Journal of Machine Learning Research 22 (2021) 1-32 Submitted 9/19; Revised 6/21; Published 8/21

Method of Contraction-Expansion (MOCE) for Simultaneous Inference in Linear Models

Fei Wang WAFEI@UMICH.EDU
CarGurus, Cambridge, MA 02141, USA and Tencent, Shenzhen, Guangdong 518057, China

Ling Zhou ZHOULING@SWUFE.EDU.CN
Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China

Lu Tang LUTANG@PITT.EDU
University of Pittsburgh, Pittsburgh, PA 15261, USA

Peter X.K. Song PXSONG@UMICH.EDU
University of Michigan, Ann Arbor, MI 48109, USA

Editor: Jie Peng

Abstract

Simultaneous inference after model selection is of critical importance to address scientific hypotheses involving a set of parameters. In this paper, we consider a high-dimensional linear regression model in which a regularization procedure such as LASSO is applied to yield a sparse model. To establish a simultaneous post-model selection inference, we propose a method of contraction and expansion (MOCE) along the line of debiasing estimation in that we investigate a desirable trade-off between model selection variability and sample variability by the means of forward screening. We establish key theoretical results for the inference from the proposed MOCE procedure. Once the expanded model is properly selected, the theoretical guarantees and simultaneous confidence regions can be constructed by the joint asymptotic normal distribution. In comparison with existing methods, our proposed method exhibits stable and reliable coverage at a nominal significance level and enjoys substantially less computational burden. Thus, our MOCE approach is trustworthy in solving real-world problems.

Keywords: Debiasing, Forward screening, LASSO, Simultaneous inference.

1. Introduction

We consider the linear model with a response vector $\mathbf{y} = (y_1, ..., y_n)^T$ and an $n \times p$ design matrix X,

$$\mathbf{y} = X\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \cdots, \beta_p^*)^T \in \mathbb{R}^p$ denotes an unknown p-dimensional true regression coefficients, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is an n-dimensional *i.i.d.* random errors with mean zero and variance $\sigma^2 I_n$, where I_n is the $n \times n$ identity matrix. All columns in X are normalized to have mean zero and ℓ_2-norm 1. The sample covariance matrix of the p predictors and its corresponding population covariance matrix are denoted by $S = \frac{1}{n} X^T X$ and Σ, respectively. Let $\mathcal{A} = \{j : \beta_j^* \neq 0, j = 1, \ldots, p\}$ be the support of $\boldsymbol{\beta}^*$ with cardinality $s_0 = |\mathcal{A}|$. In this paper, assuming $p \to \infty$ as $n \to \infty$, we focus on simultaneous statistical inferences on a subset of $\boldsymbol{\beta}^*$ when $p \gg n$.

Arguably, in the setting of $p \gg n$, a simultaneous inference for the entire set of p parameters, i.e. $\boldsymbol{\beta}^*$, is generally not tractable due to the issue of model identification. A key assumption widely adopted in the current literature to facilitate statistical inference is the sparsity of $\boldsymbol{\beta}^*$, namely $s_0 \ll n$, in addition to some regularity conditions on the design matrix; see for example Meinshausen (2015); van de Geer et al. (2014); Zhang and Zhang (2014), among others. The sparsity assumption of the true signals necessitates variable selection, which has been extensively studied in the past two decades or so. Being one of the most celebrated variable selection methods, Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) has gained great popularity in both theory and applications. Specifically, a LASSO estimator is obtained by minimizing the following penalized objective function:

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big(\frac{1}{2n}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\Big), \tag{1.2}$$

where $\|\cdot\|_r$ is the ℓ_r-norm of a vector, $r = 1, 2$, and $\lambda > 0$ is the tuning parameter. Based on LASSO estimator, $\hat{\boldsymbol{\beta}}_\lambda$, given in (1.2), statistical inferences for parameters in $\boldsymbol{\beta}^*$ in the aspects of hypothesis test and confidence region construction have recently received considerable attention in the literature because statistical inference has been always playing a central role in the statistical theory and providing one of the most effective ways for the transition from data to knowledge.

Some progresses in post-model selection inferences have been reported in the literature. The method LASSO+mLS proposed in Liu and Yu (2013) first performs LASSO model selection and then draws statistical inferences based on the selected model. This approach requires model selection consistency and some incoherence conditions on the design matrix (Zhao and Yu, 2006; Meinshausen and Yu, 2009; Bühlmann and van de Geer, 2011). Inference procedures built upon those strong conditions have been noted as being impractical and exhibited poor performances due to the lack of uniform validity of inferential procedures over sequences of models; see for example, Leeb and Pötscher (2008); Chernozhukov et al. (2015).

To overcome the reliance on the oracle asymptotic distribution in inference, many solutions have been proposed in recent years. Among those, three methods are so far known for a valid post-model selection inference. (i) The first kind is sample splitting method (Wasserman and Roeder, 2009; Meinshausen et al., 2009; Meinshausen and Bühlmann, 2010) and resampling method (Minnier et al., 2011). A key drawback of the sample splitting method is its requirement of a random split of the data and the results may be sensitive to the sample splitting, while the resampling approach entails a strong restrictive exchangeability condition on the design matrix. (ii) The second kind is group inference proposed in Meinshausen (2015). Unfortunately, this approach fails to show desirable power to detect individual signals, and thus it is not useful in practical studies. (iii) The third kind is low-dimensional projection (LDP) (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014). Such an inferential method is rooted in a seminal idea of debiasing, resulting from the use of penalized objective function that causes estimation shrinkage. This method will be adopted in this paper for a new paradigm of post-model selection inference. Following the debiasing approach proposed by Zhang and Zhang (2014), Cai and Guo (2017) investigates both adaptivity and minimax rate of the debiased estimation, which provides useful insights on the rate of model contraction and expansion considered in this paper. Specifically, an LDP estimator, $\hat{\boldsymbol{b}}$, takes a debiasing step under an operation of this form: $\hat{\boldsymbol{b}} = \hat{\boldsymbol{\beta}}_\lambda + \frac{1}{n}\hat{\Theta}X^T(\mathbf{y} - X\hat{\boldsymbol{\beta}}_\lambda)$, where $\hat{\Theta}$ is a sparse estimate of precision matrix Σ^{-1}. When the matrix $\hat{\Theta}$ is properly constructed, the bias term, $\Delta = \sqrt{n}(\hat{\Theta}S - I_p)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$, would become asymptotically negligible. In this case, statistical inference can be conducted using the above debiased estimator $\hat{\boldsymbol{b}}$. It is known that obtaining a desirable $\hat{\Theta}$ is not a trivial task due to the singularity of sample covariance S. For examples, van de Geer et al. (2014) proposes to use node-wise LASSO to get $\hat{\Theta}$, while Javanmard and Montanari (2014) adopts a convex optimization algorithm to obtain $\hat{\Theta}$. It is worth noting that the number of parameters estimated in $\hat{\Theta}$ using the approaches proposed by van de Geer et al. (2014) and Javanmard and Montanari (2014) is p^2, which can be computationally expensive. In a setting similar to that of the LDP estimator, Zhang and Cheng (2017) proposes a bootstrap-based simultaneous inference for a group, *say* $\mathcal{G}$, of parameters in $\boldsymbol{\beta}^*$ via the distribution of quantity $\max_{j\in\mathcal{G}} \sqrt{n}|\hat{b}_j - \beta^*_j|$, where the bootstrap resampling, unfortunately, demands much more computational power than a regular LDP estimator based on the node-wise LASSO estimate $\hat{\Theta}$.

Overcoming the excessive computational cost on acquiring $\hat{\Theta}$ motivates us to consider a ridge type of approximation to the precision matrix Σ^{-1}, in a similar spirit to the approach proposed by Ledoit and Wolf (2004) for the estimation of a high-dimensional covariance matrix. Note that the LASSO estimator $\hat{\boldsymbol{\beta}}_\lambda$ satisfies the following Karush-Kuhn-Tucker (KKT) condition:

$$-\frac{1}{n}X^T\boldsymbol{\epsilon} + S(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*) + \lambda\boldsymbol{\kappa} = \mathbf{0}, \tag{1.3}$$

where $\boldsymbol{\kappa} = (\kappa_1, \cdots, \kappa_p)^T$ is the subdifferential of $\|\hat{\boldsymbol{\beta}}_\lambda\|_1$ whose jth component is $\kappa_j = 1$ if $\hat{\beta}_{\lambda,j} > 0$, $\kappa_j = -1$ if $\hat{\beta}_{\lambda,j} < 0$, and $\kappa_j \in [-1, 1]$ if $\hat{\beta}_{\lambda,j} = 0$. Let $\boldsymbol{\tau}$ be a $p \times p$ diagonal matrix $\mathrm{diag}(\tau_1, \cdots, \tau_p)$ with all positive element $\tau_j > 0$, $j = 1, \cdots, p$. We propose to add a term $\boldsymbol{\tau}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)$, and then multiply $\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}$ on the both sides of (1.3), leading to an equivalent expression of (1.3),

$$-\frac{1}{n}\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}X^T\boldsymbol{\epsilon} + \left\{(\hat{\boldsymbol{\beta}}_\lambda + \lambda\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\kappa}) - \boldsymbol{\beta}^*\right\} - \hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*) = \mathbf{0}, \tag{1.4}$$

where $\hat{\boldsymbol{\beta}}_\lambda + \lambda\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\kappa}$ defines a debiased estimator, with $\hat{\Sigma}_{\boldsymbol{\tau}} = S + \boldsymbol{\tau}$ being a ridge-type sample covariance matrix. It is easy to see that on the basis of (1.4), establishing a valid inference on $\boldsymbol{\beta}^*$ becomes straightforward if (i) $\hat{\Sigma}_{\boldsymbol{\tau}}$ is nonsingular and (ii) the bias term $\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)$ is asymptotically negligible under a properly tuned matrix $\boldsymbol{\tau}$. The associated technical treatments are of theoretical interest but methodologically challenging. To address such challenges, we propose a new approach, with some mild regularity conditions, termed as Method of Contraction and Expansion (MOCE).

