CONTRIBUTIONS TO PENALIZED ESTIMATION

Sunyoung Shin

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research in the College of Arts and Sciences.

Chapel Hill
2014

Approved by:

Jason P. Fine

Yufeng Liu

Michael R. Kosorok

J.S. Marron

Kai Zhang

ABSTRACT

SUNYOUNG SHIN: Contributions to Penalized Estimation
(Under the direction of Jason P. Fine and Yufeng Liu)

Penalized estimation is a useful statistical technique to prevent overfitting problems. In penalized methods, the common objective function is in the form of a loss function for goodness of fit plus a penalty function for complexity control. In this dissertation, we develop several new penalization approaches for various statistical models. These methods aim for effective model selection and accurate parameter estimation.

The first part introduces the notion of partially overlapping models across multiple regression models on the same dataset. Such underlying models have at least one overlapping structure sharing the same parameter value. To recover the sparse and overlapping structure, we develop adaptive composite M-estimation (ACME) by doubly penalizing a composite loss function, as a weighted linear combination of the loss functions. ACME automatically circumvents the model misspecification issues inherent in other composite-loss-based estimators.

The second part proposes a new refit method and its applications in the regression setting through model combination: ensemble variable selection (EVS) and ensemble variable selection and estimation (EVE). The refit method estimates the regression parameters restricted to the selected covariates by a penalization method. EVS combines model selection decisions from multiple penalization methods and selects the optimal model via the refit and a model selection criterion. EVE considers a factorizable likelihood-based model whose full likelihood is the multiplication of likelihood factors. EVE is shown to have asymptotic efficiency and computational efficiency.

The third part studies a sparse undirected Gaussian graphical model (GGM) to explain conditional dependence patterns among variables. The edge set consists of conditionally dependent variable pairs and corresponds to nonzero elements of the inverse covariance matrix under the Gaussian assumption. We propose a consistent validation method for edge selection (CoVES) in the penalization framework. CoVES selects candidate edge sets along the solution path and finds the optimal set via repeated subsampling. CoVES requires simple computation and delivers excellent performance in our numerical studies.

ACKNOWLEDGMENTS

First of all, I would like to express my deepest gratitude to my two advisors, Dr. Jason Fine and Dr. Yufeng Liu. Without their guidance and support, this dissertation would not be possible. Their keen insight and inspiring intuition have motivated me to focus on this dissertation and pursue more research opportunities. The fruitful academic discussions with them have made my research experience more enjoyable and their invaluable advices have made my career path more clear.

Next I would like to thank my committee members, Dr. Michael Kosorok, Dr. J. S. Marron, and Dr. Kai Zhang. I really appreciate that they have taken time out from their busy schedule to serve as members of my committee. Their helpful advices and comments have made this dissertation significantly better.

Last but not least, my sincere gratitude goes out to my friends and family. They always believe in me and have encouraged me through my Ph.D. journey. With their belief, support and encouragement, I was able to successfully complete this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER1: INTRODUCTION

1.1 Background on Penalization

In the past two decades, there have been significant developments in penalization techniques, both in terms of methodology and applications. One of the most popular examples is the least absolute shrinkage and selection operator (Lasso) proposed by Tibshirani (1996), which is closely related to nonnegative garrote (Breiman 1995). Other examples include smoothly clipped absolute deviation (SCAD), elastic net and adaptive Lasso. See Fan and Li (2001), Zou and Hastie (2005), Zou (2006) and Hastie, Tibshirani, and Friedman (2001) and references therein for details.

We consider a training dataset with n independently and identically distributed random samples $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$, $i = 1, \cdots, n$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is a vector of predictors and $y_i \in \mathbb{R}$ is the response variable. Our interest is to identify the underlying relationship between the predictors and the response. Such a relationship is commonly learned through a loss function, $L(\boldsymbol{z}, (\alpha, \boldsymbol{\beta}^T))$, where $\alpha \in \mathbb{R}$ is an intercept parameter, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a parameter vector of interest. In classical statistics, the estimator of the parameters is the minimizer of the empirical loss function as below:

$$\underset{(\alpha, \boldsymbol{\beta}^T)^T}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(\boldsymbol{z}_i, (\alpha, \boldsymbol{\beta}^T)). \tag{1.1}$$

The loss function is used to measure the goodness of fit of the model on the data. Some common examples of the loss term include the squared error loss in ordinary least squares regression and the negative log-likelihood in maximum likelihood estimation.

Penalized methods add a penalty term to (1.1), which controls the model complexity to avoid overfitting. Many penalized methods can be cast as optimization problems. The common objective function for optimization in a penalization method is in the form of *loss+penalty* as follows:

$$\min_{(\alpha,\boldsymbol{\beta}^T)^T} \frac{1}{n}\sum_{i=1}^{n} L(\boldsymbol{z}_i, (\alpha, \boldsymbol{\beta}^T)) + \lambda p(\boldsymbol{\beta}), \tag{1.2}$$

where $p(\cdot)$ is the penalty function and $\lambda \geq 0$ is the regularization parameter. The regularization parameter determines the amount of penalty on the model complexity(Hastie, Tibshirani, and Friedman 2001). A number of penalty functions have been developed for sparse and structured estimation in numerous statistical models. For example, the penalty function for Lasso is the L_1-norm penalty, $\sum_{j=1}^{p} |\beta_j|$.

This chapter first discusses some loss functions and briefly examines penalization methods. Section 1.1.1 reviews various loss functions and their corresponding statistical models. Section 1.1.2 explores intuitions, properties, and computational algorithms of penalization techniques.

1.1.1 Loss Functions in Penalized Estimation

Many loss functions are available for penalization methods. To address a statistical problem, we may choose a suitable loss function. Several examples include least squares loss, check loss, asymmetric least squares loss, composite loss and negative log-likelihood loss for various statistical models. We first review them and briefly introduce related penalization methods.

A simple and popular choice of loss functions is the following least squares loss in a

linear regression setup:

$$\sum_{i=1}^{n} L(\boldsymbol{z}_i, (\alpha, \boldsymbol{\beta}^T)) = \sum_{i=1}^{n} (y_i - \alpha - \boldsymbol{x}_i^T \boldsymbol{\beta})^2. \tag{1.3}$$

Many studies on penalized methods started with this loss function and extended the methods to other loss functions. Breiman (1995) and Tibshirani (1996) introduced nonnegative Garrote and Lasso for the least squares loss function. These penalization techniques have been adapted for likelihood-based regression, quantile regression, and etc.

Koenker and Bassett (1978) introduced the quantile regression model to provide a complete picture on the conditional distribution of the response. The τth conditional quantile function, $f_\tau(\boldsymbol{x})$, is defined as $P(y \leq f_\tau(\boldsymbol{x})|\boldsymbol{x}) = \tau$, for $0 < \tau < 1$ (Wu and Liu 2009). We estimate the τth quantile as a linear function of the predictors with the check loss function:

$$\sum_{i=1}^{n} L(\boldsymbol{z}_i, (\alpha, \boldsymbol{\beta}^T)) = \sum_{i=1}^{n} \{\tau(y_i - \alpha - \boldsymbol{x}_i^T \boldsymbol{\beta})_+ + (1-\tau)(y_i - \alpha - \boldsymbol{x}_i^T \boldsymbol{\beta})_-\}, \tag{1.4}$$

where $t_+ = tI(t \geq 0)$ and $t_- = tI(t < 0)$. Some penalized methods for quantile regression were studied by Wu and Liu (2009) and Wang, Li, and Jiang (2007a).

Motivated by quantile regression, Newey and Powell (1987) proposed asymmetric least squares regression. The check function is replaced with the asymmetric least squares loss function:

$$\sum_{i=1}^{n} L(\boldsymbol{z}_i, (\alpha, \boldsymbol{\beta}^T)) = \sum_{i=1}^{n} \{\tau(y_i - \alpha - \boldsymbol{x}_i^T \boldsymbol{\beta})_+^2 + (1-\tau)(y_i - \alpha - \boldsymbol{x}_i^T \boldsymbol{\beta})_-^2\}. \tag{1.5}$$

The τth expectile is defined as $\mu_\tau(\boldsymbol{x}) = \hat{\alpha}_{aLS} + \boldsymbol{x}^T \hat{\boldsymbol{\beta}}_{aLS}$, where $(\hat{\alpha}_{aLS}, \hat{\boldsymbol{\beta}}_{aLS}^T)^T$ is the

minimizer of (1.5). It has the interpretation that the average distance from the responses, y_i below $\mu_\tau(\boldsymbol{x})$ to $\mu_\tau(\boldsymbol{x})$ is $100\tau\%$ (Fan and Gijbels 1996). To our knowledge, penalization methods for the asymmetric least squares have not been studied.

Recent studies on penalization have introduced a composite loss function, a weighted linear combination of multiple loss functions. When we combine K loss functions, the composite loss function has the intercept parameters, $\boldsymbol{\alpha}^T = (\alpha_1, \cdots, \alpha_K)$ and K parameter vectors of interest, $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K$. Most existing work assumes the same regression slope across the multiple losses, that is $\boldsymbol{\beta} = \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_K \in \mathbb{R}^p$. Zou and Yuan (2008) proposed the equally weighted composite quantile regression (EWCQR) based on the following loss function:

$$\sum_{i=1}^{n} L(\boldsymbol{z}_i, (\alpha, \boldsymbol{\beta}^T)) = \sum_{i=1}^{n} \{ \sum_{k=1}^{K} \{ \tau_k (y_i - \alpha_k - \boldsymbol{x}_i^T \boldsymbol{\beta})_+ + (1 - \tau_k)(y_i - \alpha_k - \boldsymbol{x}_i^T \boldsymbol{\beta})_- \} \}, \quad (1.6)$$

where $0 < \tau_1 < \cdots < \tau_K < 1$. They developed the penalized EWCQR estimator with the adaptively weighted L_1 penalty. Bradic, Fan, and Wang (2011) introduced a composite quasi-likelihood (CQ), a more general composite loss function. The CQ is a weighted combination of K convex loss functions, $\rho_k(y_i - \alpha - \boldsymbol{x}_i^T \boldsymbol{\beta})$, $k = 1, \cdots, K$ with weights $\boldsymbol{w} = (w_1, \cdots, w_K)$. The corresponding loss function is written as follows:

$$\sum_{i=1}^{n} L(\boldsymbol{z}_i, (\alpha, \boldsymbol{\beta}^T)) = \sum_{i=1}^{n} \{ \sum_{k=1}^{K} w_k \rho_k (y_i - \alpha - \boldsymbol{x}_i^T \boldsymbol{\beta}) \}. \quad (1.7)$$

They proposed a robust and efficient penalized CQ estimator with theoretically optimal weights.

Generalized linear model (GLM) is one of the well-known likelihood-based approaches. Suppose that y_i has a density $f(g(\alpha + \boldsymbol{x}_i^T \boldsymbol{\beta}), y_i)$ conditioning on $\boldsymbol{x}_i$, where g is a known link function. The negative log-likelihood loss function is used for the

model as follows:

$$\sum_{i=1}^{n} L(\boldsymbol{z}_i, (\alpha, \boldsymbol{\beta}^T)) = -\sum_{i=1}^{n} \log f(g(\alpha + \boldsymbol{x}_i^T \boldsymbol{\beta}), y_i). \tag{1.8}$$

GLM includes linear regression model, logistic regression model and poisson regression model. Logistic regression is used for binary response modelling and poisson regression is commonly used for count response modeling. For such models, Fan and Li (2001) and Zou (2006) proposed SCAD and adaptive Lasso penalty functions.

Cox proportional hazards model is a popular semi-parametric model for survival data (Cox 1972). The Cox model has a parameter of interest and a nuisance parameter, $(\boldsymbol{\beta}, \Lambda)$. We first consider a simple model with right censoring. Denote T_i as the survival time of ith observation and C_i as the subject's right censoring time. Assume that T_i and C_i are independent given $\boldsymbol{x}_i$. We observe n independently and identically distributed samples of the triplet $(Y_i, \delta_i, \boldsymbol{x}_i)$, $i = 1, \cdots, n$, where $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. Furthermore, denote $t_1 < t_2 < \cdots < t_N$ as N ordered observed event times and (j) as the subject's index corresponding to t_j (Fan and Li 2002). The loss function for the right censored data is the partial likelihood for the parameters of interest:

$$\sum_{i=i}^{n} L(\boldsymbol{z}_i, (\alpha, \boldsymbol{\beta}^T)) = -\sum_{j=1}^{N} [\boldsymbol{x}_{(j)}^T \boldsymbol{\beta} - \log\{\sum_{i \in R_j} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})\}], \tag{1.9}$$

where R_i is the risk set at time t_i, $R_i = \{j : Y_j \geq t_i\}$. Tibshirani (1997) and Fan and Li (2002) studied penalization methods for the partial likelihood-based Cox model.

Some survival models do not have an explicit partial likelihood form, such as Cox frailty model and Cox models for interval or doubly censored data. We consider their profile likelihood as an alternative, where the nuisance parameter is profiled out (Murphy and Van der Vaart 2000). Fan and Li (2002) imposed direct penalization for the

profile likelihoods for the frailty model. Similarly, we can regularize the profile likelihood with interval or doubly censored data. Note that these are challenging problems since the corresponding profile likelihoods do not have closed form expressions (Fan and Li 2002).

Undirected graphical models are known to be useful for explaining association structure in multivariate random variables (Lauritzen 1996, Drton and Perlman 2007). We denote a graph as $G = (V, E)$, where $V = \{x_1, \cdots, x_p\}$ is the set of vertices and E is the set of edges between vertices. Each vertex corresponds to a variable and an edge between vertices identifies their conditional dependence given all the other vertices. Figure 1.1 shows a graphical model with five vertices, $(x_1, x_2, x_3, x_4, x_5)$ and four edges, $\{(x_1, x_3),\ (x_2, x_3),\ (x_3, x_4)\ (x_4, x_5)\}$. The first edge, (x_1, x_3) implies that x_1 and x_3 are conditionally dependent given (x_2, x_4, x_5). The other edges can be interpreted in the same manner. Gaussian graphical models (GGM) impose a multivariate Gaussian distribution to the p-dimensional vector, $\boldsymbol{x} = (x_1, \cdots, x_p)$. Denote the distribution as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is a mean vector and $\boldsymbol{\Sigma}$ is a nonsingular covariance matrix. The corresponding loss function is the negative log-likelihood function:

$$-\log|\boldsymbol{\Theta}| + \mathrm{tr}(S\boldsymbol{\Theta}), \tag{1.10}$$

where $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix, $S = \dfrac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$ is the sample covariance matrix, and $\bar{\boldsymbol{x}} = \dfrac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$. The maximum likelihood estimator of $\boldsymbol{\Theta}$ exists and is unique with probability one if $n > p$, and Buhl (1993) studied the case of $n < p$. Estimating the structure of GGM is equivalent to recovering the support of the precision matrix (Lauritzen 1996). Specifically, non-zero off-diagonal elements in the precision matrix correspond to the edge elements of E. Friedman, Hastie, and Tibshirani (2008) and Yuan and Lin (2007a) proposed a L_1 regularization framework for GGM to recover

the support of the precision matrix.

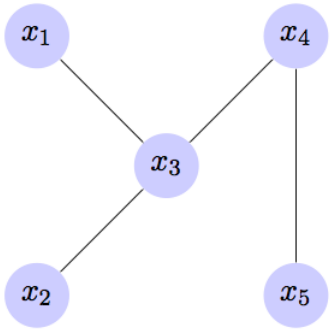


Figure 1.1: Simple Undirected Graph Example (Lee 2013)

1.1.2 Properties and Computational Issues of Penalized Estimation

Penalized estimation can perform simultaneous variable selection and estimation with a proper choice of the penalty function. Tibshirani (1996) gave an intuitive explanation on the sparse estimation for the Lasso. Assume that each predictor is standardized to have mean zero and variance one. The Lasso intercept estimate is $\sum_{i=1}^{n} y_i/n$ and the Lasso estimate for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, is determined by the following constrained optimization problem

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t, \tag{1.11}$$

where t is the tuning parameter. Note that the loss term in (1.11) can be rewritten as $\frac{1}{n}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ls})^T \mathbf{X}^T \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{ls})$ plus a constant, where $\hat{\boldsymbol{\beta}}_{ls}$ is the ordinary least squares estimate and $\mathbf{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^T$. Figure 1.2 illustrates its elliptical contours and the constraint as the black square for $p = 2$. The Lasso estimate is the coordinate that the contours first touch the square. It will be sometimes on the axes, and hence a zero coefficient can be obtained via the Lasso.

Many penalization methods have the general formulation of penalized estimation in (1.2). The Lasso problem in (1.11) can be reformulated as the equivalent optimization

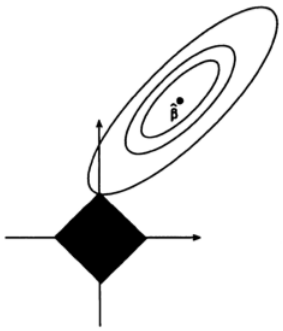


Figure 1.2: Geometry of Lasso ($p = 2$) (Tibshirani 1996)

problem:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{d} |\beta_j|. \tag{1.12}$$

It is the summation of a least squares loss function and L_1 norm penalty. Another example is the graphical Lasso (*glasso*) for sparse inverse covariance estimation (Friedman et al. 2008). It is known to be useful for explaining association structure in high-dimensional data such as gene expression data and microRNA data. We regularize the negative log-likelihood for GGM in (1.10) with the L_1-penalty over a positive definite constraint:

$$\min_{\boldsymbol{\Theta}>0} - \log|\boldsymbol{\Theta}| + \mathrm{tr}(S\boldsymbol{\Theta}) + \lambda||\boldsymbol{\Theta}||_1, \tag{1.13}$$

where $||\boldsymbol{\Theta}||_1$ is the L_1-norm, the sum of the absolute values of the elements of $\boldsymbol{\Theta}$. We estimate the true edge set of the GGM with a proper choice of tuning parameter, and then obtain a sparse GGM.

The penalty functions in penalized estimation can be roughly categorized into two classes: convex penalty functions and nonconvex penalty functions. The convex penalties such as Lasso and adaptive Lasso have computational advantage since the corresponding optimization problems have convex objective functions. The nonconvex penalties might have theoretical advantage over the convex penalties, but their computation might be challenging due to the nonconvexity of the objective function. The

SCAD penalty, the minimax concave penalty (MCP), and the folded concave penalties are common examples of the nonconvex penalties. Further details on these penalty functions can be found in Fan and Li (2001), Zhang (2010) and Fan, Xue, Zou, et al. (2014).

The theoretical properties of penalized estimation have been studied in the literature. Certain penalization methods such as SCAD and adaptive Lasso satisfy the desirable theoretical properties (Fan and Li 2001, Zou 2006). These are known as the oracle properties since the methods asymptotically perform as well as the oracle estimator, which knows the true model in advance (Donoho and Johnstone 1994). The oracle procedures have consistency in variable selection and asymptotic normality of the nonzero coefficients with the same efficiency as the oracle estimator.

Computational algorithms have been intensively studied for many penalization methods. Efron, Hastie, Johnstone, and Tibshirani (2004) proposed a powerful least angle regression (LARS) algorithm, an advanced version of forward selection with the least squares loss. The computational cost for the entire solution path is of the same order as the full ordinary least squares. Its simple modification calculates Lasso and adaptive Lasso estimates with the least squares loss. Zou and Li (2008) developed a unified algorithm based on local linear approximation (LLA) for the nonconvex penalties with negative log-likelihood loss. The proposed one-step LLA estimator from the algorithm reduces the computational burden. Friedman, Hastie, and Tibshirani (2010) suggested a coordinate-wise descent algorithm for the convex penalized least squares regression and GLM. Later, Breheny and Huang (2011) studied its applications to the nonconvex penalties for the least squares regression and the logistic regression. Given the tuning parameter, each step of the algorithm is applied to a single parameter with the remaining parameters fixed, and the updated solution is used as a warm start for the next step. The algorithm can be also used to solve iterative modified Lasso problems

in GGM (Friedman, Hastie, and Tibshirani 2008).

1.2 New Contributions and Outline

The contribution of this dissertation is to give new insights on penalized estimation in various statistical models. We propose some new methods with theoretical investigation and extensive numerical studies. The outline of the proposal is as follows:

- Chapter 2 introduces the notion of overlapping structure in a composite loss function and defines partially overlapping models for several models of interest. We develop the oracle M-estimator for partially overlapping models and establish its theoretical properties. Furthermore, we suggest adaptive composite M-estimation, regularized estimation for the sparsity and overlapping structure recovery of the overlapping models. The method is theoretically justified and numerically demonstrated as competitive against several existing methods with composite loss functions.

- Chapter 3 first introduces the refit method, a simple two-step procedure based on a penalization method. Based on the refitting, we propose ensemble variable selection (EVS) and ensemble variable selection and estimation (EVE). EVS obtains candidate refit estimators according to voting results from several penalization methods and chooses the optimal one by a certain information criterion or cross-validation. Numerical studies illustrate that EVS can often identify the best penalized method in each scenario. Next, EVE is studied for a factorizable likelihood-based model in the penalization framework. In such a model, the full likelihood can be factorized into distinct likelihood factors. EVE is a multi-step procedure based on information combination across the factors, the refitting, and the least squares approximation (LSA) penalization method in Wang and Leng

(2007). We perform numerical studies for simulated prospective doubly censored data and analyze Multicenter AIDS Cohort Study (MACS) data with EVE.

- Chapter 4 studies the edge selection for sparse high-dimensional undirected GGM. We develop consistent validation for edge selection (CoVES) motivated by consistent cross-validation for generalized linear models in Feng and Yu (2013). Its underlying target is a sparse graph model, where a small number of variables are conditionally dependent. CoVES first obtains the candidate edge structures from the entire *glasso* solution path. For each selected graph structure, CoVES computes the empirical negative log-likelihood via repeated random subsampling validation. Finally, CoVES selects the edge structure having the smallest negative log-likelihood as the optimal structure. We study its asymptotic property under growing sample size with fixed dimension and show its competitive performance to conventional selection methods from numerical studies.

CHAPTER2: ADAPTIVE ESTIMATION FOR PARTIALLY OVERLAPPING MODELS

2.1 Introduction

Regression modeling has been a popular statistical tool to explain the association between a response variable and covariates in a dataset. A statistical regression model targets a profile of the conditional distribution of the response given the predictors. We estimate conditional mean of response as a linear function of predictors in classical linear regression while we estimate conditional median as a linear function of predictors in median regression. It is of great interest to consider several linear models to describe a more complete picture of the conditional distribution. We may simultaneously fit the models on the dataset and estimate the parameters. Such joint estimation borrows information across the models and is referred as to composite estimation.

The composite estimation may be based on combing loss functions as weighted averages of loss functions tailored to individual models. Given n independent identically distributed samples, $\boldsymbol{z}_1 = (\boldsymbol{x}_1, y_1), \cdots, \boldsymbol{z}_n = (\boldsymbol{x}_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, consider the following K different empirical convex loss functions to each model:

$$\frac{1}{n}\sum_{i=1}^{n} L_k(\boldsymbol{z}_i, (\alpha_k, \boldsymbol{\beta}_k)) \equiv \frac{1}{n}\sum_{i=1}^{n} L_k(y_i, \alpha_k + \boldsymbol{x}_i^T\boldsymbol{\beta}_k),\ k = 1, \cdots, K, \tag{2.1}$$

where α_k's are the different intercept terms across the models and $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K \in \mathbb{R}^p$ are the parameter vectors for all models of interest. We employ distinct parameter vectors

for the loss functions. Our composite loss function is formulated as:

$$L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)) \equiv \sum_{k=1}^{K} w_k L_k(y_i, \alpha_k + \boldsymbol{x}_i^T \boldsymbol{\beta}_k), \tag{2.2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_K^T)^T \in \mathbb{R}^{K \times p}$, and $\boldsymbol{w} = (w_1, \cdots, w_K)^T$ is a positive weight vector. Note that minimizing (2.2) without further assumptions on parameter overlap is equivalent to minimizing the loss functions separately. The loss functions may have the same or different forms. For example, in composite quantile regression (CQR), each L_k is a check function with the arguments to L_k being used to fit models to different quantiles (Zou and Yuan 2008). For the τ-th quantile, $L^\tau(t) = \tau t_+ + (1-\tau)t_-$, where $t_+ = tI(t \geq 0)$ and $t_- = tI(t < 0)$ respectively. Combining the check function for median regression with the usual least squares loss function yields an example of composite loss functions derived from different L_k.

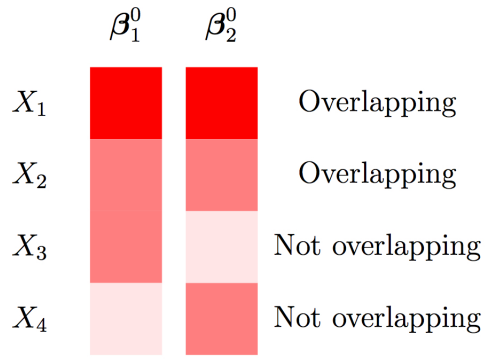


Figure 2.1: Partially Overlapping Models

Composite estimation is useful when the underlying parameter structures are partially overlapped. In the partially overlapping models, some parameters are the same across loss functions, while others are different. Overlap may occur between two or more loss functions. Figure 2.1 shows a simple example of the partially overlapping models. Each parameter vector corresponds to both loss functions (β_1 and β_2). The first and second covariates (X_1 and X_2) have rows of the same color, which impart the

same parameter values (β_1 and β_2) to both loss functions. We call this arrangement overlapping structure. According to the definition of overlapping structure, the third and fourth covariates in this example do not overlap across the models. The fourth element of β_1 and the third element of β_2 demonstrate sparse structure. They appear white-shaded, which indicates that they are zero-valued parameters. Both CQR and L_1-L_2 loss functions may have overlapping parameter vectors for different quantiles or median and expectation, depending on the effects of the covariates on the variance function.

A complete overlapping structure is one extreme of partially overlapping structures, where all parameters are common to all loss functions. For the completely overlapping models, Bradic, Fan, and Wang (2011) and Zou and Yuan (2008) used the composite loss functions, with the goal of improving efficiency of the regression parameter estimators. Their composite loss function has the following form:

$$L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)) \equiv \sum_{k=1}^{K} w_k L_k(y_i, \alpha_k + \boldsymbol{x}_i^T \boldsymbol{\beta}), \tag{2.3}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^K$ and $\boldsymbol{\beta} \in \mathbb{R}^p$. The composite loss in (2.3) is identical to that in (2.2), except that the regression slopes are the same for different k. Such M-estimation has been studied for efficient and sparse estimation when the underlying model follows a classical linear model. The assumption leads to completely overlapping models, where the individual loss functions have a common parameter vector. Note that the existing methods do not consider each loss as a model, but rather consider the composite loss function as an approximation of the unknown log-likelihood function of the error distribution (Bradic, Fan, and Wang 2011).

The completely overlapping modeling in composite loss estimation (Bradic et al. 2011) may limit flexibility in statistical modelling. Consider a linear location-scale

model whose several covariates affect the scale of response and its error is centered to zero but not symmetric. Different loss functions estimate different parameters defined both by the mean and variance of the response. The parameters are the same for the covariates which have no effect on the variance function (Carroll and Ruppert 1988). Parameter vector for L_2 is the same as the regression parameter vector of the model while parameter vector for L_1 is the weighted sum of the regression parameter vector and the scale parameter vector. Other examples in which different loss functions may correspond to models with partially overlapping parameters include composite quantile regression (CQR), in which multiple L_1 loss functions are linked to different quantiles.

In this chapter, we aim for the efficient composite estimation under weaker assumptions on the overlapping structure, the partially overlapping structure. To adapt such overlapping structure in the models, we incorporate penalization into (2.2). The penalty is applied to all absolute pairwise differences between coefficients corresponding to each covariate. In addition to this grouping penalty, we also employ a penalty for sparse estimation, as in Bradic et al. (2011). The objective function for our empirical composite loss function with double penalties is

$$\sum_{k=1}^{K}\sum_{i=1}^{n} w_k L_k(y_i, \alpha_k + \boldsymbol{x}_i^T\boldsymbol{\beta}_k) + n\sum_{k=1}^{K}\sum_{j=1}^{p} p_{\lambda_{1n}}(|\beta_{kj}|) + n\sum_{k<k'}\sum_{j=1}^{p} p_{\lambda_{2n}}(|\beta_{k'j} - \beta_{kj}|). \quad (2.4)$$

The penalty terms in (2.4) applied to the difference in the coefficients enable recovery of the overlapping structure by shrinking small differences towards zero. The penalty term applied to each coefficient encourages sparsity by shrinking small coefficients towards zero. One should recognize that the penalization of the differences is used not for variable selection, but for selecting the overlapping structure across the multiple loss functions. The fused lasso (Tibshirani, Saunders, Rosset, Zhu, and Knight 2004) also has a sparse penalty term combined with a penalty term for pairwise differences. Their

pairwise penalty serves a different purpose, that of identifying local consistency of coefficients in a single model.

In the sequel, we propose and study adaptive composite M-estimation (ACME) based on (2.4) which simultaneously shrinks towards the true overlapping model structure while estimating the shared coefficients in that structure. As in Bradic et al. (2011), our procedure yields estimators with improved efficiency by information combination across the models. Our procedure correctly selects both the true overlap structure and the true non-zero parameters in the true model structure with probability 1 in large samples. The parameter estimators hereby obtained are oracle in the sense that they have the same distribution as the oracle estimator based on knowing the true model structure a prior, both the true overlapping parameters and the true non-zero parameters.

The rest of the chapter is organized as follows. In Section 2.2, we introduce notation for the distinct parameter vector across models, based on overlap in the $\boldsymbol{\beta}_k$'s, and define the oracle estimator. The large sample properties of the oracle estimator are established under partially overlapping models. Section 2.3 presents ACME for partially overlapping models and describes its implementation along with a rigorous discussion of its theoretical properties. Section 2.4 contains numerical results from an extensive simulation study and Section 2.5 reanalyzes a well known dataset on the annual salaries of professional baseball players. All proofs are relegated to Section 2.7.

2.2 Oracle M-estimator for Overlapping Models

Before discussing our procedure, it would be helpful to understand the underlying model and the oracle estimator. Oracle procedures estimate the parameters of interest when the underlying parameter structure is known in advance (Fan and Li 2002). For partially overlapping models, we define the oracle estimation as the unpenalized

estimation with constraints on the sparsity and overlapping structure. We first introduce notations and settings for partially overlapping models and investigate theoretical properties of the oracle estimation.

2.2.1 Models and Notations

We first consider the K separate models with their corresponding loss functions in (2.1). The risk function for the kth model is defined as the expectation of kth loss function, $R_k(\alpha_k, \boldsymbol{\beta}_k) = \mathbb{E}_{\boldsymbol{z}}[L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k)]$ for $\boldsymbol{\beta}_k \in \mathbb{R}^p$, $k = 1, \cdots, K$. The true parameter vector for the kth model is the minimizer of the corresponding risk function, $R_k(\alpha_k, \boldsymbol{\beta}_k)$, with $(\alpha_k^0, \boldsymbol{\beta}_k^{0T})^T = \underset{(\alpha_k, \boldsymbol{\beta}_k^T)^T \in \Theta \subset \mathbb{R}^{p+1}}{\operatorname{argmin}} R_k(\alpha_k, \boldsymbol{\beta}_k)$. We estimate the parameter vector of each model by minimizing its corresponding loss function. Consider a stack of all parameter vectors across all models, and define the $K \cdot (p+1)$-dimensional true parameter vector as $(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T = (\alpha_1^0, \cdots, \alpha_K^0, \boldsymbol{\beta}_1^{0T}, \cdots, \boldsymbol{\beta}_K^{0T})^T$.

Next we describe the underlying parameter structure across the multiple models with set notations. We can identify the underlying sparse and overlapping structure with sparsity sets and overlap sets. Denote $\mathcal{A}_k = \{j \in \{1, \ldots, p\} : \beta_{kj}^0 \neq 0\}$ as the index set of the non-zero parameters to the kth model and $\mathcal{A}_k^c = \{1, \ldots, p\} \backslash \mathcal{A}_k$ as its complement. This set notation implies $\boldsymbol{\beta}_{\mathcal{A}_k^c}^0 = \mathbf{0} \in \mathbb{R}^{|\mathcal{A}_k^c|}$, $k = 1, \cdots, K$, and thus describes the sparse structure of the model k. Note that the underlying sparse structure for all models can be obtained from the collection of the nonzero parameter index sets, $\mathcal{A}^0 \equiv \{\mathcal{A}_k\}_{k=1}^K$.

