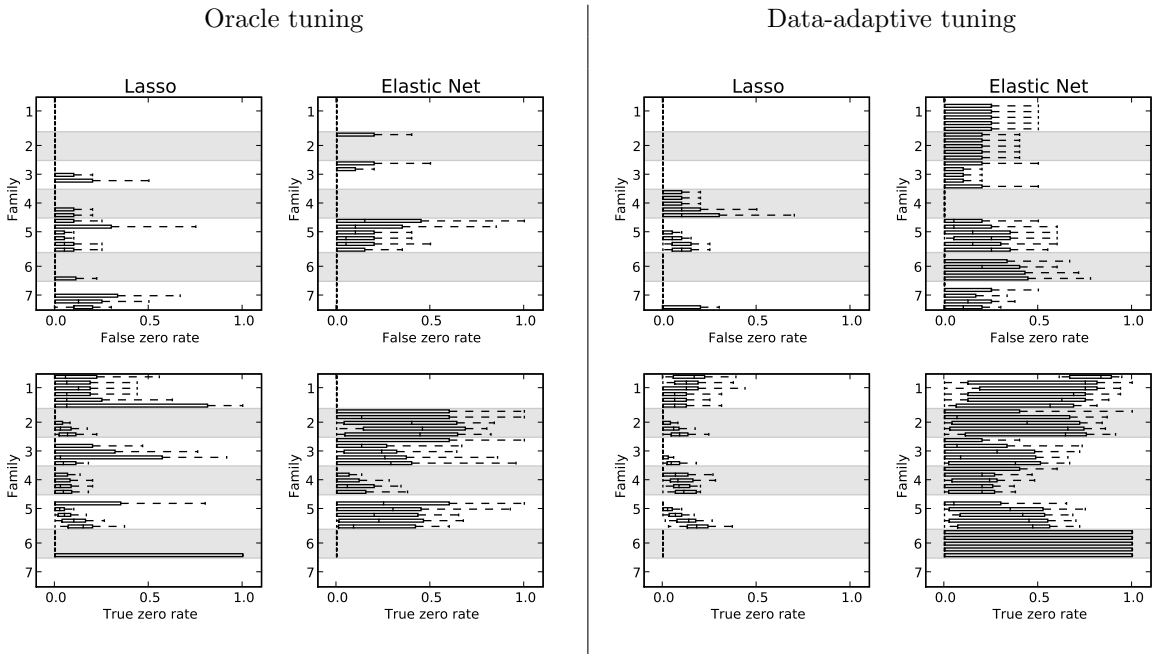


Figure 3.12: The proportions of zero estimates among truly nonzero coefficients (false zero rates) and among truly zero coefficients (true zero rates) when $R^2 = 0.1$ for AR(1) data. The true zero rates for Family 7 are not available because the population models in Family 7 are not sparse.

mance of ridge regression is not due to deficiencies in the model selection statistics or in the approximation to the degrees of freedom.

We also considered the sparsity of the selected estimates for the Lasso and the elastic net. The Lasso and the elastic net can produce sparse solutions when having some amount of L_1 regularization on coefficients. We initially thought that having sparse estimates can be beneficial for ranking when the difference between coefficients is small because the sparse solution could let the smaller coefficient to be estimated as zero, which would possibly increase CS. Otherwise, it would be hard to estimate the coefficients of nearly the same sizes in the correct order. Figure 3.12 shows the proportions of zero estimates among truly nonzero coefficients (false zero rates, the first row) and among truly zero estimates (true zero rates, the second row) in the simulation results using AR(1) data. Note that the true zero rates for Family 7

are not available because the population models in Family 7 do not contain zero coefficients. The left panel shows the results under oracle tuning and the right panel shows the results under data-adaptive tuning based on AIC. Under oracle tuning, both the Lasso and the elastic net performs best when some of the estimates for truly nonzero coefficients are zero in a few population models. When there are groups of highly correlated predictors, the Lasso tends to select one variable from a group [72]. Thus, it is not surprising that a fraction of the false zero rates are positive for the Lasso. The elastic net, however, practically involves ridge regression and the Lasso as special cases, and we found that it can improve the ranking performance by letting some of estimates for truly nonzero coefficients be zero. For example, for the models in Family 5, the elastic net may improve the CS by estimating the coefficients of -0.5 or 0.5 as zero, while it may be hard to get the correct ranking for the pairs of 1 and ± 0.5 using non-sparse solutions. Comparing the results under oracle tuning and data-adaptive tuning, we found that both false zero rates and true zero rates for the elastic net become unstable when using data-adaptive tuning, which would decrease the ranking performance of the elastic net under data-adaptive tuning. As discussed above, this is presumably because it is difficult to tune the two-dimensional regularization parameters in data-adaptive way. We note that the results using GWASimulator data are similar to Figure 3.12.