MOCE offers a practically feasible way to perform a valid simultaneous post-model selection inference under suitable conditions, in which the ridge type matrix $\boldsymbol{\tau}$ is properly tuned to establish desirable theoretical guarantees. As seen later in the paper, the ridge matrix $\boldsymbol{\tau}$ plays a key role in determining the length of confidence intervals, which can vary according to signal strengths as desired. We provide checkable conditions for the magnitude of $\boldsymbol{\tau}$ in MOCE to efface the estimation bias asymptotically so to obtain the valid large-sample theory for post-model selection inferences. Following a suggestion given by one of the reviewers, we implement a cross-validation procedure to select $\boldsymbol{\tau}$, which appears to work well in our numerical studies. When $\boldsymbol{\tau}$ is properly determined, to achieve proper coverage, MOCE is able to provide a wider confidence interval for a signal parameter, while a narrower one for a null signal parameter. This is because a null signal is known with zero coefficient (i.e., no need for estimation once being identified), whereas a signal is only known with non-zero coefficient, whose parameter value needs to be further estimated in order to construct its confidence interval. Thus, the latter estimation step incurs extra variability in inference, resulting naturally in a wider confidence interval. MOCE takes on an expanded model, denoted by $\tilde{\mathcal{A}}$, that is enlarged from an initial model selected by the LASSO estimator, denoted by $\hat{\mathcal{A}}$, so that $\hat{\mathcal{A}} \subseteq \tilde{\mathcal{A}}$ surely. Implementing the idea of model expansion is practically feasible. In this paper, we adopt an intelligent model expansion procedure along the line of the forward screening method proposed by Wang (2009) to construct an expanded model. With such an expanded model, we rewrite the original KKT condition, where the precision matrix Σ^{-1} is estimated accordingly. Under the sparsity assumption $s_0 = o(n/\log p)$ and some additional regularity conditions, the bias term in (1.4) vanishes asymptotically, and consequently the confidence region for a set of regression parameters is readily constructed in the paradigm of MOCE.

This paper makes new contributions to the following five domains. (i) MOCE is established under different sparsity conditions required for valid simultaneous inference in comparison to those given in the current literature. That is, MOCE assumes the sparsity condition $s_0 = o(n/\log p)$, instead of the popular sup-sparsity assumption, $s_0 = o(\sqrt{n}/\log p)$; also, MOCE does not demand additional sparsity assumptions required by the node-wise LASSO to obtain sparse estimate of the precision matrix. (ii) MOCE is shown to achieve a smaller error bound in terms of mean squared error (MSE) in comparison to the seminal LDP debiasing method. In effect, MOCE estimator has the MSE rate $\|\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} - \boldsymbol{\beta}^*\|_2 = O_p(\sqrt{\tilde{s}\log(\tilde{s})/n})$ with $\tilde{s}$ being the size of the expanded model, lower than $O_p(\sqrt{p/n})$, the rate of the LDP estimator. (iii) MOCE enjoys both reproducibility and numerical stability in inferences because the model expansion leaves little ambiguity for post-selection inference as opposed to many existing methods based on a selected model that may vary substantially due to different tuning procedures (Berk et al., 2013). (iv) MOCE is advantageous for its fast computation, because of the use of the ridge-type matrix inverse, which is known to be conceptually simple and computationally efficient. It is shown that the computational complexity of MOCE is of order $O(n(p-\tilde{s})^2)$, in comparison to the order $O(2np^2)$ of the LDP method. (v) MOCE enables us to construct a new simultaneous test similar to the classical Wald test for a set of parameters based on its asymptotic normal distribution. The proposed hypothesis test method is computationally superior to the bootstrap-based test (Zhang and Cheng, 2017) based on the sup-norms of individual estimation errors. All these improvements above make MOCE ready to be applied in real-world applications.

The rest of the paper is organized as follows. Section 2 introduces notation and Section 3 provides preliminary results that are used in the proposed method. In Section 4 we introduce MOCE and discuss its general theoretical properties. Section 5 introduces specific algorithms for model expansion along the line of the forward screening, as well as some discussion of computational complexity. Through simulation experiments, Section 6 illustrates performances of MOCE, with comparison to existing methods. Section 7 contains some concluding remarks. Some lengthy technical proofs are included in the Appendix.

2. Notation

For a vector $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_p)^T \in \mathbb{R}^p$, the ℓ_0-norm is $\|\boldsymbol{\nu}\|_0 = \sum_{j=}^{p} 1\{|\nu_j| > 0\}$; the ∞-norm is $\|\boldsymbol{\nu}\|_\infty = \max\limits_{1 \le j \le p} |\nu_j|$; and the ℓ_2-norm is $\|\boldsymbol{\nu}\|_2^2 = \sum_{j=1}^{p} \nu_j^2$. For a $p \times p$ matrix $W = (w_{ij})_{1 \le i \le j \le p} \in \mathbb{R}^{p \times p}$, the ∞-norm is $|W|_\infty = \max\limits_{1 \le j, j' \le p} |w_{jj'}|$ and the Frobenius norm is $\|W\|_F^2 = \mathrm{tr}(W^T W)$ where $\mathrm{tr}(W)$ is the trace of matrix W. Let $\rho^+_{\min}(W)$ and $\rho^+_{\max}(W)$ denote the smallest and largest nonzero singular values of a positive semi-definite matrix W (Horn and Johnson, 2012), respectively.

With a given index subset $\mathcal{B} \subseteq \{1, \ldots, p\}$, vector $\boldsymbol{\nu} \in \mathbb{R}^p$ and matrix $W \in \mathbb{R}^{p \times p}$ can be partitioned as $\boldsymbol{\nu} = (\boldsymbol{\nu}_{\mathcal{B}}^T,\ \boldsymbol{\nu}_{\mathcal{B}^c}^T)^T$ and $W = \left(\begin{smallmatrix} W_{\mathcal{B}\mathcal{B}} & W_{\mathcal{B}\mathcal{B}^c} \\ W_{\mathcal{B}^c\mathcal{B}} & W_{\mathcal{B}^c\mathcal{B}^c} \end{smallmatrix}\right)$. For two positive definite matrices W_1 and W_2, their Löewner order $W_1 \succ W_2$ indicates that $W_1 - W_2$ is positive definite. For two sequences of real numbers $\{u_n\}$ and $\{v_n\}$, the expression $u_n \asymp v_n$ means that there exist positive constants c and C such that $c \le \liminf_n (u_n/v_n) \le \limsup_n (u_n/v_n) \le C$.

