Statistica Sinica Preprint No: SS-2019-0081	
Title	Penalized Interaction Estimation for Ultrahigh Dimensional Quadratic Regression
Manuscript ID	SS-2019-0081
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0081
Complete List of Authors	Cheng Wang Binyan Jiang and Liping Zhu
Corresponding Author	Liping Zhu
E-mail	zhu.liping@ruc.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

Statistica Sinica

Penalized Interaction Estimation for Ultrahigh Dimensional Quadratic Regression

Cheng Wang, Binyan Jiang and Liping Zhu

Shanghai Jiao Tong University, Hong Kong Polytechnic University, Renmin University of China

Abstract: Quadratic regressions extend linear models by simultaneously including the main effects and the interactions between the covariates. As such, estimating interactions in high-dimensional quadratic regressions has received extensive attention. Here, we introduce a novel method that allows us to estimate the main effects and the interactions separately. Unlike existing methods for ultrahigh-dimensional quadratic regressions, our proposal does not require the widely used heredity assumption. In addition, our proposed estimates have explicit formulae and obey the invariance principle at the population level. We estimate the interactions in matrix form under a penalized convex loss function. The resulting estimates are shown to be consistent, even when the covariate dimension is an exponential order of the sample size. We develop an efficient alternating direction method of multipliers algorithm to implement the penalized estimation. This algorithm fully exploits the cheap computational cost of the matrix multiplication and is much more efficient than existing penalized methods, such as the all-pairs LASSO. We demonstrate the promising performance of the proposed method

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

using extensive numerical studies.

Key words and phrases: High dimension, interaction estimation, quadratic regression, support recovery.

1. INTRODUCTION

A fundamental problem in scientific research is understanding how the features under investigation interact. Interaction estimation has been shown to be very attractive in both parameter estimation and model prediction (Bien et al., 2013; Hao et al., 2018), especially for data sets with complicated structures. Efron et al. (2004) pointed out that for Boston housing data, the prediction accuracy can be improved significantly if interactions are included in addition to all main effects. In general, ignoring interactions by considering main effects alone may lead to an inaccurate or biased estimation, resulting in a poor prediction of an outcome of interest. In contrast, considering both interactions and main effects can improve model interpretability and prediction substantially, thus achieving a better understanding of how the outcome depends on the predictive features (Fan et al., 2015). While it is important to identify interactions that may reveal real relationships between the outcome and the predictive features, the number of parameters scales squarely with the number of predictive features.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

This makes parameter estimation and model prediction very challenging for problems with large, or even moderate dimensionality.

1.1 Interaction Estimation, Feature Selection, and Screening

Estimating interactions is a challenging problem because the number of pairwise interactions increases quadratically with the number of covariates. In the past decade, there has been a surge of interest in interaction estimation in quadratic regression. Roughly speaking, existing interaction estimation procedures can be classified into three categories. In the first category of low- or moderate-dimensional settings, standard techniques such as the ordinary least squares can be readily used to estimate all pairwise interactions and main effects. However, this simple one-stage strategy becomes impractical or even infeasible for moderate- or high-dimensional problems, owing to the rapid increase in the dimensionality caused by the interactions. In the second category of moderate- or high-dimensional settings, where feature selection becomes imperative, several one-stage regularization methods have been proposed, and some of which require either the strong or the weak heredity assumption; see, for example, Yuan et al. (2009), Choi et al. (2010), Bien et al. (2013), Lim and Hastie (2015), and Haris et al. (2016). These regularization methods are computationally feasible

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

and the theoretical properties of the resulting estimates are well understood for moderate- or high-dimensional problems. However, in the third category of ultrahigh-dimensional problems, these regularization methods are no longer feasible because their implementation requires storing and manipulating a large-scale design matrix and solving complex constrained optimization problems. The memory and computational costs are usually extremely expensive and prohibitive. Several two-stage approaches have been proposed for both ultrahigh-dimensional regression and classification problems, including Hao and Zhang (2014), Fan et al. (2015), Hao et al. (2018), and Kong et al. (2017). Two-stage approaches estimate the main effects and interactions in two separate stages, which significantly reduces their computational complexity. However, these two-stage approaches hinge heavily on either the strong or the weak heredity assumption. These methods are computationally scalable, but may break down when the heredity assumption is violated.

1.2 Heredity Assumption and Invariance Principle

Adding an extra layer of flexibility to linear models, quadratic regressions include both the main effects and the pairwise interactions between the covariates. Denote Y as the outcome variable and $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

as the covariate vector. For notational clarity, we define $\mathbf{u} \stackrel{\text{def}}{=} \mathrm{E}(\mathbf{x}) \in \mathbb{R}^p$. In general, a quadratic regression has the form

$$\mathrm{E}(Y \mid \mathbf{x}) = \alpha + (\mathbf{x} - \mathbf{u})^{\mathrm{T}}\boldsymbol{\beta} + (\mathbf{x} - \mathbf{u})^{\mathrm{T}}\boldsymbol{\Omega}(\mathbf{x} - \mathbf{u}), \tag{1.1}$$

where $\alpha \in \mathbb{R}^1$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$, and $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_{k,l})_{p\times p} \in \mathbb{R}^{p\times p}$ are all unknown parameters. To ensure model identifiability, we further assume that $\boldsymbol{\Omega}$ is symmetric; that is, $\boldsymbol{\Omega}^{\mathrm{T}} = \boldsymbol{\Omega}$. Our goal is to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$, which characterize the main effects and the interactions, respectively. The intercept α is also useful for prediction.

In the literature, heredity structures (Nelder, 1977; Hamada and Wu, 1992) are widely imposed to avoid the quadratic computational cost of searching over all pairs of interactions. The heredity structures assume that the support of $\boldsymbol{\Omega}$ can be inferred from the support of $\boldsymbol{\beta}$. The strong heredity assumption requires that an interaction between two covariates be included in the model only if both main effects are important. The weak assumption relaxes this constraint, stating that at least one main effect must be important. The strong and weak heredity structures are defined as follows:

$$\text{strong heredity:} \quad \boldsymbol{\Omega}_{k,l} \neq 0 \Rightarrow \beta_k^2 > 0 \text{ and } \beta_l^2 > 0,$$

$$\text{weak heredity:} \quad \boldsymbol{\Omega}_{k,l} \neq 0 \Rightarrow \beta_k^2 + \beta_l^2 > 0.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

Using the heredity assumptions, one can first seek a small number of important main effects, and then only consider interactions involving these discovered main effects. However, the main effects corresponding to important interactions may be difficult to detect. An example is $Y = (1+X_1)(1+X_2)+\varepsilon$, where X_1 and X_2 are drawn independently from $\mathcal{N}(-1, 1)$ and ε is standard normal. In this example, $\text{cov}(X_1, Y) = \text{cov}(X_2, Y) = 0$. The main effects X_1 and X_2 are thus unlikely to be detectable by a working linear model $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \epsilon$, indicating that the heredity assumptions do not necessarily facilitate finding interactions by first searching for the main effects. Ritchie et al. (2001) provided a real-data example to demonstrate the existence of pure interaction models in practice. Cordell (2009) raised concerns that many existing methods that depend on the heredity assumption may miss pure interactions in the absence of main effects.

