Two-Stage Penalized Least Squares Method for Constructing Large Systems of Structural Equations

By:

Chen Chen, Min Zhang, and Dabao Zhang *

Technical Report #15-03

*This work was partially supported by NSF CAREER award IIS-0844945 and the Cancer Care Engineering project at the Oncological Science Center of Purdue University.

Department of Statistics
Purdue University

November 2, 2015

Two-Stage Penalized Least Squares Method for Constructing Large Systems of Structural Equations

Chen Chen, Min Zhang, and Dabao Zhang *
Department of Statistics, Purdue University, West Lafayette, IN 47906

October 30, 2015

Abstract

Linear systems of structural equations have been recently investigated to reveal the structures of genome-wide gene interactions in biological systems. However, building such a system usually involves a huge number of endogenous variables and even more exogenous variables, and hence demands a powerful statistical method which limits memory consumption and avoids intensive computation. We propose a two-stage penalized least squares method to build large systems of structural equations. Fitting one linear model for each endogenous variable at each stage, the method employs the L_2 penalty at the first stage to obtain consistent estimation of a set of well-defined surrogate variables, and the L_1 penalty at the second stage to consistently select regulatory endogenous variables among massive candidates. Without fitting a full information model, the method is computationally fast and allows for parallel implementation. The resultant estimates of the regulatory effects enjoy the oracle properties, that is, they perform as well as if the true regulatory endogenous variables were known. We also demonstrated the effectiveness of the method by conducting simulation studies, showing its improvements over other methods. Our method was applied to construct a yeast gene regulatory network with a genetical genomics data.

*This work was partially supported by NSF CAREER award IIS-0844945 and the Cancer Care Engineering project at the Oncological Science Center of Purdue University.

Keywords: Graphical model; High-dimensional data; Reciprocal graphical model; Simultaneous equation model; Structural equation model

1 INTRODUCTION

We consider a linear system with p endogenous and q exogenous variables. With a sample of n observations from this system, we denote by $\mathbf{Y}_{n\times p} = (\mathbf{Y}_1, \cdots, \mathbf{Y}_p)$ and $\mathbf{X}_{n\times q} = (\mathbf{X}_1, \cdots, \mathbf{X}_q)$ the observed values of endogenous and exogenous variables, respectively. The interactions between endogenous variables and direct causal effects by exogenous variables can be described by a system of structural equations,

$$\mathbf{Y} = \mathbf{Y}\boldsymbol{\Gamma} + \mathbf{X}\boldsymbol{\Psi} + \boldsymbol{\epsilon}, \tag{1}$$

where the $p \times p$ matrix $\boldsymbol{\Gamma}$ has zero diagonal elements and contains regulatory effects, the $q \times p$ matrix $\boldsymbol{\Psi}$ contains causal effects, and $\boldsymbol{\epsilon}$ is an $n \times p$ matrix of error terms. We assume that $\mathbf{X}$ and $\boldsymbol{\epsilon}$ are independent of each other, and each component of $\boldsymbol{\epsilon}$ is independently distributed as normal with zero mean while rows of $\boldsymbol{\epsilon}$ are identically distributed.

With gene expression levels and genotypic values as endogenous and exogenous variables, respectively, model (1) has been used to represent a gene regulatory network with each equation modeling the regulatory effects of other genes and causal effects of cis-eQTL (i.e., expression quantitative trait loci located within the region of their target gene) on a given gene, see Xiong *et al.* (2004), Liu *et al.* (2008), Logsdon and Mezey (2010), and Cai *et al.* (2013), among others. Genetical genomics experiments (Jansen and Nap 2001) have been widely undertaken to obtain genome-wide gene expressions and genotypic values (Schadt *et al.* 2003). However, fitting a system of structural equations in (1) to genetical genomics data for the purpose of revealing a whole-genome gene regulatory network is still hindered by lack of an effective statistical method which addresses issues brought by large numbers of endogenous and exogenous variables.

Several efforts have been put to construct the system (1) with genetic genomics data. Xiong *et al.* (2004) proposed to use a genetic algorithm to search for genetic networks which minimize the Akaike information criterion (AIC; Akaike 1974), and Liu *et al.* (2008) instead proposed to minimize the Bayesian information criterion (BIC; Schwartz 1978) and its modification (Broman and Speed 2002) for the optimal genetic networks. Both AIC and BIC are applicable to inferring networks for only a small number of endogenous variables. For a large system with many endogenous and exogenous variables, Cai *et al.* (2013) proposed to maximize a penalized likelihood to construct a sparse system. However, it is computationally prohibitive to fit a large system based on the likelihood function of the complete model. Logsdon and Mezey (2010) instead proposed to apply the adaptive lasso (Zou 2006) to fitting each structural equation separately, then recover the network relying on additional assumption on unique exogenous variables. However, Cai *et al.* (2013) demonstrated its inferior performance via simulation studies, which is consistent with our conclusion.

Instead of the full information model specified in (1), we here seek to establish the large system via constructing a large number of limited information models, each for one endogenous variable (Schmidt 1976). For example, when k-th endogenous variable is concerned, we can focus on the k-th structural equation in (1) which models the regulatory effects of other endogenous variables and direct causal effects of exogenous variables, however, the system structures contained in other structural equations are skipped, leading to the following limited-information model,

$$\begin{cases} \mathbf{Y}_k = \mathbf{Y}_{-k}\boldsymbol{\gamma}_k + \mathbf{X}\boldsymbol{\psi}_k + \boldsymbol{\epsilon}_k, \\ \mathbf{Y}_{-k} = \mathbf{X}\boldsymbol{\pi}_{-k} + \boldsymbol{\xi}_{-k}. \end{cases} \tag{2}$$

Here $\mathbf{Y}_{-k}$ refers to $\mathbf{Y}$ excluding the k-th column, $\boldsymbol{\gamma}_k$ refers to the k-th column of $\boldsymbol{\Gamma}$ excluding the diagonal zero, and $\boldsymbol{\psi}_k$ and $\boldsymbol{\epsilon}_k$ refer to the k-th columns of $\boldsymbol{\Psi}$ and $\boldsymbol{\epsilon}$ respectively. The second part of the model (2) is from the following reduced model by excluding the k-th equation, with $\boldsymbol{\pi} = \boldsymbol{\Psi}(\mathbf{I} - \boldsymbol{\Gamma})^{-1}$ and $\boldsymbol{\xi} = \boldsymbol{\epsilon}(\mathbf{I} - \boldsymbol{\Gamma})^{-1}$,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\pi} + \boldsymbol{\xi}. \tag{3}$$

In a classical low-dimensional setting, it is well known that a two-stage least squares (2SLS) method can produce consistent estimates of the parameters when the system is identifiable (Theil 1953; Basmann 1957). However, as in a typical genetical genomics experiment, we here are interested in constructing a large system with the number of endogenous variables p much larger than the sample size n. Such a high-dimensional and small sample size data set makes it infeasible to directly apply 2SLS method. Indeed, $p \gg n$ results in perfect fits of reduced form equations at the first stage, which implies to regress against the observed values of endogenous variables at the second stage and therefore obtain ordinary least squares estimates of the parameters. It is well known that such ordinary least squares estimates are inconsistent. Furthermore, constructing a large system demands, at the second stage, selecting regulatory endogenous variables among massive candidates, i.e., variable selection in fitting high-dimensional linear models.

Here we propose a two-stage penalized least squares (2SPLS) method to address the challenges in establishing system (1) in the case $p \gg n$. The method fits one regularized linear model for each endogenous variable at each stage. At the first stage, the L_2 penalty is employed to obtain consistent estimates of a set of well-defined surrogate variables which allow to separately investigate individual structural models and consistently estimate all regulatory effects for each endogenous variable. At the second stage, each endogenous variable is regressed against the estimates of surrogate variables, and the L_1 penalty is employed to identify regulatory variables among massive candidates. The use of regularization techniques helps avoid overfitting at the first stage and allows to exploit sparse structure of the system at the second stage. We show that the resultant estimates of regulatory effects enjoy the oracle properties.

The proposed method addresses three challenging issues in constructing a large system of structural equations, i.e., memory capacity, computational time, and statistical power. First, the limited information models are employed to develop the algorithm so as to avoid managing the full information models which may consist of many subnetworks and involve a massive number of endogenous variables. Second, allowing to fit one linear model for each endogenous variable at each stage makes the algorithm computationally fast. It also makes it feasible to parallel the large number of model fittings at each stage. Third, the oracle properties of the resultant estimates show that the proposed method can achieve optimal power in identifying and estimating regulatory effects. Furthermore, the efficient computation makes it feasible to use the bootstrap method to evaluate the significance of regulatory effects for small data sets.

The rest of this paper is organized as follows. We first state an identifiable model in the next section. Provided in Section 3 is a new view on the classical 2SLS method, which motivates our development of the 2SPLS method in Section 4. In Section 5, we show that the estimates from 2SPLS have the oracle properties with the proof included in the Appendix. Simulation studies are carried out in Section 6 to evaluate the performance of 2SPLS. An application to a real data set to infer a yeast gene regulatory network is presented in Section 7. We conclude this paper with a discussion in Section 8.