For the self-containedness, we introduce the compatibility condition and sparse eigenvalue condition that are widely adopted in the literature; refer to Bickel et al. (2009) and Bühlmann and van de Geer (2011) for more details. For a given subset $\mathcal{J} \subseteq \{1, \cdots, p\}$ and a constant $k \ge 1$, define the following subspace $\mathbb{R}(\mathcal{J}, k)$ in $\mathbb{R}^p$: $\mathbb{R}(\mathcal{J}, k) = \{\boldsymbol{\nu} \in \mathbb{R}^p : \|\boldsymbol{\nu}_{\mathcal{J}^c}\|_1 \le k\|\boldsymbol{\nu}_{\mathcal{J}}\|_1\}$. A sample covariance matrix $S = \frac{1}{n} X^T X$ is said to satisfy the compatibility condition if for $1 \le s \le p$ and $k > 0$ there exists a constant $\phi_0 > 0$ such that

$$\min_{\substack{\mathcal{J} \subset \{1,\ldots,p\} \\ |\mathcal{J}| \le s}} \ \min_{\boldsymbol{\nu} \in \mathbb{R}(\mathcal{J},k)} \frac{\|X\boldsymbol{\nu}\|_2^2 s}{n\|\boldsymbol{\nu}\|_1^2} \ge \phi_0. \tag{2.1}$$

A sample covariance matrix S is said to satisfy the sparse eigenvalue $SE(s)$ condition if for any $\boldsymbol{\nu} \in \mathbb{R}^p$ with $1 \le \|\boldsymbol{\nu}\|_0 \le s$, it holds that

$$0 < \lambda_{\min}(s) \le \lambda_{\max}(s) < \infty \tag{2.2}$$

where

$$\lambda_{\min}(s) = \min_{1 \le \|\boldsymbol{\nu}\|_0 \le s} \frac{\|X\boldsymbol{\nu}\|_2^2}{n\|\boldsymbol{\nu}\|_2^2},\ \lambda_{\max}(s) = \max_{1 \le \|\boldsymbol{\nu}\|_0 \le s} \frac{\|X\boldsymbol{\nu}\|_2^2}{n\|\boldsymbol{\nu}\|_2^2}.$$

3. Preliminary Results

As discussed above, when the bias term $\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)$ in (1.4) is asymptotically negligible, the modified KKT (1.4) enables us to establish an asymptotic distribution for the proposed debiased estimator of the form:

$$\tilde{\boldsymbol{\beta}}_{\boldsymbol{\tau}} = \hat{\boldsymbol{\beta}}_\lambda + \lambda\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\kappa}. \tag{3.1}$$

This section presents some finite-sample bounds for the estimation bias when the ridge-type covariance matrix is invoked for the matrix inversion. We begin with the first regularity condition on the design matrix X given as follows.

Assumption 1 *The design matrix X in the linear model* (1.1) *satisfies the compatibility condition in* (2.1) *for $k = 3$ and $s = s_0$, where s_0 is the number of true signals.*

Assumption 1 is routinely assumed for the design matrix X in a high-dimensional linear model; see for example, Bickel et al. (2009); Zhang and Zhang (2014), among others. It is primarily used to ensure the ℓ_1-norm convergence for the LASSO estimator $\hat{\boldsymbol{\beta}}_\lambda$ (Bühlmann and van de Geer, 2011); that is, $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*\|_1 = O_p\left(s_0\sqrt{\frac{\log p}{n}}\right)$.

Lemma 3.1 below assesses both the Frobenius norm and ∞-norm of the factor $\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}$, a key term in the estimation bias $\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)$.

Lemma 3.1 *Consider the sample covariance $S = \frac{1}{n}X^TX$. Let the ridge matrix $\boldsymbol{\tau} = diag(\tau_1, \cdots, \tau_p)$ with $\tau_j > 0$ for $j = 1, \cdots, p$, $\tau_{\min} = \min_{1\le j\le p}\tau_j$ and $\tau_{\max} = \max_{1\le j\le p}\tau_j$. Let $\hat{\Sigma}_{\boldsymbol{\tau}} = S + \boldsymbol{\tau}$. Then, the Frobenius norm and ∞-norm of $\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}$ are given as follows, respectively:*

$$\begin{aligned}\max(p-n,0) + \frac{\min(n,p)}{\{\rho^+_{max}(\boldsymbol{\tau}^{-1/2}S\boldsymbol{\tau}^{-1/2})+1\}^2} &\le \|\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}\|_F^2 \\ &\le \max(p-n,0) + \frac{\min(n,p)}{\{\rho^+_{min}(\boldsymbol{\tau}^{-1/2}S\boldsymbol{\tau}^{-1/2})+1\}^2};\end{aligned}$$

$$\frac{\tau_{\min}}{\rho^+_{max}(S)+\tau_{\max}} \le |\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}|_\infty \le \begin{cases} \frac{\tau_{\max}}{\tau_{\min}}, & \text{if } p > n; \\ \frac{\tau_{\max}}{\rho^+_{min}(S)+\tau_{\min}}, & \text{if } p \le n. \end{cases}$$

The proof of Lemma 3.1 is given in Appendix A.1. According to Lemma 3.1, when $p \le n$, it is interesting to note that the ∞-norm $|\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}|_\infty$ is bounded above by $\frac{\tau_{\max}}{\rho^+_{\min}(S)+\tau_{\min}}$. This upper bound may converge to 0 if $\tau_{\max} = o(1)$ and $\rho^+_{\min}(S) = O(1)$. On the other hand, when $p > n$, its upper bound is $\tau_{\max}/\tau_{\min}$, which is always greater than or equal to 1. Hence, when $p < n$ the bias term $\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*)$ can be controlled by an appropriately small $\boldsymbol{\tau}$, leading to a simultaneous inference on $\boldsymbol{\beta}$ by the means of debiasing. In contrast, the case "$p > n$" presents the difficulty of bias reduction for $\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}$. Such insight motivates us to seek for an alternative solution in the framework of post-model selection inference, resulting in our proposed MOCE.

The proposed MOCE mimics the well-known physical phenomenon of thermal contraction and expansion for materials with the tuning parameter λ being an analog to temperature. Specifically, MOCE reduces LASSO estimation bias in two steps as shown in Figure 1. In the step of contraction, LASSO selects a model $\hat{\mathcal{A}}$, represented by the small circle in Figure 1, which may possibly miss some true signals contained in the set $\mathcal{A}$. In the step of expansion, MOCE enlarges $\hat{\mathcal{A}}$ to form an expanded model $\tilde{\mathcal{A}}$, indicated by the large circle in Figure 1. As a result, we have surely $\hat{\mathcal{A}} \subseteq \tilde{\mathcal{A}}$, and hope that under some mild conditions, with probability approaching to 1, the set of true signals $\mathcal{A}$ is completely contained by the expanded model $\tilde{\mathcal{A}}$. In other words, MOCE begins with an initial model $\hat{\mathcal{A}}$ through the LASSO regularization which contains most of true signals, and then expands $\hat{\mathcal{A}}$ into a bigger model $\tilde{\mathcal{A}}$ that with probability approaching to 1, embraces all true signals in $\mathcal{A}$. Refer to

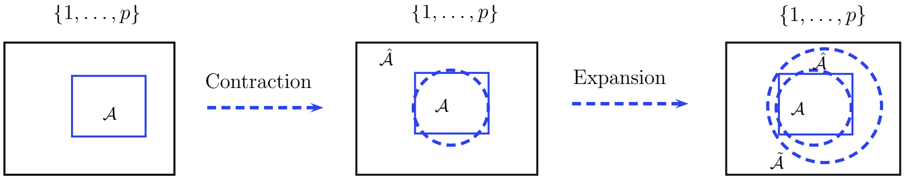

Figure 1: A schematic diagram for MOCE. The inner and outer rectangles respectively represent the true model $\mathcal{A}$ and the full model with all p predictors $\{1,\ldots,p\}$. The small and large circles denote the LASSO selected model $\hat{\mathcal{A}}$ and the expanded model $\tilde{\mathcal{A}}$, respectively.

Section 5 where specific conditions and schemes are discussed to carry out the model expansion. The key advantage of the model expansion step is to greatly mitigate the influence of subjective decision on variable selection in post-model selection inference. In effect, it is very hard, if it is not possible, to quantify this type of uncertainty incurred by human actions, which in addition to the sampling uncertainty, is needed for a valid post-model selection inference.

We now introduce notations necessary for a further discussion on the model expansion. Let $\hat{\mathcal{A}} = \{j : |\hat{\beta}_{\lambda,j}| > 0,\ j = 1,\cdots,p\}$ be a LASSO selected model, whose cardinality is denoted by $\hat{s} = |\hat{\mathcal{A}}|$. Here, both $\hat{\mathcal{A}}$ and $\hat{s}$ are dependent on the tuning parameter λ, which is suppressed in the rest of this paper for the sake of simplicity, unless necessary. Similarly, let $\tilde{\mathcal{A}}$ be an expanded model from $\hat{\mathcal{A}}$ with cardinality denoted by $\tilde{s} = |\tilde{\mathcal{A}}|$. In this paper, our model expansion scheme ensures $\hat{\mathcal{A}} \subseteq \tilde{\mathcal{A}}$ surely. Given $\mathcal{A}$ and $\tilde{\mathcal{A}}$, a model expansion leads to four disjoint subsets of predictors, $\mathcal{A} \cap \tilde{\mathcal{A}}$, $\mathcal{A} \cap \tilde{\mathcal{A}}^c$, $\mathcal{A}^c \cap \tilde{\mathcal{A}}$ and $\mathcal{A}^c \cap \tilde{\mathcal{A}}^c$. Among these subsets, two are of primary interest, namely, set $\mathcal{B}_{fn}$ of false negatives and set $\mathcal{B}_{tn}$ of true negatives. They are defined by, respectively,

$$\mathcal{B}_{fn} = \mathcal{A} \cap \tilde{\mathcal{A}}^c, \quad \mathcal{B}_{tn} = \mathcal{A}^c \cap \tilde{\mathcal{A}}^c. \tag{3.2}$$

Let their cardinalities be $b_{fn} = |\mathcal{B}_{fn}|$ and $b_{tn} = |\mathcal{B}_{tn}|$, respectively. $\mathcal{B}_{fn}$ collects signals missed by the expanded model $\tilde{\mathcal{A}}$ (i.e., false negatives), while $\mathcal{B}_{tn}$ collects all null signals that the expanded model $\tilde{\mathcal{A}}$ does not contain (i.e., true negatives).