We further introduce notations between two models for the overlapping structure illustration. Denote $\mathcal{O}_{kk'} = \{j \in \{1, \ldots, p\} : \beta_{kj}^0 = \beta_{k'j}^0 \neq 0\}$ as the index set of the same valued non-zero parameters between $\boldsymbol{\beta}_k^0$ and $\boldsymbol{\beta}_{k'}^0$ for $k \neq k'$. Note that elements of $\mathcal{O}_{kk'}$ corresponds to non-zero same valued parameters to the model k and the model k'. We can obtain the overlapping structure information across all models from the sets

for all model pairs, $\mathcal{O}_{12}, \cdots, \mathcal{O}_{1K}, \mathcal{O}_{23}, \cdots, \mathcal{O}_{K-1,K}$. In other words, the underlying overlapping structure can be illustrated from the collection of the overlapping index sets, $\mathcal{G}^0 \equiv \{O_{kk'}\}_{k \neq k'}$. Consider a collection of all possible overlappings, $\Gamma = \{\mathcal{G}\}_{\mathcal{G} \in \Gamma}$. The true grouping, $\mathcal{G}^0$, is an element of Γ.

With the true sparse structure, $\mathcal{A}^0$, we can decompose the parameters into two parts for partially overlapping models. The first part is for the entire true zero parameters, $\boldsymbol{\beta}_{\mathcal{A}_k^c} = [\beta_{kj}]_{j \in \mathcal{A}_k^c} \in \mathbb{R}^{|\mathcal{A}_k^c|}$, and the second part is the entire true non-zero and intercept parameters, $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}_{\mathcal{A}}^T)^T = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}_{\mathcal{A}_1}^T, \boldsymbol{\beta}_{\mathcal{A}_2}^T, \ \cdots, \ \boldsymbol{\beta}_{\mathcal{A}_K}^T)^T$, where $\boldsymbol{\beta}_{\mathcal{A}_k} = [\beta_{kj}]_{j \in \mathcal{A}_k}$. Note that the true parameter vector for all models, $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$, corresponds to the union of the two parts, $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}_{\mathcal{A}}^T)^T$ and $\boldsymbol{\beta}_{\mathcal{A}_k^c}$, $k = 1, \cdots, K$.

For joint estimation, we define the composite loss function as the linear combination of all loss functions with weights in (2.2). The composite risk function is the expectation of the composite loss function as $R(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T) = \mathbb{E} \sum_{k=1}^{K} w_k L_k(\alpha_k, \boldsymbol{\beta}_k) = \sum_{k=1}^{K} w_k R_k(\alpha_k, \boldsymbol{\beta}_k)$. Note that the composite risk function is a weighted linear combination of all risk functions and is separable into K risk functions. Hence, the minimizer of the composite risk function, $R(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)$, is the true parameter vector for all K models: $(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T = \underset{(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T \in \Theta \subset \mathbb{R}^{K \cdot (p+1)}}{\operatorname{argmin}} R(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)$. The composite risk function can be viewed as the risk function of the parameter vector across all models. Note that the true non-zero and intercept parameter vector is the minimizer of the composite risk function restricted to the non-zero parameters with the overlapping constraint:

$$
\begin{aligned}
(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}_{\mathcal{A}}^{0T})^T = & \underset{(\boldsymbol{\alpha}^T, \boldsymbol{\beta}_{\mathcal{A}}^T)^T}{\operatorname{argmin}} \sum_{k=1}^{K} w_k \mathcal{R}_k(\alpha_k, \boldsymbol{\beta}_{\mathcal{A}_k}) \\
& \text{subject to } \beta_{A_k j} = \beta_{A_{k'} j} \ \forall j \in \mathcal{O}_{kk'}, \ \forall k < k',
\end{aligned}
\tag{2.5}
$$

where $\mathcal{R}_k(\alpha_k, \boldsymbol{\beta}_{\mathcal{A}_k}) = \mathbb{E}_{\boldsymbol{z}} L_k(y, \alpha_k + \boldsymbol{x}^{kT} \boldsymbol{\beta}_{\mathcal{A}_k})$ and $\boldsymbol{x}_i^k = [\boldsymbol{x}_{ij}]_{j \in \mathcal{A}_k}$.

The oracle M-estimator of $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ for partially overlapping models is the unpenalized M-estimator obtained under the assumption that the sparsity and overlapping structure is known in advance. Denote the oracle estimator as $(\hat{\boldsymbol{\alpha}}^{oT}, \hat{\boldsymbol{\beta}}^{oT})^T$. Similar to the true parameters, we have the decomposition of the oracle estimator into the zero parameter part and the non-zero parameter part:

$$\hat{\boldsymbol{\beta}}^o_{\mathcal{A}_k^c} = [\beta^o_{kj}]_{j \in \mathcal{A}_k^c} = \mathbf{0}_{|\mathcal{A}_k^c|} \in \mathbb{R}^{|\mathcal{A}_k^c|}$$
$$(\hat{\boldsymbol{\alpha}}^{oT}, \hat{\boldsymbol{\beta}}^{oT}_{\mathcal{A}})^T = (\hat{\boldsymbol{\alpha}}^{oT}, \hat{\boldsymbol{\beta}}^{oT}_{\mathcal{A}_1}, \cdots, \hat{\boldsymbol{\beta}}^{oT}_{\mathcal{A}_K})^T \in \mathbb{R}^{K + \sum_{k=1}^K |\mathcal{A}_k|}, \text{ where } \hat{\boldsymbol{\beta}}^o_{\mathcal{A}_k} = [\beta^o_{kj}]_{j \in \mathcal{A}_k}.$$

The first part estimates the true zero parameters of all models and the second part estimates the true non-zero and intercept parameters. Since we know the sparsity pattern of the models, $\mathcal{A}_1^c, \cdots, \mathcal{A}_K^c$, we estimate the corresponding parameters as zeros. Analogous to the definition of the true parameters in (2.5), the oracle estimator to the non-zero parameters minimizes the empirical weighted multiple loss functions with the overlapping structure constraint:

$$(\hat{\boldsymbol{\alpha}}^{oT}, \hat{\boldsymbol{\beta}}^{oT}_{\mathcal{A}})^T = \underset{(\boldsymbol{\alpha}^T, \boldsymbol{\beta}_{\mathcal{A}}^T)^T}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K w_k L_k(y_i, \alpha_k + \boldsymbol{x}_i^{kT} \boldsymbol{\beta}_{\mathcal{A}_k})$$
$$\text{subject to } \beta_{A_k j} = \beta_{A_{k'} j} \; \forall j \in \mathcal{O}_{kk'}, \text{ for any } k < k'.$$

2.2.2 Distinct Parametrization and Distinct Oracle M-estimator

The common parametrization in Section 2.2.1 includes the duplication of the same valued parameters from overlapping structures. That is, the parametrization is redundant for partially overlapping models. The left panel of Figure 2.2 shows an example of such redundant parametrization. We use two 4-dimensional parameter vectors, $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^4$, to describe the models from Figure 2.1. The first parameter pair, β_{11} and β_{21}, has the same value, and the second parameter pair, β_{12} and β_{22}, also has another

same value. We can use one parameter, θ_{11}, for β_{11} and β_{21}, and another parameter, θ_{21}, for β_{12} and β_{22} as in the right panel of Figure 2.2. Furthermore, this parametrization excludes the zero-valued parameters, β_{23} and β_{14}. We call such parametrization distinct parametrization or non-redundant parametrization. The underlying sparse and overlapping structure is imposed on the non-redundant parametrization for the true non-zero and intercept parameters. The distinct parametrization is used for a lower dimensional formulation of the oracle M-estimator.

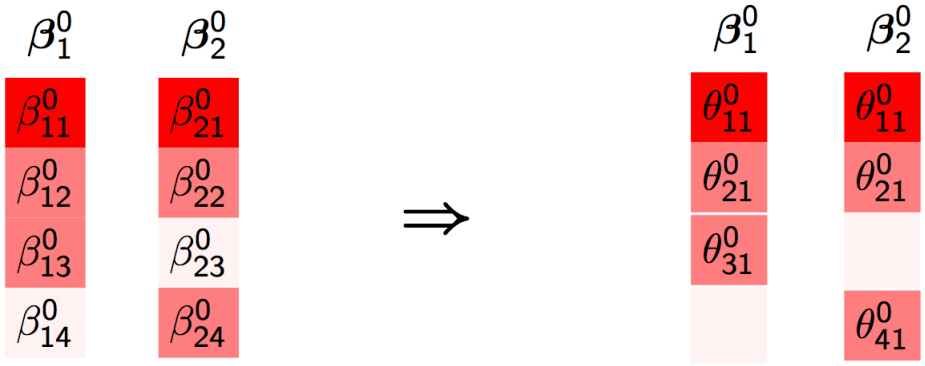


Figure 2.2: Illustration of Distinct Parametrization with $\beta_{14}^0 = \beta_{23}^0 = 0$

To define our distinct oracle estimator, we borrow the notations from Bondell and Reich (2007) for the parametrization. Consider the union of the index sets of the non-zero parameters of all models, $\bigcup_{k=1}^{K} \mathcal{A}_k = \{j_1, \cdots, j_Q\}$. It corresponds to the index set of covariates with a non-zero true parameter in at least one model. Denote its cardinality as $Q = |\bigcup_{k=1}^{K} A_k|$, which is less than or equal to the number of covariates, p. Given a variable, x_{j_q}, $j_q \in \bigcup_{k=1}^{K} \mathcal{A}_k$, consider the unique true non-zero parameter values among the elements of $\{\beta^0_{\mathcal{A}_k j_q} : \forall k \ s.t. \ j_q \in \mathcal{A}_k\}$. They are called the true distinct parameters to the variable, x_{j_q}. For example, we have $\bigcup_{k=1}^{K} A_k = \{1, 2, 3, 4\}$ for the models in Figure 2.2. The first two covariates, x_1 and x_2, have one true non-zero parameter value from $\{\beta^0_{11}, \beta^0_{21}\}$ and $\{\beta^0_{12}, \beta^0_{22}\}$ respectively since we have $\beta^0_{11} = \beta^0_{21}$ and $\beta^0_{12} = \beta^0_{22}$. For the third and fourth covariates, x_3 and x_4, each has one true non-zero parameter value, β^0_{13}

and β_{24}^0, respectively.

Suppose we have the $G_q(\leq K)$ true distinct parameters denoted as $\theta_{q1}^0, \cdots, \theta_{qG_q}^0$ for $q = 1, \cdots, Q$. We denote the true distinct parameter vector across all covariates as

$$\begin{aligned}\boldsymbol{\theta}^0 &= (\boldsymbol{\theta}_0^0, \boldsymbol{\theta}_1^0, \cdots, \boldsymbol{\theta}_Q^0)^T \\ &= (\theta_{01}^0,\ \cdots,\ \theta_{0K}^0,\ \theta_{11}^0,\ \cdots,\ \theta_{1G_1}^0\ \cdots,\ \theta_{Q1}^0,\ \cdots,\ \theta_{QG_Q}^0)^T \in \mathbb{R}^{K+\sum_{q=1}^Q G_q},\end{aligned}$$

where $\boldsymbol{\theta}_0^0 = (\theta_{01}^0,\ \cdots,\ \theta_{0K}^0)^T$ is the true intercept vector, $\boldsymbol{\alpha}^0$. The true distinct parameter vector is the non-redundant enumeration of the true parameters in terms of overlapping structure for all models along the predictors.

We can define the distinct composite loss function with the non-redundant parametrization as $\mathcal{L}(\boldsymbol{z}_i, \boldsymbol{\theta}) = \sum_{k=1}^K w_k \mathcal{L}_k(y_i, \boldsymbol{\theta}_{0k} + \boldsymbol{x}_i^{kT} \boldsymbol{\beta}_{\mathcal{A}_k}(\boldsymbol{\theta}))$, where $[\boldsymbol{\beta}_{\mathcal{A}_k}(\boldsymbol{\theta})]_j$ is an element of $\boldsymbol{\theta}$ corresponding to $\boldsymbol{\beta}_{\mathcal{A}_k j}, j \in \mathcal{A}_k$. The distinct composite loss function is a random convex function on $\mathbb{R}^{K+\sum_{q=1}^Q G_q}$. The distinct composite risk function is the expectation of the distinct composite loss function with $\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z}}[\mathcal{L}(\boldsymbol{z}, \boldsymbol{\theta})] = \sum_{k=1}^K w_k \mathcal{R}_k(\theta_{0k}, \boldsymbol{\beta}_{\mathcal{A}_k}(\boldsymbol{\theta}))$. Note that the minimizer of the distinct composite risk function is the true distinct parameter vector.

The distinct oracle M-estimator of $\boldsymbol{\theta}$ is defined as the minimizer of the distinct loss function as follows:

$$\begin{aligned}\hat{\boldsymbol{\theta}}^o &= (\hat{\theta}_{01}^o,\ \cdots,\ \hat{\theta}_{0K}^o, \hat{\theta}_{11}^o,\ \cdots,\ \hat{\theta}_{1G_1}^o,\ \cdots,\ \hat{\theta}_{Q1}^o,\ \cdots,\ \hat{\theta}_{QG_Q}^o)^T \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\boldsymbol{z}_i, \boldsymbol{\theta}) \in \mathbb{R}^{K+\sum_{q=1}^Q G_q}.\end{aligned}$$

We assume that the dimension of the distinct oracle M-estimator, $K + \sum_{q=1}^Q G_q$, is less than the sample size, n. The distinct oracle M-estimator can be viewed as the non-redundant enumeration of the oracle M-estimator, $(\hat{\boldsymbol{\alpha}}^{oT}, \hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oT})^T$, in terms of overlaps.

Specifically, every element of $\hat{\boldsymbol{\theta}}_q^o$ $(q = 1, \cdots, Q)$ corresponds to some nonzero elements among $\hat{\beta}_{1j_q}^o, \cdots, \hat{\beta}_{Kj_q}^o$ when they are overlapped. Conversely, every nonzero element among $\hat{\beta}_{1j_q}^o, \cdots, \hat{\beta}_{Kj_q}^o$ corresponds to one element of $\hat{\boldsymbol{\theta}}_q^o$.

2.2.3 Asymptotic Properties of Distinct Oracle M-estimator

Before introducing ACME in Section 2.3, we establish the asymptotic properties of the distinct oracle M-estimator in Section 2.2.3. For the theoretical properties, the following assumptions on all K separate loss functions are required.

A1. $(\alpha_k^0, \boldsymbol{\beta}_k^{0T})^T = \underset{(\alpha_k, \boldsymbol{\beta}_k^T)^T \in \Theta \subset \mathbb{R}^{p+1}}{\operatorname{argmin}} \mathbb{E}L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k),\ k = 1, \cdots, K$ are bounded and unique.

A2. $\mathbb{E}L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k) < \infty$ for each $(\alpha_k, \boldsymbol{\beta}_k^T) \in \mathbb{R}^{p+1},\ k = 1, \cdots, K$.

A3. a) $L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k)$ is differentiable w.r.t. $(\alpha_k, \boldsymbol{\beta}_k^T)^T$ at $(\alpha_k^0, \boldsymbol{\beta}_k^0)$ for $\mathbb{P}_{\boldsymbol{z}}$-almost every $\boldsymbol{z} = (\boldsymbol{x}, y)$ with derivative $\nabla_{(\alpha_k, \boldsymbol{\beta}_k^T)^T} L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k^0)$ and

$$J_k(\alpha_k^0, \boldsymbol{\beta}_k^0) \equiv \mathbb{E}[\nabla_{(\alpha_k, \boldsymbol{\beta}_k^T)^T} L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k^0) \cdot \nabla_{(\alpha_k, \boldsymbol{\beta}_k^T)^T} L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k^0)^T] < \infty.$$

b) The risk function $R_k(\alpha_k, \boldsymbol{\beta}_k) = \mathbb{E}[L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k)]$ is twice differentiable w.r.t. $(\alpha_k, \boldsymbol{\beta}_k^T)^T$ at $(\alpha_k^0, \boldsymbol{\beta}_k^{0T})^T$ with a positive definite Hessian matrix, $H_k(\alpha_k^0, \boldsymbol{\beta}_k^0)$.

A4. The loss function, $L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k)$, is convex with respect to $(\alpha_k, \boldsymbol{\beta}_k^T)^T$ for $\mathbb{P}_{\boldsymbol{z}}$-almost every $\boldsymbol{z}$.

Similar conditions can be found for one model setting in Section 2.1 of Rocha, Wang, and Yu (2009). The assumption, *A1*, ensures that the parameter for the kth model, $(\alpha_k^0, \boldsymbol{\beta}_k^{0T})^T$, is well defined. The second assumption, *A2*, guarantees that the pointwise limit of the loss function is the risk function. From *A3*, we can consider local quadratic asymptotic approximations to the risk function around the parameter. Note that we

approximate the loss function to the risk function at each point near the parameter. The last assumption, *A4*, is used to apply Convexity Lemma (Pollard 1991) for the uniformity of approximation.

Lemma 2.1 shows that the composite loss function of (2.3) satisfies the same assumptions as *A1-A4* if all loss functions, $L_1, \cdots, L_K$, satisfy the assumptions. In other words, the composite loss function automatically satisfies the desirable properties for such approximation.

Lemma 2.1. *If all loss functions, $L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k)$, $k = 1, \cdots, K$, satisfy the assumptions, A1, $\cdots$, A4, then the composite loss function, $L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T))$ also satisfies the same assumptions.*

Next we present Lemma 2.2 under the same assumptions for theoretical investigation of the oracle M-estimator and ACME. We prove consistency and asymptotic normality of the distinct oracle M-estimator and $\sqrt{n}$-consistency, selection and overlapping consistency, and asymptotic normality of ACME in Section 2.3.2.

Lemma 2.2. *If each loss function, $L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k)$, $k = 1, \cdots, K$, satisfies the assumptions, A1-A4, then*

(a) There exists a $K \cdot (p+1)$ dimensional random vector $\boldsymbol{W} \sim N(0, J(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}))$ such that

$$\sum_{i=1}^{n}[L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) + \frac{\boldsymbol{u}^T}{\sqrt{n}}) - L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}))] - [\frac{1}{2}\boldsymbol{u}^T \cdot H(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) \cdot \boldsymbol{u} + \boldsymbol{W}^T \cdot \boldsymbol{u}] \xrightarrow{p} 0$$

for each $\boldsymbol{u} \in \mathbb{R}^{K\cdot(p+1)}$

(b) For every compact set $K \subset \mathbb{R}^{K\cdot(p+1)}$,

$$\sup_{\boldsymbol{u}\in K} || \sum_{i=1}^{n}[L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) + \frac{\boldsymbol{u}^T}{\sqrt{n}}) - L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}))]$$

$$-[\frac{1}{2}\boldsymbol{u}^T \cdot H((\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})) \cdot \boldsymbol{u} + \boldsymbol{W}^T \cdot \boldsymbol{u}]|| \xrightarrow{p} 0$$

Lemma 2.2 shows the pointwise convergence and the uniform convergence of the loss, $\sum_{i=1}^{n}[L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) + \frac{\boldsymbol{u}^T}{\sqrt{n}}) - L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}))]$. It is a generalization of Lemma 2 of Rocha et al. (2009), which considers the setting of a single loss function.

The distinct oracle M-estimator is a special type of M-estimators based on the distinct loss function. Its asymptotic properties are established using M-estimation theories.

Lemma 2.3. *If the loss assumptions, A1-A4, are satisfied for all K separate loss functions, then $\hat{\boldsymbol{\theta}}^o$ converges in probability to $\boldsymbol{\theta}^0$ as $n \to \infty$.*

Lemma 2.3 shows the consistency of the distinct oracle M-estimator, which is used for Theorem 2.1. It states that the distinct oracle M-estimator has the asymptotic normality.

Theorem 2.1. *If the loss assumptions, A1-A4, are satisfied for all K separate loss functions, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^o - \boldsymbol{\theta}^0) \xrightarrow{d} N(0, \mathcal{H}(\boldsymbol{\theta}^0)^{-1}\mathcal{J}(\boldsymbol{\theta}^0)\mathcal{H}(\boldsymbol{\theta}^0)^{-1})), \text{ as } n \to \infty$$

$$\text{where } [\mathcal{H}(\boldsymbol{\theta}^0)]_{ij} = \frac{\partial^2 \mathcal{R}(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}, \text{ and } \mathcal{J}(\boldsymbol{\theta}^0) = \mathbb{E}[\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{z}, \boldsymbol{\theta}^0)\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{z}, \boldsymbol{\theta}^0)^T].$$

The non-redundant oracle estimator across models asymptotically follows a normal distribution, similar to some oracle estimators based on a single model. We can extend the results for the original estimators as shown in Corollary 2.1.

Corollary 2.1. *If the above assumptions are satisfied, then $\sqrt{n}(\hat{\boldsymbol{\beta}}^o_{\mathcal{A}_k} - \boldsymbol{\beta}^0_{\mathcal{A}_k}) = O_p(1)$ for all $k = 1, \cdots, K$.*

From Corollary 2.1, we also have $\sqrt{n}-$consistency of the composite oracle estimator, $\hat{\boldsymbol{\beta}}^{o}_{\mathcal{A}}$. The asymptotic property is preserved because the oracle estimator for each model is a subset of the distinct oracle estimator.

2.3 Adaptive Composite M-estimation for Overlapping Structure

The joint estimation procedure, ACME, improves the performance of all models as it shares the information across the multiple models. The penalized estimation recovers the true parameter structure in terms of the sparsity and overlapping. The two penalty terms in ACME objective function, (2.4), control the sparsity and overlapping level.

2.3.1 Choice of Penalty Functions

For the two penalty terms, $p_{\lambda_{1n}}(|t|)$ and $p_{\lambda_{2n}}(|t|)$, we consider folded concave penalty functions and weighted L_1 penalty functions. First, the general folded concave penalty functions on $t \in [0, \infty)$ satisfy the conditions below (Fan, Xue, Zou, et al. 2014)

(i) $p_\lambda(t)$ is increasing and concave in $t \in [0, \infty)$;

(ii) $p_\lambda(t)$ is differentiable in $t \in (0, \infty)$ with $p'_\lambda(0) := p'_\lambda(0+) \geq a_1\lambda$;

(iii) $p'_\lambda(t) \geq a_1\lambda$ for $t \in (0, a_2\lambda]$;

(iv) $p'_\lambda(t) = 0$ for $t \in [a\lambda, \infty)$ with the pre-specified constant $a > a_2$,

where a_1 and a_2 are some fixed positive constants. The penalty function is differentiable on $t \in (0, \infty)$ and right differentiable at zero, thus it can produce sparse solutions. The penalty functions are flat for $t \in [a\lambda, \infty)$ to reduce the estimation bias. The SCAD and MCP penalty functions are typical examples of the folded concave penalty functions.

For $\theta > 0$, the first derivative of the SCAD penalty is

$$p'_\lambda(\theta) = \lambda I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda),$$

where $a > 2$ and $\lambda > 0$ are tuning parameters (Fan and Li 2001). We commonly select $a = 3.7$. Note that SCAD has $a_1 = 1$ and $a_2 = 1$ in the form of the folded concave penalty functions. MCP is defined as

$$p_\lambda(\theta) = \lambda \int_0^\theta (1 - \frac{x}{a\lambda})^+ dx,$$

with a tuning parameter $a > 1$. MCP has $a_1 = 1 - a^{-1}$ and $a_2 = 1$ in the form of the folded concave penalty functions (Zhang 2010).

The weighted L_1 penalties take the form of $p'_\lambda(|\hat{\theta}^{(0)}|)|\theta|$, where $\hat{\theta}^{(0)}$ is a consistent estimator of θ^0. The weighted L_1 penalty function provides a one-step local linear approximation estimator (Zou and Li 2008). We consider two types of the preliminary penalty functions for the weighted L_1 penalty functions, $p_\lambda(t)$. The first function is the folded concave penalty and the second one is $p_\lambda(t) = \lambda p(t)$, where $p'(t)$ is continuous on $(0, \infty)$ and there is some $s > 0$ such that $p'(t) = O(t^{-s})$ as $t \to 0+$. Additionally, the adaptive Lasso penalty is obtained by letting $p'_\lambda(\hat{\theta}^{(0)}) \equiv \lambda |\hat{\theta}^{(0)}|^{-s}$, where $s > 0$ (Zou 2006). We adopt the one-step SCAD penalty in numerical studies for both steps in Section 2.4.

2.3.2 Theoretical Results

We establish the theoretical properties of ACME under the assumptions, *A1-A4*, on all models. We develop the asymptotic theories based on the objective function in (2.4), which is denoted as $Q_n(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)$. In particular, we focus on the oracle properties of ACME for partially overlapping models.

Lemma 2.4. *If $\lambda_{1n} \to 0$, $\lambda_{2n} \to 0$ for folded concave, one-step folded concave penalty functions, and $\sqrt{n}\lambda_{1n} \to \infty$, $\sqrt{n}\lambda_{2n} \to \infty$ for weighted L_1 penalty functions, there is a local minimizer of $Q_n(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)$ such that*

$$\sqrt{n}|(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T)^T - (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T| = O_p(1).$$

If both $p_{\lambda_{1n}}(t)$ and $p_{\lambda_{2n}}(t)$ are weighted L_1 penalty functions, then $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T)^T$ is the unique global minimizer.

Lemma 2.4 demonstrates the existence of a $\sqrt{n}$-consistent penalized M-estimator with a proper choice of λ_n. We control the magnitude of $Q_n((\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T) + \boldsymbol{u}^T/\sqrt{n}) - Q_n(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)$ for a sufficiently large $|\boldsymbol{u}|$ to show the selection and overlapping consistency in Theorem 2.2. The notion of overlapping consistency is analogous with that of selection consistency. We achieve the overlapping consistency and both $\hat{\beta}_{kj}$ and $\hat{\beta}_{k'j}$ have the exactly same values for any index $j \in \mathcal{O}_{kk'}$ with probability tending to 1.

Theorem 2.2. *Suppose that $\lambda_{1n} \to 0$, $\lambda_{2n} \to 0$, $\sqrt{n}\lambda_{1n} \to \infty$, $\sqrt{n}\lambda_{2n} \to \infty$ for folded concave, one-step folded concave penalty functions. For weighted L_1 penalty functions, suppose $\sqrt{n}\lambda_{1n} \to 0$, $\sqrt{n}\lambda_{2n} \to 0$, $n^{\frac{s+1}{2}}\lambda_{1n} \to \infty$, $n^{\frac{s+1}{2}}\lambda_{2n} \to \infty$. If there exists at least one $j \in \mathcal{O}_{kk'}$ for some $k < k'$, then $P(\bigcap_{k=1}^{K} \bigcap_{j \in \mathcal{A}_k^c} \{\hat{\beta}_{kj} = 0\} \cap \bigcap_{k<k'} \bigcap_{j \in \mathcal{O}_{kk'}} \{\hat{\beta}_{kj} = \hat{\beta}_{k'j}\}) \to 1$ as $n \to \infty$.*

Theorem 2.2 implies that the ACME achieves selection consistency and overlapping consistency. Let $\hat{\mathcal{A}}_k = \{j \in \{1, \cdots, p\} : \hat{\beta}_{kj} \neq 0\}$ denote as the non-zero coefficient index set corresponding to the kth loss function. Denote $\hat{\mathcal{G}}$ as the estimated grouping. The selection and overlapping consistency can be written as $P(\{\hat{\mathcal{A}}_k = \mathcal{A}_k, k = 1, \cdots, K\} \cap \{\hat{\mathcal{G}} = \mathcal{G}_0\}) \to 1$.

Let $\hat{\boldsymbol{\theta}}_{\mathcal{A}^0}(\mathcal{G}^0)$ denote our distinct ACME from (2.4) provided we know the true overlapping structure, $\mathcal{G}^0$, and the true sparse structure, $\mathcal{A}^0$. We focus on the asymptotic

distribution of $\hat{\boldsymbol{\theta}}_{\mathcal{A}^0}(\mathcal{G}^0)$ since our estimator selects the true K models and has the true overlapping structure with probability tending to one. Note that its dimension is same as the dimension of the distinct oracle estimator.

Theorem 2.3. *If the assumptions in Theorem 2.2 are satisfied, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}}(\mathcal{G}^0) - \boldsymbol{\theta}^0) \xrightarrow{d} N(0, \mathcal{H}(\boldsymbol{\theta}^0)^{-1}\mathcal{J}(\boldsymbol{\theta}^0)\mathcal{H}(\boldsymbol{\theta}^0)^{-1})).$$

Theorem 2.3 states that the distinct estimator has the same asymptotic distribution as the distinct oracle estimator in Theorem 2.1. The ACME across the multiple models follows a normal distribution in terms of non-zero non-redundant enumeration as the penalized estimators of a single model for the non-zero parameters follow a normal distribution (Fan and Li 2001).

2.3.3 Choice of Weights and Tuning Parameters

The asymptotic distribution of the distinct ACME in Theorem 2.3 leads to theoretical optimal weights to achieve the efficiency across the multiple models. The theoretical criterion for the choice of weights is to maximize the efficiency of the estimator (Bradic et al. 2011). We can use the determinant of the asymptotic covariance matrix of the estimator or its trace as the criterion. Note that its asymptotic covariance is a function of the unknown matrices of $\mathcal{J}(\boldsymbol{\theta}^0)$ and $\mathcal{H}(\boldsymbol{\theta}^0)$, and both depend on the weight vector, $\boldsymbol{w}$. Similarly, completely overlapping models also have the asymptotic normal distribution and their asymptotic covariance depends on the weight vector (Bradic et al. 2011). In this underlying classical linear model setup, the asymptotic covariance matrix can be simplified as the multiplication of a scalar function and a function of predictors. The scalar function has the weight vector and the random errors of the model as its variables, thus the weight vector can be decoupled from the asymptotic covariance

matrix. Bradic et al. (2011) chooses the weight vector by minimizing the function. However, such decoupling cannot be obtained for partially overlapping models, due to the complex form of the asymptotic covariance.

To address the problem, we suggest a data dependent approach to select weights. We first obtain the separate penalized M-estimators as the initial separate estimators with

$$\hat{\boldsymbol{\beta}}_k^{(0)} = \underset{(\alpha_k, \boldsymbol{\beta}_k^T)^T \in \Theta \subset \mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^{n} L_k(y_i, \alpha_k + \boldsymbol{x}_i^T \boldsymbol{\beta}_k) + n \sum_{j=1}^{p} p_{\lambda_{1n}}(|\beta_{kj}|), k = 1, \cdots, K.$$

The preliminary M-estimator achieves sparse estimation, but does not attain overlapping estimation. Note that zero-estimated parameters can be estimated as non-zero in the ACME procedure. Next we calculate data-driven weights, $\boldsymbol{w} = (w_1, \cdots, w_K)^T$ based on the preliminary estimators. We set w_k to be proportional to the reciprocal of the empirical loss function of the initial estimators with

$$w_k \propto [\frac{1}{n} \sum_{i=1}^{n} L_k(y_i, \alpha_k + \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_k^{(0)})]^{-1}, \ k = 1, \cdots, K.$$

We recommend this weight ratio for the same leverage of each loss function to the composite loss function. For computational efficiency, they are rescaled to have sum to one as $\sum_{k=1}^{K} w_k = 1$. We adopt this choice of weights in numerical studies of Section 2.4, which yields excellent performance. We assume positive weights because the presence of a zero weight automatically removes the parameter vector of the corresponding model.

To obtain the optimal tuning parameters for λ_{1n} and λ_{2n}, we use 5-fold cross-validation (Fan and Li 2001). Denote the full dataset by $T = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n\}$. We randomly divide T into the five test sets, $T_1, \ \cdots, \ T_5$. Then, their corresponding training sets are $T - T_1, \ \cdots, \ T - T_5$. We obtain the ACME from the vth training set $T - T_v$ as $(\hat{\boldsymbol{\alpha}}^{(v)T}, \hat{\boldsymbol{\beta}}^{(v)T})^T = (\hat{\boldsymbol{\alpha}}^{(v)T}, \hat{\boldsymbol{\beta}}_1^{(v)T}, \ \cdots, \ \hat{\boldsymbol{\beta}}_K^{(v)T})^T$. We choose the optimal tuning

parameter pairs by minimizing the following cross-validation criterion:

$$CV(\lambda_{1n}, \lambda_{2n}) = \sum_{v=1}^{5} \sum_{\boldsymbol{z}_i \in T_v} \sum_{k=1}^{K} \frac{L_k(y_i, \hat{\alpha}_k^{(v)} + \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_k^{(v)})}{L_k(y_i, \hat{\alpha}_k^{(0)} + \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_k^{(0)})}.$$

A two dimensional grid search is performed for the selection of $(\lambda_{1n}, \lambda_{2n})$.