2 THE IDENTIFIABLE MODEL

We follow the practice of constructing system (1) in analyzing genetic genomics data, and assume that each endogenous variable is affected by a unique set of exogenous variables. That is, the structural equation in (2) has known zero elements of $\boldsymbol{\psi}_k$. Explicitly, we use $\mathcal{S}_k$ to denote the set of row indices of known nonzero elements in $\boldsymbol{\psi}_k$. Then we have known sets

$\mathcal{S}_k, k = 1, 2, \cdots, p$, which dissect the set $\{1, 2, \cdots, q\}$. We explicitly state this assumption in the below.

Assumption A. $\mathcal{S}_k \neq \emptyset$ for $k = 1, \cdots, p$, but $\mathcal{S}_j \cap \mathcal{S}_k = \emptyset$ as long as $j \neq k$.

The above assumption indeed satisfies the rank condition (Schmidt 1976), which is a sufficient condition for model identification. Since each $\boldsymbol{\psi}_k$ has a set of known zero components, hereafter we ignore them and simply rewrite the structural equation in model (2) as,

$$\mathbf{Y}_k = \mathbf{Y}_{-k}\boldsymbol{\gamma}_k + \mathbf{X}_{\mathcal{S}_k}\boldsymbol{\psi}_k + \boldsymbol{\epsilon}_k, \qquad \boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \sigma_k^2\mathbf{I}_n). \tag{4}$$

3 NEW VIEW OF THE TWO-STAGE LEAST SQUARES METHOD

Because $\mathbf{Y}_{-k}$ and $\boldsymbol{\epsilon}_k$ are correlated, fitting solely model (4) results in biased estimates of $\boldsymbol{\gamma}_k$ and $\boldsymbol{\psi}_k$. However, we notice that the following two sets of variables are independent,

$$\begin{cases} \mathbf{Z}_{-k} = E[\mathbf{Y}_{-k}|\mathbf{X}] = \mathbf{X}\boldsymbol{\pi}_{-k}, \\ \boldsymbol{\varepsilon}_k = \boldsymbol{\epsilon}_k + \boldsymbol{\xi}_{-k}\boldsymbol{\gamma}_k. \end{cases}$$

Consequently, consistent estimates of $\boldsymbol{\gamma}_k$ and $\boldsymbol{\psi}_k$ can be obtained by applying least squares method to the following model,

$$\mathbf{Y}_k = \mathbf{Z}_{-k}\boldsymbol{\gamma}_k + \mathbf{X}_{\mathcal{S}_k}\boldsymbol{\psi}_k + \boldsymbol{\varepsilon}_k. \tag{5}$$

When regulatory effects are considered, $\{\mathbf{Z}_j = E[\mathbf{Y}_j|\mathbf{X}] = \mathbf{X}\boldsymbol{\pi}_j : j = 1, 2, \cdots, p\}$ serves as a set of surrogate variables which can help estimate both $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$ in model (1).

In practice, $\mathbf{Z}_{-k}$ is unknown as it involves unknown $\boldsymbol{\pi}_{-k}$. Suppose we instead have a consistent estimate $\hat{\boldsymbol{\pi}}_{-k}$ of $\boldsymbol{\pi}_{-k}$, i.e.,

$$\hat{\boldsymbol{\pi}}_{-k} = \boldsymbol{\pi}_{-k} + o_P(1).$$

Let $\hat{\mathbf{Z}}_{-k} = \mathbf{X}\hat{\boldsymbol{\pi}}_{-k}$, and further take the following assumption.

Assumption B. $n^{-1}\mathbf{X}^T\mathbf{X} \to \mathbf{C}$, where $\mathbf{C}$ is a positive definite matrix.

It is easy to see that

$$\begin{cases} \frac{1}{n}\hat{\mathbf{Z}}_{-k}^T\hat{\mathbf{Z}}_{-k} = \frac{1}{n}\mathbf{Z}_{-k}^T\mathbf{Z}_{-k} + o_P(1), \\ \frac{1}{n}\hat{\mathbf{Z}}_{-k}^T\mathbf{X}_{\mathcal{S}_k} = \frac{1}{n}\mathbf{Z}_{-k}^T\mathbf{X}_{\mathcal{S}_k} + o_P(1), \\ \frac{1}{n}\hat{\mathbf{Z}}_{-k}^T\mathbf{Y}_k = \frac{1}{n}\mathbf{Z}_{-k}^T\mathbf{Y}_k + o_P(1). \end{cases} \tag{6}$$

When replacing $\mathbf{Z}_{-k}$ with $\hat{\mathbf{Z}}_{-k}$ in model (5), we obtain the following least squares estimators of $\boldsymbol{\gamma}_k$ and $\boldsymbol{\psi}_k$,

$$\begin{pmatrix} \hat{\boldsymbol{\gamma}}_k \\ \hat{\boldsymbol{\psi}}_k \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\hat{\mathbf{Z}}_{-k}^T\hat{\mathbf{Z}}_{-k} & \frac{1}{n}\hat{\mathbf{Z}}_{-k}^T\mathbf{X}_{\mathcal{S}_k} \\ \frac{1}{n}\mathbf{X}_{\mathcal{S}_k}^T\hat{\mathbf{Z}}_{-k} & \frac{1}{n}\mathbf{X}_{\mathcal{S}_k}^T\mathbf{X}_{\mathcal{S}_k} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{n}\hat{\mathbf{Z}}_{-k}^T\mathbf{Y}_k \\ \frac{1}{n}\mathbf{X}_{\mathcal{S}_k}^T\mathbf{Y}_k \end{pmatrix}.$$

Certainly, the properties in (6) imply that the above estimators approach to the least squares estimators of fitting model (5), which are also consistent.

Theorem 1 *Suppose Assumptions A and B are satisfied for the system (1) with fixed $p \ll n$ and $q \ll n$. When there exists a consistent estimator $\hat{\boldsymbol{\pi}}_{-k}$ of $\boldsymbol{\pi}_{-k}$, the ordinary least squares estimators of $(\boldsymbol{\gamma}_k, \boldsymbol{\psi}_k)$ obtained by regressing $\mathbf{Y}_k$ against $(\mathbf{X}\hat{\boldsymbol{\pi}}_{-k}, \mathbf{X}_{\mathcal{S}_k})$ are also consistent.*

When a $\sqrt{n}$-consistent least squares estimator of $\boldsymbol{\pi}_j$ is obtained by fitting each equation in (3) for $j = 1, \cdots, p$, the resultant estimators of $\boldsymbol{\gamma}_k$ and $\boldsymbol{\psi}_k$ are exactly the 2SLS estimators by Theil (1953) and Basmann (1957). In the following, we consider to construct the system (1) in the case that $p \gg n$. Such a high-dimensional and small sample size data set makes it infeasible to directly apply the 2SLS method.

4 THE TWO-STAGE PENALIZED LEAST SQUARES METHOD

4.1 The Method

To construct the limited-information model (2), we can obtain consistent estimates of the surrogate variables by fitting high-dimensional linear models, and then conduct a high-dimensional variable selection following our view on model (5). Hence we propose a two-stage penalized least squares (2SPLS) procedure to construct each model in (2) so as to establish the large system (1).

At the first stage, we use the ridge regression to fit each reduced-form model in (3) to obtain consistent estimates of the surrogate variables. That is, for each $j = 1, 2, \cdots, p$, we obtain the ridge regression estimator of $\boldsymbol{\pi}_j$ by minimizing the following penalized sum of squares

$$\|\mathbf{Y}_j - \mathbf{X}\boldsymbol{\pi}_j\|^2 + \tau_j||\boldsymbol{\pi}_j||^2, \tag{7}$$

where $\tau_j > 0$ is a tuning parameter that controls the strength of the penalty. The solution to the minimization problem is $\hat{\boldsymbol{\pi}}_j = (\mathbf{X}^T\mathbf{X} + \tau_j\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}_j$, which leads to a consistent estimate of $\mathbf{Z}_j$,

$$\hat{\mathbf{Z}}_j = \mathbf{P}_{\tau_j}\mathbf{Y}_j,$$

where $\mathbf{P}_{\tau_j} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \tau_j\mathbf{I})^{-1}\mathbf{X}^T$. With a proper choice of τ_j, ridge regression has very good prediction performance as shown in the next section.

At the second stage, we replace $\mathbf{Z}_{-k}$ with $\hat{\mathbf{Z}}_{-k}$ in model (5) to derive estimates of $\boldsymbol{\gamma}_k$ and $\boldsymbol{\psi}_k$. Specifically, we minimize the following penalized error squares to obtain estimates of $\boldsymbol{\gamma}_k$ and $\boldsymbol{\psi}_k$,

$$\frac{1}{2}\|\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\boldsymbol{\gamma}_k - \mathbf{X}_{\mathcal{S}_k}\boldsymbol{\psi}_k\|^2 + \lambda_n \boldsymbol{\omega}_k^T|\boldsymbol{\gamma}_k|, \tag{8}$$

where $|\boldsymbol{\gamma}_k|$ implies to componentwisely take absolute values of $\boldsymbol{\gamma}_k$, $\boldsymbol{\omega}_k$ is a known weight vector, and $\lambda_n > 0$ is a tuning parameter.

Minimizing for $\boldsymbol{\psi}_k$ in (8) leads to

$$\hat{\boldsymbol{\psi}}_k = (\mathbf{X}_{\mathcal{S}_k}^T\mathbf{X}_{\mathcal{S}_k})^{-1}\mathbf{X}_{\mathcal{S}_k}^T(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\boldsymbol{\gamma}_k),$$

where $\mathbf{X}_{\mathcal{S}_k}$ is usually of low dimension, and the above least squares estimator of $\boldsymbol{\psi}_k$ is easy to obtain.