For the expanded model $\tilde{\mathcal{A}}$, we introduce Assumption 2, which is the regularity condition for the expanded model.

Assumption 2 *Let $\mathcal{A}$ and $\tilde{\mathcal{A}}$ be the set of true signals and the set of variables selected in the model expansion, respectively. The expanded model $\tilde{\mathcal{A}}$ satisfies the following regularity conditions:*

(a) $P(\mathcal{A} \subset \tilde{\mathcal{A}}) \to 1$ as $n \to \infty$;

(b) Matrix X in model (1.1) *satisfies the sparse eigenvalue $SE(\tilde{s})$ condition in* (2.2) *where $\tilde{s} = |\tilde{\mathcal{A}}|$.*

Assumption 2.(a) is a condition of expansion consistency, which may hold if the expanded model $\tilde{\mathcal{A}}$ is constructed by some well-behaved variable screening methods that are of screening consistency, such as the forward screening method by Wang (2009). Thus, this is a mild condition, which is partially controlled by the choice of a certain model expansion strategy. Assumption 2.(b) requires that any $\tilde{s} \times \tilde{s}$ main diagonal sub-matrices of the sample covariance matrix $S = X^T X/n$ has finite positive minimum and maximum singular values, so that the selected expanded model has a well-defined Hessian matrix. This condition is similar to that given by Zhang and Zhang (2014) in that the sparse eigenvalue condition is assumed for the initial LASSO model $\hat{\mathcal{A}}$ for their LDP inference. This condition is natural in MOCE because we use a larger model for inference. We will revisit Assumption 2 in Section 5.1 when a specific model expansion scheme is considered, where we show that this regularity condition holds under some checkable and interpretable conditions.

4. Method of Contraction and Expansion (MOCE)

In this section we first present the definition of MOCE estimator, and then establish several key large-sample properties for the proposed MOCE estimator under an expanded model $\tilde{\mathcal{A}}$ that satisfies Assumptions 1-2. Specific algorithms for the model expansion will be discussed later in Section 5.

4.1 MOCE

An expanded model $\tilde{\mathcal{A}}$ is said to be *viable* if it satisfies Assumptions 1-2. For a viable expanded model $\tilde{\mathcal{A}}$, we partition a LASSO estimator $\hat{\boldsymbol{\beta}}_\lambda$ given in (1.2) as $\hat{\boldsymbol{\beta}}_\lambda = (\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}}^T, \hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c}^T)^T$. Rewrite the KKT condition (1.3) according to this partition, respectively, for $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{A}}^c$:

$$-\frac{1}{n} X_{\tilde{\mathcal{A}}}^T (\mathbf{y} - X_{\tilde{\mathcal{A}}} \hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}} - X_{\mathcal{B}_{fn}} \hat{\boldsymbol{\beta}}_{\mathcal{B}_{fn}} - X_{\mathcal{B}_{tn}} \hat{\boldsymbol{\beta}}_{\mathcal{B}_{tn}}) + \lambda \boldsymbol{\kappa}_{\tilde{\mathcal{A}}} = \mathbf{0}, \tag{4.1}$$

$$-\frac{1}{n} X_{\tilde{\mathcal{A}}^c}^T (\mathbf{y} - X_{\tilde{\mathcal{A}}} \hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}} - X_{\tilde{\mathcal{A}}^c} \hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c}) + \lambda \boldsymbol{\kappa}_{\tilde{\mathcal{A}}^c} = \mathbf{0}. \tag{4.2}$$

It follows from (4.1) that

$$S_{\tilde{\mathcal{A}}\mathcal{B}_{fn}} (\hat{\boldsymbol{\beta}}_{\mathcal{B}_{fn}} - \boldsymbol{\beta}^*_{\mathcal{B}_{fn}}) + S_{\tilde{\mathcal{A}}\mathcal{B}_{tn}} \hat{\boldsymbol{\beta}}_{\mathcal{B}_{tn}} - \frac{1}{n} X_{\tilde{\mathcal{A}}}^T \boldsymbol{\epsilon} + S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} (\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}}) + \lambda \boldsymbol{\kappa}_{\tilde{\mathcal{A}}} = \mathbf{0}. \tag{4.3}$$

In regard to the expanded model $\tilde{\mathcal{A}}$, the corresponding $\boldsymbol{\tau}$-matrix is an $\tilde{s} \times \tilde{s}$ positive diagonal matrix, denoted by $\boldsymbol{\tau}_a$, and the corresponding ridge sample covariance submatrix is denoted by $\hat{\Sigma}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} = S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} + \boldsymbol{\tau}_a$. Adding $\boldsymbol{\tau}_a (\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}})$ and multiplying $\boldsymbol{\tau}_a$ on both sides of equation (4.3), we have

$$\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}} = \frac{1}{n} \hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} X_{\tilde{\mathcal{A}}}^T \boldsymbol{\epsilon} + \boldsymbol{r}_a, \tag{4.4}$$

where the debiased estimator $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a}$ of $\boldsymbol{\beta}^*_{\tilde{\mathcal{A}}}$ takes the form:

$$\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a} = \hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}} + \lambda \hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} \boldsymbol{\kappa}_{\tilde{\mathcal{A}}}, \tag{4.5}$$

and the remainder $\boldsymbol{r}_a$ is given by

$$\boldsymbol{r}_a = \hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} \boldsymbol{\tau}_a (\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}}) + \hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} S_{\tilde{\mathcal{A}}\mathcal{B}_{tn}} \hat{\boldsymbol{\beta}}_{\mathcal{B}_{tn}} + \hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} S_{\tilde{\mathcal{A}}\mathcal{B}_{fn}} (\hat{\boldsymbol{\beta}}_{\mathcal{B}_{fn}} - \boldsymbol{\beta}^*_{\mathcal{B}_{fn}}) \overset{def}{=} I_{11} + I_{12} + I_{13}. \tag{4.6}$$

If $\rho^+_{\max}(\boldsymbol{\tau}_a) = o(\sqrt{\log p}/n)$ holds, Lemma 4.1 below shows that $\|\boldsymbol{r}_a\|_2 = o_p(1/\sqrt{n})$, which is a higher order term than the parametric rate of convergence with respective to the sampling uncertainty. Thus, as stated in Theorem 4.1, equation (4.4) implies that the debiased estimator $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a}$ of the parameters within the expanded model $\tilde{\mathcal{A}}$ is consistent and follows asymptotically a normal distribution.

Now, consider the complementary model $\tilde{\mathcal{A}}^c$. Following similar steps of deriving equation (4.4), we rewrite (4.2) as follows:

$$S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}}(\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}}) + \hat{\Sigma}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}(\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}^c}) + \lambda\boldsymbol{\kappa}_{\tilde{\mathcal{A}}^c} = \frac{1}{n}X^T_{\tilde{\mathcal{A}}^c}\boldsymbol{\epsilon} + \boldsymbol{\tau}_c(\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}^c}),$$

where the corresponding ridge sample covariance submatrix is $\hat{\Sigma}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c} = S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c} + \boldsymbol{\tau}_c$ and $\boldsymbol{\tau}_c$ is a $(p - \tilde{s}) \times (p - \tilde{s})$ matrix of positive diagonals. Plugging (4.4) and (4.5) into the above equation, we obtain

$$\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}^c} = \frac{1}{n}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}(X^T_{\tilde{\mathcal{A}}^c} - S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}X^T_{\tilde{\mathcal{A}}})\boldsymbol{\epsilon} + \boldsymbol{r}_c, \tag{4.7}$$

where $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c}$ is the debiased estimator of $\boldsymbol{\beta}^*_{\tilde{\mathcal{A}}^c}$, which takes the following form:

$$\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c} = \hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c} + \lambda\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\boldsymbol{\kappa}_{\tilde{\mathcal{A}}^c} - \lambda\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}\boldsymbol{\kappa}_{\tilde{\mathcal{A}}}. \tag{4.8}$$

Moreover, the associated remainder term $\boldsymbol{r}_c$ is

$$\boldsymbol{r}_c = \hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\boldsymbol{\tau}_c(\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}^c}) - \hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}}\boldsymbol{r}_a \overset{def}{=} I_{21} + I_{22}\boldsymbol{r}_a. \tag{4.9}$$

If $\rho^+_{\min}(\boldsymbol{\tau}_c) = O\big(\sqrt{\lambda_{\max}(p - \tilde{s})}\big)$ holds, we can show $\|\boldsymbol{r}_c\|_2 = o_p(1/\sqrt{n})$ in Lemma 4.1. Once again, this reminder $\boldsymbol{r}_c$ is asymptotically ignorable in comparison to the parametric convergence rate with respect to the sampling uncertainty.