2.4 Simulation Studies

We first perform simulation studies under a classical linear model and a linear location-scale model. Each dataset in Sections 2.4.1-2.4.2 is generated from both of these two models. We obtain ACME for both least absolute deviations (LAD) regression and least squares (LS) regression with a composite L_1-L_2 loss function. We compare it with separate LAD and LS estimators such as ordinary unpenalized LAD and LS estimators (Ordinary), adaptive Lasso penalized LAD and LS estimators (AdLasso), and one-step SCAD penalized LAD and LS estimators (SCAD). We also compare with penalized composite quasi-likelihood (PCQ) in Bradic et al. (2011), which is developed for a classical linear model. PCQ assumes the completely overlapping structure across all loss functions.

For comparison, we report the median of model errors (MME), the standard error of model errors (SE), the number of correctly classified non-zero estimators (TP), and the number of incorrectly classified zero estimators (FP). The model error of each estimator is defined as $ME(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T \mathbb{E}(\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)$. We also evaluate the overlapping performance across the LAD and LS models. The overlapping structures are categorized into four types: truly grouped estimators, truly grouped non-zero estimators, truly grouped zero estimators, and truly ungrouped estimators. Denote the index set of each category as TG, NG, ZG, and UG respectively. Since TG is partitioned into NG and ZG, TG ratio is the weighted average of NG ratio and ZG ratio with the weights of

|NG|/|TG| and |ZG|/|TG|.

2.4.1 Classical Linear Regression Model

In this section, we consider the classical linear model from Fan and Li (2001):

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}^0 + \epsilon_i,$$

where $\boldsymbol{\beta}^0 = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)$. The covariate $\boldsymbol{x}_i$ is multivariate normal with zero mean and covariance, $\text{Cov}(x_{ij_1}, x_{ij_2}) = 0.5^{|j_1 - j_2|}$, $1 \leq j_1, j_2 \leq 12$. Suppose that the error term, $\epsilon_1, \cdots, \epsilon_n$, follows a normal distribution ($N(0, 3)$), a double exponential distribution (DE), and a t distribution with d.f. 4 ($t(4)$). We consider both LAD regression and LS regression. In this case, the true models are completely overlapped since the true parameter vector of the LS regression is the same as the true parameter vector of the LAD regression. For these models, both PCQ and ACME use the composite L_1-L_2 loss function. Our choice of weight for ACME is $(w_1, w_2) \propto (1/MAE(\hat{\alpha}_{lad}^{SCAD}, \hat{\boldsymbol{\beta}}_{lad}^{SCAD}), 1/MSE(\hat{\alpha}_{ls}^{SCAD}, \hat{\boldsymbol{\beta}}_{ls}^{SCAD}))$, where $MAE(\hat{\alpha}_{lad}^{SCAD}, \hat{\boldsymbol{\beta}}_{lad}^{SCAD})$ is the mean of absolute errors of the SCAD-LAD estimator and $MSE(\hat{\alpha}_{ls}^{SCAD}, \hat{\boldsymbol{\beta}}_{ls}^{SCAD}))$ is the mean of squared errors of the SCAD-LS estimator. The results are obtained from 100 simulated datasets with $n = 100$ and $n = 500$. We use 5-fold cross-validation for the tuning parameter selection.

From the first three columns of Tables 2.1 and 2.2, the performance of ACME is the best for both L_1 and L_2 under DE error with $n = 100, 500$ and under $t(4)$ with $n = 100$ in terms of MME. Under $N(0, 3)$ with $n = 100, 500$, the MMEs of the PCQ are smaller than those of ACME, but ACME outperforms the others. In this setting, PCQ is generally comparable to ACME because PCQ achieves the oracle overlapping structure. All the estimators successfully select the significant variables, $\beta_1^0, \beta_2^0, \beta_5^0$, as

evidenced by TP. ACME performs the best in terms of FP in most cases.

	Estimation	N(0,3) MME (TP, FP)	DE MME (TP, FP)	t(4) MME (TP, FP)	LLS MME (TP, FP)
LAD	Oracle	0.1192	0.0484	0.0482	0.4853
		(3 , 0)	(3 , 0)	(3 , 0)	(10 , 0)
	Ordinary	0.5643	0.34	0.2493	0.9383
		(3 , 9)	(3 , 9)	(3 , 9)	(10 , 8)
	AdLasso	0.2713	0.1115	0.1008	0.7472
		(3 , 2.52)	(3 , 1.84)	(3 , 2.44)	(9.97 , 2.42)
	SCAD	0.2632	0.091	0.1014	0.6476
		(3 , 2.48)	(3 , 1.59)	(3 , 2.17)	(9.96 , 1.56)
	PCQ oracle	0.0738	0.067	0.0386	6.8094
		(3 , 0)	(3 , 0)	(3 , 0)	(7 , 0)
	PCQ	0.1395	0.1356	0.0981	14.3802
		(3 , 1.97)	(3 , 3.72)	(3 , 3)	(9.59 , 7.1)
	ACME oracle	0.0786	0.0642	0.0411	0.6278
		(3 , 0)	(3 , 0)	(3 , 0)	(10 , 0)
	ACME	0.1761	0.085	0.0694	0.6717
		(3 , 1.62)	(3 , 1.16)	(3 , 1.4)	(9.78 , 1.03)
LS	Oracle	0.0727	0.0881	0.0428	2.866
		(3 , 0)	(3 , 0)	(3 , 0)	(7 , 0)
	Ordinary	0.3892	0.3871	0.2794	10.0807
		(3 , 9)	(3 , 9)	(3 , 9)	(7 , 11)
	AdLasso	0.1569	0.1647	0.1054	6.0877
		(3 , 1.79)	(3 , 1.88)	(3 , 1.83)	(6.88 , 3.26)
	SCAD	0.1436	0.1719	0.1038	6.4209
		(3 , 1.96)	(3 , 2.11)	(3 , 2.06)	(6.88 , 4.82)
	PCQ oracle	0.0738	0.067	0.0386	1.6273
		(3 , 0)	(3 , 0)	(3 , 0)	(7 , 0)
	PCQ	0.1395	0.1356	0.0981	8.3698
		(3 , 1.97)	(3 , 3.72)	(3 , 3)	(7 , 9.69)
	ACME oracle	0.0786	0.0642	0.0411	1.48
		(3 , 0)	(3 , 0)	(3 , 0)	(7 , 0)
	ACME	0.1434	0.1238	0.0802	5.3363
		(3 , 1.63)	(3 , 1.41)	(3 , 1.51)	(6.85 , 2.38)

Table 2.1: Simulation Results with Model Errors and Numbers of Correct Non-Zeros/Incorrect Zeros (n=100)

In this setting, we have TG= $\{1, 2, \cdots, 11, 12\}$, NG=$\{1, 2, 5\}$, ZG=$\{3, 4, 6, \cdots, 12\}$ and UG= $\emptyset$. In the first three rows of Table 2.3, ACME has reasonable ratios of the NG as well as the ZG. Most ZGs are higher than NGs since the two penalty terms for overlapping and sparsity encourage to increase the ZG ratio. We can view that the NG ratio is a more accurate measure on the performance of the overlapping penalization

	Estimation	N(0,3) MME (TP, FP)	DE MME (TP, FP)	t(4) MME (TP, FP)	LLS MME (TP, FP)
LAD	Oracle	0.0255	0.0072	0.0072	0.0453
		(3 , 0)	(3 , 0)	(3 , 0)	(10 , 0)
	Ordinary	0.1074	0.0409	0.0403	0.0589
		(3 , 9)	(3 , 9)	(3 , 9)	(10 , 7.99)
	AdLasso	0.0453	0.0134	0.0148	0.0544
		(3 , 1.69)	(3 , 1.52)	(3 , 1.79)	(10 , 1.17)
	SCAD	0.0393	0.0126	0.0132	0.0489
		(3 , 1.53)	(3 , 1.42)	(3 , 1.58)	(10 , 0.85)
	PCQ oracle	0.014	0.0082	0.0074	5.6941
		(3 , 0)	(3 , 0)	(3 , 0)	(7 , 0)
	PCQ	0.0174	0.0224	0.0148	8.8911
		(3 , 1.12)	(3 , 3.38)	(3 , 2.56)	(9.99 , 7.85)
	ACME oracle	0.0156	0.0088	0.0071	0.059
		(3 , 0)	(3 , 0)	(3 , 0)	(10 , 0)
	ACME	0.0311	0.0108	0.01	0.0542
		(3 , 0.82)	(3 , 1.17)	(3 , 1.14)	(10 , 0.3)
LS	Oracle	0.0135	0.0133	0.0096	0.6803
		(3 , 0)	(3 , 0)	(3 , 0)	(7 , 0)
	Ordinary	0.0712	0.0671	0.0471	1.7359
		(3 , 9)	(3 , 9)	(3 , 9)	(7 , 11)
	AdLasso	0.0229	0.0238	0.0178	1.0036
		(3 , 1.16)	(3 , 1.27)	(3 , 1.39)	(7 , 2.31)
	SCAD	0.0191	0.024	0.012	1.1313
		(3 , 1.22)	(3 , 1.56)	(3 , 1)	(7 , 3.39)
	PCQ oracle	0.014	0.0082	0.0074	1.4777
		(3 , 0)	(3 , 0)	(3 , 0)	(7 , 0)
	PCQ	0.0174	0.0224	0.0148	1.5568
		(3 , 1.12)	(3 , 3.38)	(3 , 2.56)	(7 , 10.84)
	ACME oracle	0.0156	0.0088	0.0071	0.2633
		(3 , 0)	(3 , 0)	(3 , 0)	(7 , 0)
	ACME	0.0189	0.0206	0.0132	0.7471
		(3 , 0.92)	(3 , 1.32)	(3 , 1.28)	(7 , 1.01)

Table 2.2: Simulation Results with Model Errors and Numbers of Correct Non-Zeros/Incorrect Zeros (n=500)

than the ZG ratio. The ZG ratio of ACME is almost 30% higher than that of all separate estimators under the both $n = 100$ and $n = 500$. ACME has almost two thirds NG ratio except for the normal distribution with $n = 100$. Note that Ordinary, AdLasso, and SCAD have zero NG ratios because the separate estimation does not involve any overlapping penalization. PCQ possesses complete overlapping because the dataset is assumed to be generated from a classical linear model. Hence, PCQ successfully recovers the overlapping structure.

2.4.2 Linear Location-Scale Model

Under linear location-scale models, both LS regression and LAD regression are partially overlapping models as some covariates affect the scale of the response. Our dataset is generated from the following linear location-scale model:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}^0 + \boldsymbol{x}_i^T \boldsymbol{\gamma}^0 \epsilon_i,$$

where $\boldsymbol{\beta}^0 = (3, 3, 3, 3, 3, 3, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)^T$ and $\boldsymbol{\gamma}^0 = (0, 0, 0, 0, 3, -3, 3, -3, 3, -3, 0, 0, 0, 0, 0, 0, 0, 0)^T$. The covariate, $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{i18})^T$, is generated from a multivariate standard normal distribution, $N(\mathbf{0}, I_{18\times 18})$. Assume that the error term, ϵ_i, follows a shifted gamma distribution, $\Gamma(0.25,\ 2) - 0.5$. Note that the distribution is skewed to the right and centered to mean 0. The true parameter vector of the LS regression model is $\boldsymbol{\beta}_{ls}^0 = \boldsymbol{\beta}^0$ and the true parameter vector of LAD regression model is $\boldsymbol{\beta}_{lad}^0 = (3, 3, 3, 3, 1.762, 4.238, 1.762, 1.238, -1.238, 1.238, 0, 0, 0, 0, 0, 0, 0, 0)^T$. Similar to Section 2.4.1, we use the composite L_1-L_2 loss function. We implement the simulation with 100 repetitions under $n = 100$ and $n = 500$.

From the last columns of Tables 2.1 and 2.2, the ACME has the second smallest MME for LAD regression, and the smallest MME for LS regression with $n = 100, 500$.

		n=100				n=500			
	Category	TG	NG	ZG	UG	TG	NG	ZG	UG
N(0,3)	Oracle	0.75	0	1		0.75	0	1	
	Ordinary	0	0	0		0.0008	0	0.0011	
	AdLasso	0.4883	0	0.6511		0.56	0	0.7467	
	SCAD	0.4758	0	0.6344		0.57	0	0.76	
	PCQ oracle	1	1	1		1	1	1	
	PCQ	1	1	1		1	1	1	
	ACME oracle	1	1	1		1	1	1	
	ACME	0.78	0.5567	0.8544		0.8692	0.69	0.9289	
t(4)	Oracle	0.75	0	1		0.75	0	1	
	Ordinary	0	0	0		0	0	0	
	AdLasso	0.4767	0	0.6356		0.5333	0	0.7111	
	SCAD	0.4725	0	0.63		0.5683	0	0.7578	
	PCQ oracle	1	1	1		1	1	1	
	PCQ	1	1	1		1	1	1	
	ACME oracle	1	1	1		1	1	1	
	ACME	0.8508	0.6867	0.9056		0.8575	0.7033	0.9089	
DE	Oracle	0.75	0	1		0.75	0	1	
	Ordinary	0	0	0		0	0	0	
	AdLasso	0.5008	0	0.6678		0.5567	0	0.7422	
	SCAD	0.5117	0	0.6822		0.5392	0	0.7189	
	PCQ oracle	1	1	1		1	1	1	
	PCQ	1	1	1		1	1	1	
	ACME oracle	1	1	1		1	1	1	
	ACME	0.8333	0.6667	0.8889		0.8408	0.6833	0.8933	
LLS	Oracle	0.6667	0	1	0	0.6667	0	1	0
	Ordinary	0	0	0	0	0	0	0	0
	AdLasso	0.3458	0	0.5187	0.0033	0.45	0	0.675	0
	SCAD	0.3017	0	0.4525	0.0017	0.4125	0	0.6188	0
	PCQ oracle	1	1	1	1	1	1	1	1
	PCQ	1	1	1	1	1	1	1	1
	ACME oracle	1	1	1	0	1	1	1	0
	ACME	0.7458	0.53	0.8538	0.2217	0.8642	0.6925	0.95	0.005

Table 2.3: Simulation Results with Grouping Ratios

The SCAD has the smallest MME for LAD and the SCAD has the second smallest MME for LS. Both the separate estimators and ACME show much better performance for the LAD regression than the LS regression due to the skewed error distribution. From this point of view, it is desirable to have a trade-off between LAD and LS estimation performance as in ACME. The ACME sacrifices the LAD estimation performance about 5% with $n = 100$ and 10% with $n = 500$ while it gains the LS estimation performance almost 15% with $n = 100$ and 30% with $n = 500$. Overall, ACME has very competitive performance in terms of MME, sparsity and overlapping structure recovery. The performance of PCQ is poor as expected because both LAD and LS regression models are assumed to be completely overlapped.

The grouping performance results under this model is summarized at the bottom of Table 2.3. We have TG= $\{1, 2, 3, 4, 11, \cdots, 18\}$, NG= $\{1, 2, 3, 4\}$, ZG= $\{11, \cdots, 18\}$ and UG= $\{5, 6, \cdots, 10\}$. ACME has much higher TG, NG, ZG ratios than separate estimation. Both NG and ZG ratios increase as the sample size increases. ACME also has higher UG ratio, whose oracle target is zero. However, the ratio drastically drops to 0.005 from 0.2217 as the sample size is increased to $n = 500$ from $n = 100$. PCQ shows successful performance for underlying grouped variables (TG, NG, ZG), while it groups the variables which are not truly overlapped (UG).

2.5 Baseball Data Analysis

We analyze the major league baseball (MLB) players' annual salary dataset. We are interested in the salary determinants of low-paid, median-paid, and highly-paid players respectively. We obtain ACME for three quantile regression models to the quantiles, 0.25, 0.5, 0.75. The baseball dataset was obtained from *http://lib.stat.cmu.edu*. The dataset consists of the records and information on 263 North American MLB players in 1986 season and their salary in 1987 season. This dataset was previously studied

by He, Ng, and Portnoy (1998) and Li, Liu, and Zhu (2007). They assumed that the salary is a function of only the number of home runs in the previous year (HR) and the number of years in MLB (YEARS).

In addition to HR, YEARS, we consider all covariates such as their performance in the previous years and their league, division, and position information. The response is the annual salary on opening day in 1987 in thousands of dollars. The first seven predictors are as follows: the number of hits (HIT), the number of runs (RUN), the number of runs batted in (RBI), the number of walks (WALK), the number of put outs (PUTOUT), the number of assists (ASSIST), and the number of errors (ERROR). We employ seven dummy variables for league & division and position information: National East (NE), National West (NW), American East (AE), Infielder (IN), Outfielder (OUT), Catcher (CC), and Designated Hitter (DH). We treat American West (AW) and Utility Players (UP) as the base groups of the league & division and the position respectively. Note that we dropped the players' number of batting in 1986 (BAT) and performance records in their career. The BAT is highly correlated with the other variables such as HIT, HR, RUN, RBI, and WALK. Especially, the correlation between the BAT and the HIT is 0.9640. Most of the correlations between the performance records during their career are almost 0.9, which indicates severe collinearity.

Our goal is to determine important covariates on the first, second, and third quantiles of the players' salaries. We use a CQR loss function for the analysis with the quantile vector, $\tau = (0.25, 0.5, 0.75)$. Each quantile corresponds to the low-paid, median-paid and highly-paid players. We perform separate quantile regression estimation methods, PCQ, and ACME. The separate regression methods include ordinary, adaptive Lasso and one-step SCAD penalized quantile regression estimation. We use the 5-fold cross-validation for the tuning parameter selection.

ACME provides interpretable results by grouping the similar effects across the different quantiles. In Table 2.4, ACME selects HIT, YEARS, PUTOUT, league & division and positions across the three quantiles. The second quantile regression model is partially overlapped with the third quantile regression for the three covariates: HIT, YEARS, and PUTOUT. In other words, they are seen to have the same strength of impact on the median-paid and highly-paid baseball players' salary. Note that their effects are weaker in the low-paid players' salaries. It is interesting that HR is found to be significant only for the highly-paid players. The other coefficients such as RUN and RBI shrink to zero across all quantiles. Both WALK and ASSIST are non-zero in the preliminary estimator for the third quantile, but they shrink to zero in the ACME procedure.

The players' position is shown to be another important factor on the annual salary. Across all quantiles, the outfielders (OUT) are seen as the most-paid position. The catchers' (CC) and the infielders' (IN) salaries are the second and third highest, and the designated hitters (DH) and the utility players (UP) have the second-lowest and lowest salaries. Similar to the position, we can analyze the league & division factor on the players' salaries. Table 2.4 also reports the standard errors of the ordinary coefficients and their significance. They are obtained from the Markov chain marginal bootstrap (MCMB) with 500 repetitions (He and Hu 2002, Kocherginsky, He, and Mu 2005). ACME selects all variables known to be significant by MCMB under the significance level of 0.1.

Table 2.5 shows the test errors for all estimation procedures from 10 repetitions. In each iteration, randomly selected 28 data points are assigned as a test set and the remaining 235 data points are assigned as a training set. ACME is shown to have the best performance across all quantiles. It outperforms the ordinary quantile regression models to all quantiles. Compared with the other estimators, ACME has

	Ordinary (SE)	Sig.	AdLasso	SCAD	PCQ	ACME
(Intercept)	-245.5120 (73.4387)		3.5418	-219.1371	-515.5512	-222.0246
HIT	0.7907 (1.7183)		0	1.2864	2.9716	1.2815
HR	-5.3061 (4.9069)		0	0	2.0697	0
RUN	1.8274 (2.7044)		1.2953	0	0	0
RBI	2.4403 (2.6514)		0.2118	0	0	0
WALK	0.7804 (1.5287)		0	0	2.4083	0
YEARS	30.2551 (4.3717)	(**)	25.0385	31.0540	34.7556	31.2286
PUTOUT	-0.0890 (0.0978)		0	0.0015	0.1878	0.0118
ASSIST	-0.1639 (0.2459)		0	0	-0.0423	0
ERROR	-4.0178 (4.5298)		0	0	-5.4835	0
NE	-0.3179 (50.4264)		0	0	119.9565	0
NW	14.4817 (46.9023)		0	24.2768	49.2665	19.6613
AE	45.4914 (48.4438)		0	38.5924	94.8061	40.6199
IN	158.2192 (70.4252)	(**)	0	131.8874	146.3136	130.0462
OUT	103.3899 (71.0636)		0	163.0241	104.6079	160.9292
CC	192.0264 (75.5660)	(**)	0	144.9067	180.0394	147.7828
DH	-79.9613 (122.5664)		0	-10.3131	-37.9423	-11.7313
(Intercept)	-433.8376 (70.6211)		-377.3501	-350.5207	-389.9337	-345.8087
HIT	4.0231 (1.5517)	(**)	2.9242	2.9508	2.9716	2.9707
HR	6.6351 (6.2462)		2.5825	0	2.0697	0
RUN	-1.8305 (2.7047)		0	0	0	0
RBI	-1.4046 (2.5405)		0	0	0	0
WALK	2.0973 (1.3878)		1.7366	0	2.4083	0
YEARS	40.8095 (4.6872)	(**)	38.4487	42.1105	34.7556	42.5428
PUTOUT	0.2477 (0.1416)	(*)	0.2641	0.3109	0.1878	0.2662
ASSIST	-0.2267 (0.2770)		-0.0258	0	-0.0423	0
ERROR	-1.8804 (4.0841)		-0.5691	0	-5.4835	0
NE	108.5532 (52.4478)	(**)	93.8747	128.5615	119.9565	130.8570
NW	12.7587 (47.3871)		0	29.9324	49.2665	32.3740
AE	40.8497 (45.3921)		23.4914	81.1657	94.8061	73.9340
IN	190.6089 (78.3024)	(**)	89.7357	54.3756	146.3136	66.6862
OUT	136.6861 (62.7354)	(**)	95.4291	104.7711	104.6079	103.4506
CC	145.0478 (81.5529)	(*)	103.9739	80.8829	180.0394	90.0636
DH	-1.8963 (133.4392)		0	0	-37.9423	0
(Intercept)	-391.8350 (81.0963)		-361.7759	-399.4956	-245.9810	-374.7126
HIT	4.8975 (2.1460)	(**)	4.1554	3.4490	2.9716	2.9707
HR	13.3862 (7.9316)	(*)	12.4493	9.6505	2.0697	13.0354
RUN	-2.4222 (3.7428)		-1.4637	0	0	0
RBI	-1.9237 (3.7097)		-1.6779	0	0	0
WALK	3.2575 (1.9991)		3.5655	1.9914	2.4083	0
YEARS	39.3092 (6.4817)	(**)	41.4364	40.8961	34.7556	42.5428
PUTOUT	0.2982 (0.1529)	(*)	0.3053	0.2727	0.1878	0.2662
ASSIST	-0.6020 (0.3831)		-0.5430	-0.3295	-0.0423	0
ERROR	-1.7205 (6.3196)		-0.4648	0	-5.4835	0
NE	172.1072 (61.4199)	(**)	151.9045	156.2564	119.9565	183.7244
NW	46.0431 (60.8648)		33.0716	54.2641	49.2665	66.9276
AE	112.6242 (70.0346)		95.4571	101.6325	94.8061	82.9392
IN	224.1558 (100.9911)	(**)	164.4403	137.2256	146.3136	120.8592
OUT	62.4650 (87.4832)		42.4966	86.4714	104.6079	149.6180
CC	49.1510 (106.0594)		17.4022	63.3216	180.0394	91.7776
DH	-129.9760 (216.2998)		-174.4692	-69.4182	-37.9423	7.7997

Note. (**) indicates significant level 0.05 and (*) indicates significant level 0.1.

Table 2.4: Regression Coefficients of Baseball Dataset

better performance in two of the three quantiles. For example, ACME has smaller errors than SCAD in the second and third quantiles. Note that the performance of PCQ is substantially biased in the first quantile. Because PCQ assumes complete overlapping models, the first quantile regression modeling is dragged upward to the other two quantiles.

	Ordinary	AdLasso	SCAD	PCQ	ACME
Q1	75.9326	75.9482	72.7219	81.9500	74.2914
Q2	106.4342	105.3914	106.0978	103.6183	105.4999
Q3	92.7157	92.3668	93.7098	93.7860	91.9224

Table 2.5: Test Errors of Baseball Data for Three Quantiles

2.6 Discussion

In this chapter, we have proposed adaptive composite estimation for partially overlapping models. We have first introduced the notion of partially overlapping regression models on a given dataset. The overlapping structure is the same effect of a covariate on the response across multiple models. Partially overlapping models have at least one overlapping structure. We have also considered the sparse structure of the regression parameters for all models. ACME achieves both goals with a doubly penalized composite loss function. Its regular penalty function encourages the sparse structure recovery and the other penalty function induces the overlapping structure recovery. The arguments of the second penalty function are all pairwise differences of the coefficients for each covariate across the models. We have showed its selection and overlapping consistency under the proper choice of the tuning parameters. We have also established the asymptotic normality of non-redundant ACME, given the true sparse and overlapping structure. In the numerical studies, ACME have outperformed the separate penalized M-estimation and the composite M-estimation under the complete overlapping structure assumption.

2.7 Proofs

2.7.1 Proof of Lemma 2.1

From *A1*, the minimizer of the composite risk function, $\boldsymbol{\beta}^0$ is bounded and unique. The composite risk function is finite for each $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T \in \mathbb{R}^{K\cdot(p+1)}$ since it is a weighted linear combination of the finite separate risk functions from *A2*. The composite loss function, $L(\boldsymbol{z}, (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T))$, is also differentiable with respect to $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ at $(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T$ for $\mathbb{P}_{\boldsymbol{z}}$-almost every $\boldsymbol{z}$ with derivative

$$\begin{aligned}
&\nabla_{(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T} L(\boldsymbol{z}, (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)) \\
=&(w_1 \nabla_{(\alpha_1, \boldsymbol{\beta}_1)} L_1(y, \alpha_1 + \boldsymbol{x}^T\boldsymbol{\beta}_1)^T, \ \cdots, w_K \nabla_{(\alpha_K, \boldsymbol{\beta}_K)} L_K(y, \alpha_K + \boldsymbol{x}^T\boldsymbol{\beta}_K)^T)^T.
\end{aligned}$$

The variance of the score function at the true parameters is

$$\begin{aligned}
J(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) &\equiv \mathbb{E}[\nabla_{(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T} L(\boldsymbol{z}, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})) \cdot \nabla_{(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T} L(\boldsymbol{z}, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}))^T] \\
&= \mathbb{E}[w_k \nabla_{(\alpha_k, \boldsymbol{\beta}_k^T)^T} L_k(y, \alpha_k + \boldsymbol{x}^T\boldsymbol{\beta}_k^0) \cdot w_l \nabla_{(\alpha_l, \boldsymbol{\beta}_l^T)^T} L_l(y, \alpha_l + \boldsymbol{x}^T\boldsymbol{\beta}_l^0)^T]_{k,l=1}^K.
\end{aligned}$$

Note that the $J(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})$ is a $K(p+1) \times K(p+1)$ block matrix with K^2 blocks of $(p+1) \times (p+1)$ submatrices, denoted as $[J_{kl}(\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0)]_{k,l=1}^K$. All the on-diagonal block matrices are finite since $J_{kk}(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) = w_k^2 J_k(\alpha_k^0, \boldsymbol{\beta}_k^0) < \infty$ from *A3* a). The finiteness of the off-diagonal blocks is elementwise shown by Cauchy-Schwarz inequality.

The gradient vector and the Hessian matrix of the composite risk function are as follows:

$$\begin{aligned}
\nabla_{(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T} R(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T) &= (w_1 \nabla_{(\alpha_1, \boldsymbol{\beta}_1^T)^T} R_1(\alpha_1, \boldsymbol{\beta}_1)^T, \ \cdots, \ w_K \nabla_{(\alpha_K, \boldsymbol{\beta}_K^T)^T} R_K(\alpha_K, \boldsymbol{\beta}_K)^T), \\
H(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T) &= diag(w_1 H_1(\alpha_1, \boldsymbol{\beta}_1), \cdots, w_K H_K(\alpha_K, \boldsymbol{\beta}_K)).
\end{aligned}$$

The Hessian matrix at the true parameters, $H(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})$, is also positive definite from *A3* b). The composite risk function also has the same assumption on its twice differentiability and the positive definiteness of Hessian matrix. Lastly, the composite loss function is a linear combination of the convex functions with respect to $(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$. Hence, the composite loss function achieves the assumption, *A4*.

2.7.2 Proof of Lemma 2.3

By definition, both $\hat{\boldsymbol{\theta}}^o$ and $\boldsymbol{\theta}^0$ are the unique minimizers of the empirical distinct loss function and the distinct risk function respectively. We obtain the pointwise convergence of the empirical distinct loss function to the distinct risk function by the weak law of large numbers for any $\boldsymbol{\theta}$. The uniform convergence of the empirical distinct loss function to the distinct risk function can be verified by Convexity Lemma from Pollard (1991). The conditions on Theorem 5.7 of Van der Vaart (2000) are satisfied, thus this completes the proof.

2.7.3 Proof of Theorem 2.1

The distinct loss function and risk function satisfy the conditions for the asymptotic normality of an M-estimator. See Theorem 5.23 of Van der Vaart (2000) for further details. The distinct loss function, $\mathcal{L}(\boldsymbol{z}, \boldsymbol{\theta})$, is differentiable with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^0$ for $\mathbb{P}_{\boldsymbol{z}}$-almost every $\boldsymbol{z}$ with derivative $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{z}, \boldsymbol{\theta}^0)$ and $\mathbb{E}[\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{z}, \boldsymbol{\theta}^0) \cdot \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{z}, \boldsymbol{\theta}^0)^T] < \infty$. The distinct risk function is twice differentiable with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^0$ with the positive definite Hessian matrix $\mathcal{H}(\boldsymbol{\theta}^0)$.

2.7.4 Proof of Corollary 2.1

Note that the $\sqrt{n}$-consistency of distinct oracle estimator is equivalent to the $\sqrt{n}$-consistency of separate oracle estimator:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}^o - \boldsymbol{\theta}^0) = O_p(1) \Leftrightarrow \sqrt{n}(\hat{\boldsymbol{\beta}}^o_{A^0} - \boldsymbol{\beta}^0_{A^0}) = O_p(1) \Leftrightarrow \sqrt{n}(\hat{\boldsymbol{\beta}}^o_{A^0_k} - \boldsymbol{\beta}^0_{A^0_k}) = O_p(1)\ \forall k = 1, \cdots, K.$$

The "If" part of the first equivalence is obtained from $\sqrt{n}|\hat{\boldsymbol{\theta}}^o - \boldsymbol{\theta}^0| \leq \sqrt{n}|\hat{\boldsymbol{\beta}}^o_{A^0} - \boldsymbol{\beta}^0_{A^0}|$. The "Only if" part is from $\sqrt{n}|\hat{\boldsymbol{\beta}}^o_{A^0} - \boldsymbol{\beta}^0_{A^0}| = \sqrt{n}\sum_{k=1}^{K}|\hat{\boldsymbol{\beta}}^o_{A^0_k} - \boldsymbol{\beta}^0_{A^0_k}| \leq \sqrt{n}K|\hat{\boldsymbol{\theta}}^o - \boldsymbol{\theta}^0|$. The second equivalence is straightforward as $\sqrt{n}|\hat{\boldsymbol{\beta}}^o_{A^0} - \boldsymbol{\beta}^0_{A^0}| = (\sum_{k=1}^{K}\sqrt{n}|\hat{\boldsymbol{\beta}}^o_{A^0_k} - \boldsymbol{\beta}^0_{A^0_k}|^2)^{\frac{1}{2}}$.