Plugging $\hat{\boldsymbol{\psi}}_k$ into (8), we can solve the following minimization problem to obtain an estimate of $\boldsymbol{\gamma}_k$,

$$\hat{\gamma}_k = \arg\min_{\gamma_k}\left\{\frac{1}{2}(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\boldsymbol{\gamma}_k)^T\mathbf{H}_k(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\boldsymbol{\gamma}_k) + \lambda_n\boldsymbol{\omega}_k^T|\boldsymbol{\gamma}_k|\right\}. \tag{9}$$

This is equivalent to a variable selection problem in regressing $\mathbf{H}_k\mathbf{Y}_k$ against high-dimensional $\mathbf{H}_k\hat{\mathbf{Z}}_{-k}$. We will resort to adaptive lasso to select nonzero components of $\boldsymbol{\gamma}_k$ and estimate them. Specifically, picking up a $\delta > 0$ and obtaining $\tilde{\boldsymbol{\gamma}}_k$ as a $\sqrt{n}$-consistent estimate of $\boldsymbol{\gamma}_k$, we calculate the weight vector $\boldsymbol{\omega}_k$ with components inversely proportional to components of $|\tilde{\boldsymbol{\gamma}}_k|^\delta$. The above minimization problem (9) is a convex optimization problem which is computationally efficient.

4.2 Tuning Parameter Selection

In this method, we need to select tuning parameters at each stage. At the first stage, we propose to choose each τ_j in (7) by the method of generalized cross-validation (GCV; Golub *et al.* 1979), that is,

$$\tau_j = \arg\min_{\tau>0} G_j(\tau) = \arg\min_{\tau>0} \frac{(\mathbf{Y}_j - \mathbf{P}_\tau \mathbf{Y}_j)^T (\mathbf{Y}_j - \mathbf{P}_\tau \mathbf{Y}_j)}{(n - \mathrm{tr}\{\mathbf{P}_\tau\})^2}.$$

It is a rotation-invariant version of ordinary cross-validation, and leads to an approximately optimal estimate of the surrogate variable $\mathbf{Z}_j$. At the second stage, the tuning parameter λ_n in (9) is obtained via K-fold cross validation.

5 THEORETICAL PROPERTIES

As an extension of the classical 2SLS method to high dimensions, the proposed 2SPLS method also has some good theoretical properties. In this section, we will show that the 2SPLS estimates enjoy the oracle properties. As the second-stage estimation replies on the ridge estimates $\hat{\mathbf{Z}}_{-k}$ obtained from the first stage, we first discuss the theoretical properties on $\hat{\mathbf{Z}}_{-k}$, which provide guarantee for the oracle properties of our proposed estimates.

As aforementioned, each τ_j in (7) is obtained by the method of generalized cross-validation. Interestingly, as stated by Golub *et al.* (1979), τ_j obtained by GCV is closely related to the one minimizing

$$T_j(\tau) = (\mathbf{Z}_j - \mathbf{P}_\tau \mathbf{Y}_j)^T (\mathbf{Z}_j - \mathbf{P}_\tau \mathbf{Y}_j).$$

Indeed, the following result follows Theorem 2 of Golub *et al.* (1979).

Theorem 2 *Suppose that all components of $\boldsymbol{\pi}_j$ are i.i.d. with mean zero and variance $\sigma^2_{\boldsymbol{\pi}}$, then*

$$\arg\min_{\tau>0} E\left[E[G_j(\tau)|\boldsymbol{\pi}_j]\right] = \arg\min_{\tau>0} E\left[E[T_j(\tau)|\boldsymbol{\pi}_j]\right] = \frac{\sigma^2_{\boldsymbol{\xi}_j}}{\sigma^2_{\boldsymbol{\pi}}},$$

where $\sigma^2_{\boldsymbol{\xi}_j}$ is the variance component of $\boldsymbol{\xi}_j$ in model (2).

This theorem implies that the GCV estimate $\hat{\mathbf{Z}}_j = \mathbf{P}_{\tau_j}\mathbf{Y}_j$ is approximately the optimal estimate of the surrogate variable $\mathbf{Z}_j$. Furthermore, as the optimal tuning parameter approximates a constant determined by the variance components ratio, hereafter we take the following assumption on τ_j.

Assumption C. $\tau_j/\sqrt{n} \to 0$ as $n \to \infty$, for $j = 1, \cdots, p$.

Denote $\mathbf{H}_k = \mathbf{I} - \mathbf{X}_{\mathcal{S}_k}(\mathbf{X}_{\mathcal{S}_k}^T\mathbf{X}_{\mathcal{S}_k})^{-1}\mathbf{X}_{\mathcal{S}_k}^T$, we then have the following properties on $\hat{\mathbf{Z}}_{-k}$.

Theorem 3 *For $k = 1, \ldots, p$, let $\mathbf{M}_k = \boldsymbol{\pi}_{-k}^T(\mathbf{C} - \mathbf{C}_{\bullet\mathcal{S}_k}\mathbf{C}_{\mathcal{S}_k,\mathcal{S}_k}^{-1}\mathbf{C}_{\mathcal{S}_k\bullet})\boldsymbol{\pi}_{-k}$ where each $\mathbf{C}_{\mathcal{S}_r\mathcal{S}_c}$ is a submatrix of $\mathbf{C}$ identified with row indices in $\mathcal{S}_r$ and column indices in $\mathcal{S}_c$ (the dot implies all rows or columns). Then, under Assumptions A, B, and C,*

a. $n^{-1}\hat{\mathbf{Z}}_{-k}^T\mathbf{H}_k\hat{\mathbf{Z}}_{-k} \to_p \mathbf{M}_k$, as $n \to \infty$;

b. $n^{-1/2}(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\boldsymbol{\gamma}_k)^T\mathbf{H}_k\hat{\mathbf{Z}}_{-k} \to_d N(\mathbf{0}, \sigma_k^2\mathbf{M}_k)$, as $n \to \infty$.

Since $n^{-1}\mathbf{Z}_{-k}^T\mathbf{H}_k\mathbf{Z}_{-k} \to \mathbf{M}_k$, Theorem 3.a states that $\hat{\mathbf{Z}}_{-k}^T\mathbf{H}_k\hat{\mathbf{Z}}_{-k}$ is a good approximation to $\mathbf{Z}_{-k}^T\mathbf{H}_k\mathbf{Z}_{-k}$. On the other hand, $\mathbf{H}_k(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\boldsymbol{\gamma}_k)$ is the error term in regressing $\mathbf{H}_k\mathbf{Y}_k$ against $\mathbf{H}_k\hat{\mathbf{Z}}_{-k}$, and Theorem 3.b implies that $n^{-1}(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\boldsymbol{\gamma}_k)^T\mathbf{H}_k\hat{\mathbf{Z}}_{-k} \to_d 0$. Thus $\hat{\mathbf{Z}}_{-k}$ results in regression errors with good properties, i.e., the error effects on the 2SPLS estimators will vanish when the sample size gets sufficiently large.

In summary, the above theorem indicates that $\hat{\mathbf{Z}}_{-k}$ behaves the same way as $\mathbf{Z}_{-k}$ asymptotically, which makes it feasible to replace $\mathbf{Z}_{-k}$ with $\hat{\mathbf{Z}}_{-k}$ at the second stage. The crucial properties of $\hat{\mathbf{Z}}_{-k}$ in Theorem 3, together with the good theoretical properties of adaptive lasso, will lead to the oracle properties of our proposed estimates. We denote the j-th elements of $\boldsymbol{\gamma}_k$ and $\hat{\boldsymbol{\gamma}}_k$ as γ_{kj} and $\hat{\gamma}_{kj}$, respectively. Then, with a proper choice of λ_n, the proposed method enjoys the following oracle properties.

Theorem 4 *(Oracle Properties) Let $\mathcal{A}_k = \{j : \gamma_{kj} \neq 0\}$, $\hat{\mathcal{A}}_k = \{j : \hat{\gamma}_{kj} \neq 0\}$, and $\mathbf{M}_{k,\mathcal{A}_k}$ be the submatrix of $\mathbf{M}_k$ identified with both row and column indices in $\mathcal{A}_k$. Suppose that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\delta-1)/2} \to \infty$. Then, under Assumptions A, B, and C, the estimates from the proposed 2SPLS method satisfy the following properties,*

a. Consistency in variable selection: $\lim_{n\to\infty} P(\hat{\mathcal{A}}_k = \mathcal{A}_k) = 1$;

b. Asymptotic normality: $\sqrt{n}(\hat{\boldsymbol{\gamma}}_{k,\mathcal{A}_k} - \boldsymbol{\gamma}_{k,\mathcal{A}_k}) \to_d N(\mathbf{0}, \sigma_k^2 \mathbf{M}_{k,\mathcal{A}_k}^{-1})$, *as* $n \to \infty$.

It is worthwhile to mention that Theorem 3 plays an essential role in establishing the oracle properties of 2SPLS. In fact, as long as the properties in Theorem 3 hold true for the first-stage estimates of $\mathbf{Z}_{-k}$, we can generalize the second-stage regularization to a wide class of regularization methods, all the theoretical properties of which can be inherited by our proposed two-stage method.