An ideal quantification of the importance of the main effects and interactions should satisfy the invariance principle with respect to a location-scale transformation of the covariates. It is a natural and common strategy to quantify the importance of the main effects and interactions through the supports of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ in model (1.1). In a conventional linear model, where only the main effects are present and interactions are absent (i.e., $\boldsymbol{\Omega} = \mathbf{0}_{p\times p}$ in model (1.1)), the invariance principle is satisfied. In contrast,

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

in a quadratic regression (1.1) with a general $\mathbf{\Omega}$, the invariance principle is very likely violated. To demonstrate this issue, we recast model (1.1) as

$$\mathrm{E}(Y \mid \mathbf{x}) = (\alpha - \mathbf{u}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{u}^{\mathrm{T}}\mathbf{\Omega}\mathbf{u}) + \mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - 2\mathbf{\Omega}\mathbf{u}) + \mathbf{x}^{\mathrm{T}}\mathbf{\Omega}\mathbf{x}. \tag{1.2}$$

In this model, the importance of the main effects and interactions is naturally characterized through the support of $(\boldsymbol{\beta} - 2\mathbf{\Omega}\mathbf{u})$ and $\mathbf{\Omega}$, respectively. This indicates that the interactions are invariant, whereas the main effects are sensitive to a location transformation. The heredity condition and the invariance principle are discussed in detail by Hao and Zhang (2017). In an ultrahigh-dimensional quadratic regression, using a one-stage approach which simultaneously estimate the main effects and the interactions under the heredity assumption, or a two-stage approach which searches for main effects prior to searching for interactions, in model (1.1) or model (1.2) may lead to quite different conclusions. It is thus desirable to estimate the interactions directly, without knowing the main effects in advance. Direct interaction estimation without heredity constraints is, however, to the best of our knowledge, much more challenging and still unsolved in the literature. If both $\boldsymbol{\beta}$ and $\mathbf{\Omega}$ in model (1.1) are treated as random rather than fixed, then the strong heredity condition is satisfied almost surely. In this case, however, the main effects are too weak to be used to search for interactions.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

1.3 Our Contributions

We consider interaction estimation in ultrahigh-dimensional quadratic regressions without the heredity assumption. We make at least the following two important contributions to the literature:

1. We obtain a general and explicit expression for a quadratic regression with as minimal assumptions as possible. Surprisingly, it turns out that such an explicit solution relies only on certain moment conditions on the ultrahigh-dimensional covariates, which is satisfied by the widely used normality assumption. Explicit forms can be derived for both the main effects and the interactions. Thus, the quadratic regression can be implemented separately as two independent tasks relating to the main effects and interactions. Under weaker moment assumptions, our approach is still valid in detecting the direction of the true interactions. Our proposed method differs from existing one-step and two-step procedures in that we do not require the heredity assumption, and we give explicit forms for both the main effects and the interactions. Estimating the main effects through a separate working linear model ensures that the resulting estimate satisfies the desirable invariance principle. We show that our approach to detecting interactions is robust to the estimation of the main effects. Even

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

when the main effects are not estimated precisely, we are still able to detect the interactions accurately.

2. We show that the interaction inference is equivalent to a particular matrix estimation at the population level. We estimate the interactions in matrix form under a penalized convex loss function, which yields a sparse solution. We establish the consistency of our proposed estimates when the covariate dimension p grows, approximately, in an exponential order of the sample size n; specifically, $p = o\left\{\exp(ns_p^{-2})\right\}$, where s_p is the size of the underlying true model. Compared with the conventional penalized least squares approach, the penalization of the matrix form is appealing in terms of both memory storage and computation cost. An efficient algorithm is developed to implement our procedure. This algorithm fully exploits the cheap computational cost for the matrix multiplication, and is more efficient than existing penalized methods. For example, the algorithm can handle the case with $p = 10000$ covariates. The developed R package "PIE" is available at `https://github.com/cescwang85/PIE`. More details can be found in the package and the simulation part.

This remainder of this paper is organized as follows. We begin with a quadratic regression model in Section 2, and derive closed forms for both

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

the main effects and the interactions. We propose a direct penalized estimation for a high-dimensional sparse quadratic model. To implement the proposed method, an efficient alternating direction method of multipliers (ADMM) algorithm is provided. We also study the theoretical properties of our proposed estimates. We illustrate the performance of the proposed method, using simulations in Section 3, and an application to a real-world problem in Section 4. Section 5 concludes the paper. All technical details and additional simulations are deferred to the Supplementary Material.

The following notation is used throughout the remainder of the paper. For a real $p \times q$ matrix $\mathbf{A}_{p\times q} = (\mathbf{A}_{k,l})_{p\times q}$, let $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ denote its maximum and minimum singular values, respectively. Let $\|\mathbf{A}\|_F \stackrel{\text{def}}{=} \{\text{tr}(\mathbf{A}^{\text{T}}\mathbf{A})\}^{1/2}$ be the Frobenius norm, $\|\mathbf{A}\|$ be the spectral norm, and $\text{tr}(\cdot)$ be the trace operator of $\mathbf{A}$. We further define

$$\|\mathbf{A}\|_\infty \stackrel{\text{def}}{=} \max_{1\le k\le p, 1\le l\le q} |\mathbf{A}_{k,l}|, \|\mathbf{A}\|_1 \stackrel{\text{def}}{=} \sum_{k=1}^{p}\sum_{l=1}^{q} |\mathbf{A}_{k,l}|, \text{ and } \|\mathbf{A}\|_L \stackrel{\text{def}}{=} \max_{1\le k\le p} \sum_{l=1}^{q} |\mathbf{A}_{k,l}|.$$

2. THE ESTIMATION PROCEDURE

2.1 The Rationale

In this section, we discuss how to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$, which characterize the main effects and the interactions, respectively, in model (1.1). Note that $\boldsymbol{\beta} = E\{\partial E(Y \mid \mathbf{x})/(\partial \mathbf{x})\}$ and $\boldsymbol{\Omega} = E\{\partial^2 E(Y \mid \mathbf{x})/(\partial \mathbf{x}\partial \mathbf{x}^{\text{T}})\}/2$. There-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

fore, estimating $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ amounts to estimating $E\{\partial E(Y \mid \mathbf{x})/(\partial \mathbf{x})\}$ and $E\{\partial^2 E(Y \mid \mathbf{x})/(\partial \mathbf{x}\partial \mathbf{x}^{\mathrm{T}})\}$, respectively. However, this is not straightforward, especially when $\mathbf{x}$ is ultrahigh dimensional. To illustrate the rationale of our proposal, we assume for now that $\mathbf{x}$ follows $\mathcal{N}(\mathbf{u}, \boldsymbol{\Sigma})$. It follows immediately from Stein's Lemma (Stein, 1981; Li, 1992) that

$$E\{\partial E(Y \mid \mathbf{x})/(\partial \mathbf{x})\} = \boldsymbol{\Sigma}^{-1}\mathrm{cov}(\mathbf{x}, Y) \text{ and}$$

$$E\{\partial^2 E(Y \mid \mathbf{x})/(\partial \mathbf{x}\partial \mathbf{x}^{\mathrm{T}})\} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_y\boldsymbol{\Sigma}^{-1},$$

where $\boldsymbol{\Lambda}_y \stackrel{\text{def}}{=} E\Big[\{Y - E(Y)\}(\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^{\mathrm{T}}\Big]$. Define $r \stackrel{\text{def}}{=} Y - E(Y) - (\mathbf{x}-\mathbf{u})^{\mathrm{T}}\boldsymbol{\beta}$, which is the residual obtained by regressing Y on $\mathbf{x}$ linearly. The Hessians of $E(Y \mid \mathbf{x})$ and $E(r \mid \mathbf{x})$ are equal. Accordingly, we have

$$E\left\{\partial^2 E(Y \mid \mathbf{x})/(\partial \mathbf{x}\partial \mathbf{x}^{\mathrm{T}})\right\} = E\left\{\partial^2 E(r \mid \mathbf{x})/(\partial \mathbf{x}\partial \mathbf{x}^{\mathrm{T}})\right\}.$$

By Stein's Lemma, we can obtain that

$$E\left\{\partial^2 E(r \mid \mathbf{x})/(\partial \mathbf{x})(\partial \mathbf{x}^{\mathrm{T}})\right\} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_r\boldsymbol{\Sigma}^{-1},$$

where $\boldsymbol{\Lambda}_r \stackrel{\text{def}}{=} E\{r(\mathbf{x}-\mathbf{u})(\mathbf{x}-\mathbf{u})^{\mathrm{T}}\}$. This indicates that, if $\mathbf{x}$ is normal, we have explicit forms for $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$. Specifically,

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\mathrm{cov}(\mathbf{x}, Y), \text{ and } \boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}/2,$$

where $\boldsymbol{\Lambda}$ stands for either $\boldsymbol{\Lambda}_y$ or $\boldsymbol{\Lambda}_r$.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

Note that the normality assumption is widely used in the literature on interaction estimation; see, for example, Hao and Zhang (2014), Simon and Tibshirani (2015), Bien et al. (2015), and Hao et al. (2018). In the present context, we show that the normality assumption can be relaxed.