Now, combining the two debiased estimators (4.5) and (4.8), namely, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} = (\hat{\boldsymbol{\beta}}^T_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a}, \hat{\boldsymbol{\beta}}^T_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c})^T$, we express the proposed MOCE estimator for $\boldsymbol{\beta}^*$ as follows,

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} = \hat{\boldsymbol{\beta}}_{\lambda} + \lambda L^{-1}_{\boldsymbol{\tau}}\boldsymbol{\kappa}, \tag{4.10}$$

where matrix $L^{-1}_{\boldsymbol{\tau}}$ is a 2×2 lower-triangular block matrix given by

$$L^{-1}_{\boldsymbol{\tau}} = \begin{pmatrix} \hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} & \mathbf{0} \\ -\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} & \hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c} \end{pmatrix}.$$

In comparison to the original naive debiased estimator in equation (3.1), the proposed new debiased estimator in (4.10) (i.e. the MOCE estimator) presents a different bias correction term, $\lambda L^{-1}_{\boldsymbol{\tau}}\boldsymbol{\kappa}$. Consequently, the inverse matrix of $L^{-1}_{\boldsymbol{\tau}}$, $L_{\boldsymbol{\tau}}$, takes the form of

$$L_{\boldsymbol{\tau}} = \begin{pmatrix} \hat{\Sigma}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} & \mathbf{0} \\ S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}} & \hat{\Sigma}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c} \end{pmatrix},$$

which is different from the simple ridge covariance matrix $\hat{\Sigma}_{\boldsymbol{\tau}} = S + \boldsymbol{\tau}$ in (3.1). The fact of $L^{-1}_{\boldsymbol{\tau}}$ being a lower triangular matrix implies that the MOCE estimator $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c}$ of the parameters outside the expanded model $\tilde{\mathcal{A}}^c$ in (4.8) has no impact on the MOCE estimator $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a}$ of the parameters inside the expanded model $\tilde{\mathcal{A}}$ in (4.5). This is a consequence of the proposed sequential operation of contraction and expansion.

Lemma 4.1 *Consider a viable expanded model that satisfies Assumptions 1-2. Assume* $s_0 = o(n/\log p)$*,* $\lambda \asymp \sqrt{\log p/n}$ *and*

$$\rho^{+}_{max}(\boldsymbol{\tau}_a) = o(\sqrt{\log p}/n), \quad \rho^{+}_{min}(\boldsymbol{\tau}_c) = O(\sqrt{\lambda_{\max}(p-\tilde{s})}). \tag{4.11}$$

Then, the reminders in (4.4) and (4.7) $\|\boldsymbol{r}_a\|_2 = o_p(1/\sqrt{n})$ *and* $\|\boldsymbol{r}_c\|_2 = o_p(1/\sqrt{n})$*.*

The proof of Lemma 4.1 is given in Appendix A.2. The lemma establishes respective ℓ_2-norm bounds for the error terms $\boldsymbol{r}_a$ and $\boldsymbol{r}_c$ under positive diagonal matrices $\boldsymbol{\tau}_a$ and $\boldsymbol{\tau}_c$. Because of the condition (4.11), it suffices to implement MOCE with $\boldsymbol{\tau}_a = \tau_a I$ and $\boldsymbol{\tau}_c = \tau_c I$, where τ_a and τ_c are two scalars. Thus, in the remaining sections, we only consider these special forms of $\boldsymbol{\tau}_a$ and $\boldsymbol{\tau}_c$.

4.2 ASN under Gaussian Errors

Assumption 3 *Error terms in model* (1.1)*,* $\epsilon_1, \ldots, \epsilon_n$*, are independent and identically distributed Gaussian random variables with mean zero and variance* σ^2*,* $0 < \sigma^2 < \infty$*.*

We are interested in simultaneous inference for a parameter vector that contains at most m parameters, where m is a fixed constant smaller than n. To set up the framework, we consider a p-dimensional vector $\boldsymbol{d} = (d_1, \ldots, d_p)^T$ in a parameter space $\mathcal{M}_m$ defined as follows:

$$\mathcal{M}_m = \Big\{\boldsymbol{d} \in \mathbb{R}^p : \|\boldsymbol{d}\|_2 = 1,\ \|\boldsymbol{d}\|_0 \leq m\Big\}. \tag{4.12}$$

Theorem 4.1 *Consider a viable expanded model that satisfies Assumptions 1-2 and 3. Suppose* $s_0 = o(n/\log p)$*,* $\tau_a = o(\sqrt{\log p}/n)$*,* $\tau_c = O(\sqrt{\lambda_{\max}(p-\tilde{s})})$*, and* $\lambda \asymp \sqrt{\log p/n}$*. Then, for any* $\boldsymbol{d} \in \mathcal{M}_m$ *such that* $v^2 = \sigma^2 \boldsymbol{d}^T L_{\boldsymbol{\tau}}^{-1} S (L_{\boldsymbol{\tau}}^{-1})^T \boldsymbol{d} > 0$*, the MOCE estimator* $\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}$ *in* (4.10) *satisfies*

$$\sqrt{n} v^{-1} \boldsymbol{d}^T (\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} - \boldsymbol{\beta}^*) = \frac{1}{\sqrt{n}} v^{-1} \boldsymbol{d}^T L_{\boldsymbol{\tau}}^{-1} X^T \boldsymbol{\epsilon} + o_p(1),$$

where $\frac{1}{\sqrt{n}} v^{-1} \boldsymbol{d}^T L_{\boldsymbol{\tau}}^{-1} X^T \boldsymbol{\epsilon}$ *follows* $N(0,1)$ *distribution.*

Proof Combining (4.4) and (4.7) with partition $\boldsymbol{d} = (\boldsymbol{d}_{\tilde{\mathcal{A}}}^T, \boldsymbol{d}_{\tilde{\mathcal{A}}^c}^T)^T$ gives

$$\sqrt{n} \boldsymbol{d}^T (\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} - \boldsymbol{\beta}^*) = \frac{1}{\sqrt{n}} \boldsymbol{d}^T L_{\boldsymbol{\tau}}^{-1} X^T \boldsymbol{\epsilon} + \sqrt{n} \boldsymbol{d}_{\tilde{\mathcal{A}}}^T \boldsymbol{r}_a + \sqrt{n} \boldsymbol{d}_{\tilde{\mathcal{A}}^c}^T \boldsymbol{r}_c.$$

Assumptions 1, 2 and Lemma 4.1 imply that $\|\sqrt{n} \boldsymbol{d}_{\tilde{\mathcal{A}}}^T \boldsymbol{r}_a\|_2 = o_p(1)$ and $\|\sqrt{n} \boldsymbol{d}_{\tilde{\mathcal{A}}^c}^T \boldsymbol{r}_c\|_2 = o_p(1)$. Then, Theorem 4.1 follows immediately from Assumption 3 that $\frac{1}{\sqrt{n}} v^{-1} \boldsymbol{d}^T L_{\boldsymbol{\tau}}^{-1} X^T \boldsymbol{\epsilon}$ follows $N(0,1)$ distribution. ■

Theorem 4.1 suggests that MOCE has the following useful properties: (i) MOCE can perform a joint inference for a set of transformed parameters specified by the space $\mathcal{M}_m$ based on a relaxed sparsity assumption $s_0 = o(n/\log p)$, in comparison to the existing case of $s_0 = o(\sqrt{n}/\log p)$; and (ii) MOCE avoids the "ambiguity" issue of post-selection inference (Berk et al., 2013) caused by the instability of selected models. Besides these properties, in the following sections we further show that MOCE has a smaller mean squared error (MSE) bound than that of the LDP method. In addition, we propose a new test for a set of parameters, which is different from the bootstrap test given by Zhang and Cheng (2017).