Next we revisited everything discussed above with $R^2 = 0.5$. The results shown in Figure 3.13 show better ranking performance for ridge regression than the elastic net for most members of Family 5, and model 1.2. Ridge regression performed better than the Lasso for a number of models, and was comparable for all others. There is no model for which either the elastic net or the Lasso has a substantial advantage over ridge regression. The magnitudes of the differences among the methods tended

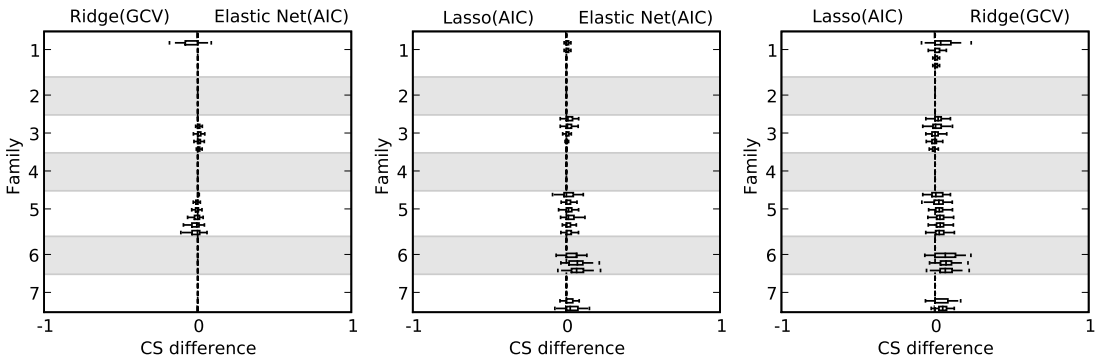


Figure 3.13: Pairwise comparisons of CS among the three methods using data-adaptive tuning for AR(1) data when $R^2 = 0.5$.

to be smaller with $R^2 = 0.5$ compared to $R^2 = 0.1$. The results shown are for the AR(1) predictor data; for the GWA predictor data all three methods performed quite similarly for most of the models. Besides, the false/true zero rates for the Lasso and elastic net estimates (not shown) were found to be similar to those with $R^2 = 0.1$, although the false zero rates are lower when $R^2 = 0.5$.

To interpret the results presented in this section, it is helpful to know that most of the CS variation within each population was due to variation in the design matrices X, rather than variation in the outcomes $Y|X$. When the genetic data are sampled from a population with high correlation between variables, any given sampled design matrix X can have anywhere from modest to severe collinearity. Since the differential performance of CS depends strongly on the structure of $X'X$, it is not surprising that sampling variation in X has a major influence on the results. On the other hand, with a sample size of $n = 500$ as used throughout our study, once X was sampled, the variance over repeated Y samples was relatively small.

3.4.4 Performance for Prediction

At present, most genetic analyses focus on identifying gene/trait associations that can be pursued to identify genetic variants that may have a mechanistic influence

on the trait. Predictive analysis is of some interest, although genetic variants for complex traits found to date generally contain too little information for making meaningful predictions. Here we consider the performance of ridge regression, the Lasso, and the elastic net for prediction using the same set of simulation populations as used for assessing variable selection performance. For simplicity, all results shown use the "tuning set" method for setting the tuning parameters as discussed above.