Proposition 1. Suppose $\mathbf{x}$ is drawn from the factor model $\mathbf{x} = \mathbf{\Gamma}_0\mathbf{z} + \mathbf{u}$, where $\mathbf{\Gamma}_0$ satisfies $\mathbf{\Gamma}_0\mathbf{\Gamma}_0^{\mathrm{T}} = \mathbf{\Sigma} > 0$ and $\mathbf{z} \stackrel{\text{def}}{=} (Z_1, \ldots, Z_q)^{\mathrm{T}}$, where $Z_1, \cdots, Z_q$ are independent and identically distributed (i.i.d.), with $\mathrm{E}(Z_k) = 0$, $\mathrm{E}(Z_k^2) = 1$, $\mathrm{E}(Z_k^3) = 0$ and $\mathrm{E}(Z_k^4) = \Delta$. We further assume either (C1): $\Delta = 3$ or (C2): $\mathrm{diag}(\mathbf{\Gamma}_0^{\mathrm{T}}\mathbf{\Omega}\mathbf{\Gamma}_0) = \mathbf{0}$. Then, the parameters α, $\boldsymbol{\beta}$, and $\mathbf{\Omega}$ in model (1.1) have the following explicit forms:

$$\alpha = E(Y) - \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{\Lambda})/2,\ \boldsymbol{\beta} = \mathbf{\Sigma}^{-1}\mathrm{cov}(\mathbf{x}, Y),\ \text{and}\ \mathbf{\Omega} = \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}/2. \quad (2.3)$$

The factor model is widely assumed in random matrix theory (Bai and Saranadasa, 1996) and high-dimensional inference (Chen et al., 2010), where higher-order moment assumptions of $\mathbf{x}$ are required. The moment conditions on $\mathbf{z}$ play an important role in deriving an explicit form for $\mathbf{\Omega}$. Condition (C1) is satisfied if $\mathbf{x}$ is normal. When $\mathbf{\Gamma}_0 = \mathbf{I}_{p\times p}$, condition (C2) requires the absence of quadratic terms of the form X_k^2 in model (1.1); that is,

$$\mathrm{E}(Y \mid \mathbf{x}) = \alpha + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{i\neq j} \mathbf{\Omega}_{i,j} X_i X_j,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

where $X_1, \cdots, X_p$ are i.i.d.

We provide two explicit forms for estimating $\mathbf{\Omega}$. One is based on the response Y, and the other is based on the residual r. The difference between $\mathbf{\Lambda}_y$ and $\mathbf{\Lambda}_r$ is that we remove the main effects in $\mathbf{\Lambda}_r$, or equivalently, the linear trend in model (1.1), before we estimate the interactions $\mathbf{\Omega}$. It is natural to expect that the residual-based $\mathbf{\Lambda}_r$ is superior to the response-based $\mathbf{\Lambda}_y$ in that the sample estimate of $\mathbf{\Lambda}_r$ has less variability than that of $\mathbf{\Lambda}_y$ (Cheng and Zhu, 2017). In effect, we can replace $\boldsymbol{\beta}$ with an arbitrary $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^p$, which yields $\widetilde{r} \stackrel{\text{def}}{=} Y - E(Y) - (\mathbf{x} - \mathbf{u})^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}$. Similarly, we can define $\mathbf{\Lambda}_{\widetilde{r}} \stackrel{\text{def}}{=} E\{\widetilde{r}(\mathbf{x} - \mathbf{u})(\mathbf{x} - \mathbf{u})^{\mathrm{T}}\}$. Under the normality assumption, $\mathbf{x}$ is symmetric about $\mathbf{u}$ and, hence, $\mathbf{\Lambda}_r = \mathbf{\Lambda}_{\widetilde{r}}$. This ensures that, to estimate $\mathbf{\Omega}$ accurately, the proposed method does not hinge on the sparsity of the main effects because we do not require $\boldsymbol{\beta}$ to be estimated consistently. Even if the main effects are not sufficiently sparse or are not estimated very accurately, we can directly use the response-based method $\mathbf{\Sigma}^{-1}\mathbf{\Lambda}_y\mathbf{\Sigma}^{-1}$ or the residual-based method $\mathbf{\Sigma}^{-1}\mathbf{\Lambda}_{\widetilde{r}}\mathbf{\Sigma}^{-1}$, which uses a lousy residual $\widetilde{r} = Y - E(Y) - (\mathbf{x} - \mathbf{u})^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}$, and $\widetilde{\boldsymbol{\beta}}$ can be a lousy estimate of $\boldsymbol{\beta}$. In effect $\mathbf{\Lambda}_y$ equals $\mathbf{\Lambda}_{\widetilde{r}}$ by setting $\widetilde{\boldsymbol{\beta}} = \mathbf{0}_{p\times 1}$ in $\widetilde{r}$. This makes the proposed method quite different from existing procedures, which assume the heredity conditions and require an accurate estimate of the main effects in order to recover the

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

interactions. In contrast, the proposed method does not require that we estimate the main effects precisely. We illustrate this phenomenon using simulations in Section 3.

2.2 Interaction Estimation

We showed that both $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ have explicit forms under moment conditions in Section 2.1. In particular, $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\text{cov}(\mathbf{x}, Y)$ and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}/2$ for $\boldsymbol{\Lambda}$ being $\boldsymbol{\Lambda}_y$ or $\boldsymbol{\Lambda}_r$. In this section, we discuss how to estimate $\boldsymbol{\Sigma}^{-1}\text{cov}(\mathbf{x}, Y)$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}$ at the sample level. Estimating $\boldsymbol{\Sigma}^{-1}\text{cov}(\mathbf{x}, Y)$ is straightforward by noting that it is a solution to the minimization problem

$$\arg\min_{\mathbf{b}} \text{E}\{Y - \text{E}(Y) - (\mathbf{x} - \mathbf{u})^{\text{T}}\mathbf{b}\}^2.$$

Therefore, we can simply estimate $\boldsymbol{\Sigma}^{-1}\text{cov}(\mathbf{x}, Y)$ using the penalized least squares by regressing $\{Y - \text{E}(Y)\}$ on the ultrahigh-dimensional covariates $(\mathbf{x} - \mathbf{u})$ linearly. We do not provide details about how to estimate $\boldsymbol{\Sigma}^{-1}\text{cov}(\mathbf{x}, Y)$ because the penalized least squares estimation is well documented (Tibshirani, 1996; Fan and Li, 2001). Throughout our numerical studies, we use the LASSO (Tibshirani, 1996) to estimate $\boldsymbol{\beta}$. The resulting solution is denoted by $\widehat{\boldsymbol{\beta}}$.

In what follows, we concentrate on how to estimate $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}/2$, where

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

$\mathbf{\Lambda}$ can be $\mathbf{\Lambda}_y$ or $\mathbf{\Lambda}_r$. For an arbitrary matrix $\mathbf{B} = (\mathbf{B}_{k,l})_{p\times p}$, we have

$$\begin{aligned}\mathbf{\Omega} &= \arg\min_{\mathbf{B}}\Big[\mathrm{tr}(\mathbf{B} - \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}/2)^{\mathrm{T}}(\mathbf{B} - \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}/2)\Big] \\ &= \arg\min_{\mathbf{B}}\Big[\mathrm{tr}(\mathbf{B} - \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}/2)^{\mathrm{T}}\mathbf{\Sigma}(\mathbf{B} - \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}/2)\mathbf{\Sigma}\Big],\end{aligned}$$

and

$$\begin{aligned}&\mathrm{tr}(\mathbf{B} - \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}/2)^{\mathrm{T}}\mathbf{\Sigma}(\mathbf{B} - \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}/2)\mathbf{\Sigma} \\ &= \mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{B}\mathbf{\Sigma}) - \mathrm{tr}(\mathbf{B}\mathbf{\Lambda}) + \mathrm{tr}(\mathbf{\Sigma}^{-2}\mathbf{\Lambda}^2)/4.\end{aligned}$$

Ignoring the constant, the term $\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{B}\mathbf{\Sigma}) - \mathrm{tr}(\mathbf{B}\mathbf{\Lambda})$ quantifies the distance between $\mathbf{B}$ and $\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}/2$. Therefore, to seek a $p\times p$ matrix $\mathbf{B}$ that can approximate $\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}/2$ very well, it suffices to consider the following minimization problem:

$$\arg\min_{\mathbf{B}}\Big[\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{B}\mathbf{\Sigma}) - \mathrm{tr}(\mathbf{B}\mathbf{\Lambda})\Big], \tag{2.4}$$

as long as we have faithful estimates of $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$. The above loss function in matrix form is convex, which guarantees that the local minimum must be a global minimum.