4.3 Length of Confidence Interval

Hypothetically, if we fitted data with the oracle model, the smallest variance among the least squares estimators of nonzero parameters would be bounded below by $\sigma^2\rho^+_{\min}(S^{-1}_{\mathcal{A}\mathcal{A}})$, while the estimators of zero parameters would be degenerated as being zero with zero variance. Thus, in this oracle model case the gap between the variances of respective estimators for zero and nonzero parameters would be at least $\sigma^2\rho^+_{\min}(S^{-1}_{\mathcal{A}\mathcal{A}})$. This is a benchmark property for the variances of estimators, which should be accommodated in a valid inference. In fact, existing approaches for post-model selection inference, including Zhang and Zhang (2014); van de Geer et al. (2014); Zhang and Cheng (2017), have not accounted for heterogeneous magnitudes in the variances of their proposed estimators. As shown in their simulation studies, variances of nonzero parameter estimators and variances of zero parameter estimators appear to have the same order because only a single tuning process is used in the determination of tuning parameters. This explains why the LDP method is more likely to reach the 95% coverage for zero parameters than for nonzero parameters in the reported simulation studies.

The proposed MOCE estimation helps alleviate the above dilemma; we show that the ridge tuning matrix with different $\boldsymbol{\tau}_a$ and $\boldsymbol{\tau}_c$ parameters leads to different lengths of confidence intervals for parameters inside and outside of a viable expanded model $\tilde{\mathcal{A}}$. Numerically, in Section 6 we demonstrate that variances between the MOCE estimators with respect to $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{A}}^c$ appear different in their magnitudes due to the use of the two tuning processes with the ridge matrices. When $\boldsymbol{\tau}_c$ is large enough, Corollary 4.1 shows that $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a}$ has a larger variance than that of $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c}$. The lower bound of $var(\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a})$ is at the order $O(1/\rho^+_{\min}(S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}))$, while the upper bound of $var(\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c})$ is at the order $O(1/\rho^+_{\max}(S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}))$. Consequently, the resulting length of confidence interval differs between parameters in $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{A}}^c$.

To present Corollary 4.1, let $\mathbf{e}_1, \ldots, \mathbf{e}_{\tilde{s}} \in \mathbb{R}^p$ be the standard basis vectors that span subspace $\mathbb{R}^{\tilde{s}} \subset \mathbb{R}^p$, and similarly let $\mathbf{e}^\perp_1, \ldots, \mathbf{e}^\perp_{p-\tilde{s}} \in \mathbb{R}^p$ be the standard basis for subspace $\mathbb{R}^{p-\tilde{s}} \subset \mathbb{R}^p$.

Corollary 4.1 *Under the same assumptions as those in Theorem 4.1, the minimal variance of $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a}$ is larger than the maximal variance of $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c}$,*

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a}) &\geq \min_{1\leq i\leq \tilde{s}} \sigma^2 \mathbf{e}_i^T L_{\boldsymbol{\tau}}^{-1} S L_{\boldsymbol{\tau}}^{-1} \mathbf{e}_i \geq c_1/\rho^+_{min}(S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}) \\ &\geq c_2/\rho^+_{max}(S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}) \geq \max_{1\leq i\leq p-\tilde{s}} \sigma^2 (\mathbf{e}_i^\perp)^T L_{\boldsymbol{\tau}}^{-1} S L_{\boldsymbol{\tau}}^{-1} \mathbf{e}_i^\perp \geq var(\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c}), \end{aligned}$$

where c_1 and c_2 are two positive constants.

Proof of Corollary 4.1 is given in Appendix A.3. Cai and Guo (2017) studied the problem about constructing an adaptive confidence interval, in which the interval has its length automatically adjusted to the true sparsity of the unknown regression vector, while maintaining a pre-specified coverage probability. They showed that it is impossible to construct a confidence interval for β_j^* adaptive to the sparsity s_0 with this range of sparsity, $\sqrt{n}/\log p \leq s_0 \leq n/\log p$. Our MOCE provides adaptive confidence intervals depending on whether variables are contained in the expanded model, but the resulting confidence interval length may not be optimal, which is worth further exploration.

4.4 ASN under Non-Gaussian Errors

When the errors ϵ_i's do not follow a Gaussian distribution, Theorem 4.2 shows that $\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}$ still converges to a Gaussian distribution when Assumption 3 is replaced by Assumption 4.

Assumption 4 *Let $w_i = \frac{1}{\sqrt{n}}\boldsymbol{d}^T L_{\boldsymbol{\tau}}^{-1}\mathbf{x}_i$, $\boldsymbol{d} \in \mathcal{M}_m$, with $\mathbf{x}_i$ being the ith column of matrix $X^T = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$. For some $r > 2$,*

$$\sup_{1\leq i\leq n} \mathbb{E}|\epsilon_i|^r < \infty \text{ and } \lim_{n\to\infty} \max_{1\leq i\leq n} \frac{w_i^2}{\sum_{i=1}^n w_i^2} = 0.$$

Theorem 4.2 *Consider a viable expanded model that satisfies Assumptions 1-2 and 4. Suppose $s_0 = o(n/\log p)$, $\tau_a = o(\sqrt{\log p/n})$, $\tau_c = O(\sqrt{\lambda_{\max}(p-\tilde{s})})$, and $\lambda \asymp \sqrt{\log p/n}$. Then, for any $\boldsymbol{d} \in \mathcal{M}_m$ such that $v^2 = \sigma^2\boldsymbol{d}^T L_{\boldsymbol{\tau}}^{-1} S (L_{\boldsymbol{\tau}}^{-1})^T \boldsymbol{d} > 0$, the MOCE estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}$ in (4.10) satisfies*

$$\sqrt{n}v^{-1}\boldsymbol{d}^T(\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} - \boldsymbol{\beta}^*) = \frac{1}{\sqrt{n}}v^{-1}\boldsymbol{d}^T L_{\boldsymbol{\tau}}^{-1} X^T \boldsymbol{\epsilon} + o_p(1),$$

where $\frac{1}{\sqrt{n}}v^{-1}\boldsymbol{d}^T L_{\boldsymbol{\tau}}^{-1} X^T\boldsymbol{\epsilon}$ follows asymptotically $N(0,1)$ distribution.

The proof of Theorem 4.2 is given in Appendix A.5.

4.5 ℓ_2-norm Error Bounds

For the popular LDP method (Zhang and Zhang, 2014), it has been shown that the debiased estimator $\hat{\boldsymbol{\beta}}_{LDP}$ satisfies

$$\|\hat{\boldsymbol{\beta}}_{LDP} - \boldsymbol{\beta}^*\|_2 = O_p(\sqrt{p/n}), \tag{4.13}$$

which is higher than $O_p(\sqrt{s_0 \log p/n})$, the order that LASSO achieves. Refer to Section 3.3 in Zhang and Zhang (2014). Below Corollary 4.2 shows that the MOCE's ℓ_2-norm error bound is of order $O_p(\sqrt{\tilde{s}\log\tilde{s}/n})$, which is lower than the LDP's order $O_p(\sqrt{p/n})$. This improved error bound is largely resulted from the fact that MOCE produces lower variances for null signals than for signals, as stated in Corollary 4.1. Assumption 5 is required to establish such improvement in the ℓ_2-norm error bound analytically. Let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$.

Assumption 5 *The error $\boldsymbol{\epsilon}$ satisfies*

$$\|\frac{1}{n}X_{\tilde{\mathcal{A}}}^T\boldsymbol{\epsilon}\|_\infty = O_p(\sqrt{\log\tilde{s}/n}), \quad \|\frac{1}{n}X_{\tilde{\mathcal{A}}^c}^T\boldsymbol{\epsilon}\|_\infty = O_p(\sqrt{\log(p-\tilde{s})/n}).$$

Assumption 5 is widely used in the literature of high-dimensional models; see for examples Bickel et al. (2009); Negahban et al. (2012). It is easy to verify that it holds in the case of sub-Gaussian random errors.

Corollary 4.2 *Consider a viable expanded model that satisfies Assumptions 1-2 and 5 where the ridge matrices are $\boldsymbol{\tau}_a = \tau_a I$ and $\boldsymbol{\tau}_c = \tau_c I$. Assume $s_0 = o(n/\log p)$, $\tau_a = o(\sqrt{\log p}/n)$, and $\tau_c = O(\sqrt{\lambda_{\max}(p-\tilde{s})})$. Then the ℓ_2-norm error bounds of the MOCE estimator $\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} = (\hat{\boldsymbol{\beta}}^T_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a}, \hat{\boldsymbol{\beta}}^T_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c})^T$ in (4.10) are given by, respectively,*

$$\|\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}}\|_2 = O_p(\sqrt{\tilde{s}\log\tilde{s}/n}), \textit{ and}$$
$$\|\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}^c}\|_2 = o_p\big(\max\{1/\sqrt{n}, \sqrt{(p-\tilde{s})\log(p-\tilde{s})/n}/\tau_c\}\big).$$

The proof of Corollary 4.2 is given in Appendix A.4. Note that when τ_c is chosen to be large enough, the ℓ_2-norm error bound of the MOCE estimator $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c}$ will be dominated by that of $\|\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}}\|_2$ on the expanded model $\tilde{\mathcal{A}}$, which is order $O_p(\sqrt{\tilde{s}\log\tilde{s}/n})$.