Figure 3.14 shows the boxplots of rMSE scores defined in (3.9) for $R^2 = 0.1$. In the rMSE formulation, MSE_1 corresponds to one method labeled on the upper left margin and and MSE_2 corresponds to the other method labeled on the upper right margin. Thus, if the rMSE has positive values, it means the method labeled on the upper right margin predicts better than the other method. For the both data sets, the elastic net performs substantially better than either ridge regression or the Lasso. The comparison between ridge regression and the Lasso is mixed, with ridge regression performing better for some models and the Lasso performing better for others. Figure 3.15 shows the results for $R^2 = 0.5$. The elastic net continues to dominate both the Lasso and ridge regression, and now the Lasso generally outperforms ridge regression.

3.5 Discussion

The analytical and numerical results suggest that regularization can substantially improve the accuracy when ranking variables according to their estimated effect sizes on the response, and that with realistic sample sizes, the regularization can be tuned reasonably well in a data-adaptive way. On the other hand, it is notable that it is possible regularization doesn't improve the ranking performance at all even when the predictor variables are highly correlated. Depending on the correlation structures of predictors and the patterns of true effects, the ranking performance with L_1 or L_2

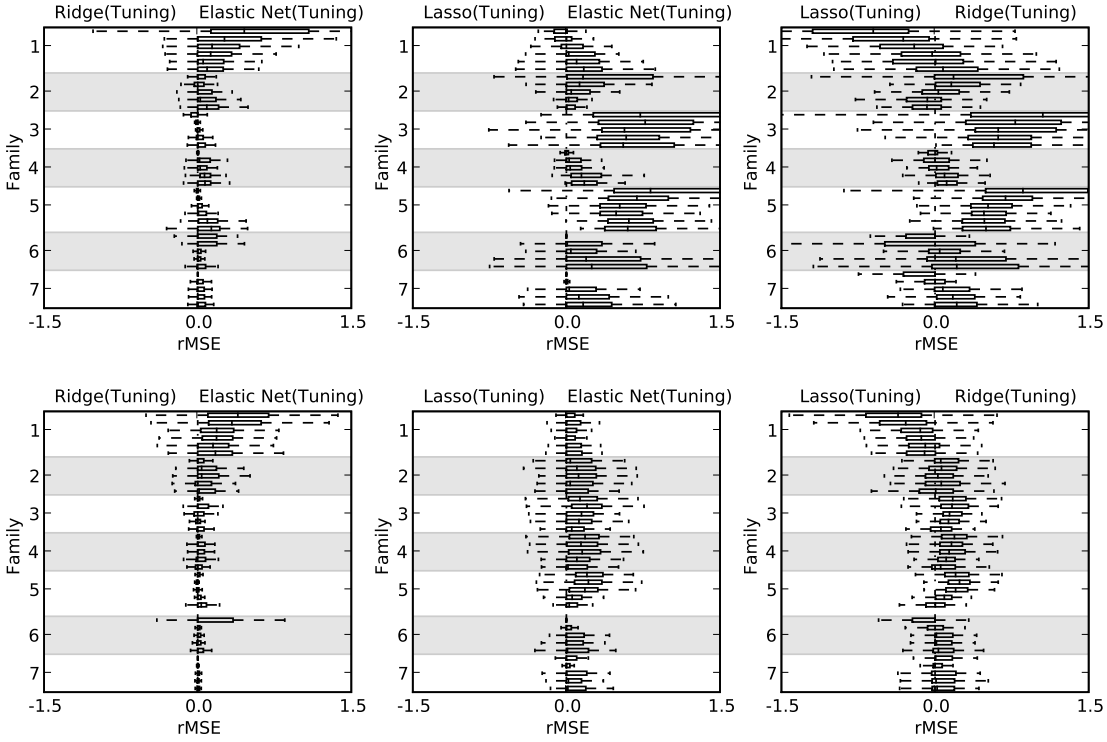


Figure 3.14: Pairwise comparisons of prediction MSE among the three methods under "tuning set" tuning for AR(1) data (top) and GWASimulator data (bottom) when $R^2 = 0.1$.