To construct faithful estimates for $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$, we suppose $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ is a random sample of $(\mathbf{x}, Y)$. Denote

$$\overline{\mathbf{x}} \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n}\mathbf{x}_i,\ \overline{Y} \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n}Y_i,\ \widehat{\mathbf{\Sigma}} \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}},$$

$$\widehat{\mathbf{\Lambda}}_y \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y})(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}, \widehat{\mathbf{\Lambda}}_r \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n}\widehat{r}_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

where $\widehat{r}_i \overset{\text{def}}{=} Y_i - \overline{Y} - (\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$. We propose the following penalized interaction estimation (PIE) to estimate $\boldsymbol{\Omega}$, for $\widehat{\boldsymbol{\Lambda}}$ being $\widehat{\boldsymbol{\Lambda}}_y$ or $\widehat{\boldsymbol{\Lambda}}_r$:

$$\text{PIE:} \quad \widehat{\boldsymbol{\Omega}} = \underset{\mathbf{B}\in\mathbb{R}^{p\times p}}{\arg\min}\, \mathrm{tr}(\mathbf{B}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\mathbf{B}\widehat{\boldsymbol{\Sigma}}) - \mathrm{tr}(\mathbf{B}\widehat{\boldsymbol{\Lambda}}) + \lambda_n\|\mathbf{B}\|_1, \tag{2.5}$$

where λ_n is a tuning parameter and $\|\mathbf{B}\|_1 = \sum_{k=1}^{p}\sum_{l=1}^{p}|\mathbf{B}_{k,l}|$. To ease subsequent illustration, we further define the following notation:

$$\text{PIE}_y: \quad \widehat{\boldsymbol{\Omega}}_y = \underset{\mathbf{B}\in\mathbb{R}^{p\times p}}{\arg\min}\, \mathrm{tr}(\mathbf{B}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\mathbf{B}\widehat{\boldsymbol{\Sigma}}) - \mathrm{tr}(\mathbf{B}\widehat{\boldsymbol{\Lambda}}_y) + \lambda_{1n}\|\mathbf{B}\|_1, \tag{2.6}$$

$$\text{PIE}_r: \quad \widehat{\boldsymbol{\Omega}}_r = \underset{\mathbf{B}\in\mathbb{R}^{p\times p}}{\arg\min}\, \mathrm{tr}(\mathbf{B}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\mathbf{B}\widehat{\boldsymbol{\Sigma}}) - \mathrm{tr}(\mathbf{B}\widehat{\boldsymbol{\Lambda}}_r) + \lambda_{2n}\|\mathbf{B}\|_1. \tag{2.7}$$

2.3 Implementation

In this section, we develop an efficient algorithm to solve (2.5), which includes (2.6) and (2.7) as special cases. We rewrite the problem as

$$\min_{\mathbf{B}\in\mathbb{R}^{p\times p}} \mathrm{tr}(\mathbf{B}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\mathbf{B}\widehat{\boldsymbol{\Sigma}}) - \mathrm{tr}(\mathbf{B}\widehat{\boldsymbol{\Lambda}}) + \lambda_n\|\boldsymbol{\Psi}\|_1, \text{ such that } \boldsymbol{\Psi} = \mathbf{B}, \tag{2.8}$$

which motivates us to form the augmented Lagrangian as

$$\begin{aligned} L(\mathbf{B}, \boldsymbol{\Psi}, \mathbf{L}) &= \mathrm{tr}(\mathbf{B}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\mathbf{B}\widehat{\boldsymbol{\Sigma}}) - \mathrm{tr}(\mathbf{B}\widehat{\boldsymbol{\Lambda}}) + \lambda_n\|\boldsymbol{\Psi}\|_1 \\ &\quad + \mathrm{tr}\left\{\mathbf{L}(\mathbf{B}-\boldsymbol{\Psi})\right\} + (\rho/2)\|\mathbf{B}-\boldsymbol{\Psi}\|_F^2, \end{aligned} \tag{2.9}$$

where $\rho > 0$ is a step size parameter. By the standard ADMM (Boyd et al., 2011), the augmented Lagrangian (2.9) can be solved by successively

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

updating $(\mathbf{B}, \mathbf{\Psi}, \mathbf{L})$:

$$\text{The } \mathbf{B} \text{ step:} \quad \mathbf{B}^{k+1} = \underset{\mathbf{B}\in\mathbb{R}^{p\times p}}{\arg\min}\, L(\mathbf{B}, \mathbf{\Psi}^k, \mathbf{L}^k), \tag{2.10}$$

$$\text{The } \mathbf{\Psi} \text{ step:} \quad \mathbf{\Psi}^{k+1} = \underset{\mathbf{\Psi}\in\mathbb{R}^{p\times p}}{\arg\min}\, L(\mathbf{B}^{k+1}, \mathbf{\Psi}, \mathbf{L}^k), \tag{2.11}$$

$$\text{The } \mathbf{L} \text{ step:} \quad \mathbf{L}^{k+1} = \mathbf{L}^k + \rho(\mathbf{B}^{k+1} - \mathbf{\Psi}^{k+1}). \tag{2.12}$$

Define the elementwise soft thresholding operator $\text{soft}(\mathbf{A}, \lambda) \overset{\text{def}}{=} \{\max(\mathbf{A}_{k,l} - \lambda, 0)\}_{p\times p}$. For the $\mathbf{\Psi}$ step, given $\mathbf{B}^{k+1}$, $\mathbf{L}^k$, ρ, and λ_n, the solution is

$$\mathbf{\Psi}^{k+1} \overset{\text{def}}{=} \text{soft}(\mathbf{B}^{k+1} + \rho^{-1}\mathbf{L}^k, \lambda_n/\rho).$$

The $\mathbf{B}$ step amounts to solving the equation

$$2\widehat{\mathbf{\Sigma}}\mathbf{B}^{k+1}\widehat{\mathbf{\Sigma}} + \rho\mathbf{B}^{k+1} = \mathbf{\Lambda}^k, \tag{2.13}$$

where $\mathbf{\Lambda}^k \overset{\text{def}}{=} \widehat{\mathbf{\Lambda}} - \mathbf{L}^k + \rho\mathbf{\Psi}^k$. We make a singular value decomposition to obtain $\widehat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{D}_0\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U} \in \mathbb{R}^{p\times m}$, $m = \min(n, p)$, and $\mathbf{D}_0 \overset{\text{def}}{=} \text{diag}(d_1, \cdots, d_m)$ is a diagonal matrix. Define $\mathbf{D} \overset{\text{def}}{=} (\mathbf{D}_{k,l})_{p\times p}$, where $\mathbf{D}_{k,l} \overset{\text{def}}{=} 2d_kd_l/(2d_kd_l + \rho)$. Given $\mathbf{\Psi}^k$, $\mathbf{L}^k$, and ρ, the solution to (2.13) is given by

$$\mathbf{B}^{k+1} = \rho^{-1}\mathbf{\Lambda}^k - \rho^{-1}\mathbf{U}\{\mathbf{D} \circ (\mathbf{U}^{\mathrm{T}}\mathbf{\Lambda}^k\mathbf{U})\}\mathbf{U}^{\mathrm{T}},$$

where $\circ$ denotes the Hadamard product.

The algorithm is summarized in Algorithm 1. This algorithm yields a symmetric estimate of $\mathbf{\Omega}$, which is denoted by $\widehat{\mathbf{\Omega}}$. The computational complexity of each iteration is no more than $\text{O}\{\min(n, p)p^2\}$, and the memory

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

requirement is no more than $O(p^2)$ because we only need to store a few $p \times p$ or $p \times \min(n, p)$ matrices in computer memory.

As a first-order method for convex problems, the convergence analysis of the ADMM algorithm under various conditions has been well documented in the recent optimization literature; see, for example, Nishihara et al. (2015), Hong and Luo (2017), and Chen et al. (2017). The following lemma states the convergence of our proposed ADMM algorithm.

Lemma 1. Given $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Lambda}}$, suppose that the ADMM algorithm (2.10)-(2.12) generates a sequence of solutions $\{(\mathbf{B}^k, \mathbf{\Psi}^k, \mathbf{L}^k), k = 1, \ldots\}$. Then, $\{(\mathbf{B}^k, \mathbf{\Psi}^k), k = 1, \ldots\}$ converges linearly to the minimizer of (2.8), and $\|\mathbf{B}^k - \mathbf{\Psi}^k\|_F$ converges linearly to zero, as $k \to \infty$.