4.6 Simultaneous Test

In this section, we consider a simultaneous test for a set of parameters $\mathcal{G} \subset \{1, \ldots, p\}$, whose cardinality $|\mathcal{G}| = g$ satisfying $g/n \to \gamma \in (0,1)$. With respect to $\mathcal{G}$, $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}$ can be partitioned accordingly as $({\boldsymbol{\beta}^*_{\mathcal{G}}}^T, {\boldsymbol{\beta}^*_{\mathcal{G}^c}}^T)^T$ and $(\hat{\boldsymbol{\beta}}^T_{\boldsymbol{\tau}\mathcal{G}}, \hat{\boldsymbol{\beta}}^T_{\boldsymbol{\tau}\mathcal{G}^c})^T$. We want to test the following hypothesis:

$$H_0 : {\boldsymbol{\beta}^*}_j = 0 \text{ for all } j \in \mathcal{G} \textit{ vs } H_a : {\boldsymbol{\beta}^*}_j \neq 0 \text{ for at least one } j \in \mathcal{G}.$$

The MOCE's asymptotic covariance matrix $\hat{\Sigma}_{LSL} = L^{-1}_{\boldsymbol{\tau}} S (L^{-1}_{\boldsymbol{\tau}})^T$ relies on two ridge parameters $\boldsymbol{\tau}_a$ and $\boldsymbol{\tau}_c$, and is constructed according to a partition induced by the expanded model. To reduce the impact from noisy elements in $\hat{\Sigma}_{LSL}$, we use Bai and Saranadasa (1996)'s test without involving the inverse of $\hat{\Sigma}_{\mathcal{G}\mathcal{G}} = \{\hat{\Sigma}_{LSL}\}_{\mathcal{G}\mathcal{G}}$. Our proposed test statistic W_{bs} takes the follows form:

$$W_{bs} = \frac{n\hat{\boldsymbol{\beta}}^T_{\boldsymbol{\tau}\mathcal{G}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}\mathcal{G}} - \sigma^2 \mathrm{tr}(\Sigma^*_{\mathcal{G}\mathcal{G}})}{\sigma^2\Big\{2\mathrm{tr}((\Sigma^*_{\mathcal{G}\mathcal{G}})^2)\Big\}^{1/2}}. \tag{4.14}$$

As stated in Theorem 4.3 below, provided two extra assumptions, the test statistic W_{bs} converges in distribution to the standard normal distribution $N(0,1)$ under the null hypothesis. Thus, the null hypothesis is rejected if W_{bs} is greater than $100(1-\alpha)\%$ upper standard normal percentile. As stated in Srivastava (2007), we calculate (4.14) by replacing $\mathrm{tr}(\Sigma^*_{\mathcal{G}\mathcal{G}})$ by $\mathrm{tr}(\hat{\Sigma}_{\mathcal{G}\mathcal{G}})$ and $\mathrm{tr}((\Sigma^*_{\mathcal{G}\mathcal{G}})^2)$ by $\frac{n^2}{(n+2)(n-1)}\big[\mathrm{tr}(\hat{\Sigma}^2_{\mathcal{G}\mathcal{G}}) - \frac{1}{n}\mathrm{tr}(\hat{\Sigma}_{\mathcal{G}\mathcal{G}})^2\big]$.

Theorem 4.3 *Under the null hypothesis, suppose the same conditions in Theorem 4.1 hold. If $\hat{\Sigma}_{\mathcal{G}\mathcal{G}}$ converges to $\Sigma^*_{\mathcal{G}\mathcal{G}}$ in probability and $\frac{g}{n} \to \gamma \in (0,1)$, then we have $W_{bs} \xrightarrow{d} N(0,1)$ as $p \to \infty$ and $n \to \infty$.*

The proof of Theorem 4.3 is given in Appendix A.6. Note that W_{bs} demonstrates a way for the simultaneous test in the high dimensional setting. However, it is not the most powerful test. It deserves a further exploration for how to construct a more powerful test based on the MOCE estimator. Bootstrap-based tests may be a promising way in this scenario to further reduce a test's sensitivity on $\boldsymbol{\tau}_a$, $\boldsymbol{\tau}_c$ and $\tilde{\mathcal{A}}$.

5. Model Expansion

To demonstrate the feasibility of the proposed MOCE framework for post-model selection inference, in this section we introduce a model expansion strategy based on the forward screening method proposed by Wang (2009). This chosen model expansion scheme may not be optimal but it suffices to show that the proposed MOCE framework is not void. A primary purpose of model expansion is to mitigate the uncertainty of model selection to a level lower than the sampling uncertainty. To do so, we present Algorithm 1 and Algorithm 2 to carry out model expansion, starting with the LASSO selected model $\hat{\mathcal{A}}$. It follows that $\hat{\mathcal{A}} \subseteq \tilde{\mathcal{A}}$ surely. Under some regularity conditions, we guarantee that the resulting expanded model $\tilde{\mathcal{A}}$ satisfies Assumption 2.

Intuitively, when an expanded model $\tilde{\mathcal{A}}$ is too small, it is likely to miss some true signal parameters; on the other hand, when an expanded model $\tilde{\mathcal{A}}$ is too large, it would include many null parameters. The perfect case would be $\tilde{\mathcal{A}} = \mathcal{A}$. The size of $\tilde{\mathcal{A}}$ pertains to a trade-off between the uncertainty of model selection and efficiency of statistical inference. Thus, we present some results related to the size of the resulting $\tilde{\mathcal{A}}$ by the model expansion algorithms.

5.1 Algorithm for Model Expansion

We invoke the method of forward screening proposed by Wang (2009) to develop our model expansion algorithm. Once again, we stress that this is just one model expansion strategy, and there may exist better ones in the derivation of an expanded model. We will show in Proposition 5.1 below that this algorithm, termed as Algorithm 1, enables us to obtain a viable expanded model that satisfies Assumption 2. This means that we can find at least one legitimate implementation to fulfill the proposed strategy of contraction and expansion.

The theoretical justification of the forward screening requires Assumption 6 originally introduced by Wang (2009) with Gaussian random variables. In this paper we relax the normality condition to that X is marginally sub-Gaussian (Kuchibhotla and Chakrabortty, 2020), so that Algorithm 1 is also applied to bounded and categorical random variables, which are pervasive in applications. A random vector $Z = (Z_1, \ldots, Z_p)^T \in \mathbb{R}^p$ is said to be marginally sub-Gaussian if $\|Z\|_{M,\psi} \overset{def}{=} \sup_{1\le i\le p} \|Z_i\|_\psi < \infty$, where $\|\cdot\|_\psi$ is the Orlicz norm with $\psi(x) = \exp(x^2) - 1$ (Kuchibhotla and Chakrabortty, 2020). According to Theorem 4.2 in Kuchibhotla and Chakrabortty (2020), when p satisfies $\log p = o(n(\log n)^{-2})$, with probability at least $1 - 6e^{-t}$ for $t > 0$, the ∞-norm of $S - \Sigma$ satisfies

$$\|S - \Sigma\|_\infty \le O\left(\sqrt{\frac{t + 2\log p}{n}}\right), \tag{5.1}$$

which is the convergence rate achieved by a Gaussian random vector Z.

Algorithm 1: Algorithm for model expansion via the method of forward screening

```
1 The size of an initial expanded model $\tilde{\mathcal{A}}_0$ is set at $\tilde{s}_0 = s_0 + \eta p$, where $s_0 = |\mathcal{A}|$ may
    be estimated by $\hat{s} = |\hat{\mathcal{A}}|$ and $\eta \in [0, 1)$ is a certain prior proportion to fix the
    upper limit of the model size.
2 Select the first $2\hat{s}$ variables into $\tilde{\mathcal{A}}_0$ by the froward screening approach and the
    other $\eta p - \hat{s}$ variables are randomly sampled from the rest of $p - 2\hat{s}$ variables.
    This step intentionally injects additional noise variables into the expanded model
    to help reduce the sensitivity of the expanded model from the variable selection
    relative to the sampling variability.
3 The final expanded model $\tilde{\mathcal{A}}$ is created by merging the initial expanded model $\tilde{\mathcal{A}}_0$
    with the set of LASSO selected variables, that is $\tilde{\mathcal{A}} = \tilde{\mathcal{A}}_0 \cup \hat{\mathcal{A}}$. This step ensures
    that $\hat{\mathcal{A}} \subseteq \tilde{\mathcal{A}}$ surely.
```

Assumption 6 *Let $X_1, \dots, X_n$ be independent and identical random vectors in $\mathbb{R}^p$ with $EX_i = 0$ and $cov(X_i) = \Sigma = (\sigma_{ij})_{ij}$.*

(a) $\max_{1 \le i \le n} \|X_i\|_{M,\psi} \le K_{n,p} < \infty$*;*

(b) $0 < 2\eta_{\min} \le \rho^+_{\min}(\Sigma) \le \rho^+_{\max}(\Sigma) < \eta_{\max}/2$*;*

(c) $\|\boldsymbol{\beta}^*\|_2 \le C_\beta$*, and* $\min_{j \in \mathcal{A}} |\beta^*_j| \ge \nu n^{-\xi_b}$ *and* $\nu > 0$*;*

(d) there exist $\xi > 0$ and $\xi_0 > 0$ such that $\log p = \nu n^{\xi} = o(n(\log n)^{-2})$*,* $s_0 = \nu n^{\xi_0}$ *and* $\xi + 6\xi_0 + 12\xi_b < 1$.