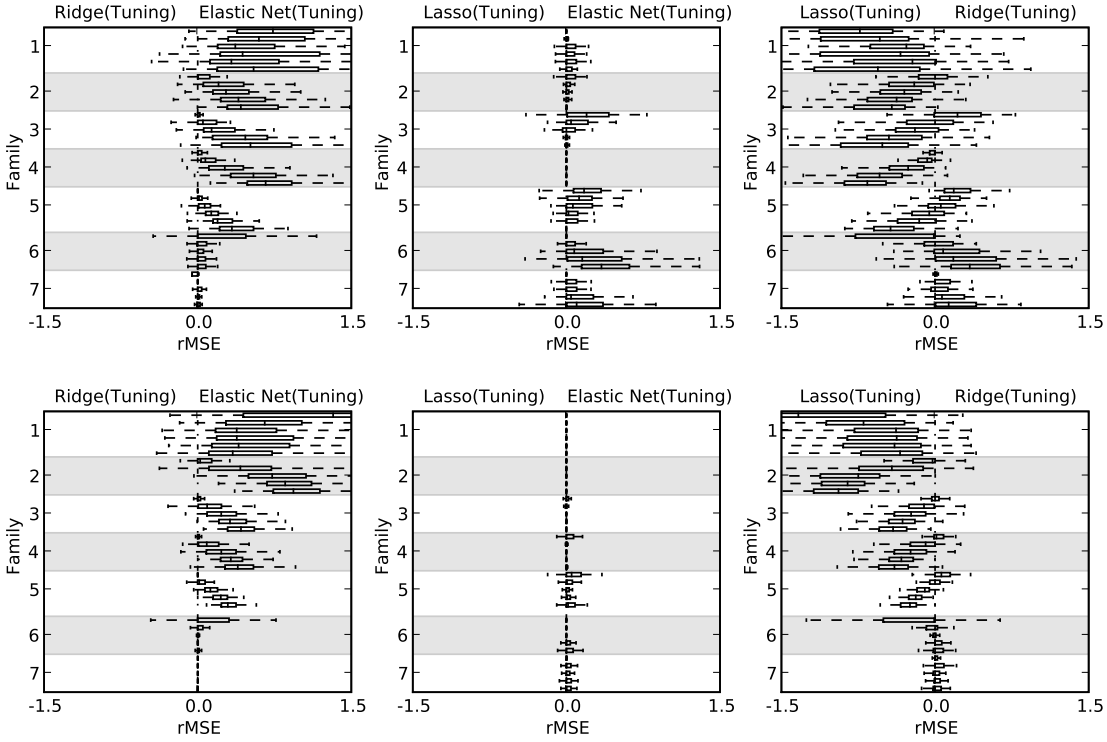


Figure 3.15: Pairwise comparisons of prediction MSE among the three methods under "tuning set" tuning for AR(1) data (top) and GWASimulator data (bottom) when $R^2 = 0.5$.

regularization can uniformly decrease as the tuning parameter λ increases, meaning that the regularization decreases the ranking accuracy compared to OLS. However, in most of the models that were considered in the simulation studies, regularization was found to have improvement on ranking performance over OLS.

By focusing on the accuracy of ranking, we focus on a somewhat different perspective on model selection performance from other previous work. A substantial amount of previous work on regularization focused on prediction or variable selection. For the purpose of variable selection, they focus on figuring out which variables have zero or nonzero effects. In previous work, it has been shown the regularization that includes L_1 penalty (the Lasso or the elastic net) performs well for prediction and variable selection by estimating some of the coefficients as zero especially when the true model is sparse. However, for the purpose of ranking, we could find that ridge regression can often estimate the ranking more accurately than the L_1 regularization, even for the model where the true coefficients include zeros. We note that for prediction performance, the L_1 regularized methods work better than the ridge regression as in previous work.

Another distinguishing aspect of this work is that, motivated by the application of genetic mapping, we focused on models with small overall R^2 values. In contrast, much of the previous discussion of regularization and model selection has focused on settings with higher R^2 values, such as 0.7 ([21]) or 0.96 ([56], [21]). In the simulation studies, it was found that the differences among the methods became smaller as the R^2 value increased.

As we found the L_2 regularization can have better ranking performance than the L_1-based methods, several possible explanations about our findings are suggested. First, the Lasso and the elastic net can shrink some of the coefficient estimates to

exactly zero resulting in a sparse model. Initially it was thought that the ability of getting a sparse model would be potentially beneficial, but it turned out that it is not necessarily true, because the criterion CS is reduced by the ties caused by zero coefficient estimates unless the true coefficients have ties as well. Second, it would be more challenging for the L_1-based methods tune the regularization, because the ranking performance varies much between variable pairs – the regularization parameter that works well for one variable pair might work poorly for other pairs. Although this can also happen for the ridge regression, it has less harmful effect for the ridge regression since the CS of the ridge regression does not deteriorate to 0.5 as λ increases. Third, if the univariate ranking is close to the true ranking, ridge regression can select the univariate ranking by having a large λ, but the L_1-based methods often cannot approximate the univariate ranking well.