It remains to choose an appropriate tuning parameter for PIE_y or PIE_r. Motivated by Efron et al. (2004), we use PIE to find a sparse model, but not to estimate the coefficients. For a given λ_n, we fit a least squares estimation on the support of $\widehat{\mathbf{\Omega}}$ estimated by PIE_y or PIE_r, which yields the residual sum of squares. We choose λ_n that minimizes the Bayesian information criterion (BIC). Our limited experience indicates that this procedure is very fast and effective.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

Algorithm 1 ADMM algorithm for solving (2.5)

Initialization:

1: Input $\{(\mathbf{x}_i, Y_i), i = 1, \cdots, n\}$, the tuning parameter λ_n and ρ;

2: Calculate $\widehat{\mathbf{\Lambda}}$ and the singular value decomposition of the centered design matrix $(\mathbf{x}_1 - \overline{\mathbf{x}}, \cdots, \mathbf{x}_n - \overline{\mathbf{x}})_{p\times n}$ to get $\widehat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{D}_0\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U} \in \mathbb{R}^{p\times m}$, $\mathbf{D}_0 = \mathrm{diag}\{d_1, \ldots, d_m\}$, and $m = \min(n, p)$;

3: Define $\mathbf{D} \stackrel{\mathrm{def}}{=} (\mathbf{D}_{k,l})_{m\times m}$, where $\mathbf{D}_{k,l} = 2d_k d_l/(2d_k d_l + \rho)$;

4: Start from $k = 0$, $\mathbf{L}^0 = \mathbf{0}_{p\times p}, \mathbf{B}^0 = \mathbf{0}_{p\times p}$.

Iteration:

5: Define $\mathbf{\Lambda}^k \stackrel{\mathrm{def}}{=} \widehat{\mathbf{\Lambda}} - \mathbf{L}^k + \rho\mathbf{B}^k$. Update $\mathbf{B}^{k+1} = \rho^{-1}\mathbf{\Lambda}^k - \rho^{-1}\mathbf{U}\{\mathbf{D} \circ (\mathbf{U}^{\mathrm{T}}\mathbf{\Lambda}^k\mathbf{U})\}\mathbf{U}^{\mathrm{T}}$;

6: Update $\mathbf{\Psi}^{k+1} \stackrel{\mathrm{def}}{=} \mathrm{soft}(\mathbf{B}^{k+1} + \rho^{-1}\mathbf{L}^k, \lambda_n/\rho)$;

7: Update $\mathbf{L}^{k+1} \stackrel{\mathrm{def}}{=} \mathbf{L}^k + \rho(\mathbf{B}^{k+1} - \mathbf{\Psi}^{k+1})$;

8: Update $k = k + 1$;

9: Repeat step 5 through step 8 until convergence.

Output: $\widehat{\mathbf{\Omega}} = \mathbf{B}^{k+1}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

2.4 Asymptotic Properties

For notational clarity, we denote the support of $\mathbf{\Omega} = (\mathbf{\Omega}_{k,l})_{p\times p}$ by $\mathcal{S} \stackrel{\text{def}}{=} \{(k,l) : \mathbf{\Omega}_{k,l} \neq 0\}$, the complement of $\mathcal{S}$ by $\mathcal{S}^c$, and the cardinality of $\mathcal{S}$ by $s_p \stackrel{\text{def}}{=} \|\mathbf{\Omega}\|_0$. Similarly, we denote by $\widehat{\mathcal{S}}_y$ and $\widehat{\mathcal{S}}_r$ the respective support of $\widehat{\mathbf{\Omega}}_y$ and $\widehat{\mathbf{\Omega}}_r$, and by $\widehat{\mathcal{S}}_y^c$ and $\widehat{\mathcal{S}}_r^c$ the respective complement of $\widehat{\mathcal{S}}_y$ and $\widehat{\mathcal{S}}_r$. We further define $\mathbf{\Gamma} \stackrel{\text{def}}{=} \mathbf{\Sigma} \otimes \mathbf{\Sigma}$, $M \stackrel{\text{def}}{=} \|\mathbf{\Gamma}_{\mathcal{S},\mathcal{S}}^{-1}\|_L$, and $\kappa \stackrel{\text{def}}{=} 1 - \|\mathbf{\Gamma}_{\mathcal{S}^c,\mathcal{S}}\mathbf{\Gamma}_{\mathcal{S},\mathcal{S}}^{-1}\|_L$, where $\mathbf{\Gamma}_{\mathcal{S},\mathcal{S}}$ is a submatrix of $\mathbf{\Gamma}$ with rows and columns indexed by $\mathcal{S}$, and $\mathbf{\Gamma}_{\mathcal{S}^c,\mathcal{S}}$ is a submatrix of $\mathbf{\Gamma}$ with rows and columns indexed respectively by $\mathcal{S}^c$ and $\mathcal{S}$. Denote $c_0, C_0, c_1, C_1, \ldots$ as a sequence of generic constants, which may take different values at various places. We assume the following regularity conditions in order to study the asymptotic properties of $\widehat{\mathbf{\Omega}}_y$ and $\widehat{\mathbf{\Omega}}_r$.

(A1): Assume $c_0^{-1} \leq \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) \leq c_0$, where $\lambda_{\min}(\mathbf{\Sigma})$ and $\lambda_{\max}(\mathbf{\Sigma})$ are the respective smallest and largest eigenvalues of $\mathbf{\Sigma}$.

(A2): Assume X_k is sub-Gaussian; that is, $\mathrm{E}\{\exp(c_0|\mathbf{e}^{\mathrm{T}}\mathbf{x}|^2)\} \leq C_0 < \infty$, for any unit-length vector $\mathbf{e}$.

(A3) Assume $\mathrm{E}\{\exp(c_1|Y|^\alpha)\} \leq C_1 < \infty$, for some $0 < \alpha \leq 2$.

(A4) Assume the irrepresentability condition holds; that is, $\kappa > 0$.

(A5) Assume $\mathbf{x}$ is symmetric about $\mathbf{u}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

Conditions (A1) and (A2) are widely assumed in the literature. Condition (A3) is assumed to control the tail behavior of Y through concentration inequalities. Condition (A4) is analog, but not identical, to the irrepresentability condition used to establish model selection consistency in the LASSO. In our context, the irrepresentability condition is imposed on $\mathbf{\Gamma} \stackrel{\text{def}}{=} \mathbf{\Sigma} \otimes \mathbf{\Sigma}$ because the interaction effects are included. This condition is also used to study the model consistency of the graphical LASSO (Ravikumar et al., 2011; Zhang and Zou, 2014; Liu and Luo, 2015). In contrast, if the linear effects are of primary interest, the irrepresentability condition is imposed on $\mathbf{\Sigma}$; see, for example, Zhao and Yu (2006) and Zou (2006). Condition (A5) is assumed to ensure the consistency of residual-based approaches.

Theorem 1. Let $\lambda_{1n} \stackrel{\text{def}}{=} c_1\{n^{-\alpha/(\alpha+1)}\log(p)\}^{1/2}$ for sufficiently large c_1, and assume $s_p\{n^{-1}\log(p)\}^{1/2} \to 0$. Under conditions (A1)-(A4),

(i) if we further assume $\min_{(k,l)\in\mathcal{S}} |\mathbf{\Omega}_{k,l}| > c_2 M\lambda_{1n}$ for sufficiently large c_2, then $\text{pr}\big(\widehat{\mathcal{S}}_y = \mathcal{S}\big) = 1 - O(p^{-1})$.

(ii) $\text{pr}\big(\|\widehat{\mathbf{\Omega}}_y - \mathbf{\Omega}\|_\infty \le c_3\lambda_{1n}M\big) = 1 - O(p^{-1})$, for sufficiently large c_3.

(iii) $\text{pr}\big(\|\widehat{\mathbf{\Omega}}_y - \mathbf{\Omega}\|_F \le c_4 s_p^{1/2}\lambda_{1n}M\big) = 1 - O(p^{-1})$, for sufficiently large c_4.

Theorem 1 shows that, as long as the signal strength of the interactions

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

is not too small, the proposed method can identify the support correctly with a very high probability. In other words, $\widehat{\boldsymbol{\Omega}}_y$ is asymptotically selection consistent. Theorem 1 also shows that $\widehat{\boldsymbol{\Omega}}_y$ is a consistent estimate of $\boldsymbol{\Omega}$ under both the infinity norm and the Frobenius norm.