2.7.5 Proof of Lemma 2.4

Our aim is to show that, for a sufficiently large constant C,

$$P\{\inf_{|\boldsymbol{u}|=C,\ \forall k} Q_n((\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) + n^{-\frac{1}{2}}\boldsymbol{u}^T) > Q(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})\} \to 1,$$

where $\boldsymbol{u} = (\boldsymbol{u}_0^T, \boldsymbol{u}_1^T, \cdots, \boldsymbol{u}_K^T)^T \in \mathbb{R}^{K(p+1)}$, $\boldsymbol{u}_0 \in \mathbb{R}^K$ and $\boldsymbol{u}_k \in \mathbb{R}^p$. That is, there is a minimizer inside the ball $|(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T - (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T| < n^{-\frac{1}{2}}C$, with probability tending to 1. It is the same argument as in the proof of Theorem 1 in Fan and Li (2001). Our objective function is (2.4). Let us define

$$\begin{aligned}
&D_n(\boldsymbol{u}) \equiv Q_n((\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) + n^{-\frac{1}{2}}\boldsymbol{u}^T) - Q_n(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) \\
=&\sum_{i=1}^{n}[L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) + \frac{\boldsymbol{u}^T}{\sqrt{n}}) - L(\boldsymbol{z}_i, (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})] \\
+&n\sum_{k=1}^{K}\sum_{j=1}^{p}(p_{\lambda_{1n}}(|\beta^0_{kj} + n^{-\frac{1}{2}}u_{kj}|) - p_{\lambda_{1n}}(|\beta^0_{kj}|)
\end{aligned}$$

$$
\begin{aligned}
&+n\sum_{k<k'}\sum_{j=1}^{p}(p_{\lambda_{2n}}(|\beta^0_{k'j}+n^{-\frac{1}{2}}u_{k'j}-\beta^0_{kj}-n^{-\frac{1}{2}}u_{kj}|)-p_{\lambda_{2n}}(|\beta^0_{k'j}-\beta^0_{kj}|))\\
&\geq\sum_{i=1}^{n}[L(\boldsymbol{z}_i,\boldsymbol{\beta}^0+\frac{\boldsymbol{u}}{\sqrt{n}})-L(\boldsymbol{z}_i,\boldsymbol{\beta}^0)]+n\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}(p_{\lambda_{1n}}(|\beta^0_{kj}+n^{-\frac{1}{2}}u_{kj}|)-p_{\lambda_{1n}}(|\beta^0_{kj}|))\\
&+n\sum_{k<k'}\sum_{j\in\mathcal{O}^c_{kk'}}(p_{\lambda_{2n}}(|\beta^0_{k'j}+n^{-\frac{1}{2}}u_{k'j}-\beta^0_{kj}-n^{-\frac{1}{2}}u_{kj}|)-p_{\lambda_{2n}}(|\beta^0_{k'j}-\beta^0_{kj}|))\\
&\equiv T_1+T_2+T_3
\end{aligned}
$$

The inequality holds because $\beta^0_{kj}=0$ if $j\in\mathcal{A}^c_k$ and $\beta^0_{k'j}=\beta^0_{kj}$ if $j\in\mathcal{O}_{kk'}$. By Lemma 2.2, the T_1 converges to $\frac{1}{2}\boldsymbol{u}^TH(\boldsymbol{\alpha}^{0T},\boldsymbol{\beta}^{0T})\boldsymbol{u}+\boldsymbol{W}^T\boldsymbol{u}$ in probability and further uniformly converges on any compact subset of $\mathbb{R}^d$. We consider the T_2 and T_3 parts with three types of penalty functions: folded concave, one-step folded concave and weighted L_1 penalty functions. We first examine the folded concave penalty functions. For a large n, if $|t|>a\lambda_{1n}$ and $\lambda_{1n}\to 0$,

$$
T_2=n\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}(p_{\lambda_{1n}}(\beta^0_{kj}+n^{-\frac{1}{2}}u_{kj})-p_{\lambda_{1n}}(\beta^0_{kj}))=0 \tag{2.6}
$$

since $p'_{\lambda_{1n}}(t)=0$. The same argument is applied to the T_3 for a large n:

$$
T_3=n\sum_{k<k'}\sum_{j\in\mathcal{O}^c_{kk'}}(p_{\lambda_{2n}}(\beta^0_{k'j}-\beta^0_{kj}+n^{-\frac{1}{2}}(u_{k'j}-u_{kj}))-p_{\lambda_{2n}}(\beta^0_{k'j}-\beta^0_{kj}))=0. \tag{2.7}
$$

For the weighted L_1 penalty, the terms T_2 and T_3 go to zero in probability. We now consider one-step folded concave penalty functions under the assumption of $\lambda_{1n}\to 0$.

$$
T_2=\sqrt{n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}p'_{\lambda_{1n}}(|\beta^{(0)}_{kj}|)\frac{|\beta^0_{kj}+n^{-\frac{1}{2}}u_{kj}|-|\beta^0_{kj}|}{1/\sqrt{n}}=o_p(1) \tag{2.8}
$$

Note that $\dfrac{|\beta^0_{kj}+n^{-\frac{1}{2}}u_{kj}|-|\beta^0_{kj}|}{1/\sqrt{n}}\to sgn(\beta^0_{kj})u_{kj}$ and $\sqrt{n}p'_{\lambda_{1n}}(|\beta^{(0)}_{kj}|)\xrightarrow{p}0$ as $|\beta^{(0)}_{kj}|\xrightarrow{p}$

$|\beta_{kj}^0| \neq 0$ and $p'_{\lambda_{1n}}(t) = 0$ for $t > a\lambda_{1n}$. For T_3,

$$T_3 = \sqrt{n} \sum_{k<k'} \sum_{j \in \mathcal{O}_{kk'}^c} p'_{\lambda_{2n}}(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|) \frac{|\beta_{k'j}^0 - \beta_{kj}^0 + n^{-\frac{1}{2}}(u_{k'j} - u_{kj})| - |\beta_{k'j}^0 - \beta_{kj}^0|}{1/\sqrt{n}} \quad (2.9)$$

Similar to T_2, we obtain $\frac{|\beta_{k'j}^0 - \beta_{kj}^0 + n^{-\frac{1}{2}}(u_{k'j} - u_{kj})| - |\beta_{k'j}^0 - \beta_{kj}^0|}{1/\sqrt{n}} \rightarrow sgn(\beta_{k'j}^0 - \beta_{kj}^0)(u_{k'j} - u_{kj})$ and $\sqrt{n}p'_{\lambda_{2n}}(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|) \xrightarrow{p} 0$. Thus, T_3 is also $o_p(1)$. For the other weighted L_1 penalty functions, we obtain

$$T_2 = \sqrt{n}\lambda_{1n} \sum_{k=1}^{K} \sum_{j \in \mathcal{A}_k} p'(|\beta_{kj}^{(0)}|) \frac{|\beta_{kj}^0 + n^{-\frac{1}{2}} u_{kj}| - |\beta_{kj}^0|}{1/\sqrt{n}}, \quad (2.10)$$

under the assumption that $\sqrt{n}\lambda_{1n} \rightarrow 0$.

Each term converges to a certain value in a probabilistic sense. $p'(|\beta_{kj}^{(0)}|) \xrightarrow{p} p'(|\beta_{kj}^0|)$ by the continuity of the derivative of the penalty function and the last term goes to $sgn(\beta_{kj}^0)u_{kj}$. As $\sqrt{n}\lambda_{1n} \rightarrow 0$, we have $T_2 = o_p(1)$. In a similar way, we can write T_3 as

$$T_3 = \sqrt{n}\lambda_{2n} \sum_{k<k'} \sum_{j \in \mathcal{O}_{kk'}^c} p'(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|) \frac{|\beta_{k'j}^0 - \beta_{kj}^0 + n^{-\frac{1}{2}}(u_{k'j} - u_{kj})| - |\beta_{k'j}^0 - \beta_{kj}^0|}{1/\sqrt{n}} \quad (2.11)$$

$p'(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|) \xrightarrow{p} p'(|\beta_{k'j}^0 - \beta_{kj}^0|)$ and the next term goes to $sgn(\beta_{k'j}^0 - \beta_{kj}^0)(u_{kj} - u_{kj})$. We have $T_3 = o_p(1)$ as $\sqrt{n}\lambda_{2n} \rightarrow 0$. The terms T_2 and T_3 converge to zero in probability under every penalty function. For the $|\boldsymbol{u}|$ equal to a sufficiently large C, $Q_n((\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T}) + n^{-\frac{1}{2}}\boldsymbol{u}^T) - Q_n(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})$ is dominated by the quadratic term, $\frac{1}{2}\boldsymbol{u}^T H(\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})\boldsymbol{u}$. Thus, the $\sqrt{n}$ consistency is achieved.

2.7.6 Lemma 2.5 and Theorem 2.2

Lemma 2.5. *Suppose that* $\lambda_{1n} \rightarrow 0$, $\lambda_{2n} \rightarrow 0$, $\sqrt{n}\lambda_{1n} \rightarrow \infty$, $\sqrt{n}\lambda_{2n} \rightarrow \infty$ *for folded concave, one-step folded concave penalty functions. For weighted* L_1 *penalty*

functions, suppose $\sqrt{n}\lambda_{1n} \to 0$, $\sqrt{n}\lambda_{2n} \to 0$, $n^{\frac{s+1}{2}}\lambda_{1n} \to \infty$, $n^{\frac{s+1}{2}}\lambda_{2n} \to \infty$. *Assume that there exists at least one* $j \in \mathcal{O}_{kk'}$ *for some* $k < k'$. *Consider a given random vector* $(\boldsymbol{\alpha}^{DT}, \boldsymbol{\beta}^{DT})^T$ *and* $\boldsymbol{c}$, *whose lengths are* $K \cdot (p+1)$. *Denote* $\boldsymbol{\beta}^{DT} = (\boldsymbol{\beta}^{D_1 T}, \cdots, \boldsymbol{\beta}^{D_K T})$, *where* $\boldsymbol{\beta}^{D_k} = [\beta_j^{D_k}]_{j=1}^p$. *Suppose that* $\beta_{kj}^D = 0 \ \forall j \in \mathcal{A}_k^c$ *for every* k *and* $\beta_{kj}^D = \beta_{k'j}^D \ \forall j \in \mathcal{O}_{kk'}$ *for all* $k < k'$. *Denote* $\boldsymbol{c}^T = (\boldsymbol{c}_0^T, \boldsymbol{c}_1^T, \cdots, \boldsymbol{c}_K^T)$, *where* $\boldsymbol{c}_0 = [c_{0k}]_{k=1}^K$, $\boldsymbol{c}_k = [c_{kj}]_{j=1}^p$ *and* $c_{kj} = 0$ *for* $j \in \mathcal{A}_k$ *and* $j \notin \mathcal{O}_{kk'} \ \forall k' \neq k$. *Define* $(\boldsymbol{\alpha}^{D'T}, \boldsymbol{\beta}^{D'T}) = (\boldsymbol{\alpha}^{DT}, \boldsymbol{\beta}^{DT}) + \boldsymbol{c}^T$ *and denote* $\boldsymbol{\beta}^{D'T} = (\boldsymbol{\beta}^{D_1' T}, \cdots, \boldsymbol{\beta}^{D_K' T})$, *where* $\boldsymbol{\beta}^{D_k'} = [\beta_j^{D_k'}]_{j=1}^p$. *Assume that* $|(\boldsymbol{\alpha}^{DT}, \boldsymbol{\beta}^{DT})^T - (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T| = O_p(n^{-1/2})$. *With probability tending to one, for any constant* C_1,

$$Q_n(\boldsymbol{\alpha}^{DT}, \boldsymbol{\beta}^{DT}) = \min_{|\boldsymbol{c}| \leq n^{-1/2} C_1} Q_n(\boldsymbol{\alpha}^{D'T}, \boldsymbol{\beta}^{D'T}).$$

Note that given a constant C_1, $\sum_{k<k'} \sum_{j \in \mathcal{O}_{kk'}} |c_{k'j} - c_{kj}| \leq n^{-1/2} C_2$, *where the constant,* C_2, *depends on all* $\mathcal{O}_{kk'}$*s,* $\mathcal{A}_k$*s,* K, *and* p.

Proof. It follows the same line as the proof of Lemma 1 of Wu and Liu (2009). We let $\boldsymbol{\gamma}^0 = (\boldsymbol{\alpha}^{0T}, \boldsymbol{\beta}^{0T})^T$, $\boldsymbol{\gamma}^D = (\boldsymbol{\alpha}^{DT}, \boldsymbol{\beta}^{DT})^T$ and $\boldsymbol{\gamma}^{D'} = (\boldsymbol{\alpha}^{D'T}, \boldsymbol{\beta}^{D'T})^T$.

$$\begin{aligned}
&Q_n(\boldsymbol{\gamma}^{DT}) - Q_n(\boldsymbol{\gamma}^{D'T}) = [Q_n(\boldsymbol{\gamma}^D) - Q_n(\boldsymbol{\gamma}^0)] - [Q_n(\boldsymbol{\gamma}^{D'}) - Q_n(\boldsymbol{\gamma}^0)] \\
=&\sum_{i=1}^n [L(\boldsymbol{z}_i, \boldsymbol{\gamma}^{DT}) - L(\boldsymbol{z}_i, \boldsymbol{\gamma}^{0T})] - \sum_{i=1}^n [L(\boldsymbol{z}_i, \boldsymbol{\gamma}^{D'T}) - L(\boldsymbol{z}_i, \boldsymbol{\gamma}^{0T})] \\
+&n \sum_{k=1}^K \sum_{j \in \mathcal{A}_k} (p_{\lambda_{n1}}(|\beta_{kj}^D|) - p_{\lambda_{n1}}(|\beta_{kj}^{D'}|)) - n \sum_{k=1}^K \sum_{j \in \mathcal{A}_k^c} p_{\lambda_{n1}}(|\beta_{kj}^{D'}|) \\
+&n \sum_{k<k'} \sum_{j \in \mathcal{O}_{kk'}^c} (p_{\lambda_{n2}}(|\beta_{k'j}^D - \beta_{kj}^D|) - p_{\lambda_{n2}}(|\beta_{k'j}^{D'} - \beta_{kj}^{D'}|)) - n \sum_{k<k'} \sum_{j \in \mathcal{O}_{kk'}} p_{\lambda_{n2}}(|\beta_{k'j}^{D'} - \beta_{kj}^{D'}|) \\
\equiv& U_1 + U_2 + U_3 + U_4 + U_5 + U_6,
\end{aligned}$$

where $\mathcal{O}_{kk'}^c = \{1, 2, \cdots, p\} \backslash \mathcal{O}_{kk'}$. Note that $|\boldsymbol{\beta}^D - \boldsymbol{\beta}^0| = O_p(n^{-1/2})$ and $|\boldsymbol{\beta}^{D'} - \boldsymbol{\beta}^0| =$

$O_p(n^{-1/2})$. It implies that $\boldsymbol{\beta}^D \xrightarrow{p} \boldsymbol{\beta}^0$ and $\boldsymbol{\beta}^{D'} \xrightarrow{p} \boldsymbol{\beta}^0$. First, from Lemma 2.2, U_1 and U_2 are bounded in probability.

$$
\begin{aligned}
U_1 + U_2 =& \sum_{i=1}^{n}[L(\boldsymbol{z}_i, \boldsymbol{\gamma}^{DT}) - L(\boldsymbol{z}_i, \boldsymbol{\gamma}^{0T})] - \sum_{i=1}^{n}[L(\boldsymbol{z}_i, \boldsymbol{\gamma}^{D'T}) - L(\boldsymbol{z}_i, \boldsymbol{\gamma}^{0T})] \\
=& \sqrt{n}(\boldsymbol{\gamma}^D - \boldsymbol{\gamma}^0)^T H(\boldsymbol{\gamma}^{0T})\sqrt{n}(\boldsymbol{\gamma}^D - \boldsymbol{\gamma}^0) + \boldsymbol{W}^T\sqrt{n}(\boldsymbol{\gamma}^D - \boldsymbol{\gamma}^0) + o_p(1) \\
& -\sqrt{n}(\boldsymbol{\gamma}^{D'} - \boldsymbol{\gamma}^0)^T H(\boldsymbol{\gamma}^{0T})\sqrt{n}(\boldsymbol{\gamma}^{D'} - \boldsymbol{\gamma}^0) - \boldsymbol{W}^T\sqrt{n}(\boldsymbol{\gamma}^{D'} - \boldsymbol{\gamma}^0) + o_p(1) \\
=& O_p(1) + \boldsymbol{W}^T\sqrt{n}\boldsymbol{c} + o_p(1) = O_p(1)
\end{aligned}
$$

Next, $U_3,\ U_4,\ U_5,\ U_6$ are considered with folded concave, one-step folded concave, and weighted L_1 penalty functions. We have the conditions such that $0 < \boldsymbol{c} \leq n^{-1/2}C_1$ and $0 < \sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}} |c_{k'j} - c_{kj}| \leq n^{-1/2}C_2$. For folded concave penalty functions, each term of U_3 is $o_p(1)$, thus $U_3 = o_p(1)$ by continuous mapping theorem and $c_{kj} \to 0$. The U_5 is also $o_p(1)$ from the same argument. We now show that bothU_4 and U_6 dominate in magnitude.

$$
\begin{aligned}
U_4 =& -n\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c} p_{\lambda_{1n}}(|c_{kj}|) = -np'_{\lambda_{1n}}(0+)\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c}|c_{kj}|(1+o(1)) \\
\leq& -a_1\sqrt{n}\lambda_{1n}\cdot\sqrt{n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c}|c_{kj}|(1+o(1))
\end{aligned}
$$

As $\sqrt{n}\lambda_{1n} \to \infty$ and $0 < \sqrt{n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c}|c_{kj}| \leq C_1$, we have $U_4 \xrightarrow{p} -\infty$. We obtain the same result for the U_6 as follows:

$$
\begin{aligned}
U_6 =& -n\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}} p_{\lambda_{2n}}(|c_{k'j} - c_{kj}|) = -np'_{\lambda_{2n}}(0+)(\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}}|c_{k'j} - c_{kj}|)(1+o(1)) \\
\leq& -a_1\sqrt{n}\lambda_{2n}\cdot\sqrt{n}\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}}|c_{k'j} - c_{kj}|(1+o(1)).
\end{aligned}
$$

With one-step folded concave and weighted L_1 penalty functions, the U_3, U_4, U_5 and U_6 are written as follows:

$$U_3 = n\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k} p'_{\lambda_{1n}}(|\beta_{kj}^{(0)}|)(|\beta_{kj}^{D}| - |\beta_{kj}^{D} + c_{kj}|) \tag{2.12}$$

$$U_4 = -n\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c} p'_{\lambda_{1n}}(|\beta_{kj}^{(0)}|)|c_{kj}| \tag{2.13}$$

$$U_5 = n\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}^c} p'_{\lambda_{2n}}(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)(|\beta_{k'j}^{D'} - \beta_{kj}^{D'}| - |\beta_{k'j}^{D'} - \beta_{kj}^{D'} + c_{k'j} - c_{kj}|) \tag{2.14}$$

$$U_6 = -n\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}} p'_{\lambda_{2n}}(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)\cdot|c_{k'j} - c_{kj}| \tag{2.15}$$

Both U_3 and U_5 converge to zero in probability in the same sense of (2.8) and (2.9). Both U_4 and U_6 are bounded by $-a_1\sqrt{n}\lambda_{1n}\sqrt{n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k}|c_{kj}|$ and $-a_1\sqrt{n}\lambda_{1n}\sqrt{n}\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}}|c_{k'j} - c_{kj}|$. Both go to the negative infinity in probability as $\sqrt{n}\lambda_{1n}\to\infty$. Now, we plug-in the weighted L_1 penalty function to (2.12)-(2.15).

$$U_3 = n\lambda_{1n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k} p'(|\beta_{kj}^{(0)}|)(|\beta_{kj}^{D}| - |\beta_{kj}^{D} + c_{kj}|)$$

$$U_4 = -n\lambda_{1n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c} p'(|\beta_{kj}^{(0)}|)|c_{kj}| = -n^{\frac{1+s}{2}}\lambda_{1n}\sum_{k=1}^{K}\sum_{j\in\mathcal{A}_k^c}(\sqrt{n}|\beta_{kj}^{(0)}|)^{-s}\frac{p'(|\beta_{kj}^{(0)}|)}{|\beta_{kj}^{(0)}|^{-s}}\sqrt{n}|c_{kj}|$$

$$U_5 = n\lambda_{2n}\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}^c} p'(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)(|\beta_{k'j}^{D'} - \beta_{kj}^{D'}| - |\beta_{k'j}^{D'} - \beta_{kj}^{D'} + c_{k'j} - c_{kj}|)$$

$$U_6 = -n\lambda_{2n}\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}} p'(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)\cdot|c_{k'j} - c_{kj}|$$

$$= -n^{\frac{1+s}{2}}\lambda_{2n}\sum_{k<k'}\sum_{j\in\mathcal{O}_{kk'}}(\sqrt{n}|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)^{-s}\frac{p'(|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|)}{|\beta_{k'j}^{(0)} - \beta_{kj}^{(0)}|^{-s}}\sqrt{n}|c_{k'j} - c_{kj}|$$

As $\sqrt{n}\lambda_{1n}\to\infty$ and $\sqrt{n}\lambda_{2n}\to\infty$, both U_3 and U_5 go to zero in probability as (2.10) and (2.11). As $n^{\frac{1+s}{2}}\lambda_{1n}\to\infty$ and $n^{\frac{1+s}{2}}\lambda_{2n}\to\infty$, both U_4 and U_6 go to the

negative infinity in probability. This term is higher order than any other terms, thus dominates the remaining terms. In other words, $Q_n(\boldsymbol{\gamma}^{DT}) - Q_n(\boldsymbol{\gamma}^{D'T}) < 0$ for a large n. Thus, the minimizer of $Q_n(\boldsymbol{\gamma}^{D'T})$ satisfies $\beta_{kj} = 0 \ \forall j \in \mathcal{A}_k^c$ for every k and $\beta_{k'j} = \beta_{kj}$ $\forall j \in \mathcal{O}_{kk'}$ for every $k < k'$ with probability tending to 1. Note that there exists at least one non-empty set of $\mathcal{O}_{kk'}$ for some $k < k'$. This extra condition is needed because the third term is zero without the condition. $\square$

From Lemma 2.5, the $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T)^T$ does not minimize the objective function, $Q_n(\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)$ if at least one of the true zero parameters is estimated as non-zero or at least one overlapping structure is estimated with different values with probability tending to one. Theorem 2.2 is the straightforward result from Lemma 2.5.

2.7.7 Proof of Theorem 2.3

Our proof follows the proof of the Theorem in Wang, Li, and Jiang (2007a). Denote $\hat{\boldsymbol{\theta}}_{\mathcal{A}^0}(\mathcal{G}_0)$ the minimizer of $Q'_n(\boldsymbol{\theta}) \equiv Q_n(\boldsymbol{\beta}_{\mathcal{A}^0}(\boldsymbol{\theta}))$, where $\boldsymbol{\beta}_{\mathcal{A}^0}(\boldsymbol{\theta})$ is written as $(\theta_{01}, \cdots, \theta_{0K}, \boldsymbol{\beta}^T_{\mathcal{A}_1}(\boldsymbol{\theta}), \cdots, \boldsymbol{\beta}^T_{\mathcal{A}_K}(\boldsymbol{\theta}),)^T$.

$$
\begin{aligned}
Q'_n(\boldsymbol{\theta}_{\mathcal{A}^0}) = & \sum_{k=1}^{K} \sum_{i=1}^{n} w_k L_k(y_i, \theta_{0k} + \boldsymbol{x}_i^{\mathcal{A}_k^{0}T} \boldsymbol{\beta}_{\mathcal{A}_k^0}(\boldsymbol{\theta})) + n \sum_{k=1}^{K} \sum_{j \in \mathcal{A}_k} p_{\lambda_{1n}}(\beta_{\mathcal{A}_k j}(\boldsymbol{\theta})) \\
& + n \sum_{k<k'} \sum_{j \in \mathcal{O}^c_{kk'}} p_{\lambda_{2n}}(\beta_{\mathcal{A}'_k j}(\boldsymbol{\theta}) - \beta_{\mathcal{A}_k j}(\boldsymbol{\theta}))
\end{aligned}
$$

Let $\Psi_n(\boldsymbol{u}) = Q'_n(\boldsymbol{\theta}^0 + \dfrac{\boldsymbol{u}}{\sqrt{n}})$, then $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}^0}(\mathcal{G}_0) - \boldsymbol{\theta}^0)$ is the minimizer of $\Psi_n(\boldsymbol{u}) - \Psi_n(0)$. For any $\boldsymbol{u} \in \mathbb{R}^{K + \sum_{q=1}^{Q} G_q}$, denote

$$
\begin{aligned}
& V_n(\boldsymbol{u}) \equiv \Psi_n(\boldsymbol{u}) - \Psi_n(0) \\
= & \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{z}_i, \boldsymbol{\theta}^0 + \frac{\boldsymbol{u}}{\sqrt{n}}) - \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{z}_i, \boldsymbol{\theta}^0) + n \sum_{k=1}^{K} \sum_{j \in \mathcal{A}_k} p_{\lambda_{1n}}(\beta^0_{\mathcal{A}_k j}(\boldsymbol{\theta}) + \frac{\tilde{u}_{kj}(\boldsymbol{u})}{\sqrt{n}}) - p_{\lambda_{1n}}(\beta^0_{\mathcal{A}_k j}(\boldsymbol{\theta}))
\end{aligned}
$$

$$+n\sum_{k<k'}\sum_{j\in\mathcal{O}^c_{kk'}} p_{\lambda_{2n}}(\beta^0_{\mathcal{A}_{k'}j}(\boldsymbol{\theta})-\beta^0_{\mathcal{A}_{kj}}(\boldsymbol{\theta})+\frac{\tilde{u}_{k'j}(\boldsymbol{u})-\tilde{u}_{kj}(\boldsymbol{u})}{\sqrt{n}})-p_{\lambda_{2n}}(\beta^0_{\mathcal{A}'_kj}(\boldsymbol{\theta})-\beta^0_{\mathcal{A}_kj}(\boldsymbol{\theta}))$$

$$\equiv V_{n1}(\boldsymbol{u})+V_{n2}(\boldsymbol{u})+V_{n3}(\boldsymbol{u}),$$

where $\tilde{\boldsymbol{u}}_k(\boldsymbol{u})=[\tilde{u}_{kj}]_{j\in\mathcal{A}_k}$ is the element of $\boldsymbol{u}$ corresponding to $\boldsymbol{\beta}^0_{\mathcal{A}_k}$. Similar to Lemma 2.2, we have

$$V_{n1}(\boldsymbol{u})\xrightarrow{d}\frac{1}{2}\boldsymbol{u}^T\mathcal{H}(\boldsymbol{\theta}^0)\boldsymbol{u}+\boldsymbol{W}^T_{\boldsymbol{\theta}}\boldsymbol{u},$$

where $\boldsymbol{W}_{\boldsymbol{\theta}}\sim N(0,\mathcal{J}(\boldsymbol{\theta}^0)$. Both $V_{n2}(\boldsymbol{u})$ and $V_{n3}(\boldsymbol{u})$ are $o_p(1)$ under any penalty function form as (2.6)-(2.11) in the proof of Lemma 2.4. Finally, we obtain

$$V_n(\boldsymbol{u})\xrightarrow{d}\frac{1}{2}\boldsymbol{u}^T\mathcal{H}(\boldsymbol{\theta}^0)\boldsymbol{u}+\boldsymbol{W}^T_{\boldsymbol{\theta}}\boldsymbol{u}.$$

Lemma 2.2 and Remark 1 of Davis, Knight, and Liu (1992) imply that if an objective function converges in distribution to a strictly convex function, its minimum converges in distribution to the unique minimum of the strictly convex function. Hence, we complete the proof.

CHAPTER3: ENSEMBLE VARIABLE SELECTION AND ESTIMATION

3.1 Introduction

Penalization is a widely used technique for simultaneous variable selection and parameter estimation. There are numerous sparse penalized variable selection techniques in the literature, including LASSO (Tibshirani 1996), SCAD (Fan and Li 2001), and adaptive LASSO (Zou 2006). In a regression setting, one theoretical goal of variable selection is oracle estimation (Fan and Li 2001), which requires both consistency in variable selection and asymptotic normality of the non-zero coefficient estimators with the same efficiency as the oracle estimator under the true model, where the non-zero coefficients are known a priori.

Certain procedures, such as SCAD and adaptive LASSO, are known to satisfy oracle properties (Fan and Li 2001, Zou 2006). However, other penalization methods may suffer deficiencies, in which selection consistency may be achieved without oracle estimation. LASSO was shown to be able to yield consistent variable selection if the underlying model satisfies some conditions, but the LASSO estimator does not have oracle efficiency (Meinshausen and Bühlmann 2006, Zhao and Yu 2007, Zou 2006, Yuan and Lin 2007b). Model selection criteria like Bayesian Information Criterion (BIC) can produce consistent model selection but suboptimal estimation (Yang 2005). A simple and general approach to variable selection was suggested in Wang and Leng (2007), which presents a unified theoretical framework for the regression setting. They proved that their least squares approximation (LSA) penalization method yields sparse and

consistent model selection, but the penalized estimators may not be oracle equivalent when the asymptotic covariance matrix of a preliminary estimator violates certain assumptions.

To address the above issue, we first propose a simple refit method based on an initial selection consistent estimator. The first step is to obtain a selection consistent estimator via a variable selection method, e.g., the LSA penalization method. In the second step, we only use the selected variables from the first step to refit the parameters using the corresponding unpenalized objective function. Regardless of whether the first step estimator satisfies the oracle property, the refit estimator has the oracle property, as long as the first step estimator is consistent in selection. For the LSA penalization method, the refit step gives an estimator having the oracle property, regardless of whether the covariance assumption holds.

We further suggest two novel methods based on the refit method: ensemble variable selection (EVS) and ensemble variable selection and estimation (EVE). Both methods perform simultaneous variable selection and estimation with penalization methods. EVS is applicable to a general regression setting, and EVE is useful for a likelihood-based model which satisfies the factorization assumption on the full likelihood function.

3.1.1 Ensemble Variable Selection (EVS)

One practical issue of a penalized method for variable selection and estimation is the choice of penalty functions from the numerous available penalty functions. The performance of each variable selection method is case-specific, that is, we cannot guarantee any universally preferable procedure. For each scenario, we may select the model chosen from the method with the smallest test error, but the error calculation via cross-validation is sometimes computationally expensive.

EVS combines the variable selection decisions from multiple candidate penalization

methods. We view each method as casting votes on important covariates and obtain nested candidate models according to the vote counts. EVS refits each candidate model without penalization, and selects the optimal model by selection criteria such as BIC and cross-validation (Schwarz 1978). We automatically avoid the worst performance and nearly perform the best in practice. Furthermore, it reduces the computational burden as the number of the candidate models is less than or equal to the number of the methods. We apply EVS to the South African Heart Disease (SAHD) dataset from Friedman et al. (2001) for risk factors analysis for myocardial infarction (MI).

3.1.2 Ensemble Variable Selection and Estimation (EVE)

Penalized method is a useful tool for variable selection in numerous likelihood-based models such as generalized linear models (Zou 2006, Fan and Li 2001) and Cox proportional hazards models (Zhang and Lu 2007). In the literature, the direct penalization techniques were shown to have the oracle properties, and their numerical algorithms were developed. However, for certain likelihood-based models, such direct penalization methods require model-specific theoretical work and may not be computationally feasible with existing software.

To tackle the problem, we propose an indirect penalization method, EVE for a factorizable likelihood-based model. In such model, the full likelihood is the multiplication of two likelihood factors. Its full estimator can be obtained by ensemble estimation, asymptotic efficient combination of the separate estimators from the likelihood factors via generalized least squares (GLS) (Cox 2001). By exploiting the ensemble estimation and the refit LSA method, EVE selects variables and estimates parameters without asymptotic efficiency loss. We analyze the Multicenter AIDS Cohort Study (MACS) dataset described in Kaslow et al. (1987) with EVE to find out risk factors strongly associated with HIV infection.

3.1.3 Outline

Section 3.2 presents the refit method based on a preliminary selection consistent estimator. We show its theoretical properties in Section 3.2.1 and illustrate the refit estimation based on the LSA penalization technique as an example in Section 3.2.2. We present the results of simulated data from linear regression and median regression with heteroscedasticity in Section 3.2.3.

Section 3.3 studies the EVS method based on the multiple penalization methods. We describe the procedure in Section 3.3.1 and demonstrate the performance of EVS from numerical studies and the SAHD data analysis in Sections 3.3.2-3.3.3.

Section 3.4 proposes the EVE method under the assumption that the full likelihood is factorized into two likelihood factors. We examine the likelihood factorization and the ensemble estimation in Section 3.4.1. In particular, we consider the Cox model for prospective doubly censored data in Section 3.4.2. We present numerical results from simulation studies and real data analysis in Sections 3.4.3-3.4.4.

3.2 Refitting for Variable Selection

Penalization techniques may suffer from the potential bias in the non-zero coefficients in finite sample studies. The two-step refit procedure eliminates the shrinkage effect of the non-zero coefficients to zero, maintaining the important variables from the penalization. We use the penalization only for variable selection, and then estimate the coefficients with the selected important variables. Classical inference is valid when the selected model includes all the important variables. The refit least squares approximation (LSA) estimator is introduced as an example of the refit method.

3.2.1 The Refit Method and Its Theoretical Properties

We consider a general regression model with independently and identically distributed random vectors, $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$, where $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})^T \in \mathbb{R}^p$ is a p-dimensional covariate and y_i is a 1-dimensional dependent variable. Let $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T \in \mathbb{R}^p$ be the parameter vector of interest and $\mathcal{L}_n(\boldsymbol{\beta})$ be the objective function without penalization such as least squares and negative log likelihood, which we want to minimize.