6 SIMULATION STUDIES

We conducted simulation studies to compare 2SPLS with the adaptive lasso based algorithm (AL) by Logsdon and Mezey (2010), and the sparsity-aware maximum likelihood algorithm (SML) by Cai *et al.* (2013). Both acyclic networks and cyclic networks were simulated, each involving 300 endogenous variables. Each endogenous variable was simulated to have, on average, one regulatory effect for sparse networks, or three regulatory effects for dense networks. The regulatory effects were independently simulated from a uniform distribution over $(-1, -0.5) \cup (0.5, 1)$. To allow the use of AL and SML, every endogenous variable in the same network was simulated to have the same number (either one or three) of known causal effects by the exogenous variables, with all effects equal to one. Each exogenous variable was simulated to take values 0, 1 and 2 with probabilities 0.25, 0.5 and 0.25, respectively, emulating genotypes of an F2 cross in a genetic genomics experiment. All error terms were independently simulated from $N(0, 0.1^2)$, and the sample size n varied from 100 to 1000. For each network setup, we simulated 100 data sets and applied all three algorithms to calculate the power and false discovery rate (FDR).

For inferring acyclic networks, the power and FDR of the three different algorithms are plotted in Figure 1. In the case that each endogenous variable has only one known exogenous effect (EE), 2SPLS has the greatest power to infer both sparse and dense acyclic networks from data sets with different sample sizes. In the case of three EEs available for each endogenous variable, 2SPLS still has greater power than the other two algorithms when the sample size is small or moderate. When the sample size is large, 2SPLS and SML are comparable for constructing both sparse and dense acyclic networks. In any case, 2SPLS and SML provide much greater power than AL. Indeed, AL provides power as low as under 10% when the sample size is not large, and its power is still under 50% even when the sample size increases to 1000. On the other hand, 2SPLS provides power over 80% for small sample sizes, and over 90% for moderate to large sample sizes.

As shown in Figure 1, 2SPLS controls the FDR under 20% except the case with three available EEs and very small sample sizes ($n = 100$). While it controls the FDR as low as under 5% for sparse acyclic networks when the sample sizes are large, SML reports large FDRs when the sample sizes are not large. Indeed, when the sample sizes are under 200, SML reports FDR over 40% for dense acyclic networks. In general, both 2SPLS and SML outperform AL in terms of FDR though AL reports FDR lower than 2SPLS when inferring sparse acyclic networks with one available EE from data sets of very large sample sizes.

Plotted in Figure 2 are the power and FDR of the three different algorithms when inferring cyclic networks. Similar to the results on acyclic networks, 2SPLS has greater power than the other two algorithms across all sample sizes and has lower FDR when

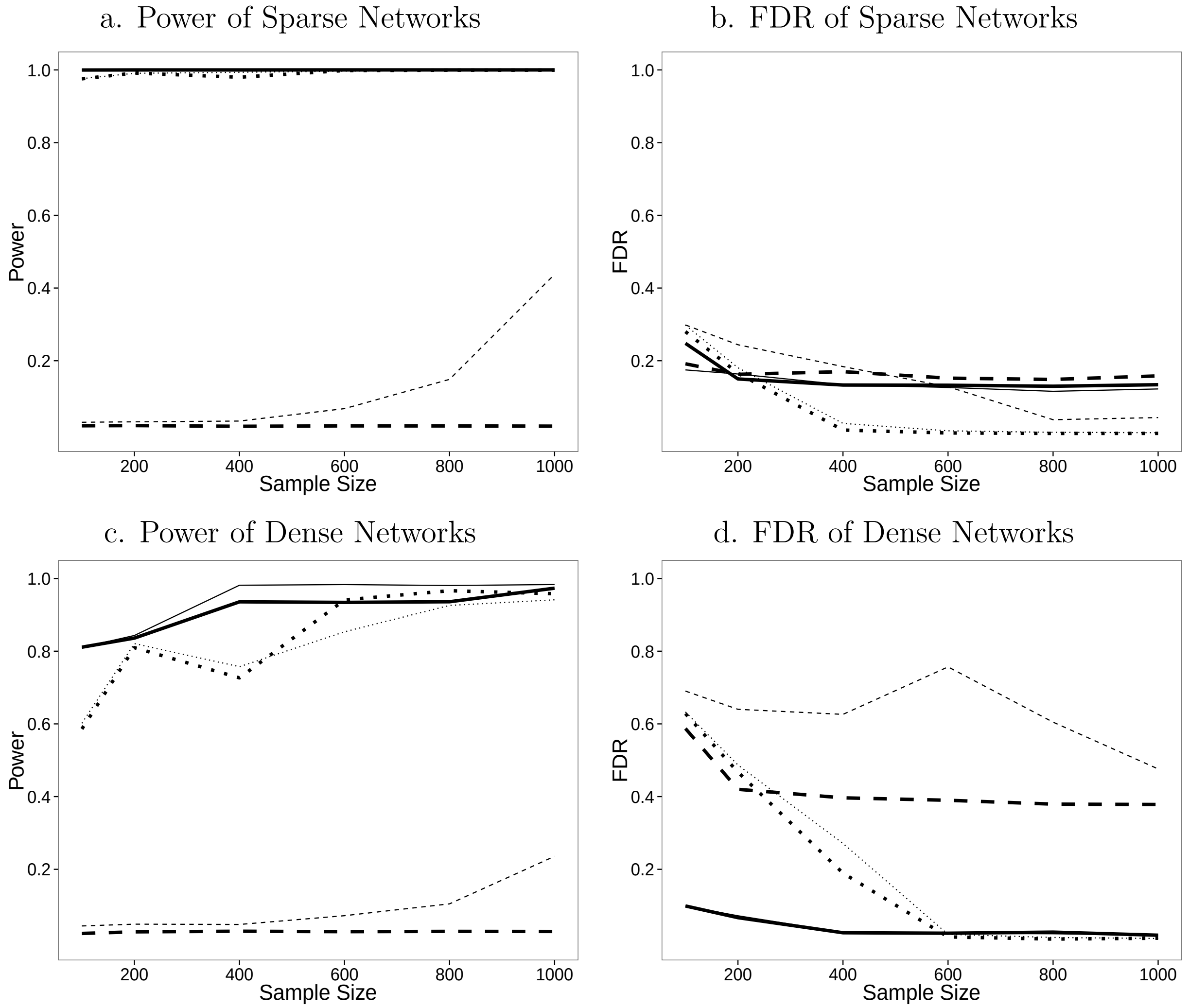

Figure 1: Performance of 2SPLS (solid lines), AL (dashed lines), and SML (dotted lines) when identifying regulatory effects in acyclic networks with one EE (thin lines) or three EEs (thick lines).

the sample size is not large. For dense cyclic networks, AL has power mostly under 20% and FDR over 30%. While it improves the FDR for sparse cyclic networks with large sample sizes, AL has power as low as under 10%. SML provides power competitive to 2SPLS for sparse cyclic networks, but its power is much lower than that of 2SPLS for dense cyclic networks. Similar to the case of acyclic networks, SML reports much higher FDR for inferring dense networks from data sets with small sample sizes though it reports very small FDR when the sample sizes are large. We also conducted simulation studies on both acyclic and cyclic networks with small to moderate number of endogenous variables (e.g., 10 to 50 endogenous variables). The performance of 2SPLS is better than AL and comparable to SML in those scenarios (results are not shown). Indeed, the power of 2SPLS exceeds 0.9 while maintaining low FDR in most of the scenarios.

While it generally reports higher power and more robust FDR than SML, 2SPLS significantly reduces the computation time in comparison to SML as it assembles the network by investigating limited-information models. To demonstrate such advantage of 2SPLS over SML, we recorded the computing time of all algorithms in inferring the same networks from small data sets ($n = 100$). Each algorithm analyzed the same data set using only one CPU in a server with Quad-Core AMD Opteron$^{\text{TM}}$ Processor 8380. Reported in Table 1 are the running times of the three algorithms for inferring different networks. Apparently, AL is the fastest, and the running time of 2SPLS usually doubles or triples that of AL. The slowest algorithm is SML which generally takes more than 40 times longer than 2SPLS to infer different networks. In particular, SML is almost 200 times slower than 2SPLS when inferring acyclic sparse networks.