Theorem 2. Let $\lambda_{2n} \stackrel{\text{def}}{=} c_5\{n^{-\alpha/(\alpha+1)}\log(p)\}^{1/2} + c_5\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1\{\log(p)/n\}^{1/2}$ for sufficiently large c_5, and assume that $s_p\{n^{-1}\log(p)\}^{1/2} \to 0$. Under the conditions (A1)-(A5),

(i) if we further assume $\min_{(k,l)\in\mathcal{S}} |\boldsymbol{\Omega}_{k,l}| > c_6 M\lambda_{2n}$ for sufficiently large c_6, then $\text{pr}\big(\widehat{\mathcal{S}}_r = \mathcal{S}\big) = 1 - O(p^{-1})$.

(ii) $\text{pr}\big(\|\widehat{\boldsymbol{\Omega}}_r - \boldsymbol{\Omega}\|_\infty \le c_7\lambda_{2n}M\big) = 1 - O(p^{-1})$, for sufficiently large c_7.

(iii) $\text{pr}\big(\|\widehat{\boldsymbol{\Omega}}_r - \boldsymbol{\Omega}\|_F \le c_8 s_p^{1/2}\lambda_{2n}M\big) = 1 - O(p^{-1})$, for sufficiently large c_8.

Theorem 2 shows that $\widehat{\boldsymbol{\Omega}}_r$, as well as $\widehat{\boldsymbol{\Omega}}_y$, possesses both selection and estimation consistency asymptotically. Moreover, the convergence rate of $\widehat{\boldsymbol{\Omega}}_r$ depends on $\widehat{\boldsymbol{\beta}}$. If $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 = o\{n^{1/(2\alpha+2)}\}$, the convergence rate term involving $\widehat{\boldsymbol{\beta}}$ will be absorbed in the first term of Theorem 2. In other words, unless the estimation error of $\widehat{\boldsymbol{\beta}}$ diverges faster than $n^{1/(2\alpha+2)}$, $\widehat{\boldsymbol{\Omega}}_r$ and $\widehat{\boldsymbol{\Omega}}_y$ share the same convergence rate.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

3. SIMULATIONS

In this section, we conduct simulations to evaluate the performance of our proposed method and compare it with that of the RAMP method (Hao et al., 2018) and the all-pairs-LASSO (Bien et al., 2013), which fits a LASSO model on all p main effects and $p(p+1)/2$ interactions. Hao et al. (2018) claimed that the RAMP outperforms other methods, such as the iFOR (Hao and Zhang, 2014) and hierNet (Bien et al., 2013), under heredity assumptions. Therefore, we do not include iFOR and hierNet in our comparison. In what follows, we refer to the RAMP method under the strong heredity condition as "RAMPs," and that under the weak heredity condition as "RAMPw." We also include the oracle estimate as a benchmark, which assumes that the main effects and the support of the interactions are known in advance. The oracle estimate simply fits the least squares estimation on the support of the interactions using the truly important main effects. We denote it as "Oracle." The RAMP and all-pairs-LASSO methods are implemented by the R packages "RAMP" and "glmnet," (Friedman et al., 2010) respectively.

To ease illustration, we denote the estimate of $\mathbf{\Omega}$ by $\widehat{\mathbf{\Omega}}$, obtained with different approaches. We evaluate the accuracy of the estimation using five criteria: the support recovery rate, denoted by "rate"; the Frobenius

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

loss, denoted by "loss"; the number of interactions estimated as nonzero, denoted by "size"; and the exact support recovery rate, denoted by "exact." Specifically, the criteria are defined as follows:

$$\begin{aligned}
\text{rate} &\stackrel{\text{def}}{=} B^{-1}\sum_{b=1}^{B}\sum_{l\le k} I(\widehat{\mathbf{\Omega}}_{k,l}^{(b)}\neq 0, \mathbf{\Omega}_{k,l}\neq 0)\Big/\sum_{l\le k} I(\mathbf{\Omega}_{k,l}\neq 0)\times 100\%,\\
\text{loss} &\stackrel{\text{def}}{=} B^{-1}\sum_{b=1}^{B}\|\widehat{\mathbf{\Omega}}^{(b)}-\mathbf{\Omega}\|_F,\ \text{size}\stackrel{\text{def}}{=} B^{-1}\sum_{b=1}^{B}\sum_{l\le k} I(\widehat{\mathbf{\Omega}}_{k,l}^{(b)}\neq 0),\ \text{and}\\
\text{exact} &\stackrel{\text{def}}{=} B^{-1}\sum_{b=1}^{B} I\big(\widehat{\mathcal{S}}^{(b)}=\mathcal{S}\big),
\end{aligned}$$

where $\mathcal{S}$ and $\widehat{\mathcal{S}}$ are the supports of $\mathbf{\Omega}$ and $\widehat{\mathbf{\Omega}}$, respectively, the superscript $^{(b)}$ stands for the bth replication, the subscript $_{k,l}$ stands for the (k,l)th entry of the associated matrix, $I(E)$ is an indicator function, equal to one if the random event E is true, and zero otherwise. The closer "rate" is to one, "loss" is to zero, "size" is to the number of truly important interactions, and "exact" is to one, the better the performance a proposal exhibits.

We consider the following four models:

$$Y = X_1+X_6+X_{10}+2X_1X_6+X_6^2+2X_6X_{10}+\varepsilon, \tag{3.14}$$

$$Y = X_6+2X_1X_6+X_6^2+2X_6X_{10}+\varepsilon, \tag{3.15}$$

$$Y = X_1+X_2+2X_1X_6+X_6^2+2X_6X_{10}+\varepsilon, \tag{3.16}$$

$$Y = 2X_1X_6+X_6^2+2X_6X_{10}+\varepsilon. \tag{3.17}$$

The strong heredity condition holds in (3.14), and the weak heredity con-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

dition holds in (3.15). Neither the strong nor the weak heredity condition holds in (3.16) or (3.17). In particular, (3.17) is a pure interaction model. We replicate each scenario $B = 100$ times to evaluate the performance of the proposals.

3.1 Estimation Accuracy

We draw $\mathbf{x}$ independently from $\mathcal{N}(\mathbf{0}_{p\times 1}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (0.5^{|k-l|})_{p\times p}$, and generate an independent error ε from $\mathcal{N}(0, 1)$. We set the sample size $n = 200$ and the dimension $p = 100$ or 200.

The simulation results for models (3.15) and (3.17) are shown in Table 1; those for models (3.14) and (3.16) are shown in Table 1 of the Supplementary Material. The proposed method exhibits stable performance across almost all scenarios. Not surprisingly, the RAMP method with the strong heredity condition, denoted by RAMPs, completely fails in models (3.15)-(3.17), where the strong heredity condition is violated. In addition, the RAMP method with weak heredity condition, denoted by RAMPw, fails in models (3.16)-(3.17), where the weak heredity condition is violated. The RAMP method exhibits satisfactory performance when the required heredity condition is satisfied. In particular, RAMPs performs quite well in model (3.14). For models (3.15)-(3.17), the oracle estimate has the smallest

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

Frobenius loss, followed by our proposed method. Our method outperforms the all-pairs-LASSO in terms of the Frobenious loss and model size. For the pure interaction model (3.17), where no main effects are present, fitting a linear regression to obtain the residuals very likely introduces some redundant bias. It is thus not surprising to see that our proposed response-based procedure (PIEy) slightly outperforms our residual-based procedure (PIEr).