The true regression parameter is written as $\boldsymbol{\beta}^0$. Denote the index set of non-zero parameters as $\mathcal{A} = \{j : \beta_j^0 \neq 0\}$ and its complement as $\mathcal{A}^c = \{j : \beta_j^0 = 0\}$. Their cardinalities are written as $|\mathcal{A}|$ and $|\mathcal{A}^c|$ respectively. We let $\boldsymbol{\beta}^0_{\mathcal{A}} = [\beta_j^0]_{j\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ and $\boldsymbol{\beta}^0_{\mathcal{A}^c} = [\beta_j^0]_{j\in\mathcal{A}^c} \in \mathbb{R}^{|\mathcal{A}^c|}$. If the true underlying model is known in advance, we obtain the oracle estimator by

$$\hat{\boldsymbol{\beta}}^o = \underset{\{\boldsymbol{\beta}\in\mathbb{R}^p:\ \beta_j=0,\ \forall j\in\mathcal{A}^c\}}{\operatorname{argmin}} \mathcal{L}_n(\boldsymbol{\beta}). \tag{3.1}$$

Similar to $\boldsymbol{\beta}^0$, the oracle estimator can be decomposed into $\boldsymbol{\beta}^o_{\mathcal{A}} = [\beta_j^o]_{j\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ and $\boldsymbol{\beta}^o_{\mathcal{A}^c} = [\beta_j^o]_{j\in\mathcal{A}^c} \in \mathbb{R}^{|\mathcal{A}^c|}$.

Next, we describe the refit procedure based on a selection consistent estimator, denoted by $\hat{\boldsymbol{\beta}}$. We select important coefficients based on $\hat{\boldsymbol{\beta}}$, and then derive an unpenalized estimator for the coefficients corresponding to the selected variables. Let $\hat{\mathcal{A}} = \{i : \hat{\beta}_i \neq 0\}$ denote the set of important variables in $\hat{\boldsymbol{\beta}}$ and $\hat{\mathcal{A}}^c = \{i : \hat{\beta}_i = 0\}$ denote its complement. If $|\hat{\mathcal{A}}|$ is less than n, the refit estimate is

$$\hat{\boldsymbol{\beta}}^{r(\hat{\mathcal{A}})} = \underset{\{\boldsymbol{\beta}\in\mathbb{R}^p:\ \beta_j=0,\ \forall j\in\hat{\mathcal{A}}^c\}}{\operatorname{argmin}} \mathcal{L}_n(\boldsymbol{\beta}). \tag{3.2}$$

In the following, we simply denote the refit estimate as $\hat{\boldsymbol{\beta}}^r$ and decompose it to $\hat{\boldsymbol{\beta}}^r_{\mathcal{A}} = [\hat{\beta}_j^r]_{j\in\mathcal{A}}$ and $\hat{\boldsymbol{\beta}}^r_{\mathcal{A}^c} = [\hat{\beta}_j^r]_{j\in\mathcal{A}^c}$. The refit selected model is $\hat{\mathcal{A}}^r = \{j : \hat{\beta}_j^r \neq 0\}$ and its complement is $\hat{\mathcal{A}}^{rc} = \{j : \hat{\beta}_j^r = 0\}$.

We establish general theoretical properties of the refit estimator based on an arbitrary preliminary selection consistent model estimator: $\sqrt{n}$-consistency, selection consistency, and the oracle property.

Theorem 3.1. *$\sqrt{n}$-Consistency*

If $P(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ *and* $\sqrt{n}(\hat{\boldsymbol{\beta}}^o_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) = O_p(1)$, *then* $\sqrt{n}(\hat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}^0) = O_p(1)$.

It suffices to show that $\sqrt{n}(\hat{\boldsymbol{\beta}}^r - \hat{\boldsymbol{\beta}}^o) = o_p(1)$ because $\sqrt{n}(\hat{\boldsymbol{\beta}}^o - \boldsymbol{\beta}^0) = \sqrt{n}(\hat{\boldsymbol{\beta}}^o_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) = O_p(1)$. Given for all $\epsilon > 0$, $P(\sqrt{n}|\hat{\boldsymbol{\beta}}^r - \hat{\boldsymbol{\beta}}^o| \geq \epsilon) = P(\sqrt{n}|\hat{\boldsymbol{\beta}}^r - \hat{\boldsymbol{\beta}}^o| \geq \epsilon,\ \hat{\boldsymbol{\beta}}^r \neq \hat{\boldsymbol{\beta}}^o) \leq P(\hat{\boldsymbol{\beta}}^r \neq \hat{\boldsymbol{\beta}}^o) \leq P(\hat{\mathcal{A}} \neq \mathcal{A}) \to 0$. The key point in this proof is the last inequality, which holds since $\hat{\boldsymbol{\beta}}^r$ is equivalent to $\hat{\boldsymbol{\beta}}^o$ on the set $\{\hat{\mathcal{A}} = \mathcal{A}\}$. This result applies in general, regardless the asymptotic distribution, as long as $\hat{\boldsymbol{\beta}}^o_{\mathcal{A}}$ is $\sqrt{n}$-consistent. That is, nonnormal $n^{1/2}$ limit distributions are permitted.

In Theorem 3.2, we obtain asymptotic normality of $\hat{\boldsymbol{\beta}}^r_{\mathcal{A}}$ by assuming asymptotic normality of $\hat{\boldsymbol{\beta}}^o_{\mathcal{A}}$. The proof is omitted since it follows along the same lines as that of Theorem 3.1.

Theorem 3.2. *Asymptotic Normality*

If $P(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ *and* $\sqrt{n}(\hat{\boldsymbol{\beta}}^o_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_{\mathcal{A}})$, *then* $\sqrt{n}(\hat{\boldsymbol{\beta}}^r_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_{\mathcal{A}})$.

Next, our interest is to show consistency of the refit estimator in variable selection.

Theorem 3.3. *Selection Consistency*

If $P(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ *and* $\sqrt{n}(\hat{\boldsymbol{\beta}}^o_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) = O_p(1)$, *then* $P(\hat{\mathcal{A}}^r = \mathcal{A}) \to 1$.

We only need to show that $P(\hat{\beta}^r_j \neq 0) \to 1$ for all $j \in \mathcal{A}$ and $P(\hat{\beta}^r_j = 0) \to 1$ for all $j \in \mathcal{A}^c$. Theorem 3.1 implies that $P(\hat{\beta}^r_j = 0) \to 0$ for all $j \in \mathcal{A}$. By selection consistency of the original estimator, $P(\hat{\beta}^r_j \neq 0) \to 0$ for all $j \in \mathcal{A}^c$.

Summarizing the results in Theorems 3.2 and 3.3, if the oracle estimator is asymptotically normal, then the refit estimator is also asymptotically normal with asymptotic

covariance matrix equal to that of the oracle estimator. The refit procedure is generally selection consistent, regardless of the asymptotic distribution of the refit estimator.

3.2.2 Refit Least Squares Approximation (LSA) Estimation

Establishing the oracle properties can be challenging, with the results often being model- and objective function-dependent. Moreover, the corresponding computations may need to be addressed on a case by case basis. In the LSA penalization framework, a least squares approximation replaces the unpenalized objective function based on a preliminary model fit and is regularized using the LASSO penalty. The resulting penalized objective function is intended to approximate the LASSO penalized least squares. The powerful path-finding algorithm LARS (Efron et al. 2004) can be directly applied, greatly simplifying the implementation of the original LASSO problem.

Wang and Leng (2007) demonstrated consistent variable selection and the oracle property for LSA, under an assumption about the asymptotic covariance matrix of the preliminary estimators. Although the covariance assumption holds when the preliminary estimators are asymptotically equivalent to maximum likelihood estimators, the assumption is not satisfied when the covariance matrix has a sandwich variance form. The sandwich form may arise in non-likelihood based estimation, for example, least squares estimation of heteroscedastic linear models, L_1 estimation of quantile regression with heteroscedastic errors, and generalized estimating equations for correlated data. In such applications, Wang and Leng (2007) proved that LSA yields sparse and consistent model selection, but that the penalized estimators may not be oracle.

We illustrate how the refit method is applied to the LSA estimator. First, one calculates $\boldsymbol{\beta}$ by finding $\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \mathcal{L}_n(\boldsymbol{\beta})$. Similar to $\boldsymbol{\beta}^0$, we partition $\tilde{\boldsymbol{\beta}}$ into two parts: $\tilde{\boldsymbol{\beta}}_{\mathcal{A}}$, $\tilde{\boldsymbol{\beta}}_{\mathcal{A}^c}$. The assumptions follow those of Wang and Leng (2007). One necessary condition is that $\tilde{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent and asymptotically normal, that is, $\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0) \xrightarrow{d}$

$N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the asymptotic covariance matrix of $\tilde{\boldsymbol{\beta}}$. In addition, the procedure requires a consistent estimate of the asymptotic covariance matrix of $\tilde{\boldsymbol{\beta}}$, $\boldsymbol{\Sigma}$, denoted by $\hat{\boldsymbol{\Sigma}}$. Lastly, for all $\mathcal{B} \supset \mathcal{A}$,

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}_{\mathcal{B}}^{\mathcal{B}} - \boldsymbol{\beta}_{\mathcal{B}}^{0}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_{\mathcal{B}}), \tag{3.3}$$

where $\tilde{\boldsymbol{\beta}}_{\mathcal{B}}^{\mathcal{B}}$ and $\boldsymbol{\beta}_{\mathcal{B}}^{0}$ are the subvectors of $\tilde{\boldsymbol{\beta}}^{\mathcal{B}}$ and $\boldsymbol{\beta}^{0}$ associated with the candidate model $\mathcal{B}$ respectively, and $\tilde{\boldsymbol{\beta}}^{\mathcal{B}} = \underset{\{\boldsymbol{\beta} \in \mathbb{R}^p:\ \beta_j = 0,\ \forall j \notin \mathcal{B}\}}{\operatorname{argmin}} \mathcal{L}_n(\boldsymbol{\beta})$. The LSA estimator is defined as

$$\tilde{\boldsymbol{\beta}}^{\lambda} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + \sum_{j=1}^{p} \lambda_j |\beta_j|, \tag{3.4}$$

where λ_j is a tuning parameter for β_j. The estimator consists of two components: $\tilde{\boldsymbol{\beta}}_{\mathcal{A}}^{\lambda} = [\tilde{\beta}_j^{\lambda}]_{j \in \mathcal{A}}$ and $\tilde{\boldsymbol{\beta}}_{\mathcal{A}^c}^{\lambda} = [\tilde{\beta}_j^{\lambda}]_{j \in \mathcal{A}^c}$. The LSA selected model is $\hat{\mathcal{A}}^l = \{i : \tilde{\beta}_i^{\lambda} \neq 0\}$. The convex optimization problem in (3.4) may be solved via a two step algorithm employing LARS (Efron, Hastie, Johnstone, and Tibshirani 2004). The first step is to obtain $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$, and the second step is to minimize the resulting $\mathcal{L}_1$-penalized least squares.

Wang and Leng (2007) demonstrated the oracle properties of $\tilde{\boldsymbol{\beta}}^{\lambda}$ under certain conditions: the suitable choice of tuning parameters, which guarantees selection consistency, and the covariance assumption, which ensures efficiency relative to the oracle estimator based on $\mathcal{L}_n(\boldsymbol{\beta})$. Define $a_n = \max\{\lambda_j, j \in \mathcal{A}\}$ and $b_n = \min\{\lambda_j, j \in \mathcal{A}^c\}$. If $\sqrt{n} a_n \xrightarrow{p} 0$ and $\sqrt{n} b_n \xrightarrow{p} \infty$, $\tilde{\boldsymbol{\beta}}^{\lambda}$ is $\sqrt{n}$-consistent and consistent in variable selection. One example of such tuning parameters is the inverse of the absolute values of consistent estimators for $\boldsymbol{\beta}$. The oracle property of LSA specifically requires that:

$$\boldsymbol{\Sigma}_{\mathcal{B}}^{-1} = \boldsymbol{\Sigma}_{(\mathcal{B})}^{-1}, \tag{3.5}$$

where $\boldsymbol{\Sigma}_{(\mathcal{B})}^{-1}$ is the submatrix of $\boldsymbol{\Sigma}^{-1}$ associated with $\mathcal{B}$ for any $\mathcal{B} \supset \mathcal{A}$. The assumption is violated if the asymptotic covariance matrix of the consistent estimate $\tilde{\boldsymbol{\beta}}$ has a

sandwich form, $\mathbf{\Sigma} = W^{-1}VW^{-1}$ for some matrices W and V, where $W \neq V$. In such settings, LSA estimators may not be as efficient as the oracle estimator even though they are $\sqrt{n}$-consistent, selection consistent, and asymptotically normal.

The refit LSA method does not require the covariance assumption to achieve the oracle property, while taking advantage of the easy and general implementation of the LSA method. The LSA method yields $\hat{\mathcal{A}}^l$, an estimate of the non-zero regression coefficients. One then minimizes the unpenalized loss function over those coefficients, that is, $\forall j \in \hat{\mathcal{A}}^l$. The refit LSA estimator is

$$\hat{\boldsymbol{\beta}}^{r(\hat{\mathcal{A}}^l)} = \underset{\{\boldsymbol{\beta}\in\mathbb{R}^p:\beta_j=0,\ \forall j\notin\hat{\mathcal{A}}^l\}}{\operatorname{argmin}} \mathcal{L}_n(\boldsymbol{\beta}). \tag{3.6}$$

To ease notation, we write the refit estimator as $\hat{\boldsymbol{\beta}}^{rl}$ and $\hat{\mathcal{A}}^{rl} = \{j : \hat{\beta}_j^{rl} \neq 0\}$. We decompose $\hat{\boldsymbol{\beta}}^{rl}$ into $\hat{\boldsymbol{\beta}}^{rl}_{\mathcal{A}}$ and $\hat{\boldsymbol{\beta}}^{rl}_{\mathcal{A}^c}$.

Corollary 3.1 below states the $\sqrt{n}$-consistency of the refit LSA estimator, assuming $\tilde{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimator and there exists a consistent estimate of asymptotic covariance of $\tilde{\boldsymbol{\beta}}$.

Corollary 3.1. *$\sqrt{n}$-Consistency of Refit LSA*
If $\sqrt{n}a_n \xrightarrow{p} 0$, $\sqrt{n}b_n \xrightarrow{p} \infty$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}^o_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) = O_p(1)$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}^{rl} - \boldsymbol{\beta}^0) = O_p(1)$.

Corollary 3.2 shows the consistent variable selection and the oracle property for the refit LSA method without the covariance assumption in Wang and Leng (2007).

Corollary 3.2. *Selection Consistency and Oracle Properties of Refit LSA*
If $\sqrt{n}a_n \xrightarrow{p} 0$, $\sqrt{n}b_n \xrightarrow{p} \infty$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}^o_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) \xrightarrow{d} N(0, \mathbf{\Sigma}_{\mathcal{A}})$, $\hat{\boldsymbol{\beta}}^{rl}$ satisfies:
(a) Selection consistency: $P(\hat{\mathcal{A}}^{rl} = \mathcal{A}) \to 1$, and
(b) Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\beta}}^{rl}_{\mathcal{A}} - \boldsymbol{\beta}^0_{\mathcal{A}}) \xrightarrow{d} N(0, \mathbf{\Sigma}_{\mathcal{A}})$.

3.2.3 Simulation Studies

We perform simulation studies to compare the refit method with the LSA method. We consider linear and median regression with heteroscedasticity. Note that the scenario violates the covariance assumption for the oracle property of LSA. We generated 500 datasets with sample sizes $n = 100$ and $n = 250$ for each setting.

The model error of an estimator, $\hat{\boldsymbol{\beta}}$, $\mu(\boldsymbol{x}^T\hat{\boldsymbol{\beta}})$ is defined as $ME(\hat{\boldsymbol{\beta}}) = E\{\mu(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}) - \mu(\boldsymbol{x}^T\boldsymbol{\beta}^0)\}^2$, where $\mu(\boldsymbol{x}^T\boldsymbol{\beta}) = E(y|\boldsymbol{x})$ (Zou and Li 2008). The relative model error (RME) of $\hat{\boldsymbol{\beta}}$ to the ordinary estimator, $\tilde{\boldsymbol{\beta}}$, is defined as $ME(\hat{\boldsymbol{\beta}})/ME(\tilde{\boldsymbol{\beta}})$. Median RME (MRME) is reported, along with true positives (TP) and false positives (FP). TP is the average number of coefficients set to non-zero among the true non-zero coefficients and FP is the average number of coefficients set to non-zero among the true zero coefficients (Bradic, Fan, and Wang 2011). We also summarize the ratios of simulated datasets which are underfit (UF), correctly fit (CF) or overfit (OF) relative to the true model. An underfitted model is any candidate model which fails to select at least one significant variable, while an overfitted model includes all important variables and at least one insignificant variable (Wang et al. 2007b).

Example 3.1. *(Linear Regression with Heteroscedasticity). We consider linear regression models with unequal variance assumption. We generate n observations from a linear regression model* $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}^0 + \sigma_i\epsilon_i$, *where* $\boldsymbol{\beta}^0 = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. *The covariate* $\boldsymbol{x}_i$ *is multivariate normal with mean 0 and covariance* $Cov(x_{ij_1}, x_{ij_2}) = 0.5^{|j_1 - j_2|}$. *The components of* ϵ_i *follow standard normal distribution. We denote* $\sigma_i = \sigma|\boldsymbol{x}_i^T\boldsymbol{\beta}^0|^\theta$ *and choose* $\sigma = 1,\ \theta = 1$ *and* $\sigma = 2,\ \theta = 0.5$.

Table 3.1 compares the refit estimator, the LSA estimator and the oracle estimator across the 500 simulated datasets. The results exhibit that the refit LSA may outperform the LSA method in terms of model error, under heteroscedascity, as might be expected. One should recognize that both methods have the same average number of

TP, FP and the same ratios of correctly fitted, overfitted or underfitted models. Thus, any improvements with the refit should be attributable to the method of estimation, not the model selection.

σ	θ	n	Method	MRME (SE)	TP	FP	UF	CF	OF
1	1	100	LSA	74.83 (2.19)	2.74	0.4	0.24	0.52	0.24
			R-LSA	70.94 (1.99)	2.74	0.4	0.24	0.52	0.24
			Oracle	47.82 (1.03)	3	0	0	1.00	0
1	1	250	LSA	66.71 (1.39)	2.96	0.22	0.04	0.77	0.19
			R-LSA	62.63 (1.30)	2.96	0.22	0.04	0.77	0.19
			Oracle	50.04 (0.97)	3	0	0	1.00	0
2	0.5	100	LSA	65.30 (1.69)	2.88	0.34	0.12	0.64	0.24
			R-LSA	64.81 (1.57)	2.88	0.34	0.12	0.64	0.24
			Oracle	45.62 (1.00)	3	0	0	1.00	0
2	0.5	250	LSA	59.74 (1.09)	3	0.15	0	0.85	0.15
			R-LSA	54.27 (1.04)	3	0.15	0	0.85	0.15
			Oracle	46.02 (0.96)	3	0	0	1.00	0

Table 3.1: Refit LSA for Linear Regression Models

Example 3.2. *(Median Regression with Heteroscedasticity). We consider a median regression model* $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}^0 + \sigma_i\epsilon_i$, *where* $\beta^0 = (3, 1.5, 0, 0, 2, 0, 0, 0)$. *The covariates* $\boldsymbol{x}_i$ *are generated in the same manner as above, and* ϵ_i *follows a mixture distribution, where 90% of observations come from a standard normal distribution and the other 10% come from a standard Cauchy distribution. Let* $\sigma_i = \sigma|\boldsymbol{x}_i^T\boldsymbol{\beta}^0|^\theta$. *The selection of* σ, θ *is the same as the above simulation study.*

In Table 3.2 we observe a more substantial decrease in model error of the refit method comparing to that of the LSA method versus the linear regression setting. The decrease is more notable, which agrees with our theoretical findings regarding the oracle property of the refit technique. A reduction of 10-15% in MRME is evidenced in such settings.

σ	θ	n	Method	MRME (SE)	TP	FP	CF	OF	UF
1	1	100	LSA	75.43 (1.45)	2.95	0.24	0.05	0.77	0.18
			R-LSA	65.42 (2.04)	2.95	0.24	0.05	0.77	0.18
			Oracle	51.02 (1.64)	3	0	0	1.00	0
1	1	250	LSA	70.74 (1.10)	3	0.05	0	0.96	0.04
			R-LSA	58.94 (1.63)	3	0.05	0	0.96	0.04
			Oracle	56.65 (1.64)	3	0	0	1.00	0
2	0.5	100	LSA	76.24 (1.72)	2.86	0.43	0.13	0.59	0.28
			R-LSA	68.11 (1.81)	2.86	0.43	0.13	0.59	0.28
			Oracle	43.55 (1.40)	3	0	0	1.00	0
2	0.5	250	LSA	61.67 (1.19)	2.99	0.13	0.01	0.88	0.11
			R-LSA	56.22 (1.41)	2.99	0.13	0.01	0.88	0.11
			Oracle	49.04 (1.33)	3	0	0	1.00	0

Table 3.2: Refit LSA for Median Regression Models

3.3 Ensemble Variable Selection

In this subsection, we suggest robust EVS from the multiple penalization methods to avoid the worst and have nearly the best performance. We first evaluate the number of votes for each covariate from the multiple methods. The candidate models of our interest are obtained based on the number of votes, from the model with unanimously chosen covariates to the model with at least one voted covariates. The next step selects the best model among the refitted candidate models, which is computationally cheap since we have a handful of nested candidate models. The preliminary penalized techniques should have $\sqrt{n}$-consistency and selection consistency for oracle estimation. Regardless of whether they are asymptotic efficient or not, the efficiency is finally achieved as we use the refit method in the last step. Not only EVS does reduce the bias of the non-zero coefficients as the refit procedure but also it is robust for the penalty function choice. The improvement on the model selection accuracy is evidenced by results from numerical studies.

3.3.1 Ensemble of Decisions on Variable Selection

Assume that we obtain K candidate models from K penalization methods, $P_1, \cdots, P_K$. Suppose we have six covariates, $x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6$. We aggregate the vote results of the K methods for each covariate into a frequency table in Table 3.3. If a penalization method selects a covariate, record 1 in the corresponding cell, otherwise, record 0 in the cell. We construct multiple nested models with covariates selected by at least m methods, $1 \leq m \leq K$. For example, the first and second covariates are selected by all the methods ($m = K$), hence the corresponding model has x_1 and x_2 as its covariates. When the m is $K-1$ or $K-2$, the corresponding model has x_1, x_2 and x_3 as its covariates. We obtain three possible candidate models: $(x_1,\ x_2,\ x_3,\ x_4)$, $(x_1,\ x_2,\ x_3)$, and $(x_1,\ x_2)$. Note that the first model is the union of the selected variables of each model and the last one is the intersection of them.

Method	x_1	x_2	x_3	x_4	x_5	x_6
P_1	1	1	1	1	0	0
P_2	1	1	0	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
P_{K-1}	1	1	1	0	0	0
P_K	1	1	1	0	0	0
Total	K	K	$K-1$	$K-3$	0	0

Table 3.3: K Models Votes Table

The next step is to select the best model among the nested models. We propose two selection criteria for the final model selection: BIC and K-fold cross-validation. BIC asymptotically selects the true sparse model if the nested models include the true model (Zou, Hastie, and Tibshirani 2007). K-fold cross-validation first requires the data to be split into K subsamples. The $K-1$ subsamples are used for training data and the remaining subsample is used for validation data. On the training data, we perform the entire procedures: the multiple penalization methods and the votes table construction.

We obtain the candidate models from the votes table and select the model having the smallest cross-validation error. The best model is then fitted to the whole dataset. Since the models are nested, we assign one single value to each model, τ, which implies that at least τ procedures select the covariates in the model.

Some penalties may be more competitive under certain settings while other penalties may be more competitive under other settings. Averaging out the results is the compromise between the performance of the penalty functions. It is not always the best but tends to work much better than the worst choice, further it can be almost the best in some numerical studies. Most studies of the penalty functions numerically demonstrate the performance of their penalization method by test error comparison. The comparison procedure may require intensive computation while the multiple penalization methods may only require simple computation. EVS is a simple method taking advantage of the information from the penalization methods. It is the combination of covariates screening from the penalization methods and the best subset selection procedure. The best subset selection can be computationally feasible in high dimensions due to the screening step.

3.3.2 Simulation Studies

In this section, we investigate the performance of the refit method and EVS under various scenarios such as linear and median regression with homoscedasticity and logistic regression. We first consider four penalty methods: adaptive Lasso, SCAD, MCP, and LSA. They are used as preliminary procedures for the refit methods and EVS. We consider both EVS with BIC (EVS-BIC) and with K-fold cross-validation (EVS-CV), discussed in Section 3.3.1. As in Section 3.2.3, we evaluate the performance of the refitting and EVS in terms of MRME, variable selection performance, and the ratios of correct model fitting. We simulate 500 datasets for each setting.

We can easily implement Adaptive Lasso, MCP and SCAD method for linear and logistic regression using *R packages* such as *glmnet* and *ncvreg* (Breheny and Huang 2011, Friedman, Hastie, and Tibshirani 2010). Each package develops coordinate descent algorithms for Lasso type penalties and concave penalties such as SCAD and MCP respectively. For SCAD and MCP penalized median regressions, we employ local linear approximation algorithms. All of the tuning parameters are selected by 5-fold cross-validation.

Example 3.3. *(Linear Regression with Homoscedasticity). We consider linear regression model of Example 1 in (Zou and Li 2008). We generate n observations from a linear regression model* $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}^0 + \sigma\epsilon_i$, *where* $\boldsymbol{\beta}^0 = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)^T \in \mathbb{R}^{12}$. *The covariate* $\boldsymbol{x}_i$ *is multivariate normal with mean 0 and covariance* $Cov(x_{ij_1}, x_{ij_2}) = 0.5^{|j_1-j_2|}$. *The components of* ϵ_i *follow standard normal distribution and* σ*=3. Sample sizes are* $n = 100$ *or* $n = 200$.

The model error is written as $ME(\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)^T E(\boldsymbol{x}\boldsymbol{x}^T)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)$. The RME of $\hat{\boldsymbol{\beta}}$ to $\tilde{\boldsymbol{\beta}}$ is $ME(\hat{\boldsymbol{\beta}})/ME(\tilde{\boldsymbol{\beta}})$. Table 3.4 shows the results for median RME (MRME), variable selection performance measures (TP, FP), and model fitting accuracies (UF, CF, OF). EVS-BIC may outperform the other penalization methods and refit methods except SCAD and R-LSA in terms of model error. EVS-BIC has almost the same variable selection performance as that of LSA. Thus, both are shown to be the most accurate fitting procedure. EVS-CV does not perform as well as EVS-BIC but is acceptable in terms of variable selection and parameter estimation. As the sample size increases, EVS-BIC tends to be closer to R-LSA. The EVS method may not be the best but does avoid the worst case. Moreover, the method reduces the variability on the results caused by the choice of penalty function.

Next, we focus on the refit methods based on the four penalty functions. They have advantage over the regular penalized methods only for LSA and SCAD with $n = 200$.

		MRME (SE)	TP	FP	UF	CF	OF
n=100	Oracle	0.19 (0.01)	3.00	0.00	0.00	1.00	0.00
	Ordinary	1.00 (0.00)	3.00	9.00	0.00	0.00	1.00
	AdLasso	0.45 (0.01)	2.99	1.62	0.01	0.36	0.63
	MCP	0.33 (0.01)	2.97	1.12	0.02	0.48	0.49
	SCAD	0.28 (0.02)	2.98	1.39	0.02	0.37	0.61
	LSA	0.34 (0.01)	2.96	0.33	0.03	0.72	0.24
	R-AdLasso	0.60 (0.01)	2.99	1.62	0.01	0.36	0.63
	R-MCP	0.46 (0.02)	2.97	1.12	0.02	0.48	0.49
	R-SCAD	0.51 (0.01)	2.98	1.39	0.02	0.37	0.61
	R-LSA	0.28 (0.01)	2.96	0.33	0.03	0.72	0.24
	EVS-CV	0.36 (0.02)	2.97	0.86	0.03	0.62	0.35
	EVS-BIC	0.30 (0.01)	2.96	0.35	0.03	0.71	0.26
n=200	Oracle	0.21 (0.01)	3.00	0.00	0.00	100.00	0.00
	Ordinary	1.00 (0.00)	3.00	9.00	0.00	0.00	100.00
	AdLasso	0.45 (0.01)	3.00	1.48	0.00	43.00	57.00
	MCP	0.31 (0.01)	3.00	0.76	0.20	68.00	31.80
	SCAD	0.28 (0.01)	3.00	0.89	0.20	64.20	35.60
	LSA	0.31 (0.01)	3.00	0.21	0.20	83.60	16.20
	R-AdLasso	0.60 (0.01)	3.00	1.48	0.00	43.00	57.00
	R-MCP	0.37 (0.01)	3.00	0.76	0.20	68.00	31.80
	R-SCAD	0.39 (0.01)	3.00	0.89	0.20	64.20	35.60
	R-LSA	0.28 (0.01)	3.00	0.21	0.20	83.60	16.20
	EVS-CV	0.32 (0.01)	3.00	0.65	0.20	75.80	24.00
	EVS-BIC	0.28 (0.01)	3.00	0.22	0.20	82.80	17.00

Table 3.4: Simulation Results for Linear Regression (Gaussian Error)

The refit method has improvement in parameter estimation when the original method attains reasonable variable selection results.

Example 3.4. *(Median Regression with Homoscedasticity). This median regression model is similar to the model of Example 3.2 in Section 3.2.3. We generate a sample of size n from a median regression model,* $y_i = \boldsymbol{x}_i\boldsymbol{\beta}^0 + \sigma\epsilon_i$*, where* $\boldsymbol{\beta}^0 = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)^T \in \mathbb{R}^{12}$ *and* $\sigma = 3$*. The covariates* $\boldsymbol{x}_i$ *are generated in the same manner as above, and* ϵ_i *follows a mixture distribution, where 90% of observations come from a standard normal distribution and the other 10% come from a standard Cauchy distribution. Consider sample sizes of* $n = 100$ *or* $n = 200$*.*

For this model, we use $n\log(\hat{\sigma}^2) + d\log n$ as the BIC (Hurvich and Tsai 1990). Table 3.5 summarizes the MRME, variable selection performance, and model fitting accuracies. The refit and EVS methods have a significant 30-40% decrease in MRME compared to those of the penalized methods. R-LSA and EVS-BIC show the best performance in terms of model error and model selection. EVS-BIC tends to select a sparser model than EVS-CV in the setting of $n = 100$ as expected, but this tendency is reduced under $n = 200$.