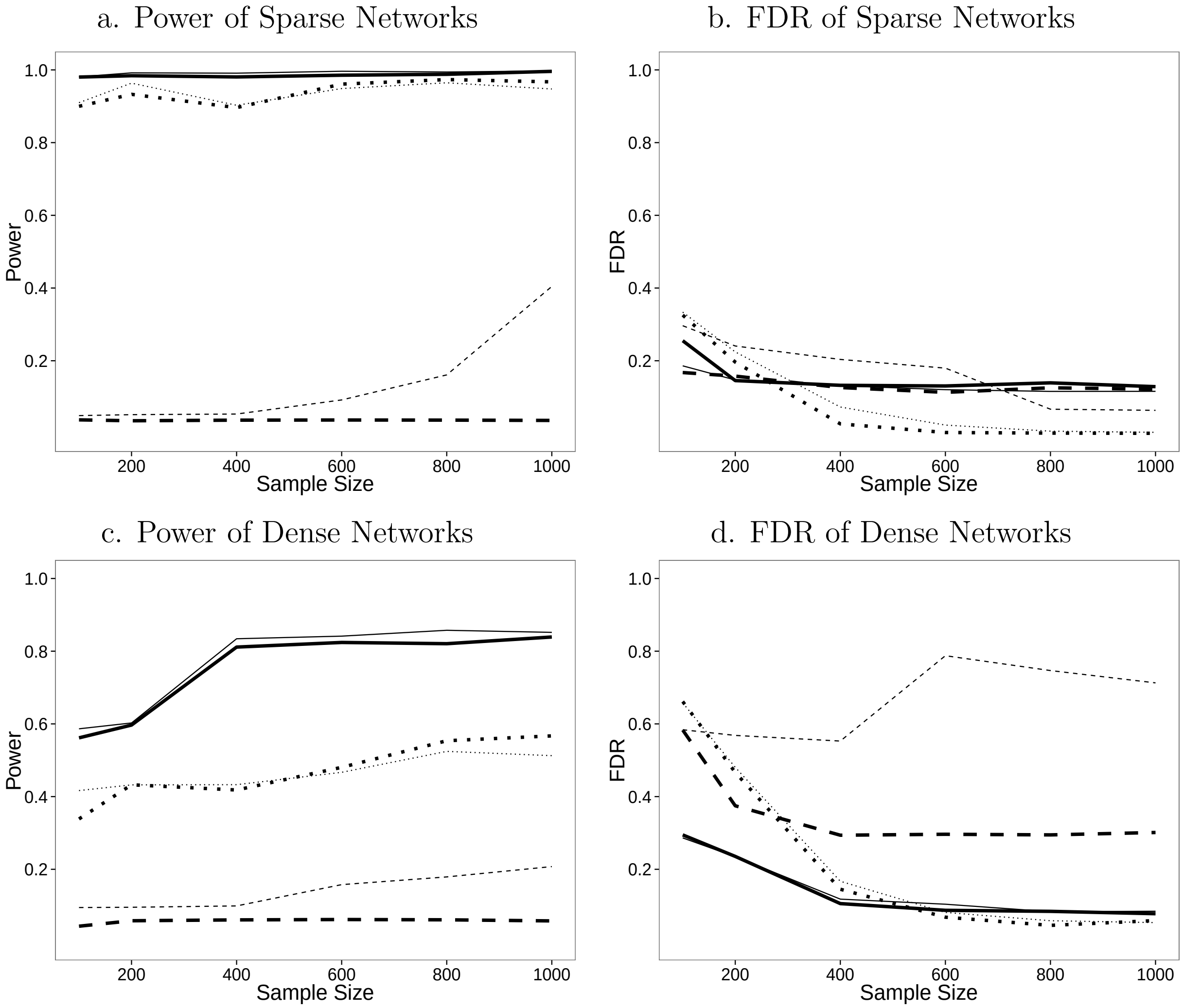

Figure 2: Performance of 2SPLS (solid lines), AL (dashed lines) and SML (dotted lines) when identifying regulatory effects in cyclic networks with one EE (thin lines) or three EEs (thick lines).

Table 1: The running time (in seconds) of inferring networks from a data set with $n = 100$.

	Acyclic				Cyclic			
	Sparse		Dense		Sparse		Dense	
	1 EE	3 EEs	1 EE	3 EEs	1 EE	3 EEs	1 EE	3 EEs
2SPLS	1303	1332	1127	1112	1297	1337	1125	1165
AL	405	652	404	637	443	659	430	781
SML	258875	195739	58509	43118	49393	58716	67949	68081

7 REAL DATA ANALYSIS

We analyzed a yeast data set with 112 segregants from a cross between two strains BY4716 and RM11-la (Brem and Kruglyak 2005). A total of 5,727 genes were measured for their expression values, and 2,956 markers were genotyped. Each marker within a genetic region (including 1kb upstream and downstream regions) was evaluated for its association with the corresponding gene expression, yielding 722 genes with marginally significant cis-eQTL (p-value < 0.05). The set of cis-eQTL for each gene was filtered to control the pairwise correlation under 0.90, and then further filtered to keep up to three cis-eQTL which have the strongest association with the corresponding gene expression.

With 112 observations of 722 endogenous variables and 732 exogenous variables, we applied 2SPLS to infer the gene regulatory network in yeast. The constructed network includes 7,300 regulatory effects in total. To evaluate the reliability of constructed gene regulations, we generated 10,000 bootstrap data sets (each with $n = 112$) by randomly sampling the original data with replacement, and applied 2SPLS to each data set to infer the gene regulatory network. Among the 7,300 regulatory effects, 323 effects were repeatedly identified in more than 80% of the 10,000 data sets, and Figure 3 shows the three largest subnetworks formed by these 323 effects. Specifically, the largest subnetwork consists of 22 endogenous variables and 26 regulatory effects, the second largest one includes 14 endogenous variables and 18 regulatory effects, and the third largest one has 11 endogenous variables and 16 regulatory effects.

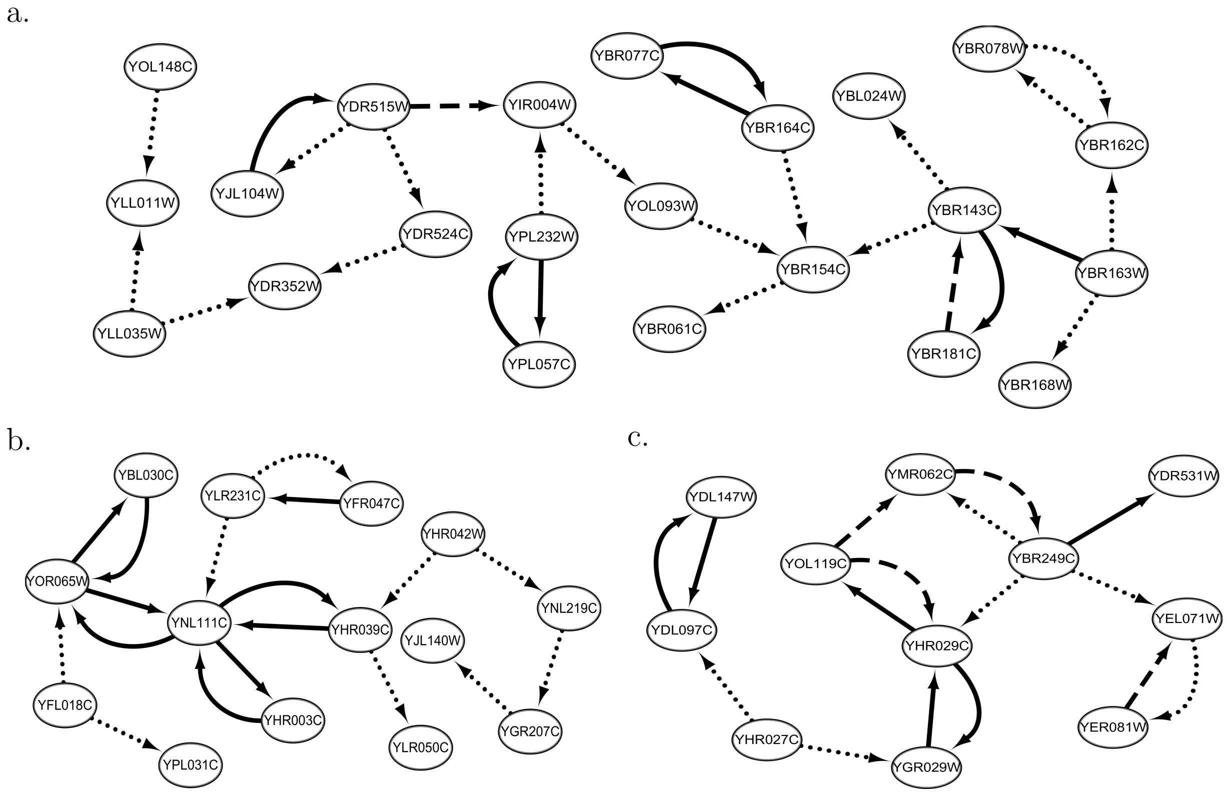

Figure 3: Three gene regulatory subnetworks in yeast (the dotted, dashed, and solid arrows implied that the corresponding regulations were constructed respectively from over 80%, 90%, and 95% of the bootstrap data sets).

A gene-enrichment analysis with DAVID (Huang *et al.* 2009) showed that the three subnetworks are enriched in different gene clusters (controlling p-values from Fisher's exact tests under 0.01). A total of six gene clusters are enriched with genes from the first subnetwork, and four of them are related to either methylation or methyltransferase. Six of 22 genes in the first subnetwork are found in a gene cluster which is related to none-coding RNA processing. The second subnetwork is enriched in nine gene clusters. While three of them are related to electron, one cluster includes half of the genes from the second subnetwork and is related to oxidation reduction. The third subnetwork is also enriched in nine different gene clusters, with seven clusters related to proteasome.

A total of 18 regulations were constructed from each of the 10,000 bootstrap data sets, and are shown in Figure 4. There are seven pairs of genes which regulate each other. It is interesting to observe that all regulatory genes up regulate the target genes except two genes, namely, YCL018W and YEL021W.