3.2 Estimation of Main Effects

In this section, we evaluate how the estimation of the main effects affects the estimation of the interactions. Both our proposed residual-based penalized interaction estimation and the RAMP method are relevant to estimating the main effects. To fix the signal-to-noise ratio for all settings, we simply draw the covariates $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{p\times p})$, and consider the following quadratic model:

$$\begin{aligned} Y = \; & d^{-1/2}\left(X_1 + X_6 + X_{10} + X_{k_1} + \cdots + X_{k_{d-3}}\right) \\ & + 2X_1X_6 + X_6^2 + 2X_6X_{10} + \varepsilon. \end{aligned}$$

The number of main effects is increased from $d = 3$ to 48. We always include X_1, X_6, and X_{10} to ensure that the strong heredity condition holds. We also randomly choose $X_{k_1}, \ldots, X_{k_{d-3}}$ from $X_{11}, \ldots, X_p$. Figure 1 reports

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

Table 1: The averages (and standard deviations) of the support recovery rate ("rate"), Frobenius loss ("loss"), model size ("size"), and exact support recovery rate ("exact") for models (3.15) and (3.17). Simulation results for models (3.14) and (3.16) are given in the Supplementary Material.

p		PIEy	PIEr	RAMPs	RAMPw	all-pairs-LASSO	Oracle
model (3.15) where the weak heredity condition is satisfied							
100	rate	98.67(6.56)	99.33(4.69)	40.67(26.20)	91.33(27.88)	100.00(0.00)	100.00(0.00)
	size	4.19(3.03)	3.73(2.11)	1.67(1.51)	3.91(3.39)	7.38(6.40)	3.00(0.00)
	loss	0.24(0.20)	0.18(0.16)	1.81(0.55)	0.30(0.68)	0.41(0.11)	0.09(0.05)
	exact	0.57(0.50)	0.70(0.46)	0.06(0.24)	0.75(0.43)	0.16(0.37)	1.00(0.00)
200	rate	99.00(5.71)	99.00(5.71)	29.67(23.16)	77.67(41.32)	100.00(0.00)	100.00(0.00)
	size	3.99(2.44)	3.42(1.08)	1.12(1.22)	4.09(4.34)	6.08(4.59)	3.00(0.00)
	loss	0.23(0.21)	0.19(0.19)	1.98(0.40)	0.62(0.97)	0.45(0.11)	0.09(0.04)
	exact	0.65(0.48)	0.73(0.45)	0.03(0.17)	0.68(0.47)	0.29(0.46)	1.00(0.00)
model (3.17) is a pure interaction model where the heredity conditions are violated							
100	rate	99.67(3.33)	100.00(0.00)	11.67(24.33)	31.67(44.79)	100.00(0.00)	100.00(0.00)
	size	4.18(4.24)	4.24(4.22)	0.71(1.58)	3.00(4.92)	5.57(3.86)	3.00(0.00)
	loss	0.13(0.12)	0.13(0.09)	2.11(0.41)	1.64(1.01)	0.42(0.11)	0.09(0.04)
	exact	0.72(0.45)	0.72(0.45)	0.03(0.17)	0.23(0.42)	0.27(0.45)	1.00(0.00)
200	rate	100.00(0.00)	100.00(0.00)	9.67(20.26)	24.33(41.26)	100.00(0.00)	100.00(0.00)
	size	3.45(1.00)	3.49(0.99)	0.51(1.21)	2.95(5.49)	5.46(5.27)	3.00(0.00)
	loss	0.11(0.06)	0.12(0.07)	2.15(0.21)	1.78(0.91)	0.44(0.11)	0.09(0.04)
	exact	0.72(0.45)	0.69(0.46)	0.00(0.00)	0.18(0.39)	0.45(0.50)	1.00(0.00)

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

the support recovery rate of $\widehat{\mathbf{\Omega}}$ and the Frobenius loss of $\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_F$.

It can be clearly seen that, as the number of main effects increases from $d = 3$ to 48, both RAMPs and RAMPw deteriorate gradually in terms of both criteria, indicating that the RAMP method requires an accurate estimate of the main effects. For the all-pairs-LASSO, the support recovery rate appears very stable, while the Frobenius loss becomes worse when d increases. In contrast, our method is very robust to the number of main effects. When the number of main effects increases, PIEy is slightly better than PIEr in terms of the Frobenius loss. These findings confirm our theoretical results in Theorem 2 because $\widehat{\boldsymbol{\beta}}$ becomes worse when d increases.

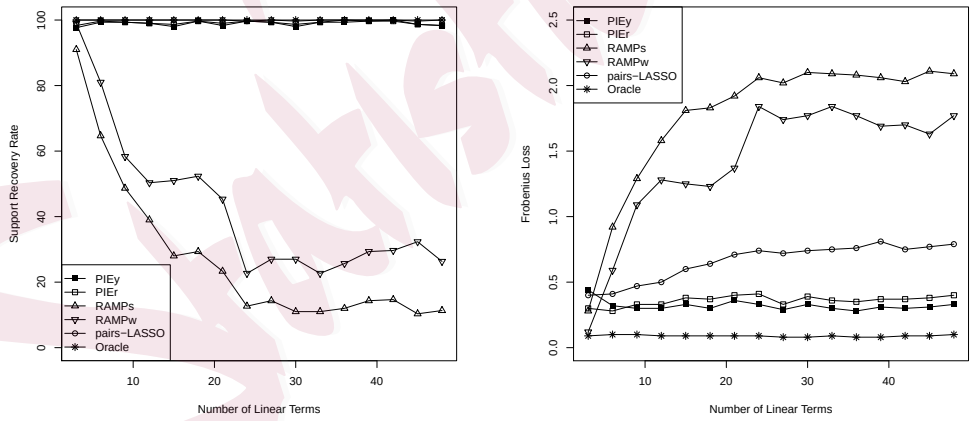


Figure 1: *The vertical axis is the support recovery rate (left) and the Frobenius loss (right) of* $\widehat{\mathbf{\Omega}}$*, and the horizontal axis is the number of main effects.*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

4. AN APPLICATION

In this section, we apply our method to the red wine data set that is publicly available at `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`. The data consist of 11 measurements of several chemical constituents, including the determination of the density, alcohol, or pH values for 1599 red wine samples from the northwest region of Portugal. The response variable is the median of the scores evaluated by human experts, and each score ranges from 0 (very bad) to 10 (very excellent). The same data set was analyzed by Cortez et al. (2009). In their analysis, interactions are found to be very helpful for prediction. The sample size $n = 1599$ and the covariate dimension $p = 11$. Following Radchenko and James (2010), we standardize all variables and conduct two experiments:

- **Experiment 1.** In addition to the original 11 covariates $X_1, \ldots, X_{11}$, we add 100 noise variables $X_{12}, \ldots, X_{111}$, among which the first 50 are generated from the standard normal distribution, and the rest are generated from the uniform distribution on the interval $[-\sqrt{3}, \sqrt{3}]$.

- **Experiment 2.** We generate the covariates in the same way as in Experiment 1 and modify the response variable Y by adding two more interactions: $Y + 0.5X_{12}X_{13} + 0.5X_{61}X_{62}$. In this experiment, both

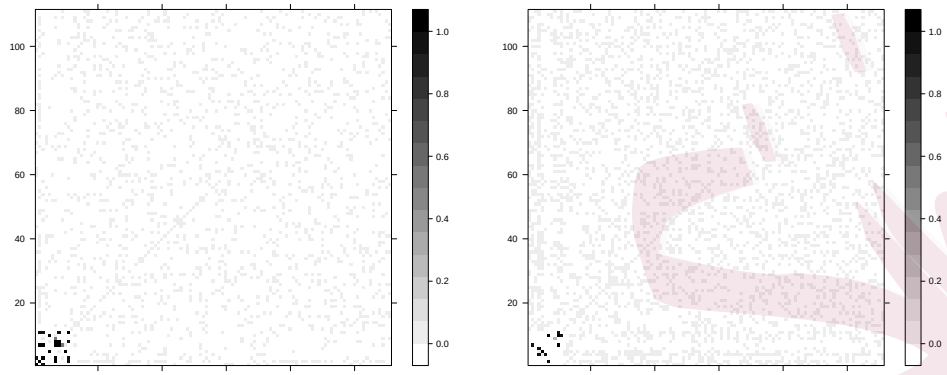
Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

the strong and the weak heredity conditions are violated.