Example 3.5. *(Logistic Regression). We simulate the data from a logistic regression model similar to Example 1 in Zhang, Li, and Tsai (2010). Consider the model* $y_i \sim$ *Bernoulli*$\{p(x_i^T\boldsymbol{\beta}^0\}$*, where* $p(u) = \dfrac{exp(u)}{1 + exp(u)}$ *and* $\boldsymbol{\beta}^0 = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0)^T \in \mathbb{R}^{12}$*. The first nine components of* $\boldsymbol{x}_i$ *are generated from the multivariate normal distribution with mean 0 and* $Cov(x_{ij_1}, x_{ij_2}) = 0.5^{|j_1 - j_2|}$*. The last three components identically and independently follow an independent Bernoulli distribution with* $p = 0.5$*. Sample sizes are* $n = 200, 400$*.*

For this model, we estimate model error by Monte Carlo simulation because it does

		MRME (SE)	TP	FP	UF	CF	OF
n=100	Oracle	0.22 (0.01)	3.00	0.00	0.00	1.00	0.00
	Ordinary	1.00 (0.00)	3.00	9.00	0.00	0.00	1.00
	AdLasso	0.59 (0.02)	2.98	2.92	0.01	0.21	0.78
	MCP	0.71 (0.02)	2.96	2.91	0.02	0.24	0.72
	SCAD	0.73 (0.01)	2.96	3.01	0.02	0.20	0.77
	LSA	0.48 (0.01)	2.94	1.11	0.03	0.42	0.52
	R-AdLasso	0.42 (0.01)	2.98	2.92	0.01	0.21	0.78
	R-MCP	0.43 (0.01)	2.96	2.91	0.02	0.24	0.72
	R-SCAD	0.44 (0.01)	2.96	3.01	0.02	0.20	0.77
	R-LSA	0.28 (0.01)	2.94	1.11	0.03	0.42	0.52
	EVS-CV	0.39 (0.01)	2.97	2.38	0.02	0.31	0.66
	EVS-BIC	0.28 (0.01)	2.95	1.03	0.03	0.44	0.51
n=200	Oracle	0.22 (0.01)	3.00	0.00	0.00	100.00	0.00
	Ordinary	1.00 (0.00)	3.00	9.00	0.00	0.00	100.00
	AdLasso	0.49 (0.01)	3.00	2.14	0.40	36.20	63.40
	MCP	0.58 (0.02)	3.00	2.12	0.20	40.60	59.00
	SCAD	0.59 (0.01)	3.00	2.22	0.20	37.40	62.20
	LSA	0.36 (0.01)	2.99	0.37	0.80	72.20	26.80
	R-AdLasso	0.34 (0.01)	3.00	2.14	0.40	36.20	63.40
	R-MCP	0.34 (0.01)	3.00	2.12	0.20	40.60	5.90
	R-SCAD	0.35 (0.01)	3.00	2.22	0.20	37.40	62.20
	R-LSA	0.20 (0.01)	2.99	0.37	0.80	72.20	26.80
	EVS-CV	0.31 (0.01)	2.99	1.67	0.40	51.60	47.80
	EVS-BIC	0.21 (0.01)	3.00	0.41	0.20	71.40	28.20

Table 3.5: Simulation Results for Median Regression (Mixture Error)

not have a closed form (Zou and Li 2008). The BIC for the logistic regression is

$$\{-2y^T\boldsymbol{x}^T\hat{\boldsymbol{\beta}} + 2\sum_{i=1}^{n}(\log(1 + exp(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}))\} + d\log n,$$

whose the first term is the binomial deviance. Table 3.6 shows that SCAD and MCP outperform LSA in terms of model error while their corresponding refit methods perform worse than the R-LSA. The model error of EVS-BIC is the smallest for the both sample sizes and its selection performance has a comparative advantage over any other procedures. With a larger sample size, EVS-BIC has more similar performance to R-LSA. The numerical studies confirm that the EVS methods are robust to the penalty function choices with competitive performance.

		MRME (SE)	TP	FP	UF	CF	OF
n=200	Oracle	0.22 (0.01)	3.00	0.00	0.00	1.00	0.00
	Ordinary	1.00 (0.00)	3.00	9.00	0.00	0.00	1.00
	AdLasso	0.41 (0.01)	3.00	1.36	0.00	0.36	0.63
	MCP	0.36 (0.01)	2.99	0.77	0.01	0.50	0.49
	SCAD	0.35 (0.01)	3.00	1.52	0.00	0.23	0.76
	LSA	0.43 (0.01)	2.74	0.14	0.17	0.71	0.12
	R-AdLasso	0.57 (0.01)	3.00	1.36	0.00	0.36	0.63
	R-MCP	0.45 (0.01)	2.99	0.77	0.01	0.50	0.49
	R-SCAD	0.58 (0.01)	3.00	1.52	0.00	0.23	0.76
	R-LSA	0.33 (0.01)	2.74	0.14	0.17	0.71	0.12
	EVS-CV	0.41 (0.01)	2.95	0.67	0.05	0.60	0.35
	EVS-BIC	0.33 (0.01)	2.94	0.33	0.05	0.71	0.24
n=400	Oracle	0.25 (0.01)	3.00	0.00	0.00	100.00	0.00
	Ordinary	1.00 (0.00)	3.00	9.00	0.00	0.00	1.00
	AdLasso	0.44 (0.01)	3.00	1.16	0.00	47.60	52.40
	MCP	0.36 (0.01)	3.00	0.64	0.20	0.64	35.80
	SCAD	0.34 (0.01)	3.00	1.12	0.20	40.20	59.60
	LSA	0.39 (0.01)	2.99	0.18	0.60	83.60	15.80
	R-AdLasso	0.54 (0.01)	3.00	1.16	0.00	47.60	52.40
	R-MCP	0.42 (0.01)	3.00	0.64	0.20	64.00	35.80
	R-SCAD	0.54 (0.01)	3.00	1.12	0.20	40.20	59.60
	R-LSA	0.31 (0.01)	2.99	0.18	0.60	83.60	15.80
	EVS-CV	0.35 (0.01)	3.00	0.39	0.40	76.40	23.20
	EVS-BIC	0.31 (0.01)	3.00	0.19	0.40	83.40	16.20

Table 3.6: Simulation Results for Logistic Regression

Table 3.7 shows the optimal τ from cross-validation. EVS-CV usually selects the

variables chosen by all penalization methods.

	Mean	Median	SE	Mean	Median	SE
Linear	3.354	4.000	0.047	3.536	4.000	0.044
Median	2.684	3.000	0.048	2.872	3.000	0.052
Logistic	3.252	4.000	0.042	3.650	4.000	0.036

Table 3.7: Optimal τ for Linear, Median, Logistic Regression

3.3.3 South African Heart Disease Data Analysis

The South African heart disease data set has been analyzed with logistic regression in many literatures (Friedman, Hastie, and Tibshirani 2001, Park and Hastie 2007, Wang and Leng 2007). The dataset is a part of the Coronary Risk-Factor Study baseline survey conducted in three rural areas of the Western Cape, South Africa (Rossouw, Du Plessis, Benadé, Jordaan, Kotze, Jooste, and Ferreira 1983). The response is the presence or absence of myocardial infarction (MI) at the time of the survey. There are 462 subjects and nine predictors: systolic blood pressure (*sbp*); cumulative tobacco (kg) (*tobacco*); low density lipoprotein cholesterol (*ldl*); adiposity (*adiposity*); family history of heart disease (*famhist*), type-A behavior (*typea*); obesity (*obesity*); current alcohol consumption (*alcohol*); and, age at onset (*age*).

Table 3.8 presents the estimators and their standard errors from the ordinary, penalized, refit, and EVS methods. Note that the refit adaptive Lasso has the same performance as EVS-CV and the refit LSA performs as well as EVS-BIC. EVS-BIC exactly selects all significant variables in the ordinary logistic regression. The non-zero coefficients of the refit LSA have larger magnitudes than those of the LSA. The 5-fold cross-validation is used for tuning parameter selection for adaptive Lasso, MCP and SCAD. We compare the three penalization methods in terms of variable selection. The adaptive Lasso has the smallest model, SCAD selects one more variable, *sbp*, and MCP

additionally selects *adiposity*.

	Ordinary (SE)	AdLasso	MCP	SCAD	LSA
(Intercept)	-6.1507 (1.31)	-5.3048	-6.1501	-6.337	-5.508
sbp	0.0065 (0.01)	0	0.0065	0.0025	0
tobacco	0.0794 (0.03)	0.0779	0.0795	0.08	0.065
ldl	0.1739 (0.06)	0.1748	0.1738	0.1717	0.1306
adiposity	0.0186 (0.03)	0	0.0186	0	0
famhist	0.9254 (0.23)	0.9136	0.9258	0.9135	0.7893
typea	0.0396 (0.01)	0.0308	0.0396	0.0379	0.0277
obesity	-0.0629 (0.04)	-0.0249	-0.063	-0.02	0
alcohol	0.0001 (0.00)	0	0	0	0
age	0.0452 (0.01)	0.0462	0.0452	0.05	0.0473
	R-MCP (SE)	R-SCAD (SE)	EVS-CV (SE) R-AdLasso	EVS-BIC (SE) R-LSA	Sig.
(Intercept)	-6.1501 (1.31)	-6.4169 (1.24)	-5.7027 (1.08)	-6.4464 (0.92)	
sbp	0.0065 (0.01)	0.0068 (0.01)	0 (0.00)	0 (0.00)	
tobacco	0.0795 (0.03)	0.0799 (0.03)	0.08 (0.03)	0.0804 (0.03)	**
ldl	0.1738 (0.06)	0.1821 (0.06)	0.1837 (0.06)	0.162 (0.06)	**
adiposity	0.0186 (0.03)	0 (0.00)	0 (0.00)	0 (0.00)	
famhist	0.9258 (0.23)	0.9245 (0.23)	0.9161 (0.23)	0.9082 (0.23)	**
typea	0.0396 (0.01)	0.039 (0.01)	0.0383 (0.01)	0.0371 (0.01)	**
obesity	-0.063 (0.04)	-0.0422 (0.03)	-0.0376 (0.03)	0 0.00	
alcohol	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	
age	0.0452 (0.01)	0.0489 (0.01)	0.0521 (0.01)	0.0505 (0.01)	**

Table 3.8: Estimates and Standard Deviations for South African Heart Data

Table 3.9 reports the test error of all the methods based on 100 repetitions. We randomly split the dataset into training and test datasets with a ratio of 9 to 1. We fit each model on the training data, and obtain the binomial deviance of the model on the test set as its test error. All the refit estimators have smaller test errors than their corresponding penalized methods except the refit adaptive Lasso. Both EVS methods, EVS-CV and EVS-BIC, show better performance than the other procedures. EVS-BIC has the smallest test error, which is the same as that of the refit LSA. This implies that the EVS-BIC identifies the refit LSA in every repetition. Further, EVS-CV has better performance than the refit MCP and the refit SCAD.

	Ordinary	AdLasso	MCP	SCAD	LSA
Mean	49.3975	49.0492	49.4069	49.4188	49.1468
SE	0.7009	0.672	0.6968	0.6924	0.6224
	R-AdLasso	R-MCP	R-SCAD	EVS-CV	EVS-BIC
Mean	49.1082	49.3523	49.2917	49.2682	48.7897
SE	0.6989	0.7012	0.6963	0.7007	0.6917

Table 3.9: Test Error for South African Heart Data

3.4 Ensemble Variable Selection and Estimation

In this section, we suggest EVE, the variable selection and estimation technique for a factorizable likelihood-based model when the direct penalization on the full likelihood is intractable. EVE is a multi-layer procedure, which incorporates the ensemble estimation via GLS and the refit LSA method. Cox (2001) showed that the ensemble estimation is asymptotically efficient based on the combination of information across the likelihood factors. The first step of EVE is the ensemble estimation on the full likelihood to obtain the ensemble estimator and its covariance estimate. The LSA method uses these estimates as the preliminary estimators and the refit method is applied to each likelihood factor-based model. We finally obtain the EVE estimator via the ensemble estimation to the refit estimators. We illustrate the procedure on the Cox proportional hazards model for the prospective doubly censored data. Simulation studies and MACS data analysis confirm that EVE is competitive with other methods.

3.4.1 Likelihood Factorization and Ensemble Estimation

We consider the full likelihood for parameter vector $\boldsymbol{\theta}$ based on observations, $(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$. Suppose that the likelihood is decomposed into two parts:

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{\theta}) = \mathcal{L}_1(\boldsymbol{y}|\boldsymbol{\theta})\mathcal{L}_2(\boldsymbol{y}|\boldsymbol{\theta}). \tag{3.7}$$

Cox (2001) suggests an asymptotically efficient estimation of the common parameter, $\boldsymbol{\theta}$. Denote $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ as separate maximum likelihood estimates and $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$ as associated observed information matrices. The efficient estimation is a combination of information via the generalized least squares estimator:

$$\sum_{i=1}^{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_i)^T\hat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_i). \tag{3.8}$$

The ensemble estimator is the minimizer of (3.8), $(\hat{\boldsymbol{\Sigma}}_1^{-1}+\hat{\boldsymbol{\Sigma}}_2^{-1})^{-1}(\hat{\boldsymbol{\Sigma}}_1^{-1}\hat{\boldsymbol{\theta}}_1+\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\theta}}_2)$ and its covariance estimate is $(\hat{\boldsymbol{\Sigma}}_1^{-1}+\hat{\boldsymbol{\Sigma}}_2^{-1})^{-1}$. The efficiency loss of ensemble estimator is $O_p(1/n)$, which implies no asymptotic efficiency loss.

We propose a variable selection and estimation procedure incorporating ensemble estimation, the refit method, and LSA. It performs variable selection and estimation even when direct penalization on the full likelihood or the likelihood factors is intractable. We first obtain the ensemble estimator and its covariance matrix from the likelihood factorization. LSA method is applied with the preliminary estimator to select important variables. We separately refit the model of each likelihood on the selected model. Finally, the refit estimators are combined into the ensemble refit estimator in the same manner as above.

The procedure only requires a maximum likelihood estimation of each factor. Ensemble estimation is performed twice to obtain the preliminary estimators for LSA and the final refit estimator for the selected model. The proposed method is indirect but feasible when the full likelihood can be factorized into separate parts. Further, its computation is simple with existing programmings. It is not applied directly to the $n << p$ situation, but we have an alternative method similar to the modified LSA of Wang and Leng (2007). Suppose two likelihood factors have a feasible direct penalization method. Two penalized estimators select two models, whose cardinality is less than n.

We perform the refit method for each likelihood factor under the union of the models. The ensemble estimator and its covariance matrix is obtained by combining the refit estimators and their observed information matrices. The regularized ensemble estimator and covariance matrix replace the ordinary ensemble estimate and its covariance.

3.4.2 The Cox Proportional Hazards Model with Prospective Doubly Censored Data

The Cox proportional hazards model is a popular likelihood based technique to examine the effect of covariates on the survival time (Cox 1972). Several approaches are suggested for the Cox model for prospective doubly censored data (Cai and Cheng 2004, Kim, Kim, and Jang 2010; 2013). To our knowledge, there is no study on the penalized proportional hazards model for prospective doubly censored data. Moreover, the existing works have focused on estimation rather than variable selection due to the complexity of the We study sparse estimation of the Cox model for prospective doubly censored data via ensemble estimation and refit LSA estimation. Assume that the prospective doubly censored dataset has information on the left censoring time for all the observations. It is an extra information for prospective doubly censored data, but plays a key role in the likelihood factorization. With the left censoring time information, the likelihood function is factorized to the likelihood of interval censoring data and the likelihood of left-truncated right-censored (LTRC) data.

First, we describe the study design of the prospective doubly censored data. The study monitors n independent individuals and each individual has the random monitoring time, C_i and the failure time, T_i. The patients are from a cross-sectional sample at baseline and have the covariates, $\boldsymbol{x}_i,\ i = 1, \cdots, n$ influencing on their failure time. We follow the subjects who have not had the event until they have an event

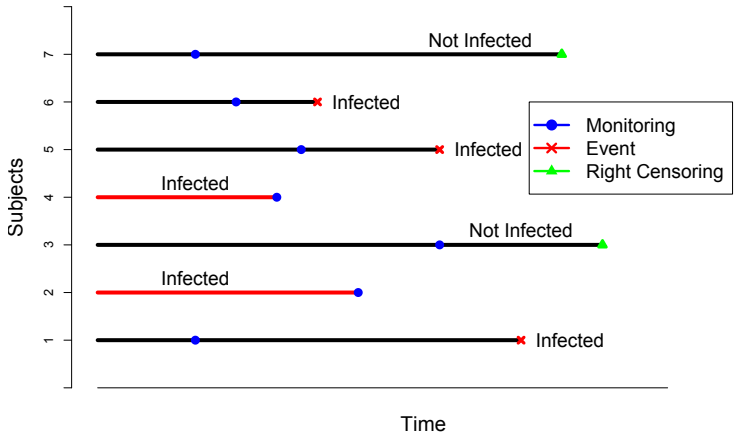
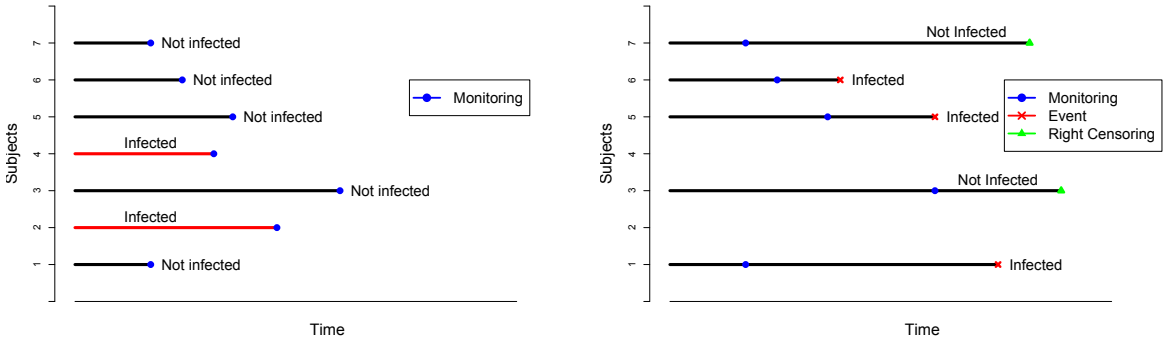


Figure 3.1: Prospective Doubly Censored Data

or are lost to follow-up. They are doubly censored because the failure time is unknown before the monitoring time (C_i) begins or after the right censoring time R_i but known between the time points. The subsamples are left-truncated right-censored (LTRC) data while the original cross sectional samples at the baseline are current status data. The current status data have the observed monitoring time and consist of the triplets, $(C_i,\ \delta_i = I(T_i \leq C_i),\ \boldsymbol{x}_i)$, $i = 1, \cdots, n$. The LTRC data observe the minimum between the event time and the right censoring time and consist of the triplets, $(Y_i = min(T_i, R_i),\ \nu_i = I(T_i \leq R_i),\ \boldsymbol{x}_i)$, $i = 1, \cdots, n$.

We further examine the details of the likelihood function and its likelihood factorization. Suppose T_i follows this distribution $F(T_i = t|\boldsymbol{x}_i)$. The likelihood for the current status data conditional on $C_i, \boldsymbol{x}_i$ is as follows:

$$\prod_{i=1}^{n} L(\delta_i|C_i, \boldsymbol{x}_i) = \prod_{i=1}^{n} F(C_i|\boldsymbol{x}_i)^{\delta_i}(1 - F(C_i|\boldsymbol{x}_i))^{1-\delta_i}. \tag{3.9}$$



(a) Current status data: $\mathcal{L}^1$ (b) Left truncated right censored data: $\mathcal{L}^2$

Figure 3.2: Information Decomposition of Prospective Doubly Censored Data

The likelihood for the LTRC conditional on $\delta_i, C_i, \boldsymbol{x}_i$ is written as below:

$$\prod_{i=1}^{n} L(Y_i, \nu_i | \delta_i, C_i, \boldsymbol{x}_i) = \prod_{i=1}^{n} \{ [\frac{f(Y_i|\boldsymbol{x}_i)}{1 - F(C_i|\boldsymbol{x}_i)}]^{\nu_i} [\frac{1 - F(Y_i|\boldsymbol{x}_i)}{1 - F(C_i|\boldsymbol{x}_i)}]^{1-\nu_i} \}^{1-\delta_i}. \tag{3.10}$$

It is divided by the truncation probability $1 - F(C_i|\boldsymbol{x}_i)$ because the observations contributing this likelihood survive beyond C_i. Right censored observations at R_i contribute the likelihood with the probability to survive beyond R_i given that they have already survived beyond C_i. Both likelihood functions are the factors of the following likelihood function for prospective doubly censored data:

$$\prod_{i=1}^{n} L(Y_i, \nu_i, \delta_i | C_i, \boldsymbol{x}_i) = \prod_{i=1}^{n} F(C_i|\boldsymbol{x}_i)^{\delta_i} f(Y_i|\boldsymbol{x}_i)^{\nu_i(1-\delta_i)} (1 - F(Y_i|\boldsymbol{x}_i))^{(1-\nu_i)(1-\delta_i)}. \tag{3.11}$$

Note that the dataset has an additional information on the baseline C_i for $i = 1, \cdots, n$. By the virtue of the information, we decompose the doubly censored likelihood into the two likelihood factors and analyze them separately.

The Cox proportional hazards model of (3.11) has $\boldsymbol{\beta}$ as the vector of regression coefficients and Λ as the cumulative hazard function. We use the profile likelihood of

$\boldsymbol{\beta}$, where Λ has been profiled out as a full likelihood for $\boldsymbol{\beta}$ (Murphy and Van der Vaart 2000). We can easily implement the procedure of Section 3.4.1 with publicly available codes. We conjecture that the profile likelihood has an asymptotically efficient ensemble estimator under the likelihood factorization.

3.4.3 Simulation Studies

Prospective doubly censored data are generated following the example of Fan and Li (2002). The event time follows the exponential hazard model:

$$h(t|\boldsymbol{x}) = exp(\boldsymbol{x}^T\boldsymbol{\beta}), \tag{3.12}$$

where $\boldsymbol{\beta}^0 = (0.8, 0, 0, 1, 0, 0, 0.6, 0, 0, 0)$. The left censoring time (C_i) follows $Exp(6.9)$ and right censoring time (R_i) is from $R_i = C_i + Exp(0.163)$. The parameters of the censoring distributions are chosen according to the specified censoring rates, which are as follows: left censoring and right censoring rates are $20.26\%, 19.70\%$ for $n = 250$, and $19.93\%, 20.01\%$ for $n = 500$. All results are obtained from 100 simulated datasets.

We assess the estimation performance of estimators in terms of mean squared error (MSE). The estimators are given from 100 simulated datasets, $\hat{\boldsymbol{\beta}}^1, \cdots, \hat{\boldsymbol{\beta}}^{100}$. MSE of the first component of estimator, $\hat{\beta}_1$ is $MSE(\hat{\beta}_1) = \sum_{i=1}^{100}(\hat{\beta}_1^i - \beta_{01})^2/100$ and MSE of $\hat{\boldsymbol{\beta}}$ is the summation of MSE of all the elements. We follow the criteria of Section 3.2.3 to measure performance of variable selection and model fitting. We fit the LTRC part using *coxph* in R and fit the current status data using *intcox* in R. For current status data, we obtain parameter estimation for full model and selected model with *intcox* R package, and estimate covariance matrix using bootstrap with replication of $B = 1000$. *coxph* function in *survival* R package gives parameter estimators and covariance estimators of full model and selected model for LTRC data. LSA estimator

is easily obtained from *lsa* function based on the preliminary estimators.

Tables 3.10-3.11 present mean squared error of estimators for the simulated dataset for $n = 250, 500$. We compare ensemble estimation with several estimators on current status data and LTRC data: oracle, ordinary, LSA, and refit LSA estimators. For comparison, we also include ensemble oracle estimator, which is optimal combination of two oracle estimators. MSE Comparison is considered first across estimators to each dataset and oracle estimator to each dataset performs best. Among the other three estimators, refit LSA estimator outperforms not only for current status data and but for LTRC data. Across all the estimators, the ensemble oracle estimator performs best. Note that ensemble oracle, ordinary, LSA, refit estimators have smaller MSE than those for current status data and LTRC data. It demonstrates that more efficient estimation is possible via combining information from two likelihoods. Double ensemble refit estimation decreases MSE even if the base refit LSA estimators have worse performance than LSA estimator under $n = 500$. MSE of double ensemble refit estimator is the smallest and is closer to that of ensemble oracle estimator as the sample size increases.

We further investigate MSE for each component of estimators. Refit LSA estimation mainly contributes to a reduction in MSE of non-zero coefficients, such as β_1, β_4, and β_7. The analysis of the current status data yields the noticeable result that MSE of refit LSA is 40% less than that of LSA for β_4 under $n = 100$ and for β_1 and β_7 under $n = 500$. In the LTRC data, MSE to β_7 is decreased by 35% from LSA to refit LSA under $n = 100$. Non-zero coefficients of refit procedure attains a smaller MSE than ordinary estimation in the LTRC data and the combination. On the other hand, the MSE of the zero coefficients is increased as a trade-off, but it is negligible compared to the non-zero coefficients. As the LTRC data analysis is stabilized with a larger sample size, such improvements are less obvious under $n = 500$.

		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	
CS	Oracle	0.0395	0	0	0.0369	0	
	Ordinary	0.0563	0.0679	0.0547	0.0555	0.0592	
	LSA	0.1697	0.0068	0.0034	0.1391	0.0008	
	R-LSA	0.1427	0.0171	0.0097	0.0829	0.0021	
LTRC	Oracle	0.0115	0	0	0.0111	0	
	Ordinary	0.0161	0.016	0.015	0.0205	0.0124	
	LSA	0.0168	0.0022	0.003	0.0148	0.0004	
	R-LSA	0.0133	0.0031	0.0047	0.0132	0.0007	
ENS	Oracle	0.009	0	0	0.0086	0	
	Ordinary	0.012	0.0129	0.0126	0.0135	0.0101	
	LSA	0.0134	0.0005	0.0012	0.013	0.0001	
	R-CS	0.0393	0.004	0.0039	0.0388	0.0003	
	R-LTRC	0.012	0.0006	0.0023	0.0123	0.0003	
	R-ENS	0.0092	0.0008	0.0023	0.0099	0.0002	
		$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\boldsymbol{\beta}}$
CS	Oracle	0	0.027	0	0	0	0.1034
	Ordinary	0.0539	0.0471	0.0512	0.0491	0.0415	0.5365
	LSA	0.0061	0.1438	0.0021	0.0013	0	0.4732
	R-LSA	0.0083	0.139	0.0036	0.0021	0	0.4074
LTRC	Oracle	0	0.0069	0	0	0	0.0296
	Ordinary	0.0193	0.0121	0.0117	0.0142	0.0129	0.1502
	LSA	0.0025	0.0112	0.0002	0.0004	0.0011	0.0527
	R-LSA	0.0049	0.0072	0.0004	0.0009	0.0023	0.0507
ENS	Oracle	0	0.0053	0	0	0	0.0228
	Ordinary	0.0156	0.0093	0.0085	0.0101	0.0089	0.1135
	LSA	0.0011	0.0097	0.0002	0	0.0004	0.0396
	R-CS	0.0017	0.0278	0.0006	0	0.0012	0.1176
	R-LTRC	0.0023	0.0073	0.0003	0	0.0014	0.0386
	R-ENS	0.0019	0.0057	0.0003	0	0.0011	0.0314

Table 3.10: Mean Squared Error of Estimators for Simulated Prospective Doubly Censored Data ($n = 250$)

		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	
CS	Oracle	0.0174	0	0	0.021	0	
	Ordinary	0.0166	0.0254	0.0203	0.0273	0.0188	
	LSA	0.0307	0.0051	0.0018	0.0347	0.0023	
	R-LSA	0.0174	0.0065	0.002	0.0246	0.0038	
LTRC	Oracle	0.0062	0	0	0.0058	0	
	Ordinary	0.0086	0.0052	0.0063	0.0078	0.0068	
	LSA	0.0071	0	0.0003	0.007	0.0002	
	R-LSA	0.0064	0	0.0006	0.0061	0.0004	
ENS	Oracle	0.0039	0	0	0.0053	0	
	Ordinary	0.0048	0.0046	0.0048	0.0055	0.0045	
	LSA	0.0047	0.0001	0.0001	0.0067	0	
	R-CS	0.0174	0.0003	0.0001	0.0211	0	
	R-LTRC	0.0064	0.0002	0.0003	0.006	0	
	R-ENS	0.0041	0.0002	0.0002	0.0053	0	
		$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	$\hat{\beta}_9$	$\hat{\beta}_{10}$	$\hat{\boldsymbol{\beta}}$
CS	Oracle	0	0.0144	0	0	0	0.0529
	Ordinary	0.0247	0.024	0.0185	0.0191	0.0142	0.2088
	LSA	0.0049	0.0324	0.002	0.0012	0.0018	0.1169
	R-LSA	0.0101	0.0201	0.0031	0.002	0.0027	0.0925
LTRC	Oracle	0	0.0046	0	0	0	0.0166
	Ordinary	0.0054	0.0075	0.0074	0.007	0.0054	0.0674
	LSA	0.0005	0.0063	0.0007	0.0002	0.0002	0.0226
	R-LSA	0.0011	0.005	0.0011	0.0005	0.0003	0.0214
ENS	Oracle	0	0.003	0	0	0	0.0122
	Ordinary	0.0044	0.0051	0.006	0.005	0.0042	0.0489
	LSA	0.0004	0.0047	0.0006	0.0002	0	0.0174
	R-CS	0.0003	0.0151	0.0014	0.0005	0	0.0561
	R-LTRC	0.0008	0.0045	0.0007	0.0004	0	0.0193
	R-ENS	0.0007	0.0032	0.0007	0.0004	0	0.0148

Table 3.11: Mean Squared Error of Estimators for Simulated Prospective Doubly Censored Data ($n = 500$)

		TP	FP	UF	CF	OF	
n=250	CS	Oracle	3.0000	0.0000	0	100	0
	Ordinary	3.0000	7.0000	0	0	100	
	LSA	2.4300	0.1400	34	52	10	
	R-LSA	2.4300	0.1400	34	52	10	
	LTRC	Oracle	3.0000	0.0000	0	100	0
		Ordinary	3.0000	7.0000	0	0	100
		LSA	3.0000	0.3000	0	76	24
		R-LSA	3.0000	0.3000	0	76	24
	Ensemble	Oracle	3.0000	0.0000	0	100	0
		Ordinary	3.0000	7.0000	0	0	100
		LSA	3.0000	0.1300	0	92	8
		R-CS	3.0000	0.1300	0	92	8
		R-LTRC	3.0000	0.1300	0	92	8
		R-ENS	3.0000	0.1300	0	92	8
n=500	CS	Oracle	3.0000	0.0000	0	100	0
		Ordinary	3.0000	7.0000	0	0	100
		LSA	2.9900	0.2800	0	76	23
		R-LSA	2.9900	0.2800	0	76	23
	LTRC	Oracle	3.0000	0.0000	0	100	0
		Ordinary	3.0000	7.0000	0	0	100
		LSA	3.0000	0.1200	0	90	10
		R-LSA	3.0000	0.1200	0	90	10
	Ensemble	Oracle	3.0000	0.0000	0	100	0
		Ordinary	3.0000	7.0000	0	0	100
		LSA	3.0000	0.0800	0	93	7
		R-CS	3.0000	0.0800	0	93	7
		R-LTRC	3.0000	0.0800	0	93	7
		R-ENS	3.0000	0.0800	0	93	7

Table 3.12: Variable Selection Performance for Simulated Prospective Doubly Censored Data

Table 3.12 summarizes ratios of the true positives/false positives and ratios of the underfitted model, the correct model, and the overfitted model with 100 repetitions. The LSA estimation for the current status data tends to select a sparser model under $n = 250$ and a redundant model under $n = 500$. Thus, it increases the ratio of the underfitted model to 34% and the ratio of the overfitted model to 25% respectively. Ensemble estimation not only eliminates such tendencies by borrowing strength from both likelihoods but is very likely to find the true model. The finite sample studies supports that the ensemble estimation performs fairly well even for the case where the sample size is relatively small. Note that LSA is a benchmark for the variable selection of refit LSA and ensemble estimation.

3.4.4 Multicenter AIDS Cohort Study (MACS) Data Analysis

The Multicenter AIDS center study (MACS) is the first and the largest study to examine the natural history of AIDS (Cole et al. 2012). The study participants are 5619 homosexual and bisexual men enrolled in four cities across the United States, beginning in 1984: Baltimore, Maryland; Chicago, Illinois; Pittsburgh, Pennsylvania; and Los Angeles, California. Every 6 months, the participants underwent a physical exam and provided a blood sample. At each visit, they completed a questionnaire on demographics, habits, disease history, and sexual activities. The seropositivity for HIV type 1 is determined by positive enzyme-linked immunosorbent assays with confirmatory Western blots.

We use the participants' birth date as the study baseline and the time to HIV infection as the response. The midpoint between the last negative seroconversion visit and the first positive seroconversion visit is chosen as the surrogate time-to-event endpoint for HIV infection (Cole et al. 2012). It is considered as a reasonable surrogate for the infection time as long as the time difference between the two visits is less than

or equal to 4 years. We drop 686 observations with missing information, incomplete information and record errors. Table 3.13 summarizes the exclusion criteria with their corresponding number of the subjects. The remaining 4801 subjects are of interest for the analysis. In the group, 1869 subjects were HIV infected prior to the first monitoring time, and 2497 subjects were not infected until the study ended. We have the event time information of 435 subjects since they were infected within the time window of the study.

Obs. No.	Description
652	Missing values in the risk factors
8	Missing HIV infection time
125	No follow-ups after the monitoring time for HIV non-infection subjects
30	Time gap between the HIV infection determining visits $>$ 4 years
4	Record errors

Table 3.13: Analysis Exclusion Criteria of Subjects

For simplicity, we constrain possible risk factors of interest to 21 variables from the participants' information at their first visit. Table 3.14 gives a detailed description on the risk factors on sexual behaviors, drug usage, sexual disease, medical histories, and demographics. We categorize *Race* and *Education* into four classes respectively: *WHITE, BLACK, HISPA, OTHER*; *HIGH, PRECOL, COL, POSTCOL*. Dummy variables are used to model the effects, and *WHITE* and *HIGH* are their base categories respectively. The final model has 27 covariates. *NDRNK* is a continuous variable and and the rest of them are binary variables.