We note that the data-adaptive tuning methods used in this work are better motivated for the purpose of prediction, not for the purpose of ranking. Also, GCV has been extensively studied for tuning parameters in the ridge regression, but it is not clear if AIC or BIC is the best criterion for the Lasso or the elastic net. So one might suspect that the better performance of the ridge regression is due to the tuning issues. While this can be partly true when using the data-adaptive tuning, different types of regularization remains to be a major factor when using the tuning set approach and the oracle tuning. The oracle tuning can be seen as a conservative upper bound for data-adaptive tuning and the tuning set approach should approximate efficient data-adaptive tuning. Under both tuning approaches, ridge regression still performed as well or better than the Lasso. This lends weight to an explanation that the differential performance results from differences in the penalty functions themselves. However it is possible that tuning methods targeted

to the variable ranking problem could be developed that perform better than tuning methods developed for prediction or estimation of the mean response.

More broadly, there are other challenges for using regression techniques to sort out unique genetic effects using observational data. An important aspect of genetic association analysis is that in general, we should not expect the true causal variant or variants to be directly measured, even for high density genotyping. If a causal variant were included in the model, ideally the non-causal linked variants would show minimal effects. This ideal situation illustrates the potential advantage of using a multiple regression approach to consider the effects of several linked variants. However in practice, we cannot expect things to work out as in the ideal setting when there are unmeasured environmental factors, and measurement errors in the genotypes and trait values. Nevertheless, consideration of the unique effects of genetic variables as estimated using multiple regression analysis has the potential to be informative at identifying potential causal variants in one or more regions of interest.

CHAPTER IV

Future Work

In Chapter III, we discussed using penalized regression methods for ranking variables. When using penalized regression methods, choosing tuning parameters is essential in practice. In the simulation studies, we considered GCV, AIC, BIC and a tuning set approach for tuning regularization parameters. However, those criteria are originally designed for prediction performance and variable selection performance. Therefore, minimizing those criteria does not guarantee the optimal ranking performance, although we found that those criteria perform fairly well for selecting the estimates with good ranking performance.

Nevertheless, a tuning method could be designed for optimizing the ranking performance instead of the prediction performance. It becomes more obvious that the tuning for prediction is not optimal for ranking when we consider the fact that ranking by univariate analysis is close to optimal in many situations. Assuming we use ridge regression for ranking variables, the ranking performance would be close to optimal when $\lambda = \infty$. As λ approaches infinity, the ridge regression estimates will be shrunken to very small values in which ranking by the differences between those small values is equivalent to ranking by univariate analysis. However, GCV would not be able to select those small estimates, because they are not likely to have good

prediction performance. Similarly, tuning based on GCV, AIC, BIC and a tuning set approach would be able to select the model that has the best ranking.

We found L_2 regularization performs well for ranking variables when the effects are weak. However, when some of the effects are strong and the others are weak, applying L_2 regularization would result in large bias in estimating the large effects, which might potentially decrease the ranking accuracy. To remedy this drawback of L_2 regularization, one could consider using a hybrid of L_1 and L_2 penalties in which small coefficients are regularized with L_2 penalty and large coefficients are regularized with L_1 penalty to reduce the bias. This type of penalty function is similar to Huber function [32], and requires additional tuning for choosing the location at which small and large coefficient are divided for different penalties. This hybrid penalty function will have the good property of ridge regression for ranking, but will not over-penalize large coefficients. We note that [46] proposes using the reversed version of Huber function ("Berhu" function) for a penalty function. Regularization based on Berhu penalty behaves like the Lasso in the sense that the solutions can be sparse, but unlike the Lasso, it does not zero out the estimates of highly correlated predictors and it can select more than n variables when $n < p$. However, this penalty [46] was not found to be useful for the purpose of ranking variables, as the sparsity of the estimates does not improve the ranking performance.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] M. Alley, D. Scudiero, A. Monks, M. Hursey, M. Czerwinski, D. Fine, B. Abbott, J. Mayo, R. Shoemaker, and M. Boyd. Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay. *Cancer Research*, 48(3):589–601, Feb 1988.