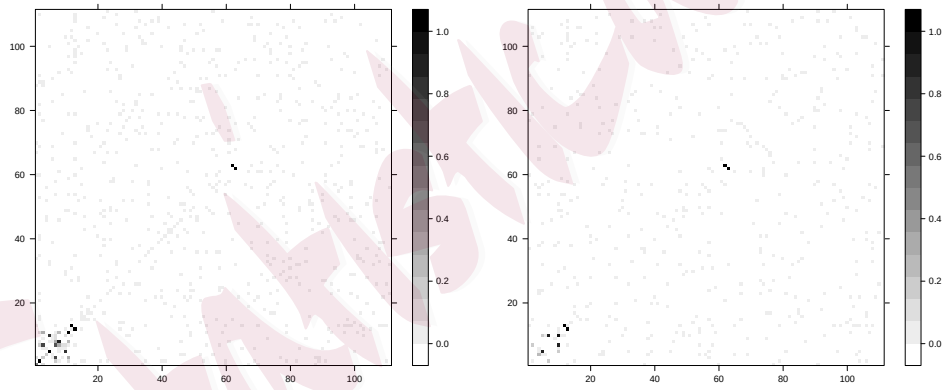
In both experiments the covariate dimension is updated to $p = 111$, leading to $111 \times 100/2 = 6,105$ possible interactions. We randomly select 400 observations as the sample, and the procedure is repeated 100 times. The heat map of the frequencies of the identified interactions are summarized in Figure 2. It can be clearly seen that, in Experiment 1, the selected interactions mainly occur among the first 11 covariates collected in the original data set while the interactions related to the remaining 100 noisy covariates are rarely detected. This indicates that both PIEy and PIEr are able to exclude irrelevant interactions. In Experiment 2, both methods are able to exclude irrelevant interactions with high probability. In addition, the interactions $X_{12}X_{13}$ and $X_{61}X_{62}$ are successfully identified throughout.

We further compare our proposed PIEy and PIEr with the all-pairs-LASSO in terms of prediction. We randomly split the observations into two halves. We use the first half as a training sample and the second as a test sample. We fit quadratic regressions using the training sample and perform a prediction using the test sample. To implement the PIEy and PIEr, we follow Example 1 and generate 100 additional noise covariates. To implement the all-pairs-LASSO, we use the original 11 covariates only. We record the averages of the squared prediction errors for each random

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081



(a) PIEy (b) PIEr



(c) PIEy (d) PIEr

Figure 2: *Heat maps of the frequency of the interactions identified out of 100 replications using PIEy and PIEr. Upper panel: Experiment 1. Lower panel: Experiment 2.*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

split. Table 2 summarizes the mean and standard deviation of the squared prediction errors and the model sizes based on 100 replications. Compared with the all-pairs-LASSO, which includes around seven interactions, the PIEy and PIEr both include fewer than four interactions and yield models that are more parsimonious. In terms of the prediction performance, the PIEy and PIEr are both comparable with the all-pairs-LASSO method.

Table 2: The prediction performance on the red wine data set. The PIEy and the PIEr are both fitted with 100 additional noise covariates, while the all-pairs-LASSO is fitted with only 11 original covariates.

	PIEy	PIEr	all-pairs-LASSO
prediction error	0.706(0.035)	0.702(0.034)	0.671(0.032)
model size	3.600(1.980)	3.640(1.580)	7.020(1.880)

5. CONCLUSION

In this paper, we have proposed a penalized estimation for detecting interactions without requiring heredity conditions. We developed an efficient ADMM algorithm to implement our estimation, and demonstrated the effectiveness of our method using numerical studies. Note that if the strong

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

or the weak heredity condition is satisfied, some existing methods, such as the RAMP method, may work relatively well, as long as the main effects are sufficiently strong. However, if the main effects are too weak to be detectable, the performance of existing methods that require the heredity assumptions may deteriorate. Our proposed method is highly robust to a violation of the heredity assumptions, because the estimation of the interaction is separable from the estimation of the main effect. Even with a lousy estimate of the main effects, we are still able to estimate the interactions consistently. When we have little prior information about whether the heredity condition holds in an application, we advocate using our method because it does not require this assumption. If the heredity condition is known to be satisfied, we can incorporate this into our method through a two-stage procedure. In the first stage, we use the penalized least squares to identify the main effects. In the second stage, we implement our procedure using only the main effects selected in the first stage. This allows us to handle ultrahigh-dimensional problems efficiently. Moreover, it would be interesting to combine our proposed method with screening procedures such as the SIRI (Jiang and Liu, 2014) in order to further improve its estimation efficiency.

Supplementary Material

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

The online Supplementary Material contains the technical proofs and additional simulation results.

Acknowledgments

We thank the Editor, Associate Editor, and two anonymous reviewers for their insightful comments. We also thank Dr. Yang Feng for his helpful discussions on the R package "RAMP." Wang's research was supported by the National Natural Science Foundation of China (11701367, 11825104) and Shanghai Sailing Program (16YF1405700). Jiang's research was supported by a PolyU Grant (153038/17P) and the Hong Kong RGC/ECS (PolyU253023/16P). Zhu is the corresponding author and his research was supported by the Natural Science Foundation of Beijing (Z19J00009) and National Natural Science Foundation of China (11731011, 11931014).

References

Bai, Z. and H. Saranadasa (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica 6*(2), 311–329.

Bien, J., N. Simon, and R. Tibshirani (2015). Convex hierarchical testing of interactions. *The Annals of Applied Statistics 9*(1), 27–42.

Bien, J., J. Taylor, and R. Tibshirani (2013). A lasso for hierarchical inter-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

actions. *The Annals of Statistics 41*(3), 1111.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning 3*(1), 1–122.

Chen, L., D. Sun, and K.-C. Toh (2017). A note on the convergence of admm for linearly constrained convex optimization problems. *Computational Optimization and Applications 66*(2), 327–343.

Chen, S., L. Zhang, and P. Zhong (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association 105*(490), 810–819.

Cheng, Q. and L. Zhu (2017). On relative efficiency of principal hessian directions. *Statistics & Probability Letters 126*, 108–113.

Choi, N. H., W. Li, and J. Zhu (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association 105*(489), 354–364.

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature reviews. Genetics 10*(6), 392.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

Cortez, P., A. Cerdeira, F. Almeida, T. Matos, and J. Reis (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems 47*(4), 547 – 553.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–499.

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

Fan, Y., Y. Kong, D. Li, and Z. Zheng (2015). Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics 43*(3), 1243–1272.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1.

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

Hao, N., Y. Feng, and H. H. Zhang (2018). Model selection for high-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

dimensional quadratic regression via regularization. *Journal of the American Statistical Association 113*(522), 615–625.

Hao, N. and H. H. Zhang (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association 109*(507), 1285–1301.

Hao, N. and H. H. Zhang (2017). A note on high-dimensional linear regression with interactions. *The American Statistician 71*(4), 291–297.

Haris, A., D. Witten, and N. Simon (2016). Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics 25*(4), 981–1004.

Hong, M. and Z.-Q. Luo (2017). On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming 162*(1-2), 165–199.

Jiang, B. and J. S. Liu (2014). Variable selection for general index models via sliced inverse regression. *The Annals of Statistics 42*(5), 1751–1786.

Kong, Y., D. Li, Y. Fan, and J. Lv (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics 45*(2), 897–922.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association 87*(420), 1025–1039.

Lim, M. and T. Hastie (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics 24*(3), 627–654.

Liu, W. and X. Luo (2015). Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of Multivariate Analysis 135*, 153–162.

Nelder, J. A. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society, Series A 140*(1), 48–77.

Nishihara, R., L. Lessard, B. Recht, A. Packard, and M. I. Jordan (2015). A general analysis of the convergence of admm. *arXiv preprint arXiv:1502.02009*.

Radchenko, P. and G. James (2010). Variable selection using adaptive non-linear interaction structures in high dimensions. *Journal of the American Statistical Association 105*(492), 1541–1553.

Ravikumar, P., M. Wainwright, G. Raskutti, and B. Yu (2011). High-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics 5*, 935–980.

Ritchie, M. D., L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics 69*(1), 138–147.

Simon, N. and R. Tibshirani (2015). A permutation approach to testing interactions for binary response by comparing correlations between classes. *Journal of the American Statistical Association 110*(512), 1707–1716.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics 9*(6), 1135–1151.

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

Yuan, M., R. Joseph, and H. Zou (2009). Structured variable selection and estimation. *The Annals of Applied Statistics 3*(4), 1738–1757.

Zhang, T. and H. Zou (2014). Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika 101*(1), 103–120.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0081

REFERENCES

Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research 7*(Nov), 2541–2563.

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

School of Mathematical Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai 200240, China.

E-mail: chengwang@sjtu.edu.cn

Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong.

E-mail: by.jiang@polyu.edu.hk

Institute of Statistics and Big Data and Center for Applied Statistics, Renmin University of China, Beijing 100872, China.

E-mail: zhu.liping@ruc.edu.cn