We apply EVE to find out important risk factors for HIV infection and estimate their effects. For purposes of comparison, we also report the four preliminary estimators of EVE: ensemble ordinary estimation, ensemble LSA, ensemble refit LTRC and ensemble refit CS. Further, we can obtain the parameter estimates using a part of the data such as the follow-up information of the non-censored subsamples at the monitoring time

(LTRC) or the infection status information at the first monitoring time (CS). Table 3.15 presents the results from the partial information (LTRC data and CS data) and Table 3.16 presents the analysis results based on EVE and its preliminary procedures.

Variables	Description
COK2Y	Used cocaine last 2 years
CON2P	Had anal receptive with condom last 2 years
CON2Y	Had anal insertive with condom last 2 years
DIABE	Diabetes diagnosed ever
GONOE	Gonorrhea ever
HAS2Y	Took marijuana/hashish last 2 years
MOU2P	Had oral receptive last 2 years
MOU2Y	Had oral insertive last 2 years
MSX2Y	Drugs used with sex last 2 years
NDRNK	Number of drinks/day since last 12 months
NEEDL	Share a needle last 5 years
OPI2Y	Took heroin/other opiates last 2 years
PIE5Y	Body part pierced in last 5 years
RADTE	Radiation therapy/treatment ever
REC2P	Had anal receptive last 2 years
REC2Y	Had anal insertive last 2 years
SMOKE	Ever smoked cigarettes
WARTE	Genital warts or anal warts ever
UNEMP	Current employment, unemployed
WHITE	White race
BLACK	Black race
HISPA	Hispanic race
OTHER	American Indian, Alaskan Native, Asian or Pacific Islander, etc.
HIGH	High school or less
PRECOL	At least one year college but no degree
COL	Four years college/got a degree
POSTCOL	Some graduate work or Post-graduate degree

Table 3.14: Description of Risk Factors

We first investigate the analysis results from the left non-censored subsamples (LTRC) at the top of Table 3.15. The first column corresponds to the ordinary estimator from the Cox model for LTRC data. Its second column is the LSA estimator based on the LTRC ordinary estimator and its third column is the refitted estimator.

The asterisks are marked for coefficient significance at the level of 0.05. All of the procedures agree in terms of the risk factors selection. We choose nine variables, and include gonorrhea (*GONOE*), drug usage during sex (*MSX2Y*), the number of drinks (*NDRNK*), needle sharing (*NEEDL*), body part piercing (*PIE5Y*), and anal receptive sex (*REC2P*). Black people are seen at to be higher risk than the other races and the risk of the HIV infection among blacks relative to the risk of HIV infection among whites is $exp(0.8087) = 2.2450$. We can analyze the effects of other risk factors in the same manner. For example, people with at least one year college but no degree (*PRECOL*) or post-graduate degree (*POSTCOL*) are at higher risk than high school graduates (*HIGH*). The risk level of the education factors can be ranked in the following order: *PRECOL* (the group at most risk), *POSTCOL* (at second-highest risk level), and *HIGH* and *COL* (at lowest risk). However, the effects of the education factors are not consistent with the natural order of the factors: *HIGH*, *PRECOL*, *COL*, and *POSTCOL*. Unfortunately, the LTRC subsample-based procedures do not select the risk factors known to be important to HIV infection such as genital warts (*WARTS*) but select a spurious variable, the number of drinks per day (*NDRNK*). Overall, each estimation procedure based on the LTRC subsamples shows poor performance in terms of variable selection.

We next focus on the estimation procedures based on the CS data at the bottom of Table 3.15. Compared to the LTRC data analysis, the ordinary estimation for the CS data selects more variables such as cocaine usage (*COK2Y*), smoking (*SMOKE*), and genital warts (*WARTE*). The HIV hazard rate is higher among black people. Also, unlike the LTRC results, here the lowest risk education group is *POSTCOL*, the next lower one is *COL*, and *PRECOL*. This corresponds to the education level and might be more convincing than the results from the LTRC data. Even though the event time information is not used, the sampling bias reduction is considered to contribute to this

LTRC	ORD (SE)	Sig.	LSA	REFIT (SE)
COK2Y	0.0306 (0.1133)		.	.
CON2P	-0.0938 (0.6149)		.	.
CON2Y	0.3767 (0.5315)		.	.
DIABE	-0.6366 (0.7162)		.	.
GONOE	0.5300 (0.1036)	(**)	0.5074	0.5333 (0.1010)
HAS2Y	-0.1757 (0.1563)		.	.
MOU2P	0.0420 (0.1068)		.	.
MOU2Y	-0.0517 (0.1239)		.	.
MSX2Y	0.5428 (0.1617)	(**)	0.4448	0.4896 (0.1430)
NDRNK	0.0928 (0.0243)	(**)	0.0841	0.0926 (0.0236)
NEEDL	0.7174 (0.2432)	(**)	0.7186	0.8108 (0.2286)
OPI2Y	0.2416 (0.2555)		.	.
PIE5Y	0.4378 (0.1170)	(**)	0.4004	0.4358 (0.1148)
RADTE	0.4011 (0.3847)		.	.
REC2P	0.4610 (0.1230)	(**)	0.3984	0.4438 (0.1176)
REC2Y	-0.1839 (0.1259)		.	.
SMOKE	0.1157 (0.1058)		.	.
WARTE	0.0713 (0.1103)		.	.
BLACK	0.8087 (0.1986)	(**)	0.6676	0.7514 (0.1891)
HISPA	0.2350 (0.2179)		.	.
OTHER	-0.6785 (0.7109)		.	.
PRECOL	0.4910 (0.1688)	(**)	0.2984	0.3708 (0.1175)
COL	0.1903 (0.1881)		.	.
POSTCOL	0.3947 (0.1821)	(**)	0.1592	0.2618 (0.1263)
UNEMP	0.1842 (0.2009)		.	.
CS	ORD (SE)		LSA	REFIT (SE)
COK2Y	0.6236 (0.0708)	(**)	0.7262	0.6039 (0.0880)
CON2P	-0.2180 (0.2999)		.	.
CON2Y	0.0144 (0.2818)		.	.
DIABE	-0.2839 (0.3314)		.	.
GONOE	0.4459 (0.1265)	(**)	0.2169	0.5007 (0.1252)
HAS2Y	0.1352 (0.1000)		.	.
MOU2P	-0.0926 (0.0572)		.	.
MOU2Y	-0.1352 (0.0899)		.	.
MSX2Y	0.2505 (0.0816)	(**)	0.3595	0.2948 (0.0978)
NDRNK	-0.0397 (0.0187)	(**)	.	.
NEEDL	0.4841 (0.1304)	(**)	0.3667	0.4888 (0.1273)
OPI2Y	-0.0527 (0.1610)		.	.
PIE5Y	0.2187 (0.0503)	(**)	0.1494	0.2222 (0.0620)
RADTE	-0.1481 (0.2205)		.	.
REC2P	0.6182 (0.0523)	(**)	0.5754	0.5885 (0.0564)
REC2Y	0.1077 (0.0717)		.	.
SMOKE	-0.2517 (0.0563)	(**)	-0.2294	-0.2603 (0.0553)
WARTE	0.4013 (0.0509)	(**)	0.4026	0.3962 (0.0485)
BLACK	0.6650 (0.1036)	(**)	0.6289	0.6962 (0.1167)
HISPA	0.3537 (0.1313)	(**)	0.3800	0.3312 (0.1164)
OTHER	0.0495 (0.2346)		.	.
PRECOL	-0.2595 (0.1090)	(**)	.	.
COL	-0.3903 (0.1024)	(**)	-0.1439	-0.2361 (0.0666)
POSTCOL	-0.6179 (0.0965)	(**)	-0.5621	-0.4557 (0.0960)
UNEMP	0.0634 (0.0917)		.	.

Note. (**) indicates significant level 0.05.

Table 3.15: MACS Analysis with LTRC Data or CS Data

improved performance. Specifically, the CS dataset considers all study participants, while the LTRC dataset approximately considers 61% of them. The LSA and refit estimators have some disagreement from the ordinary estimator in terms of variable selection. The ordinary estimation selects the number of drinks (*NDRNK*) and the education level with college entrance but no degree (*PRECOL*) as significant risk factors while the other procedures regard them having negligible effects.

The analyses with EVE and its preliminary procedures are summarized in Table 3.16. EVE is viewed as the information compromise between the LTRC analysis and the CS analysis. Note that EVE selects the same risk factors as the CS refit estimator, but the effects of the selected variables are compromised by the analysis integration. Note that oral sex is considered as a strong risk factor in the ordinary ensemble estimation at the significance level of 0.05. In the following step, oral sex is not considered strongly to be associated with HIV infection from EVE, as evidenced by the previous studies. The preliminary CS refit estimator has better performance in terms of variable selection than the preliminary LTRC refit estimator.

Next, we examine the selected risk factors via EVE in detail. First, in terms of the sexual behaviors, anal receptive sex is strongly associated with HIV infection, while anal insertive sex is not thought to be a strong risk factor. Further, condom usage during anal sex does not seem to prevent the participants from HIV infection. Our risk predictors of interest include the participants' health status and drug usage. EVE selects body part piercing and sexual diseases such as gonorrhea, and genital warts as strong factors. Diabetes and radiation treatment have little effect on HIV infection. Cocaine usage, drug usage during sex and needle sharing increase the risk of HIV infection, but users of other drugs (such as marijuana or heroin) are not shown to be at risk. As in the previous CS data analysis, the infection risk among black people is the highest and that of Hispanic people is the second highest. The infection risk of

	ORD (SE)	Sig.	LSA	EVE (SE)	Sig.
COK2Y	0.4889 (0.0532)	(**)	0.5383	0.4081 (0.0575)	(**)
CON2P	-0.3369 (0.2633)		.	.	
CON2Y	0.1332 (0.2442)		.	.	
DIABE	-0.2238 (0.2783)		.	.	
GONOE	0.5010 (0.0699)	(**)	0.4128	0.5987 (0.0694)	(**)
HAS2Y	0.1104 (0.0757)		.	.	
MOU2P	-0.093 (0.0441)	(**)	.	.	
MOU2Y	-0.1346 (0.0594)	(**)	.	.	
MSX2Y	0.3085 (0.0703)	(**)	0.3680	0.2935 (0.0841)	(**)
NDRNK	0.0020 (0.0134)		.	.	
NEEDL	0.5397 (0.1027)	(**)	0.5384	0.5442 (0.0967)	(**)
OPI2Y	0.0124 (0.1256)		.	.	
PIE5Y	0.2754 (0.0450)	(**)	0.2225	0.3105 (0.0524)	(**)
RADTE	-0.1020 (0.1883)		.	.	
REC2P	0.6013 (0.0471)	(**)	0.5472	0.5441 (0.0465)	(**)
REC2Y	0.0752 (0.0549)		.	.	
SMOKE	-0.1823 (0.0468)	(**)	-0.1040	-0.1891 (0.0466)	(**)
WARTE	0.3529 (0.0445)	(**)	0.3371	0.3474 (0.0426)	(**)
BLACK	0.7385 (0.0900)	(**)	0.7285	0.7344 (0.0937)	(**)
HISPA	0.3914 (0.1007)	(**)	0.3511	0.2610 (0.0962)	(**)
OTHER	0.0690 (0.2160)		.	.	
PRECOL	-0.1474 (0.0755)		.	.	
COL	-0.2866 (0.0796)	(**)	-0.1117	-0.2639 (0.0555)	(**)
POSTCOL	-0.3839 (0.0784)	(**)	-0.3134	-0.2852 (0.0647)	(**)
UNEMP	0.1007 (0.0815)		.	.	
	LTRC (SE)	Sig.		CS (SE)	Sig.
COK2Y	0.0640 (0.1066)			0.6039 (0.0799)	(**)
CON2P	.			.	
CON2Y	.			.	
DIABE	.			.	
GONOE	0.4936 (0.1028)	(**)		0.5007 (0.1138)	(**)
HAS2Y	.			.	
MOU2P	.			.	
MOU2Y	.			.	
MSX2Y	0.5199 (0.1482)	(**)		0.2948 (0.1138)	(**)
NDRNK	.			.	
NEEDL	0.7922 (0.2310)	(**)		0.4888 (0.1148)	(**)
OPI2Y	.			.	
PIE5Y	0.4328 (0.1155)	(**)		0.2222 (0.0603)	(**)
RADTE	.			.	
REC2P	0.4331 (0.1176)	(**)		0.5885 (0.0517)	(**)
REC2Y	.			.	
SMOKE	0.1289 (0.1034)			-0.2603 (0.0542)	(**)
WARTE	0.0722 (0.1098)			0.3962 (0.0471)	(**)
BLACK	0.7423 (0.1904)	(**)		0.6962 (0.1102)	(**)
HISPA	0.2611 (0.2165)			0.3312 (0.1148)	(**)
OTHER	.			.	
PRECOL	.			.	
COL	-0.2189 (0.1313)			-0.2361 (0.0636)	(**)
POSTCOL	-0.0272 (0.1210)			-0.4557 (0.0940)	(**)
UNEMP	.			.	

Note. (**) indicates significant level 0.05.

Table 3.16: MACS Data Analysis with Ensemble Methods

white people is similar to other races including Asians and Pacific Islanders. Another strong risk factor is education, and the risks are significantly different based on college graduation.

3.5 Discussion

In this chapter, we have proposed a general refit method and two statistical methods stemmed from the refit: EVS and EVE. The refit method eliminates the estimation bias inherent in penalization methods, and satisfies asymptotically oracle properties under the selection consistency assumptions of the preliminary penalization method. EVS selects important variables based on the voting from multiple penalization methods and refits the selected model without penalization. The oracle properties of the refit method carry over to EVS with the selection consistency assumption of the preliminary methods. EVE is based on the likelihood factorization assumption and takes advantage of the refit LSA. Its computation is efficient using existing software and its estimation is asymptotically efficient.

As a future direction, it will be interesting to compare the methods for prospective doubly censored data analysis, including EVE. We may consider performing test error calculation based on cross-validation. We first obtain the parametric part of interest and calculate the nonparametric part as a function of the regression parameter using the training data. Next, we calculate the empirical likelihood based on those estimators using the validation data. It may be a challenging problem in terms of computation, but is an important comparison tool for model fitting.

CHAPTER4: CONSISTENT VALIDATION FOR EDGE SELECTION IN HIGH DIMENSIONAL GAUSSIAN GRAPHICAL MODELS

4.1 Introduction

Undirected graphical models are known to be useful for explaining association structure in multivariate random variables (Lauritzen 1996). An edge between two variables in an undirected graphical model represents their conditional dependence given all other variables in the model. Graphical models have had many applications for complex association studies such as gene regulatory networks and social networks (Liu et al. 2010). For example, we can investigate the underlying biological relations among genes from the graph analysis of the regulatory network.

Gaussian graphical modeling (GGM) is a popular method used to learn the undirected graph structures (Lauritzen 1996, Dempster 1972). Under the Gaussian assumption, the inverse covariance matrix is of our interest and known as the precision matrix. Specifically, zero elements of the inverse covariance matrix imply conditional independence of the corresponding variables given all other variables in the model. Then, we can recast the edge selection problem of the graph as a sparsity pattern recovery of the precision matrix. In high-dimensional data, such recovery is a challenging problem due to estimation instability and computational complexity (Yuan and Lin 2007a).

To estimate the high dimensional precision matrices for Gaussian data, Friedman, Hastie, and Tibshirani (2008) and Yuan and Lin (2007a) proposed graphical LASSO (*glasso*) for the graph estimation in high-dimensional data. They regularized the negative Gaussian log-likelihood with the LASSO penalty on the off-diagonal elements of

the inverse covariance matrix. This framework provides a sparser inverse covariance estimate with a larger tuning parameter. In other words, the tuning parameter controls the sparsity level of the graph. Several tuning parameter selection methods have been developed in the literature. In particular, Foygel and Drton (2010) suggested extended BIC (eBIC), and Liu et al. (2010) proposed stability approaches for regularization parameter selection (StARS). eBIC uses an additional tuning parameter, thus needs to consider several tuning parameter values in practice. StARS only has the asymptotic sparsistency under certain assumptions, where many spurious conditional dependence patterns might be included. Furthermore, it involves the use of a cut point value requiring as empirical case by case tuning approach.

In this chapter, our aim is to construct an automatic edge selection procedure excluding such a manual tuning step for high-dimensional GGM. Specifically, the underlying graphical model is assumed to have a small number of true edges. We propose a consistent validation method for edge selection (CoVES) with a growing sample size in fixed dimensions. We recast the problem of tuning parameter selection in high-dimensional L_1 regularized GGM as the problem of graph selection from candidate GGMs along the *glasso* solution path. Next, we apply the Monte Carlo cross-validation to the candidate models for the asymptotically consistent pattern recovery. CoVES was developed from Monte Carlo cross-validation for linear models in Shao (1993) and consistent cross-validation for tuning parameter selection in penalized GLM in Feng and Yu (2013).

The rest of the chapter is organized as follows. We first introduce notations and describe the CoVES procedure in Section 4.2. Its theoretical properties are investigated in Section 4.3 and its performance is compared to other methods from simulation studies in Section 4.4. We summarize the chapter and discuss possible future directions in Section 4.5

4.2 Edge Selection in High Dimensional Gaussian Graphical Models (GGM)

In GGM, the conditional dependent relationship corresponds to the sparsity of the inverse covariance matrix, which is called the precision matrix. The solution path of the *glasso* provides sparse precision matrices along the tuning parameter as candidates. Their corresponding graph structures comprise candidate graph models of interest. CoVES performs a Monte-Carlo bootstrap among the candidate models and selects the optimal graph model.

4.2.1 Settings and Notations

First, denote a graph as $G = (V, E)$, where $V = \{x_1, \cdots, x_p\}$ is the set of vertices and E is the set of edges between vertices. Each vertex corresponds to a variable and an edge between vertices identifies their conditional dependence given all the other variables. Suppose that $\mathbf{x} = (x_1, \cdots, x_p)$ follows a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is a mean vector and $\boldsymbol{\Sigma}$ is a nonsingular covariance matrix. Without loss of generality, assume that $\boldsymbol{\mu} = \mathbf{0}$. Denote $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ as the sample mean. Define $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ as the precision matrix and $S = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$ as the sample covariance matrix. Consider the L_1 regularized negative log-likelihood as the objective function:

$$\min_{\boldsymbol{\Theta}} - \log|\boldsymbol{\Theta}| + \mathrm{tr}(S\boldsymbol{\Theta}) + \lambda||\boldsymbol{\Theta}||_1, \tag{4.1}$$

where $||\boldsymbol{\Theta}||_1$ is the L_1 norm, the sum of the absolute values of the elements of $\boldsymbol{\Theta}$.

Denote the true precision matrix as $\boldsymbol{\Theta}_0$ and the true edge set as E_0. Note that $\boldsymbol{\Theta}_{0,(j,k)} = 0$ if $(x_j, x_k) \notin E_0$ since x_j is conditionally independent of x_k given all the other variables. Following the notation of Feng and Yu (2013), we define the set of correct models, $\mathcal{E}_0$ as the set of graph models with $E \supset E_0$. We denote an optimal edge

set as $E^* \in \mathcal{E}_0$ such that for all $E \in \mathcal{E}_0$ and $E \neq E^*$, $||E^*||_0 < ||E||_0$, where $||E||_0$ is the cardinality of the edge set or the effective number of parameters in the corresponding precision matrix, $\mathbf{\Theta}$. The cardinality of E_0 is denoted as d_0. The optimal model is the graph model having the smallest cardinality among the candidate models of interest without false negatives.

4.2.2 Existing Methods

We first review several general model selection principles and present their mathematical formulations to GGM. Akaike information criterion (AIC) and Bayesian information criterion (BIC) are popular principles for important variable selection in likelihood-based models (Akaike 1974). In our framework, both AIC and BIC restrict the graph models of interest to the models along the solution path, $\mathbf{\Theta}(\lambda)$, and select a sparse graph by penalizing the effective number of parameters (Liu et al. 2010). The formulation of AIC in GGM is

$$\hat{\mathbf{\Theta}}_{AIC}(\lambda) = \underset{\mathbf{\Theta}(\lambda)>0}{\operatorname{argmin}}\{-2\log|\mathbf{\Theta}(\lambda)| + 2\operatorname{tr}(S\mathbf{\Theta}(\lambda)) + 2||\mathbf{\Theta}(\lambda)||_0\}.$$

AIC is known to be competitive under the condition that the true graph structure is complicated. Another useful model selection tool, BIC, is known to be effective when the underlying model is low-dimensional. The formulation of BIC is similar to that of AIC. The weight, 2, on the effective number of parameters in AIC is replaced with the logarithm of the sample size:

$$\hat{\mathbf{\Theta}}_{BIC}(\lambda) = \underset{\mathbf{\Theta}(\lambda)>0}{\operatorname{argmin}}\{-2\log|\mathbf{\Theta}(\lambda)| + 2\operatorname{tr}(S\mathbf{\Theta}(\lambda)) + \log n \cdot ||\mathbf{\Theta}(\lambda)||_0\}.$$

Since BIC puts a heavier penalty on the model complexity, it tends to select a sparser graph than AIC. However, both tends to overfit in high-dimensional setting.

K-fold cross-validation is also frequently used to select the best model. We split the data into K subsets and use $K-1$ of them for training and one subset for validation. In GGM, we first obtain candidate graph models using the training dataset and calculate the validation errors, the negative log-likelihoods for the models using the validation dataset. We repeat the steps for different choices of training and validation data subsets and select the graph model with the smallest validation error. It is known that the cross-validation tends to select a denser graph under the low-dimensional true structure in high dimensional data (Liu et al. 2010).

Extended BIC is an adaptation of BIC for high-dimensional GGM proposed by Foygel and Drton (2010). It adds one term for more model complexity control as follows:

$$\hat{\mathbf{\Theta}}_{eBIC}(\lambda) = \underset{\mathbf{\Theta}(\lambda)>0}{\operatorname{argmin}}\{-2\log|\mathbf{\Theta}(\lambda)| + 2\mathrm{tr}(S\mathbf{\Theta}(\lambda)) + \log n \cdot ||\mathbf{\Theta}(\lambda)||_0 + 4||\mathbf{\Theta}(\lambda)||_0 \gamma \log p\},$$

where $\gamma \in [0, 1]$. The extra term is included so that the eBIC will tend to select a sparser model than BIC would. Note that eBIC with $\gamma = 0$ is equivalent to the classical BIC. We can theoretically determine γ on the convergence rate of p with n. However, this is infeasible in practice, thus we set γ as 0.5 or tune empirically.

StARS is a random subsampling method which uses a U-statistic to measure the stability of the model across the subsamples (Liu et al. 2010). We first draw N random subsamples $s_1, \cdots, s_N$ from the sample, $x_1, \cdots, x_n$. The subsample sizes are b, smaller than the sample size n. We next construct a graph using *glasso* for each λ based on each subsample. We denote the N graphs for λ as $\hat{E_1}^b(\lambda), \cdots, \hat{E}_N^b(\lambda)$. For every edge of each graph, we obtain an instability measure from a selection probability estimate across the subsamples. The selection probability estimate to an edge (x_s, x_t) is the

average of the edge existence across the subsamples:

$$\hat{\theta}^b_{st}(\lambda) = \frac{1}{N}\sum_{j=1}^{N} \psi^{\lambda}_{st}(S_j),$$

$$\text{where } \psi^{\lambda}_{st}(S_j) = \begin{cases} 1 & \text{if the algorithm puts an edge between } (s,t), \\ 0 & \text{otherwise.} \end{cases}$$

The instability measure for the edge is

$$\hat{\xi}^b_{st}(\lambda) = 2\hat{\theta}^b_{st}(\lambda)(1 - \hat{\theta}^b_{st}(\lambda)).$$

The measure is zero under the two extreme situations where the edge is selected or excluded for every subsample. For each graph, StARS obtain the total instability over all edges, which is the average of the instability measures over all edges, $\hat{\xi}^b_{st}$, $s, t = 1, \cdots, p$. We construct the monotone total instability by taking the supremum of the instabilities up to each λ. The optimal tuning parameter is the tuning parameter whose monotone instability is not larger than a predefined cut point value. Liu et al. (2010) set 0.05 as the default cut point value. As mentioned in Section 4.1, StARS was only shown to be sparsistent in an asymptotic sense. With a large sample size in fixed dimension, the chosen edge set includes all the true important edges but may also include some spurious edges. Another tuning step is required for the cut point value selection.

4.2.3 Consistent Validation for Edge Selection (CoVES) Method

We first illustrate the CCV procedure as in Feng and Yu (2013) for GLMs. Suppose we have n independently and identically distributed observations $(\boldsymbol{x}_i, y_i)$, $i = 1, \cdots, n$, where $\boldsymbol{x}_i$ is a p-dimensional predictor and y_i is the response. We assume the conditional

distribution y given $\boldsymbol{x}$ is an exponential family with a canonical link. Its density function is written as follows: $f(y;\boldsymbol{x},\boldsymbol{\beta}) = c(y,\pi)\exp[(y\theta - b(\theta))/a(\phi)]$, where $\theta = \boldsymbol{x}\boldsymbol{\beta}$ and $\phi \in (0,\infty)$ is the dispersion parameter. Here, $\boldsymbol{\beta}$ is the parameter of interest, and $\boldsymbol{\beta}_0$ is the true parameter, with $||\boldsymbol{\beta}_0||_0 = d_0 < n$, where $||\boldsymbol{\beta}||_0 = |\{j : \beta_j \neq 0\}|$. The log-likelihood can be written as follows based on an affine transformation:

$$l(\boldsymbol{y},\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[y_i\theta_i - b(\theta_i)].$$

CCV considers sparse estimation by minimizing a penalized negative log-likelihood function with a tuning parameter, λ:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}[-l(\boldsymbol{y},\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p} p(|\beta_j|)],$$

where $p_\lambda(\cdot)$ is the penalty function. Feng and Yu (2013) considered both convex and folded-concave penalties as the penalty function for CCV. The multi-stage CCV procedure is described in Algorithm 1.

Algorithm 1. *CCV Implementation (Feng and Yu 2013)*

1. Compute the solution path using the entire dataset. A sequence of solutions $\hat{\boldsymbol{\beta}}(\lambda)$ is generated as a function of the penalty level λ.

2. Randomly split the whole dataset into $\{(\mathbf{x}_i, y_i), i \in s\}$ (the validation set) and $\{(\mathbf{x}_i, y_i), i \in s^c\}$ (the construction set) r times. The sizes of the validation set and the construction set are n_v and n_c respectively. For each split $j = 1, \cdots, r$, compute the restricted MLE path according to the active set sequence generated in the previous step.

3. Average the negative log-likelihood over the r splits for each model (from Step 1), and choose the estimator in the model with the tuning parameter corresponding to the smallest loss function value.

4. Compute the restricted MLE for the selected model.

Our proposed method follows the steps of CCV. We first consider the edge structures from the entire solution path. For each structure, CoVES computes the empirical negative log-likelihood via repeated random subsampling validation. We finally select the edge structure having the smallest negative log-likelihood. In CoVES, it is of our interest to select important edges instead of significant variables, and its corresponding likelihood is based on multivariate Gaussian distribution. Detailed algorithms are illustrated as follows in Algorithm 2.

Algorithm 2. *CoVES Implementation*

1. Calculate the solution path of the precision matrices using the entire dataset. A sequence of solutions is generated with corresponding penalty level λ *from* (4.1)*. Along the path, a sequence of candidate sets of edges are determined based on the precision estimates,* $\hat{\mathbf{\Theta}}(\lambda)$*.*

$$\hat{E}(\lambda) = \{(x_j, x_k) : \hat{\mathbf{\Theta}}(\lambda)_{(j,k)} \neq 0\}$$

2. Randomly split the dataset into a validation set, s *(size* n_v*) and a construction set,* s^c *(size* n_c*)* r *times. For each split* $j = 1, \cdots, r$*, compute the restricted MLE path according to the active edge sequence generated in Step 1.*

3. Average the negative log-likelihood over the r *splits for each active edge set in Step 1, and choose the active edge set* $\hat{E}$ *with the smallest average validation error.*

4. Compute the restricted MLE with the selected edge set $\hat{E}$ *in Step 3.*

In the second step of the CoVES algorithm, we give a detailed description for each repetition given a set of edges, E. Let $E_d = \{(1,1), \cdots, (p,p)\}$ and E be one of the estimated graphs from the full solution path. We minimize an unpenalized negative log-likelihood function with zero constraints to the unselected edges, E^c and

the corresponding optimization problem is written as follows:

$$\tilde{\mathbf{\Theta}}_{s^c,E} = \underset{\mathbf{\Theta}>0}{\operatorname{argmin}}\{\log\,\det(\mathbf{\Theta}) - \operatorname{tr}(S_{s^c}\mathbf{\Theta})\}$$
$$\text{subject to } \mathbf{\Theta}_{ij} = 0, (i,j) \in E_d \backslash E,$$

where $S_{s^c} = \dfrac{1}{n_c}\sum_{i\in s^c}(x_i - \bar{x}_{s^c})^T(x_i - \bar{x}_{s^c})$ is the empirical covariance matrix from the construction sample and $\bar{x}_{s^c} = \dfrac{1}{n_c}\sum_{i\in s^c} x_i$ is the construction sample average. Note that all the repetitions have the common set of edges, E, but may give different valued precision matrix estimators, $\tilde{\mathbf{\Theta}}_{s^c,E}$.

Next, the validation set is used to obtain the empirical negative log-likelihood for the precision matrix estimator, $\tilde{\mathbf{\Theta}}_{s^c,E}$. The corresponding log-likelihood is from the multivariate Gaussian density and is written as $l_s(\tilde{\mathbf{\Theta}}_{s^c,E})$, where $l_s(\mathbf{\Theta}) = \dfrac{1}{2}(\log\,\det(\mathbf{\Theta}) - \operatorname{tr}(S_s\mathbf{\Theta}))$. In the log-likelihood, $S_s = \dfrac{1}{n_v}\sum_{i\in s}(x_i - \bar{x}_s)^T(x_i - \bar{x}_s)$ is the empirical covariance matrix from the validation sample, and $\bar{x}_s = \dfrac{1}{n_v}\sum_{i\in s} x_i$ is the validation sample average. The negative log-likelihood evaluates how well each set of edges fits with the validation set. Note that the expected loss function evaluated at E is the expectation of the negative log-likelihood with respect to a random selection of s, $\Gamma_E = -\mathbb{E}[l_s(\tilde{\mathbf{\Theta}}_{s^c,E})]$. It is called the risk function at $\tilde{\mathbf{\Theta}}_{s^c,E}$. We take the average of the empirical negative log-likelihood across the multiple r splits to estimate the risk function, which is denoted as $\hat{\Gamma}_E$ as follows:

$$\hat{\Gamma}_E = -\frac{1}{r}\sum_{s\in\mathcal{R}} l_s(\tilde{\mathbf{\Theta}}_{s^c,E}).$$

Note that we now have numerous empirical negative log-likelihoods corresponding to the candidate sets. We choose the set with the smallest empirical negative log-likelihood. This step determines the graph structure and the edge set estimate is denoted as $\hat{E}$. In the last step, we estimate the signals of the conditional dependency in the selected

graph using the complete dataset and the estimate is denoted as $\hat{\mathbf{\Theta}}_{\hat{E}}$.

4.3 Theoretical Properties

This section describes an asymptotic property of CoVES. Feng and Yu (2013) shows that CCV consistently selects the optimal GLM among the candidate GLMs with probability tending to one. Likewise, we conjecture that CoVES recovers the true set of edges with probability tending to one.

4.3.1 Preliminary Steps

We define the edge selection consistency so that the true set of edges, E_0, is selected in an asymptotic sense. Our investigation takes place under the condition of a growing sample size with fixed dimension. Note that our method does not not have the true set as a candidate if the preliminary penalization method does not contain the true set along the tuning parameter. In order to accommodate such cases, we alternatively define the optimal edge selection consistency so that the optimal set of edges, E^*, is selected under a growing sample size with fixed dimension. In other words, the selected edge set is exactly the same as the optimal set with probability tending to one:

$$\lim_{n\to\infty} P(\hat{E} = E^*) = 1. \tag{4.2}$$

By definition, the optimal model, E^*, is unique as long as there is only one model with size d_0 among the candidate models.