[2] Y. Aulchenko, M. Struchalin, N. Belonogova, T. Axenovich, M. Weedon, A. Hofman, A. Uitterlinden, M. Kayser, B. Oostra, C. van Duijn, A. Janssens, and P. Borodin. Predicting human height by victorian and genomic methods. *European Journal of Human Genetics*, 17:1070–1075, Feb 2009.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society series B-Methodological*, 57:289–300, 1995.

[4] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188, 2001.

[5] Richard A Berk. *Regression Analysis: A Constructive Critique*. Sage Publications, 2004.

[6] P. Borst, R. Evers, M. Kool, and J. Wijnholds. A family of drug transporters: the multidrug resistance-associated proteins. *J Natl Cancer Inst*, 92(16):1295–1302, Aug 2000.

[7] L. Breiman. Better subset regression using the non-negative garrote. *Technometrics*, 37(4):373–384, 1995.

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

[9] P. Brown. Centering and scaling in ridge regression. *Technometrics*, 19(1):35–36, 1977.

[10] J. Chapman, J. Cooper, J. Todd, and D. Clayton. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered*, 56(1-3):18–31, 2003.

[11] H. Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24:17–36, 1996.

[12] The International HapMap Consortium. The international hapmap project. *Nature*, 426:789–796, 2003.

[13] J. Cook and L. Stefanski. A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association*, 89:1314–1328, 1995.

[14] D. Covell, A. Wallqvist, R. Huang, N. Thanki, A. Rabow, and X. Lu. Linking tumor cell cytotoxicity to mechanism of drug action: an integrated analysis of gene expression, small-molecule screening and structural databases. *Proteins*, 59(3):403–433, May 2005.

[15] Z. Dai, Y. Huang, W. Sadee, and P. Blower. Chemoinformatics analysis identifies cytotoxic compounds susceptible to chemoresistance mediated by glutathione and cystine/glutamate transport system xc-. *J Med Chem*, 50(8):1896–1906, Apr 2007.

[16] Hoaglin DC, Mosteller F, and Tukey JW. *Understanding robust and exploratory data analysis.* New York: Wiley, 1983.

[17] G. Dennis, B. Sherman, D. Hosack, J. Yang, W. Gao, H. Lane, and R. Lempicki. David: Database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5):P3, 2003.

[18] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression (with discussion). *Annals of Statistics*, 32(2):407–499, 2004.

[19] B. Efron, R. Tibshirani, J. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[20] G. Ellison, T. Klinowska, R. Westwood, E. Docter, T. French, and J. Fox. Further evidence to support the melanocytic origin of mda-mb-435. *Mol Pathol*, 55(5):294–299, Oct 2002.

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

[22] J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32(3):928–961, 2004.

[23] K. Frazer, S. Murray, N. Schork, and E. Topol. Human genetic variation and its contribution to complex traits. *Nat Rev Genet*, 10(4):241–251, Apr 2009.

[24] D. Freedman. Linear statistical models for causation: A critical review. *Wiley Encyclopedia of Statistics in Behavioral Science, B. Everitt and D. Howell eds.*, 2005.

[25] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

[26] W. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

[27] L. Garraway, H. Widlund, M. Rubin, G. Getz, A. Berger, S. Ramaswamy, R. Beroukhim, D. Milner, S. Granter, J. Du, C. Lee, S. Wagner, C. Li, T. Golub, D. Rimm, M. Meyerson, D. Fisher, and W. Sellers. Integrative genomic analyses identify mitf as a lineage survival oncogene amplified in malignant melanoma. *Nature*, 436(7047):117–122, Jul 2005.

[28] G. Golub, M. Heath, and G. Wahba. Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.

[29] M. Hamada and C. Wu. Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24(3):130–137, 1992.

[30] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[31] R. Huang, A. Wallqvist, N. Thanki, and D. Covell. Linking pathway gene expressions to the growth inhibition response from the national cancer institute's anticancer screen and drug mechanism of action. *Pharmacogenomics J*, 5(6):381–399, 2005.