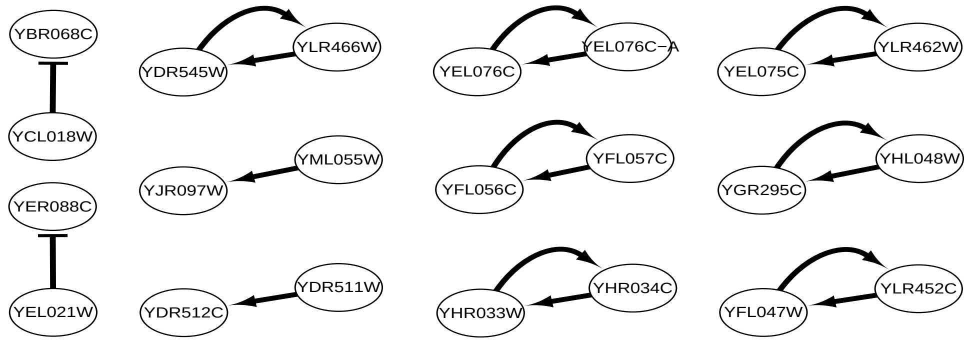

Figure 4: The yeast gene regulatory subnetworks constructed in each of 10,000 bootstrap data sets (with arrow- and bar-headed lines implying up and down regulations, respectively).

8 DISCUSSION

In a classical setting with small numbers of endogenous/exogenous variables, constructing a system of structural equations has been well studied since Haavelmo (1943, 1944). Anderson and Rubin (1949) first proposed to estimate the parameters of a single structural equation with the limited information maximum likelihood estimator. Later on, Theil (1953) and Basmann (1957) independently developed the 2SLS estimator, which is the simplest and most common estimation method for fitting a system of structural equations. However, the genetical genomics experiments usually collect data in which both the number of endogenous variables and the number of exogenous variables can be very large, invalidating the classical methods for building gene regulatory networks.

Replacing the ordinary least squares at the two stages with ridge regression and adaptive lasso respectively, the proposed 2SPLS method can consistently identify and further estimate the regulatory effects of the endogenous variables, even with a large number of endogenous variables. As a high-dimensional extension of the classical 2SLS method, the 2SPLS method is also computationally fast and easy to implement. As shown in constructing a genome-wide gene regulatory network of yeast, the high computational efficiency of 2SPLS allows us to employ the bootstrap method to calculate the p-values of regulatory effects. Meanwhile, each of the two steps, especially the second one, may be further improved by incorporating recent progresses in high-dimensional variable selection, see, for example, Chen and Chen (2008), Zhang (2010), and Lockhart *et al.* (2014).

APPENDIX A: PROOF OF THEOREM 3

a. Since $\tau_j/\sqrt{n} \to 0$ for any $1 \le j \le p$, the different choice of τ_j for each j does not affect the following asymptotic property involving τ_j,

$$n(\mathbf{X}^T\mathbf{X} + \tau_j\mathbf{I})^{-1} \to \mathbf{C}. \tag{A.1}$$

Without loss of generality, we assume $\tau_1 = \tau_2 = \cdots = \tau_p = \tau$. Then $\hat{\mathbf{Z}}_{-k} = \mathbf{P}_\tau \mathbf{Y}_{-k}$.

$$\begin{aligned}
&\frac{1}{n}\hat{\mathbf{Z}}_{-k}^T \mathbf{H}_k \hat{\mathbf{Z}}_{-k} \\
&= \frac{1}{n}(\mathbf{X}\boldsymbol{\pi}_{-k} + \boldsymbol{\xi}_{-k})^T \mathbf{P}_\tau^T \mathbf{H}_k \mathbf{P}_\tau (\mathbf{X}\boldsymbol{\pi}_{-k} + \boldsymbol{\xi}_{-k}) \\
&= \frac{1}{n}\boldsymbol{\pi}_{-k}^T \mathbf{X}^T \mathbf{P}_\tau \mathbf{H}_k \mathbf{P}_\tau \mathbf{X}\boldsymbol{\pi}_{-k} + \frac{1}{n}\boldsymbol{\xi}_{-k}^T \mathbf{P}_\tau \mathbf{H}_k \mathbf{P}_\tau \mathbf{X}\boldsymbol{\pi}_{-k} \\
&\quad + \frac{1}{n}\boldsymbol{\pi}_{-k}^T \mathbf{X}^T \mathbf{P}_\tau \mathbf{H}_k \mathbf{P}_\tau \boldsymbol{\xi}_{-k} + \frac{1}{n}\boldsymbol{\xi}_{-k}^T \mathbf{P}_\tau \mathbf{H}_k \mathbf{P}_\tau \boldsymbol{\xi}_{-k}
\end{aligned}$$

We will consider the asymptotic property of each of the above four terms.

First, $\frac{1}{n}\mathbf{X}^T\mathbf{X} \to \mathbf{C}$ implies that

$$\frac{1}{n}\mathbf{X}^T \mathbf{H}_k \mathbf{X} = \frac{1}{n}\mathbf{X}^T\{\mathbf{I} - \mathbf{X}_{\mathcal{S}_k}(\mathbf{X}_{\mathcal{S}_k}^T \mathbf{X}_{\mathcal{S}_k})^{-1}\mathbf{X}_{\mathcal{S}_k}^T\}\mathbf{X} \to \mathbf{C} - \mathbf{C}_{\bullet\mathcal{S}_k}\mathbf{C}_{\mathcal{S}_k,\mathcal{S}_k}^{-1}\mathbf{C}_{\mathcal{S}_k\bullet}. \qquad \text{(A.2)}$$

The above result and (A.1) easily lead to the following result,

$$\begin{aligned}
&\frac{1}{n}\boldsymbol{\pi}_{-k}^T \mathbf{X}^T \mathbf{P}_\tau \mathbf{H}_k \mathbf{P}_\tau \mathbf{X}\boldsymbol{\pi}_{-k} \\
&= \frac{1}{n}\boldsymbol{\pi}_{-k}^T \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \tau\mathbf{I})^{-1}\mathbf{X}^T\mathbf{H}_k\mathbf{X}(\mathbf{X}^T\mathbf{X} + \tau\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\pi}_{-k} \\
&\to \boldsymbol{\pi}_{-k}^T(\mathbf{C} - \mathbf{C}_{\bullet\mathcal{S}_k}\mathbf{C}_{\mathcal{S}_k,\mathcal{S}_k}^{-1}\mathbf{C}_{\mathcal{S}_k\bullet})\boldsymbol{\pi}_{-k} = \mathbf{M}_k. \qquad \text{(A.3)}
\end{aligned}$$

The other three terms approaching to zero directly follows that $\frac{1}{n}\boldsymbol{\xi}_{-k}^T\mathbf{X} \to_p \mathbf{0}$. Thus, $\frac{1}{n}\hat{\mathbf{Z}}_{-k}^T\mathbf{H}_k\hat{\mathbf{Z}}_{-k} \to_p \mathbf{M}_k$.

b. Since $\mathbf{H}_k(\mathbf{Y}_k - \mathbf{Y}_{-k}\boldsymbol{\gamma}_k) = \mathbf{H}_k\boldsymbol{\epsilon}_k$, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\boldsymbol{\gamma}_k)^T\mathbf{H}_k\hat{\mathbf{Z}}_{-k} \\
&= \frac{1}{\sqrt{n}}\{(\mathbf{Y}_k - \mathbf{Y}_{-k}\boldsymbol{\gamma}_k) + (\mathbf{I} - \mathbf{P}_\tau)\mathbf{Y}_{-k}\boldsymbol{\gamma}_k\}^T\mathbf{H}_k\hat{\mathbf{Z}}_{-k} \\
&= \frac{1}{\sqrt{n}}\boldsymbol{\epsilon}_k^T\mathbf{H}_k\mathbf{P}_\tau\mathbf{Y}_{-k} + \frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\{(\mathbf{I} - \mathbf{P}_\tau)\mathbf{Y}_{-k}\}^T\mathbf{H}_k\mathbf{P}_\tau\mathbf{Y}_{-k}
\end{aligned}$$

In the following, we will prove that the second term approaches to zero, and the first term asymptotically approaches to the required distribution, i.e.,

$$\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}_k^T\mathbf{H}_k\mathbf{P}_\tau\mathbf{Y}_{-k} \to_d N(\mathbf{0}, \sigma_k^2\mathbf{M}_k). \tag{A.4}$$

We notice that

$$\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}_k^T\mathbf{H}_k\mathbf{P}_\tau\mathbf{X}\boldsymbol{\pi}_{-k} \sim N(\mathbf{0}, \frac{\sigma_k^2}{n}\boldsymbol{\pi}_{-k}^T\mathbf{X}^T\mathbf{P}_\tau\mathbf{H}_k\mathbf{P}_\tau\mathbf{X}\boldsymbol{\pi}_{-k}).$$

Following (A.3), we have

$$\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}_k^T\mathbf{H}_k\mathbf{P}_\tau\mathbf{X}\boldsymbol{\pi}_{-k} \to_d N(\mathbf{0}, \sigma_k^2\mathbf{M}_k). \tag{A.5}$$

Because of (A.2) and

$$\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}_k^T\mathbf{H}_k\mathbf{X} \sim N(\mathbf{0}, \frac{\sigma_k^2}{n}\mathbf{X}^T\mathbf{H}_k\mathbf{X}),$$

we have

$$\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}_k^T\mathbf{H}_k\mathbf{X} \to_d N(\mathbf{0}, \sigma_k^2(\mathbf{C} - \mathbf{C}_{\bullet\mathcal{S}_k}\mathbf{C}_{\mathcal{S}_k,\mathcal{S}_k}^{-1}\mathbf{C}_{\mathcal{S}_k\bullet})).$$

Since $\frac{1}{n}\boldsymbol{\xi}_{-k}^T\mathbf{X} \to_p \mathbf{0}$, we can apply Slutsky's theorem and obtain that

$$\frac{1}{\sqrt{n}}\boldsymbol{\epsilon}_k^T\mathbf{H}_k\mathbf{P}_\tau\boldsymbol{\xi}_{-k} = \frac{1}{\sqrt{n}}\boldsymbol{\epsilon}_k^T\mathbf{H}_k\mathbf{X}(\mathbf{X}^T\mathbf{X} + \tau\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{\xi}_{-k} \to_p \mathbf{0}.$$

Pooling the above result and (A.5) leads to the asymptotic distribution in (A.4).