We can make some interesting remarks regarding the optimal edge selection consistency. If there are more than one model with the size of d_0, we consider the collection of the optimal edge sets, which is defined as $\{E \in \mathcal{E}_0 : ||E||_0 = d_0\}$ (Feng and Yu 2013). In this case, the optimal selection consistency implies that the edge set estimate is an

element of the optimal model set with probability tending to one. Next, we consider the situation that the true edge set E_0 lies on the solution path. In such case, we obtain a stronger theoretical property, which is the edge selection consistency since we have $E = E^*$.

Next we introduce some notations for random selection of subsamples. Denote the expectation with respect to the random selection of subsamples, $\mathcal{R}$, as $E_{\mathcal{R}}$ and the variance with respect to $\mathcal{R}$ as $V_{\mathcal{R}}$. Below are two likelihoods of Θ and $\tilde{\Theta}_E$:

$$l_s(\Theta) = \frac{n_v}{2}[\log\,\det(\Theta) - \mathrm{tr}(S_s\Theta)],$$
$$l_n(\tilde{\Theta}_E) = \frac{n}{2}[\log\,\det(\tilde{\Theta}_E) - \mathrm{tr}(S\tilde{\Theta}_E)].$$

Following the notations of Zhou et al. (2011), for any matrix $W = (w_{ij}) \in \mathbb{R}^p \times \mathbb{R}^p$, define the smallest eigenvalue of W as $\varphi_{min}(W)$ and the largest eigenvalue of W as $\varphi_{max}(W)$. Denote $||W||_F = \sqrt{\sum_i \sum_j w_{ij}^2}$ as the matrix Frobenius norm and $||W||_2 = \sqrt{\varphi_{max}(WW^T)}$ as L_2 norm. The vectorized W is denoted as vec$W \in \mathbb{R}^{p^2}$ and the Kronecker product is denoted as $\otimes$.

4.3.2 Asymptotic Results

Our conjecture is that CoVES satisfies the optimal edge selection consistency given a collection of candidate edges set, $\mathcal{E}$. In this section, we provide a sketch proof of this asymptotic property. Similar to Feng and Yu (2013), we assume that $n_c \to \infty$, $n_c/n \to 0$ as $n \to \infty$ and the number of the splits r satisfies $r^{-1}n_c^{-2}n^2 \to 0$. The collection, $\mathcal{E}$, is partitioned into two disjoint collections:

$$\mathcal{E}_0 = \{E : E \supset E_0\} \text{ and } \mathcal{E}_1 = \{E : E \not\supseteq E_0\}.$$

Each collection is called the collection of the correct edge sets and its complement respectively. We need to conduct a separate examination on each of them to obtain the theoretical property as per Shao (1993) and Feng and Yu (2013). The loss function for the edge set E is

$$\hat{\Gamma}_E = -\frac{1}{r}\sum_{s\in\mathcal{R}}\frac{1}{2}[\log\,\det(\tilde{\Theta}_{s^c}) - \mathrm{tr}(S_{s,E}\tilde{\Theta}_{s^c,E})],$$

where $\mathcal{R}$ is the collection of validation sets in different splits. Note that $\tilde{\Theta}_{s^c,E}$ is the restricted MLE on the model E using the construction set, s^c. Now, it is sufficient to show the following probabilistic arguments for the optimal edge selection consistency of (4.2):

$$P\{\exists E \in \mathcal{E}_0\backslash\{E^*\}, \hat{\Gamma}_{E^*} > \hat{\Gamma}_E\} \to 0$$
$$P\{\exists E \in \mathcal{E}_1, \hat{\Gamma}_{E^*} > \hat{\Gamma}_E\} \to 0.$$

We conjecture that the empirical loss function at E over the subsamples is written as the full likelihood evaluated at E. Hence, the difference between the empirical loss function at E and the empirical loss function at E^* over the subsamples is expected as

$$\hat{\Gamma}_E - \hat{\Gamma}_{E^*} = \frac{1}{n}\{l_n(\tilde{\Theta}_{E^*}) - l_n(\tilde{\Theta}_E)\} + O(\frac{1}{n_c}). \tag{4.3}$$

To show this, we take the expectation to the empirical loss function with respect to the random selection of validation sets, $\mathcal{R}$.

$$\begin{aligned}E_{\mathcal{R}}(\hat{\Gamma}_E) =& E_{\mathcal{R}}(\frac{1}{rn_v}\sum_{s\in\mathcal{R}} -l_s(\Theta)) + E_{\mathcal{R}}\{\frac{1}{rn_v}\sum_{s\in\mathcal{R}}[l_s(\Theta) - \frac{n_v}{2}\{\log\,\det(\tilde{\Theta}_E) - \mathrm{tr}(S_s\tilde{\Theta}_E)\}]\}\\ &+ E_{\mathcal{R}}\{\frac{1}{r}\sum_{s\in\mathcal{R}}[\frac{1}{2}\{\log\,\det(\tilde{\Theta}_E) - \mathrm{tr}(S_s\tilde{\Theta}_E)\} - \frac{1}{2}\{\log\,\det(\tilde{\Theta}_{s^c,E}) - \mathrm{tr}(S_s\tilde{\Theta}_{s^c,E})\}]\}\end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{n} l_n(\Theta) + \frac{1}{n}\{l_n(\Theta) - l_n(\tilde{\Theta}_E)\} \\
&\quad + \binom{n}{n_v}^{-1} \sum_{s \in \mathcal{AR}} \frac{1}{2}[\{\log\det(\tilde{\Theta}_E) - \log\det(\tilde{\Theta}_{s^c,E})\} - \{\mathrm{tr}(S_s\tilde{\Theta}_E) - \mathrm{tr}(S_s\tilde{\Theta}_{s^c,E}\}] \\
&\equiv -\frac{1}{n} l_n(\Theta) + A_{E1} + \binom{n}{n_v}^{-1} \sum_{s \in \mathcal{AR}} A_{E2,s}
\end{aligned}
$$

We expect that the second term, $A_{E1} = \frac{1}{n}\{l_n(\Theta) - l_n(\tilde{\Theta}_E)\}$ is the dominating term and the element of the third term, $A_{E2,s}$ can asymptotically be ignored.

4.4 Numerical Studies

In this section, we compare CoVES to other existing model selection criteria in terms of graph selection performance. The other criteria include classical model selection criteria such as AIC and BIC and high-dimensional undirected graph selection criteria such as extended BIC and StARS. See Section 4.2.2 for more details. We present the StARS results with two cut point values, 0.05 and 0.1 (*StARS 1*, *StARS 2*) and CoVES with different subsampling sizes: $\lceil n/2 \rceil$ and $\lceil 10 \cdot \sqrt{n} \rceil$(*CoVES 1*, *CoVES 2*).

The comparison criteria are true positivity rate (TPR) and false positivity rate (FPR), both of which are based on true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). TP is the number of true edges selected in the estimated graph and FP is the number of true edges not selected in the estimated graph. Similarly, TN is the number of the edges in the complement set of true edges selected in the estimated graph and FN is the number of the edges in the complement of true edges not selected in the estimated graph. TPR and FPR are calculated as follows:

$$TPR = TP/(TP + FN),\ FPR = FP/(FP + TN).$$

TPR indicates what percentage of the selected edges have true conditional dependence

and FPR indicates what percentage of the unselected edges have true conditional independence. The oracle procedure has zero negatives, that is no FP nor FN, which results in 1 TPR and 0 FPR. Thus, we expect a larger TPR and a small FPR for each method in practice.

We simulated 100 datasets of n i.i.d. p variate random samples from $N(\mathbf{0}, \mathbf{\Theta}_0)$. We consider $n = 200$ or $n = 400$ and $p = 10, 40, 50, 100$. In each scenario, a true precision matrix, $\mathbf{\Theta}_0$ is a determinant of a true graph pattern and is centered to have zero mean and variance one. We use a *glasso* package to obtain the entire solution path with the sample covariance matrix as an input. For computational simplicity, we pick 100 different equally spaced tuning parameters and obtain their corresponding edges sets. Asymmetric edge sets among them are excluded and the next procedure follows Algorithm 2. *R packages* such as *huge* and *glasso* are used to implement BIC, 5-fold cross-validation (5-fold CV), eBIC and StARS. The *huge* package also provides the visualization of the adjacency matrix, the graph pattern, the covariance matrix, and the empirical covariance matrix of the true graph structure (Zhao, Liu, Roeder, Lafferty, and Wasserman 2012).

4.4.1 Double Chain Graphs

We first investigate a double chain graph, where the vertex, x_j is conditionally dependent to $x_{j-2}, x_{j-1}, x_{j+1}$ and x_{j+2}, $j = 3, \cdot, n-2$. The rest of the vertices have the following conditional dependence pattern: x_1 is conditionally dependent with x_2 and x_3; x_2 is conditionally dependent with x_1, x_3, x_4; x_n is conditionally dependent with x_{n-1}, x_{n-2}; x_{n-1} is conditionally dependent with x_n, x_{n-2}, x_{n-3}. The true precision matrix, $\mathbf{\Theta}_0$ is tridiagonal, that is, a band matrix with five elements width. In the left panel of Figure 4.1, the adjacency matrix and the graph pattern of the true graph structure illustrates a simple double chain pattern example with 10 vertices. The white

fields of the adjacency matrix correspond to the nonzero off-diagonal true precision matrix. In the right panel of Figure 4.1, the covariance matrix is the inverse of the true precision matrix and the empirical matrix is the empirical covariance matrix estimate from the whole data. In our studies, we set the true precision matrix to have the following values based on the corresponding elements:

$$\Theta_{0,(i,j)} = \begin{cases} 1, & i = j \\ 0.6, & |i-j| = 1 \\ 0.3, & |i-j| = 2. \end{cases}$$

Figures 4.2-4.4 describe the graphs having this same double chain pattern with 40, 50, and 100 vertices respectively.

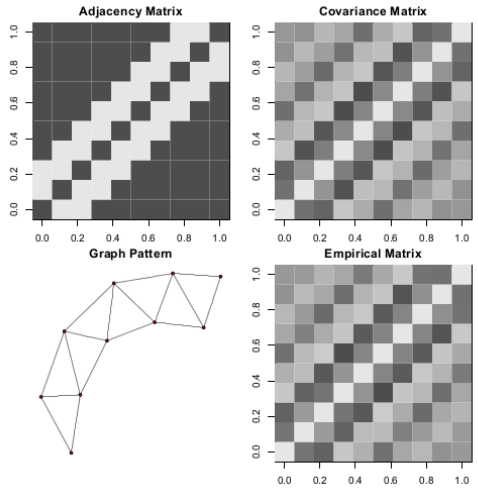


Figure 4.1: Double Chain Graph with $p = 10$

	n=200		n=400	
	TPR	FPR	TPR	FPR
CoVES 1	0.9906	0.3693	1	0.33
CoVES 2	0.99	0.3711	1	0.3264
5-fold CV	0.9959	0.4671	1	0.395
BIC	0.9959	0.4671	1	0.395
StARS 1	0.2276	0.0064	0.1953	0.0011
StARS 2	0.0782	0	0.0671	0
eBIC	0.9947	0.4529	1	0.3889

Table 4.1: Edge Selection Results for Double Chain Graph $p = 10$

We consider four scenarios from the low-dimensional case to the high-dimensional case with different numbers of variables in Tables 4.1-4.4. All the methods have an improvement in true edges selection across all scenarios as the sample size increases from $n = 200$ to $n = 400$. While the traditional model selection principles tend to

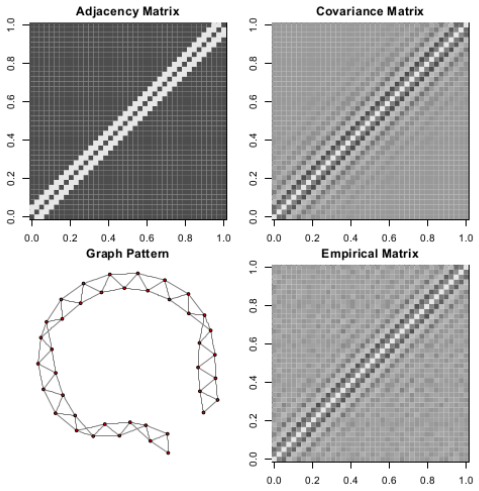


Figure 4.2: Double Chain Graph with $p = 40$

	n=200		n=400	
	TPR	FPR	TPR	FPR
CoVES 1	0.961	0.1306	0.9975	0.101
CoVES 2	0.9618	0.131	0.9981	0.1016
5-fold CV	0.9982	0.2872	1	0.2053
BIC	0.9982	0.2872	1	0.2053
StARS 1	0.7413	0.0677	0.8264	0.0586
StARS 2	0.6018	0.0555	0.437	0.0266
eBIC	0.9188	0.1182	0.9992	0.1805

Table 4.2: Edge Selection Results for Double Chain Graph $p = 40$

select a denser graph, the other methods such as eBIC, and StARS tend to select a sparser model. The underfitting issue may be inherent in the latter methods since they are developed for sparse model selection in high-dimensional data. eBIC has a smaller *TPR* and *FPR* than BIC since its penalization terms encourage a sparser model selection than that of BIC. Both StARS methods have different rates of *TP* and *FP* since they are sensitive to cut value points. This implies that the StARS cut point value should be tuned with care because such a framework only reformulates the direct tuning parameter selection problem into the indirect cut point value selection problem. However, CoVES changes the problem of a tuning parameter selection into the conventional model selection problem without a tuning parameter.

In a low-dimensional case of $p = 10$, both StARS select very sparse models, as evidenced by low TPR (0.1953 and 0.0671) and low FPR (0.0011 and 0) in Table 4.1. The cut point value should be smaller to select a denser graph. Both CoVES methods with different subsample sizes have similar performance. For $p = 40$ and $p = 50$ in Tables 4.2-4.3, the performance of CoVES are comparable with that of eBIC. Next, we consider a high-dimensional setting with $p = 100$ from Table 4.4. Note that the 5-fold

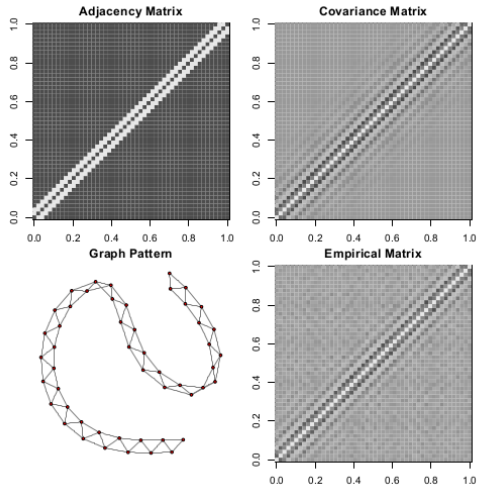


Figure 4.3: Double Chain Graph with $p = 50$

	n=200		n=400	
	TPR	FPR	TPR	FPR
CoVES 1	0.9545	0.1104	0.9973	0.085
CoVES 2	0.9539	0.1098	0.9973	0.0848
5-fold CV	0.9986	0.2735	0.9999	0.1922
BIC	0.9986	0.2735	0.9999	0.1922
StARS 1	0.7622	0.0569	0.8576	0.0477
StARS 2	0.6294	0.0458	0.6578	0.0431
eBIC	0.8574	0.0769	0.9987	0.1418

Table 4.3: Edge Selection Results for Double Chain Graph $p = 50$

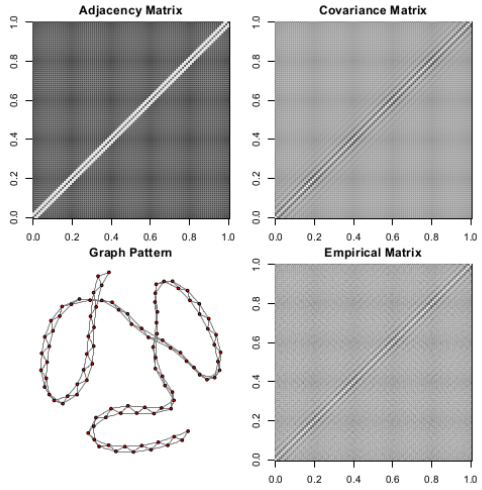


Figure 4.4: Double Chain Graph with $p = 100$

	n=200		n=400	
	TPR	FPR	TPR	FPR
CoVES 1	0.9312	0.0695	0.9966	0.0554
CoVES 2	0.9315	0.0699	0.9969	0.0557
5-fold CV	0.9988	0.2474	0.9999	0.167
BIC	0.9986	0.2444	0.9999	0.167
StARS 1	0.7948	0.0352	0.9066	0.0255
StARS 2	0.6784	0.0252	0.7651	0.022
eBIC	0.7306	0.0290	0.9901	0.0485

Table 4.4: Edge Selection Results for Double Chain Graph $p = 100$

cross-validation and BIC methods select a denser graph from high TPR and FPR and StARS and eBIC select a sparser graph from low TPR and FPR. CoVES lies in the second tier for TPR and FPR, and thus it can be viewed as a compromise between the two extremes.

4.4.2 Hub Graphs

Next, we consider a hub graph, where a vertex is conditionally dependent on multiple vertices but the vertices are only connected via the vertex. The central vertex is called a hub vertex. The setting is similar to the second example of Liu et al. (2010). Figures 4.5-4.8 show the adjacency matrices of the population precision matrices, the graph patterns, the population covariance matrices, and the sample covariance matrices. In this setting, the true precision matrix only has nonzero elements on L-shape from the hub vertex. We construct the true precision matrix as follows. Its rows and columns are partitioned into J equally-sized disjoint groups: $V_1 \cup V_2 \cup V_J = \{1, \cdots, p\}$, each group is associated with a pivotal row k. Let $|V_1| = 10$. We set $\Omega_{ik} = \Omega_{ki} = 0.5$ for $i \in V_k$ and $\Omega_{ik} = \Omega_{ki} = 0$ otherwise. The simple example is a graph with 1 hub vertex among 10 vertices from the left panel of Figure 4.5. The other three settings are illustrated at the left panel of Figure 4.6-Figure 4.8. Their graphs have 4, 5, 10 hub vertices among 40, 50, and 100 vertices respectively.

We report the results on edge selection from four different scenarios according to the number of vertices in Tables 4.5-4.8. Similar to the results from the double chain graphs in Section 4.4.1, the choice of cut point values in StARS has a significant impact on the edge selection performance and the performance of CoVES tends to be robust to the choice of the subsample size. Since the true negatives increase with the square of the number of vertices, FNR noticeably decreases with the increase in the number of vertices. Table 4.5 shows that only StARS performs poorly in selecting true edges

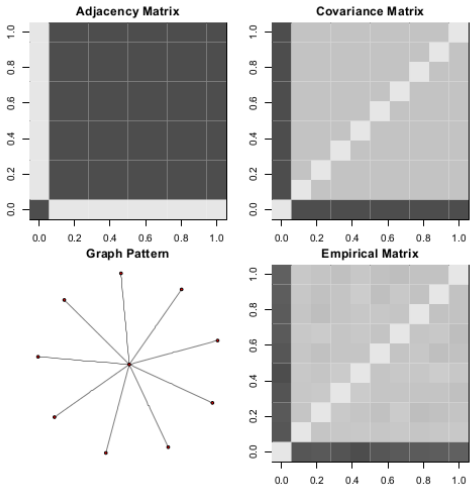


Figure 4.5: Hub Graph with $p = 10$

	n=200		n=400	
	TPR	FPR	TPR	FPR
CoVES 1	1	0.0294	1	0.0311
CoVES 2	1	0.0275	1	0.0292
5-fold CV	1	0.7572	1	0.8406
BIC	1	0.7572	1	0.8406
StARS 1	0.2933	0.0025	0.0333	0
StARS 2	0.0256	0	0.0022	0
eBIC	1	0.7572	1	0.8406

Table 4.5: Edge Selection Results for Hub Graph with $p = 10$

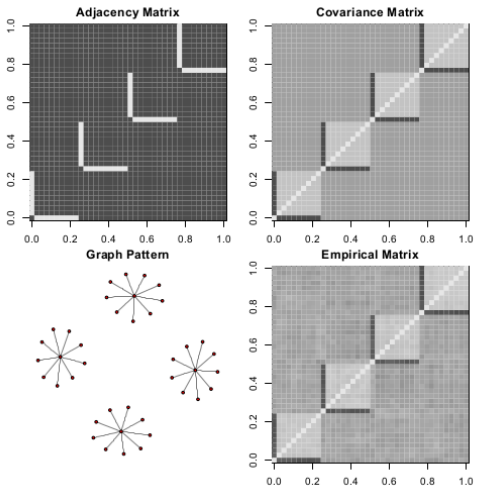


Figure 4.6: Hub Graph with $p = 40$

	n=200		n=400	
	TPR	FPR	TPR	FPR
CoVES 1	1	0.0048	1	0.0049
CoVES 2	1	0.0045	1	0.0049
5-fold CV	1	0.2608	1	0.2206
BIC	1	0.2602	1	0.2206
StARS 1	1	0.1446	1	0.0322
StARS 2	1	0.0146	1	0.0047
eBIC	1	0.2033	1	0.2136

Table 4.6: Edge Selection Results for Hub Graph with $p = 40$

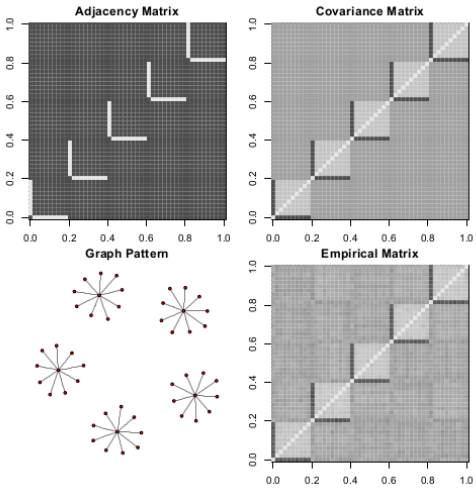


Figure 4.7: Hub Graph with $p = 50$

	n=200		n=400	
	TPR	FPR	TPR	FPR
CoVES 1	1	0.0038	1	0.0038
CoVES 2	1	0.0039	1	0.0039
5-fold CV	1	0.235	1	0.1879
BIC	1	0.2345	1	0.1879
StARS 1	1	0.1342	1	0.1173
StARS 2	1	0.0196	1	0.0073
eBIC	1	0.1598	1	0.1719

Table 4.7: Edge Selection Results for Hub Graph with $p = 50$

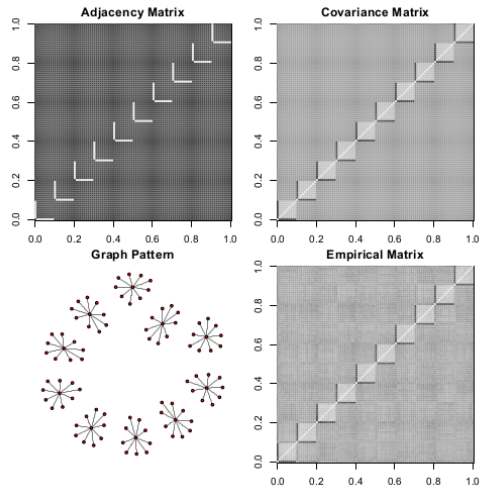


Figure 4.8: Hub Graph with $p = 100$

	n=200		n=400	
	TPR	FPR	TPR	FPR
CoVES 1	1	0.002	1	0.0016
CoVES 2	1	0.002	1	0.0017
5-fold CV	1	0.1805	1	0.126
BIC	1	0.1743	1	0.126
StARS 1	1	0.073	1	0.0704
StARS 2	1	0.0642	1	0.064
eBIC	1	0.0729	1	0.0828

Table 4.8: Edge Selection Results for Hub Graph with $p = 100$

and CoVES has the best performance for the smallest edge set selection among the sparsistent procedures. From Tables 4.6-4.8, all the methods successfully find the true edges, thus we focus on FPR for the performance comparison. CoVES outperforms other methods in terms of having a small FPR and StARS and eBIC show a comparable performance in the case of $p = 100$.

4.5 Discussion

In this chapter, we propose CoVES, a repeated subsampling method for edge selection in GGM. It is an extension of Monte Carlo cross-validation for linear models in Shao (1993) and CCV for GLM in Feng and Yu (2013). We conjecture that it can asymptotically select an optimal graph with the smallest cardinality among the graph structures including all the true edges. This will guarantee that the selected edge set is the same as the true edge set in large samples when the true edge set is contained in the solution path. For future research, it will be interesting to apply CoVES to the glioblastoma multiforme cancer dataset studied by the Cancer Genome Atlas Research Network (McLendon et al. 2008). From the gene expression data analysis, we intend to infer a gene regulatory network which may help to explain complex associations among genes.

REFERENCES

Akaike, H. (1974), "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, 19, 716–723.

Bondell, H. and Reich, B. (2007), "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR," *Biometrics*, 64, 115–123.

Bradic, J., Fan, J., and Wang, W. (2011), "Penalized composite quasi-likelihood for ultrahigh dimensional variable selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 325–349.

Breheny, P. and Huang, J. (2011), "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *The Annals of Applied Statistics*, 5, 232–253.

Breiman, L. (1995), "Better subset regression using the nonnegative garrote," *Technometrics*, 37, 373–384.

Buhl, S. L. (1993), "On the existence of maximum likelihood estimators for graphical Gaussian models," *Scandinavian Journal of Statistics*, 263–270.

Cai, T. and Cheng, S. (2004), "Semiparametric regression analysis for doubly censored data," *Biometrika*, 91, 277–290.

Carroll, R. and Ruppert, D. (1988), *Transformation and weighting in regression*, vol. 30, Chapman & Hall/CRC.

Cole, S. R., Hudgens, M. G., Tien, P. C., Anastos, K., Kingsley, L., Chmiel, J. S., and Jacobson, L. P. (2012), "Marginal structural models for case-cohort study designs to estimate the association of antiretroviral therapy initiation with incident AIDS or death," *American Journal of Epidemiology*, 175, 381–390.

Cox, D. (1972), "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 187–220.

— (2001), "Some remarks on likelihood factorization," *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, 165–172.

Davis, R., Knight, K., and Liu, J. (1992), "M-estimation for autoregressions with infinite variance," *Stochastic Processes and their Applications*, 40, 145–180.

Dempster, A. P. (1972), "Covariance selections," *Biometrics*, 157–175.

Donoho, D. L. and Johnstone, J. M. (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 425–455.

Drton, M. and Perlman, M. D. (2007), "Multiple testing and error control in Gaussian graphical model selection," *Statistical Science*, 22, 430–449.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *The Annals of Statistics*, 32, 407–499.

Fan, J. and Gijbels, I. (1996), *Local polynomial modelling and its applications*, vol. 66, Chapman & Hall/CRC.

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

— (2002), "Variable selection for Cox's proportional hazards model and frailty model," *The Annals of Statistics*, 30, 74–99.

Fan, J., Xue, L., Zou, H., et al. (2014), "Strong oracle optimality of folded concave penalized estimation," *The Annals of Statistics*, 42, 819–849.

Feng, Y. and Yu, Y. (2013), "Consistent cross-validation for tuning parameter selection in high-dimensional variable selection," *arXiv preprint arXiv:1308.5390*, Manuscript.

Foygel, R. and Drton, M. (2010), "Extended Bayesian information criteria for Gaussian graphical models." in *Advances in Neural Information Processing Systems*, pp. 604–612.

Friedman, J., Hastie, T., and Tibshirani, R. (2001), *The elements of statistical learning*, vol. 1, Springer Series in Statistics.

— (2008), "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432–441.

— (2010), "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, 33, 1–22.

Hastie, T., Tibshirani, R., and Friedman, J. J. H. (2001), *The Elements of Statistical Learning*, vol. 1, Springer New York.

He, X. and Hu, F. (2002), "Markov chain marginal bootstrap," *Journal of the American Statistical Association*, 97, 783–795.

He, X., Ng, P., and Portnoy, S. (1998), "Bivariate quantile smoothing splines," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 537–550.

Hurvich, C. and Tsai, C. (1990), "Model selection for least absolute deviations regression in small samples," *Statistics and Probability Letters*, 9, 259–265.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., Rinaldo, C. R., et al. (1987), "The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants," *American Journal of Epidemiology*, 126, 310–318.

Kim, Y., Kim, B., and Jang, W. (2010), "Asymptotic properties of the maximum likelihood estimator for the proportional hazards model with doubly censored data," *Journal of Multivariate Analysis*, 101, 1339–1351.

Kim, Y., Kim, J., and Jang, W. (2013), "An EM algorithm for the proportional hazards model with doubly censored data," *Computational Statistics and Data Analysis*, 41–51.

Kocherginsky, M., He, X., and Mu, Y. (2005), "Practical confidence intervals for regression quantiles," *Journal of Computational and Graphical Statistics*, 14, 41–55.

Koenker, R. and Bassett, G. (1978), "Regression quantiles," *Econometrica: journal of the Econometric Society*, 33–50.

Lauritzen, S. L. (1996), *Graphical models*, Oxford University Press.

Lee, W. (2013), "Some statistical learning methods for multiple high dimensional datasets," Ph.D. thesis, University of North Carolina at Chapel Hill.

Li, Y., Liu, Y., and Zhu, J. (2007), "Quantile regression in reproducing kernel Hilbert spaces," *Journal of the American Statistical Association*, 102, 255–268.

Liu, H., Roeder, K., and Wasserman, L. (2010), "Stability approach to regularization selection (stars) for high dimensional graphical models," in *Advances in Neural Information Processing Systems*, pp. 1432–1440.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., Olson, J. J., Mikkelsen, T., Lehman, N., Aldape, K., et al. (2008), "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, 455, 1061–1068.

Meinshausen, N. and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, 34, 1436–1462.

Murphy, S. and Van der Vaart, A. (2000), "On profile likelihood," *Journal of the American Statistical Association*, 95, 449–465.

Newey, W. and Powell, J. (1987), "Asymmetric least squares estimation and testing," *Econometrica: Journal of the Econometric Society*, 819–847.

Park, M. and Hastie, T. (2007), "L1-regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 659–677.

Pollard, D. (1991), "Asymptotics for least absolute deviation regression estimators," *Econometric Theory*, 7, 186–199.

Rocha, G., Wang, X., and Yu, B. (2009), "Asymptotic distribution and sparsistency for l1-penalized parametric M-estimators with applications to linear SVM and logistic regression," *arXiv preprint arXiv:0908.1940*.

Rossouw, J., Du Plessis, J., Benadé, A., Jordaan, P., Kotze, J., Jooste, P., and Ferreira, J. (1983), "Coronary risk factor screening in three rural communities. The CORIS baseline study." *The South African Medical Journal*, 64, 430–436.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Shao, J. (1993), "Linear model selection by cross-validation," *Journal of the American statistical Association*, 88, 486–494.

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

— (1997), "The lasso method for variable selection in the Cox model," *Statistics in Medicine*, 16, 385–395.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2004), "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 91–108.

Van der Vaart, A. (2000), *Asymptotic statistics*, Cambridge University Press.

Wang, H. and Leng, C. (2007), "Unified LASSO estimation by least squares approximation," *Journal of the American Statistical Association*, 102, 1039–1048.

Wang, H., Li, G., and Jiang, G. (2007a), "Robust regression shrinkage and consistent variable selection through the LAD-Lasso," *Journal of Business and Economic Statistics*, 25, 347–355.

Wang, H., Li, R., and Tsai, C.-L. (2007b), "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, 94, 553–568.

Wu, Y. and Liu, Y. (2009), "Variable selection in quantile regression," *Statistica Sinica*, 19, 801.

Yang, Y. (2005), "Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation," *Biometrika*, 92, 937–950.

Yuan, M. and Lin, Y. (2007a), "Model selection and estimation in the Gaussian graphical models," *Biometrika*, 94, 19–35.

— (2007b), "On the non-negative garrotte estimator," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 143–161.

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

Zhang, H. and Lu, W. (2007), "Adaptive Lasso for Cox's proportional hazards model," *Biometrika*, 94, 691–703.

Zhang, Y., Li, R., and Tsai, C. (2010), "Regularization parameter selections via generalized information criterion," *Journal of the American Statistical Association*, 105, 312–323.

Zhao, P. and Yu, B. (2007), "On model selection consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012), "The huge package for high-dimensional undirected graph estimation in r," *Journal of Machine Learning Research*, 13, 1059–1062.

Zhou, S., Rütimann, P., Xu, M., and Bühlmann, P. (2011), "High-dimensional covariance estimation based on Gaussian graphical models," *Journal of Machine Learning Research*, 12, 2975–3026.

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

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

Zou, H., Hastie, T., and Tibshirani, R. (2007), "On the degrees of freedom of the lasso," *The Annals of Statistics*, 35, 2173–2192.

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models," *The Annals of Statistics*, 36, 1509–1533.

Zou, H. and Yuan, M. (2008), "Composite quantile regression and the oracle model selection theory," *The Annals of Statistics*, 36, 1108–1126.