[32] P. Huber. *Robust statistics.* New York: John Wiley and Sons, 1981.

[33] R. Hung, P. Brennan, C. Malaveille, S. Porru, F. Donato, P. Boffetta, and J. Witte. Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. *Cancer Epidemiology, Biomarkers & Prevention*, 13:1013–1021, 2004.

[34] V. Joseph. A bayesian approach to the design and analysis of fractionated experiments. *Technometrics*, 48(2):219–229.

[35] A. Lee, K. Shedden, G. Rosania, and G. Crippen. Data mining the nci60 to predict generalized cytotoxicity. *J Chem Inf Model*, 48(7):1379–1388, Jul 2008.

[36] G. Lettre, A. Jackson, C. Gieger, F. Schumacher, S. Berndt, S. Sanna, S. Eyheramendy, B. Voight, J. Butler, C. Guiducci, T. Illig, R. Hackett, I. Heid, K. Jacobs, V. Lyssenko, M. Uda, Diabetes Genetics Initiative, F. U. S. I. O. N., K. O. R. A., Lung Colorectal Prostate, Ovarian Cancer Screening Trial, Nurses' Health Study, SardiN. I. A., M. Boehnke, S. Chanock, L. Groop, F. Hu, B. Isomaa, P. Kraft, L. Peltonen, V. Salomaa, D. Schlessinger, D. Hunter, R. Hayes, G. Abecasis, H. Wichmann, K. Mohlke, and J. Hirschhorn. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature Genetics*, 40(5):584–591, May 2008.

[37] C. Li and M. Li. Gwasimulator: a rapid whole-genome simulation program. *Bioinformatics*, 24(1):140–142, Jan 2008.

[38] K. Li and S. Yuan. A functional genomic study on nci's anticancer drug screen. *Pharmacogenomics J*, 4(2):127–135, 2004.

[39] R. Liu, P. Blower, A. Pham, J. Fang, Z. Dai, C. Wise, B. Green, C. Teitel, B. Ning, W. Ling, B. Lyn-Cook, F. Kadlubar, W. Sade, and Y. Huang. Cystine-glutamate transporter slc7a11 mediates resistance to geldanamycin but not to 17-(allylamino)-17-demethoxygeldanamycin. *Mol Pharmacol*, 72(6):1637–1646, Dec 2007.

[40] J. Lockwood, T. Louis, and D. McCaffrey. Uncertainty in rank estimation: implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics*, 27:255–270, 2002.

[41] N. Malo, O. Libiger, and N. Schork. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am J Hum Genet*, 82(2):375–385, Feb 2008.

[42] K. Marx, P. O'Neil, P. Hoffman, and M. Ujwal. Data mining the nci cancer cell line compound gi(50) values: identifying quinone subtypes effective against melanoma and leukemia cell classes. *J Chem Inf Comput Sci*, 43(5):1652–1667, 2003.

[43] P. McCullagh and J. Nelder. *Generalised Linear Models*. Chapman and Hall, London, 1989.

[44] R. Meyer. Value-added indicators of school performance: A primer. *Economics of Education Review*, 16:183–301, 1997.

[45] J. Nelder. The statistics of linear models: back to basics. *Statistics and Computing*, 4:221–234, 1994.

[46] A. Owen. A robust hybrid of lasso and ridge regression. Technical report, Stanford University, 2006.

[47] A. Porcari, R. Ptak, K. Borysko, J. Breitenbach, S. Vittori, L. Wotring, J. Drach, and L. Townsend. Deoxy sugar analogues of triciribine: correlation of antiviral and antiproliferative activity with intracellular phosphorylation. *J Med Chem*, 43(12):2438–2448, Jun 2000.

[48] R. Ptak, K. Borysko, A. Porcari, J. Buthod, L. Holland, C. Shipman, L. Townsend, and J. Drach. Phosphorylation of triciribine is necessary for activity against hiv type 1. *AIDS Res Hum Retroviruses*, 14(15):1315–1322, Oct 1998.

[49] B. Ring, S. Chang, L. Ring, R. Seitz, and Douglas T Ross. Gene expression patterns within cell lines are predictive of chemosensitivity. *BMC Genomics*, 9:74, 2008.