To prove that the second term asymptotically approaches to zero, we further partition it as follows,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\{(\mathbf{I} - \mathbf{P}_\tau)\mathbf{Y}_{-k}\}^T\mathbf{H}_k\mathbf{P}_\tau\mathbf{Y}_{-k} \\
&= \frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\pi}_{-k}^T\mathbf{X}^T(\mathbf{I} - \mathbf{P}_\tau)\mathbf{H}_k\mathbf{P}_\tau\mathbf{X}\boldsymbol{\pi}_{-k} + \frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\xi}_{-k}^T(\mathbf{I} - \mathbf{P}_\tau)\mathbf{H}_k\mathbf{P}_\tau\mathbf{X}\boldsymbol{\pi}_{-k} \\
&\quad + \frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\pi}_{-k}^T\mathbf{X}^T(\mathbf{I} - \mathbf{P}_\tau)\mathbf{H}_k\mathbf{P}_\tau\boldsymbol{\xi}_{-k} + \frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\xi}_{-k}^T(\mathbf{I} - \mathbf{P}_\tau)\mathbf{H}_k\mathbf{P}_\tau\boldsymbol{\xi}_{-k}.
\end{aligned}$$

It suffices to prove each of these four parts asymptotically approaches to zero.

First, notice that

$$\mathbf{X}^T(\mathbf{I}-\mathbf{P}_\tau)=\tau(\mathbf{X}^T\mathbf{X}+\tau\mathbf{I})^{-1}\mathbf{X}^T,$$

we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\pi}_{-k}^T\mathbf{X}^T(\mathbf{I}-\mathbf{P}_\tau)\mathbf{H}_k\mathbf{P}_\tau\mathbf{X}\boldsymbol{\pi}_{-k}\\ &= \frac{\tau}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\pi}_{-k}^T(\mathbf{X}^T\mathbf{X}+\tau\mathbf{I})^{-1}\mathbf{X}^T\mathbf{H}_k\mathbf{X}(\mathbf{X}^T\mathbf{X}+\tau\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\pi}_{-k}\to\mathbf{0},\end{aligned}\tag{A.6}$$

which follows (A.2) and that $\tau/\sqrt{n}\to 0$ as $n\to\infty$.

Because $\mathbf{C}_{\mathcal{S}_k\bullet}\mathbf{C}^{-1}\mathbf{C}_{\bullet\mathcal{S}_k}=\mathbf{C}_{\mathcal{S}_k\mathcal{S}_k}$, we have

$$(\mathbf{C}-\mathbf{C}_{\bullet\mathcal{S}_k}\mathbf{C}_{\mathcal{S}_k,\mathcal{S}_k}^{-1}\mathbf{C}_{\mathcal{S}_k\bullet})\mathbf{C}^{-1}(\mathbf{C}-\mathbf{C}_{\bullet\mathcal{S}_k}\mathbf{C}_{\mathcal{S}_k,\mathcal{S}_k}^{-1}\mathbf{C}_{\mathcal{S}_k\bullet})=\mathbf{C}-\mathbf{C}_{\bullet\mathcal{S}_k}\mathbf{C}_{\mathcal{S}_k,\mathcal{S}_k}^{-1}\mathbf{C}_{\mathcal{S}_k\bullet},$$

which implies that

$$\begin{aligned}&\frac{1}{n}\mathbf{X}^T\mathbf{P}_\tau^T\mathbf{H}_k^T(\mathbf{I}-\mathbf{P}_\tau)^T(\mathbf{I}-\mathbf{P}_\tau)\mathbf{H}_k\mathbf{P}_\tau\mathbf{X}\\ &= \frac{1}{n}\mathbf{X}^T\mathbf{P}_\tau\mathbf{H}_k\mathbf{P}_\tau\mathbf{X}-\frac{2}{n}\mathbf{X}^T\mathbf{P}_\tau\mathbf{H}_k\mathbf{P}_\tau\mathbf{H}_k\mathbf{P}_\tau\mathbf{X}+\frac{1}{n}\mathbf{X}^T\mathbf{P}_\tau\mathbf{H}_k\mathbf{P}_\tau^2\mathbf{H}_k\mathbf{P}_\tau\mathbf{X}\to\mathbf{0}.\end{aligned}$$

Since $\mathrm{Var}(\boldsymbol{\xi}_{-k}\boldsymbol{\gamma}_k)$ is proportional to an identity matrix, the above result leads to that

$$\mathrm{Var}\left(\frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\xi}_{-k}^T(\mathbf{I}-\mathbf{P}_\tau)\mathbf{H}_k\mathbf{P}_\tau\mathbf{X}\boldsymbol{\pi}_{-k}\right)\to\mathbf{0},$$

which implies that

$$\frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\xi}_{-k}^T(\mathbf{I}-\mathbf{P}_\tau)\mathbf{H}_k\mathbf{P}_\tau\mathbf{X}\boldsymbol{\pi}_{-k}\to_p\mathbf{0}.\tag{A.7}$$

Similarly, we can prove that, for each $\boldsymbol{\xi}_j$,

$$\mathrm{Var}\left(\frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\pi}_{-k}^T\mathbf{X}^T(\mathbf{I}-\mathbf{P}_\tau)\mathbf{H}_k\mathbf{P}_\tau\boldsymbol{\xi}_j\right)\to\mathbf{0},$$

which implies that

$$\frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\pi}_{-k}^T\mathbf{X}^T(\mathbf{I}-\mathbf{P}_\tau)\mathbf{H}_k\mathbf{P}_\tau\boldsymbol{\xi}_{-k} \to_p \mathbf{0}. \tag{A.8}$$

Note that

$$\frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\xi}_{-k}^T(\mathbf{I}-\mathbf{P}_\tau)\mathbf{H}_k\mathbf{P}_\tau\boldsymbol{\xi}_{-k} = \left\{\frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\xi}_{-k}^T(\mathbf{I}-\mathbf{P}_\tau)\mathbf{H}_k\mathbf{X}\right\}\left\{(\mathbf{X}^T\mathbf{X}+\tau\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{\xi}_{-k}\right\}.$$

Since

$$\frac{1}{n}\mathbf{X}^T\mathbf{H}_k(\mathbf{I}-\mathbf{P}_\tau)(\mathbf{I}-\mathbf{P}_\tau)\mathbf{H}_k\mathbf{X} \to \mathbf{0},$$

we have

$$\mathrm{Var}(\frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\xi}_{-k}^T(\mathbf{I}-\mathbf{P}_\tau)\mathbf{H}_k\mathbf{X}) \to \mathbf{0}.$$

Therefore,

$$\frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\xi}_{-k}^T(\mathbf{I}-\mathbf{P}_\tau)\mathbf{H}_k\mathbf{X} \to_p \mathbf{0},$$

which, together with $(\mathbf{X}^T\mathbf{X}+\tau\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{\xi}_{-k} \to_p \mathbf{0}$, leads to that

$$\frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\boldsymbol{\xi}_{-k}^T(\mathbf{I}-\mathbf{P}_\tau)\mathbf{H}_k\mathbf{P}_\tau\boldsymbol{\xi}_{-k} \to_p \mathbf{0}. \tag{A.9}$$

Pooling (A.6), (A.7), (A.8) and (A.9), we have proved that $\frac{1}{\sqrt{n}}\boldsymbol{\gamma}_k^T\{(\mathbf{I}-\mathbf{P}_\tau)\mathbf{Y}_{-k}\}^T\mathbf{H}_k\mathbf{P}_\tau\mathbf{Y}_{-k} \to_p \mathbf{0}$, which concludes our proof.

APPENDIX B: PROOF OF THEOREM 4

Let $\boldsymbol{\psi}_n(\boldsymbol{\mu}) = \|\mathbf{H}_k\mathbf{Y}_k - \mathbf{H}_k\hat{\mathbf{Z}}_{-k}(\boldsymbol{\gamma}_k+\frac{\boldsymbol{\mu}}{\sqrt{n}})\|^2 + \lambda_n\boldsymbol{\omega}_k^T|\boldsymbol{\gamma}_k+\frac{\boldsymbol{\mu}}{\sqrt{n}}|$. Let $\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}}\boldsymbol{\psi}_n(\boldsymbol{\mu})$, then $\hat{\boldsymbol{\gamma}}_k = \boldsymbol{\gamma}_k + \frac{\hat{\boldsymbol{\mu}}}{\sqrt{n}}$ or $\hat{\boldsymbol{\mu}} = \sqrt{n}(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k)$. Note that $\boldsymbol{\psi}_n(\boldsymbol{\mu}) - \boldsymbol{\psi}_n(\mathbf{0}) = V_n(\boldsymbol{\mu})$, where

$$\begin{aligned} V_n(\boldsymbol{\mu}) &= \boldsymbol{\mu}^T(\frac{1}{n}\hat{\mathbf{Z}}_{-k}^T\mathbf{H}_k\hat{\mathbf{Z}}_{-k})\boldsymbol{\mu} - \frac{2}{\sqrt{n}}(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\boldsymbol{\gamma}_k)^T\mathbf{H}_k\hat{\mathbf{Z}}_{-k}\boldsymbol{\mu} \\ &\quad + \frac{\lambda_n}{\sqrt{n}}\boldsymbol{\omega}_k^T \times \sqrt{n}(|\boldsymbol{\gamma}_k + \frac{\boldsymbol{\mu}}{\sqrt{n}}| - |\boldsymbol{\gamma}_k|). \end{aligned}$$

Denote the j-th elements of $\boldsymbol{\omega}_k$ and $\boldsymbol{\mu}$ as ω_{kj} and μ_j, respectively.