[50] N. Samani, J. Erdmann, A. Hall, C. Hengstenberg, M. Mangino, B. Mayer, R. Dixon, T. Meitinger, P. Braund, H. Wichmann, J. Barrett, I. Knig, S. Stevens, S. Szymczak, D. Tregouet, M. Iles, F. Pahlke, H. Pollard, W. Lieb, F. Cambien, M. Fischer, W. Ouwehand, S. Blankenberg, A. Balmforth, A. Baessler, S. Ball, T. Strom, I. Braenne, C. Gieger, P. Deloukas, M. Tobin, A. Ziegler, J. Thompson, H. Schunkert, W. T. C. C. C., and the Cardiogenics Consortium. Genomewide association analysis of coronary artery disease. *N Engl J Med*, 357(5):443–453, Aug 2007.

[51] K. Shedden, L. Townsend, J. Drach, and G. Rosania. A rational approach to personalized anticancer therapy: chemoinformatic analysis reveals mechanistic gene-drug associations. *Pharm Res*, 20(6):843–847, Jun 2003.

[52] R. Shoemaker. The nci60 human tumour cell line anticancer drug screen. *Nat Rev Cancer*, 6(10):813–823, Oct 2006.

[53] J. Staunton, D. Slonim, H. Coller, P. Tamayo, M. Angelo, J. Park, U. Scherf, J. Lee, W. Reinhold, J. Weinstein, J. Mesirov, E. Lander, and T. Golub. Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci U S A*, 98(19):10787–10792, Sep 2001.

[54] J. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*, 64:479–498, 2002.

[55] Y. Sun, K. Shedden, J. Zhu, N. Choi, and S. Kardia. Identification of correlated genetic variants jointly associated with rheumatoid arthritis using ridge regression. *BMC Proceedings*, 2009.

[56] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[57] A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[58] P. Visscher, W. Hill, and N. Wray. Heritability in the genomics era–concepts and misconceptions. *Nat Rev Genet*, 9(4):255–266, Apr 2008.

[59] H. Wang and C. Leng. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048, 2007.

[60] H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection via the lad-lasso. *Journal of Business & Economics Statistics*, 25(3):347–355, 2007.

[61] H. Wang, G. Li, and C. Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 69(1):63–78, 2007.

[62] H. Wang, R. Li, and C. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.

[63] M. Weedon, H. Lango, C. Lindgren, C. Wallace, D. Evans, M. Mangino, R. Freathy, J. Perry, S. Stevens, A. Hall, N. Samani, B. Shields, I. Prokopenko, M. Farrall, A. Dominiczak, Diabetes Genetics Initiative, Wellcome Trust Case Control Consortium, T. Johnson, S. Bergmann, J. Beckmann, P. Vollenweider, D. Waterworth, V. Mooser, C. Palmer, A. Morris, W. Ouwehand, Cambridge GEM Consortium, J. Zhao, S. Li, R. Loos, I. Barroso, P. Deloukas, Ma. Sandhu, E. Wheeler, N. Soranzo, M. Inouye, N. Wareham, M. Caulfield, P. Munroe, A. Hattersley, M. McCarthy, and T. Frayling. Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics*, 40(5):575–583, May 2008.

[64] G. Wei, D. Twomey, J. Lamb, K. Schlis, J. Agarwal, R. Stam, J. Opferman, S. Sallan, M. den Boer, R. Pieters, T. Golub, and S. Armstrong. Gene expression-based chemical genomics identifies rapamycin as a modulator of mcl1 and glucocorticoid resistance. *Cancer Cell*, 10(4):331–342, Oct 2006.

[65] C. Wu and M. Hamada. *Experiments: Planning, Analysis and Parameter Design Optimization*. Wiley, New York, 2000.

[66] T. Wu, Y. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, Mar 2009.

[67] M. Yuan, V. Joseph, and Y. Lin. An efficient variable selection approach for analyzing designed experiments. *Technometrics*, 49(4):430–439, 2007.

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

[69] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*. To Appear.

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

[71] H. Zou, T. Hasite, and R. Tibshirani. On the degrees of freedom of the lasso. *Annals of Statistics*, 35, 2007.

[72] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.