If $\gamma_{kj} \neq 0$, then $\omega_{kj} \to_p |\gamma_{kj}|^{-\delta}$ and $\sqrt{n}(|\gamma_{kj} + \frac{\mu_j}{\sqrt{n}}| - |\gamma_{kj}|) \to_p \mu_j \text{sign}(\gamma_{kj})$. By Slutsky's theorem, we have $\frac{\lambda_n}{\sqrt{n}}\omega_{kj}\sqrt{n}(|\gamma_{kj}+\frac{\mu_j}{\sqrt{n}}|-|\gamma_{kj}|) \to_p 0$. If $\gamma_{kj} = 0$, then $\sqrt{n}(|\gamma_{kj}+\frac{\mu_j}{\sqrt{n}}|-|\gamma_{kj}|) = |\mu_j|$ and $\frac{\lambda_n}{\sqrt{n}}\omega_{kj} = \frac{\lambda_n}{\sqrt{n}} n^{\delta/2}(|\sqrt{n}\tilde{\gamma}_{kj}|)^{-\delta}$, where $\sqrt{n}\tilde{\gamma}_{kj} = O_p(1)$. Thus,

$$\frac{\lambda_n}{\sqrt{n}}\boldsymbol{\omega}_k^T \times \sqrt{n}(|\boldsymbol{\gamma}_k + \frac{\boldsymbol{\mu}}{\sqrt{n}}| - |\boldsymbol{\gamma}_k|) \to_p \begin{cases} 0, & \text{if } \|\boldsymbol{\mu}_{\mathcal{A}_k^c}\| = 0; \\ \infty, & \text{otherwise.} \end{cases}$$

Hence, following Theorem 3 and Slutsky's theorem, we see that $V_n(\boldsymbol{\mu}) \to_d V(\boldsymbol{\mu})$ for every $\boldsymbol{\mu}$, where

$$V(\boldsymbol{\mu}) = \begin{cases} \boldsymbol{\mu}_{\mathcal{A}_k}^T \mathbf{M}_{k,\mathcal{A}_k} \boldsymbol{\mu}_{\mathcal{A}_k} - 2\boldsymbol{\mu}_{\mathcal{A}_k}^T \mathbf{W}_{k,\mathcal{A}_k}, & \text{if } \|\boldsymbol{\mu}_{\mathcal{A}_k^c}\| = 0; \\ \infty, & \text{otherwise.} \end{cases}$$

$V_n(\boldsymbol{\mu})$ is convex, and the unique minimizer of $V(\boldsymbol{\mu})$ is $(\mathbf{M}_{k,\mathcal{A}_k}^{-1}\mathbf{W}_{k,\mathcal{A}_k}, \mathbf{0})^T$. Following the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we have

$$\begin{cases} \hat{\boldsymbol{\mu}}_{\mathcal{A}_k} \to_d \mathbf{M}_{k,\mathcal{A}_k}^{-1}\mathbf{W}_{k,\mathcal{A}_k}, \\ \hat{\boldsymbol{\mu}}_{\mathcal{A}_k^c} \to_d \mathbf{0}. \end{cases}$$

Since $\mathbf{W}_{k,\mathcal{A}_k} \sim N(\mathbf{0}, \sigma_k^2 \mathbf{M}_{k,\mathcal{A}_k})$, we indeed have proved the asymptotic normality.

Now we show the consistency in variable selection. For $\forall j \in \mathcal{A}_k$, the asymptotic normality indicates that $\hat{\gamma}_{kj} \to_p \gamma_{kj}$, thus $P(j \in \hat{\mathcal{A}}_k) \to 1$. Then it suffices to show that $\forall j \notin \mathcal{A}_k$, $P(j \in \hat{\mathcal{A}}_k) \to 0$.

When $j \in \hat{\mathcal{A}}_k$, by the KKT normality conditions, we know that $\hat{\mathbf{Z}}_j^T \mathbf{H}_k(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\hat{\boldsymbol{\gamma}}_k) = \lambda_n \omega_{kj}$. Note that $\lambda_n \omega_{kj}/\sqrt{n} \to_p \infty$, whereas $\frac{1}{\sqrt{n}}\hat{\mathbf{Z}}_j^T \mathbf{H}_k(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\hat{\boldsymbol{\gamma}}_k) = \frac{1}{n}\hat{\mathbf{Z}}_j^T \mathbf{H}_k \hat{\mathbf{Z}}_{-k} \times \sqrt{n}(\boldsymbol{\gamma}_k - \hat{\boldsymbol{\gamma}}_k) + \frac{1}{\sqrt{n}}\hat{\mathbf{Z}}_j^T \mathbf{H}_k(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\boldsymbol{\gamma}_k)$. Following Theorem 3 and the asymptotic normality, $\frac{1}{\sqrt{n}}\hat{\mathbf{Z}}_j^T \mathbf{H}_k(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\hat{\boldsymbol{\gamma}}_k)$ asymptotically follows a normal distribution. Thus, $P(j \in \hat{\mathcal{A}}_k) \leq P(\hat{\mathbf{Z}}_j^T \mathbf{H}_k(\mathbf{Y}_k - \hat{\mathbf{Z}}_{-k}\hat{\boldsymbol{\gamma}}_k) = \lambda_n \omega_{kj}) \to 0$. Then we have proved the consistency in variable selection.

REFERENCES

Akaike, H. (1974), “A New Look at the Statistical Model Identification,” *IEEE Transactions on Automatic Control*, 19, 716-723.

Anderson, T. W., and Rubin, H. (1949), “Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations,” *Annals of Mathematical Statistics*, 20, 46-63.

Basmann, R. L. (1957), “A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation,” *Econometrica*, 25, 77-83.

Brem, R. B., and Kruglyak, L. (2005), “The Landscape of Genetic Complexity Across 5,700 Gene Expression Traits in Yeast,” *Proceedings of the National Academy of Sciences of the United States of America*, 102, 1572-1577.

Broman, K. W., and Speed, T. P. (2002), “A Model Selelction Approach for the Identification of Quantitative Trait Loci in Experimental Crosses,” *Journal of the Royal Statistical Society: Series B*, 64, 641-656.

Cai, X., Bazerque, J. A., and Giannakis, G. B. (2013), “Inference of Gene Regulatory Networks With Sparse Structural Equation Models Exploiting Genetic Perturbations,” *PLoS Computational Biology*, 9, e1003068.

Chen, J., and Chen, Z. (2008), “Extended Bayesian Information Criteria for Model Selection With Large Model Spaces,” *Biometrika*, 95, 759-771.

Geyer, C. (1994), “On the Asymptotics of Constrained M-Estimation,” *Annals of Statistics*, 22, 1993-2010.

Golub, G. H., Heath, M., and Wahba, G. (1979), “Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter,” *Technometrics*, 21, 215-223.

Haavelmo, T. (1943), “The Statistical Implications of a System of Simultaneous Equations,” *Econometrica*, 11, 1-12.

——(1944), “The Probability Approach in Econometrics,” *Econometrica*, 12, S1-S115.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009), “Bioinformatics Enrichment Tools: Paths Toward the Comprehensive Functional Analysis of Large Gene Lists,” *Nucleic Acids Research*, 37, 1-13.

Jansen, R. J., and Nap, J.-P. (2001), “Genetical Genomics: the Added Value From Segregation,” *Trends in Genetics*, 17, 388-391.

Knight, K., and Fu, W. (2000), “Asymptotics for Lasso-Type Estimators,” *Annals of Statistics*, 28, 1356-1378.

Liu, B., de la Fuente, A., and Hoeschele, I. (2008), “Gene Network Inference via Structural Equation Modeling in Genetical Genomics Experiments,” *Genetics*, 178, 1763-1776.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014), “A Significance Test for the Lasso,” *Annals of Statistics*, 42, 413-468.

Logsdon, B. A., and Mezey, J. (2010), “Gene Expression Network Reconstruction by Convex Feature Selection When Incorporating Genetic Perturbations,” *PLoS Computational Biology*, 6, e1001014.

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B., and Friend, S. H. (2003), “Genetics of Gene Expression Surveyed in Maize, Mouse and Man,” *Nature*, 422, 297-302.

Schmidt, P. (1976), *Econometrics*. New York: Marcel Dekker.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461-464.

Theil, H. (1953), "Repeated Least-Squares Applied to Complete Equation Systems," *Mimeo. The Hague: Central Planning Bureau.*

Xiong, M., Li, J., and Fang, X. (2004), "Identification of Genetic Networks," *Genetics*, 166, 1037-1052.

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *Annals of Statistics*, 38, 894-942.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418-1429.