Using similar arguments given in Wang (2009), we can establish Proposition 5.1.

Proposition 5.1 *Under Assumption 6, the expanded model $\tilde{\mathcal{A}}$ obtained from Algorithm 1 satisfies Assumption 2:*

$$P(\mathcal{A} \subseteq \tilde{\mathcal{A}}) \to 1, \text{ as } n \to \infty; \tag{5.2}$$

$$\eta_{\min} < \rho^+_{\min}(S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}) \le \rho^+_{\max}(S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}) < \eta_{\max}. \tag{5.3}$$

Moreover, the size of the expanded model and the size of the LASSO selected model are given by, respectively,

$$\tilde{s} = |\tilde{\mathcal{A}}| = O(n^{2\xi_0 + 4\xi_b}); \tag{5.4}$$

$$\hat{s} = |\hat{\mathcal{A}}| = o(n^{2\xi_0 + 4\xi_b}). \tag{5.5}$$

The proof of Proposition 5.1 is given in Appendix A.7.

It is worth noting that the size of the expanded model $\tilde{\mathcal{A}}$ is $O(n^{2\xi_0+4\xi_b})$ with $\xi_0 \in (0, 1/6)$ and $\xi_b \in (0, 1/12)$, which is bounded from above by $O(n^{2/3})$. This implies that the size of $\tilde{\mathcal{A}}$ cannot reach the sample size n. Thus, this forward screening algorithm for model expansion avoids undesirable scenarios in which a very large expanded model is produced to cover the true model. Also, part (d) of this assumption implies that $s_0 = o(\frac{n}{\log p})$, which is needed by

MOCE; moreover, with no surprise, we see $\hat{s} = o(\tilde{s})$. This indicates $\tilde{s}$ is much larger than $\hat{s}$, as desired. Part (c) of Assumption 6 is in fact a sufficient condition pertinent to the utility of Wang (2009)'s forward screening method for model expansion. As shown in Appendix B, when some of signal strengths are arbitrarily close to zero, the proposed MOCE method still provide valid coverage probabilities for both signal parameters and non-signal parameters.

According to Proposition 5.1 we know that s_0 satisfies $s_0 = o(n^{1/6})$ under the forward screening method for model expansion. This order of the true model size may be relaxed when certain conditions on the design matrix X are assumed. In particular, we present two sets of conditions below, both of which can relax the order to be $s_0 = o\left((\frac{n}{\log p})^{1/3} \beta_{\min}^{8/3} \right)$, and $\tilde{s} = O(s_0^2 \beta_{\min}^4)$, where $\beta_{\min} = \min_{j \in \mathcal{A}} |\beta_j^*|$.

The first set of conditions considered by Yaskov (2016) and Chafaï and Tikhomirov (2017) are given as follows. Let $X_1, \ldots, X_n$ be independent and identical random vectors in $\mathbb{R}^p$ with $\mathbb{E}X_i = 0$ and $cov(X_i) = \Sigma$. Denote $l = O(s_0^2 \beta_{min}^4)$.
(i) for $i = 1, \ldots, n$, any l-dimensional subvector of X_i, $X_i(l)$, and any $l \times l$ orthogonal projection matrices Π, $\{\|\Pi X_i(l)\|_2^2 - \text{rank}(\Pi)\}/l \to 0$ in probability as $l \to \infty$;
(ii) there exists a positive constant c such that $\mathbb{E}(x_{ij}^4) \leq c < \infty$ for $i = 1, \cdots, n, j = 1, \ldots, p$;
(iii) $0 < 2\eta_{\min} \leq \rho_{\min}^{+}(\Sigma) \leq \rho_{\max}^{+}(\Sigma) < \eta_{\max}/2$.
According to Yaskov (2016), condition (i) above characterizes a restriction on the correlation among columns of matrix X, while according to Chafaï and Tikhomirov (2017), condition (ii) above ensures the existence of an upper bound for the largest eigenvalue of the empirical matrix. It follows that under the above three conditions, (5.3) holds. Moreover, by similar arguments to those given in Wang (2009), we can show that $s_0 = o\left((\frac{n}{\log p})^{1/3} \beta_{\min}^{8/3} \right)$ and $\tilde{s} = O(s_0^2 \beta_{\min}^4)$. Refer to several examples satisfying conditions (i) and (ii) in Yaskov (2016).

The second set of conditions considered by Cai et al. (2010) assumes that any $l \times l$ submatrix $\Sigma_{l \times l} = (\sigma_{ij})$ of Σ belongs to the following parameter space:

$$\mathcal{F}_\alpha = \left\{ \Sigma_{l \times l} : \max_j \sum_i \{|\sigma_{ij}| : |i - j| > k\} \leq M k^{-\alpha} \text{ for all } k \right\}.$$

This restricts the elements σ_{ij} of the population covariance matrix Σ by a certain decay pattern. Under this condition, a tapering type of estimation may be obtained for the sample covariance matrix in an expanded model, which also gives rise to the same order $s_0 = o\left((\frac{n}{\log p})^{1/3} \beta_{\min}^{8/3} \right)$.

5.2 Algorithm for Ridge Parameter Tuning

Once the expanded model $\tilde{\mathcal{A}}$ is chosen, we determine the ridge parameters $\boldsymbol{\tau}_a$ and $\boldsymbol{\tau}_c$ to control the bias reminder terms $\boldsymbol{r}_a$ and $\boldsymbol{r}_c$, respectively defined in (4.6) and (4.9). We propose Algorithm 2 based on the means of cross-validation to determine τ_a in $\boldsymbol{\tau}_a = \tau_a I$ and τ_c in $\boldsymbol{\tau}_c = \tau_c I$ around their asymptotic orders given in Lemma 4.1. They are, $\tau_a = c_a(\sqrt{\log p}/n)^{-1}$ and $\tau_c = c_c \sqrt{\rho_{max}^{+}(S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}) \rho_{max}^{+}(S_{\tilde{\mathcal{A}}^c \tilde{\mathcal{A}}^c})}$, respectively. From the proof of Lemma 4.1, we further know with probability approaching to 1 that $\|\boldsymbol{r}_a\|_2$ is bounded by an increasing function of τ_a, while $\|\boldsymbol{r}_c\|_2$ is bounded by a decreasing function of τ_c if $\|\boldsymbol{r}_a\|_2$ is fixed, given as follows:

$$\|\boldsymbol{r}_a\|_2 \leq \frac{\tau_a}{\tau_a + \lambda_{max}(\tilde{s})}\|\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}}\|_2, \quad \|\boldsymbol{r}_c\|_2 \leq \frac{\sqrt{\lambda_{\max}(\tilde{s})\lambda_{\max}(p-\tilde{s})}}{\tau_c}\|\boldsymbol{r}_a\|_2.$$

Figure 2 show these relationships between the bias reminder terms and the ridge parameters in the setting of $p = 400$, $n = 300$, and $\alpha = 0.7$, a case considered in the simulation study in Section 6. Apparently there are no unique values of c_a and c_c to control the reminder error terms. The curvature of the two functions in Figure 2 enables us to choose some reasonably small c_a and relatively large c_c. We should not choose a very small c_a and/or a very large c_c to avoid unreliable performance of the MOCE. For example, a very large c_c often leads to the superefficiency for the parameters in $\tilde{\mathcal{A}}^c$.

Algorithm 2: Algorithm for ridge parameter selection

```
Partition the data into $k$ folds for $k$-fold cross validation
Value $c_a$ is determined by the last point before the curve rises up rapidly:
    $\hat{c}_a = \max_{c_a}\Big\{c_a > 0 : \frac{\Delta\log(\sum_{j=1}^k \|\boldsymbol{r}_a(c_a, \hat{\boldsymbol{\beta}}_{-j})\|_2)}{\Delta\log(c_a)} \leq 10^{-4}\Big\},$   (5.6)
    where $\Delta$ denotes the differencing operator so the ratio gives a numerical
    derivative; $\boldsymbol{r}_a(c_a, \hat{\boldsymbol{\beta}}_{-j})$ denotes $\boldsymbol{r}_a$ evaluated on the $j$th fold of data at $c_a$ and $\hat{\boldsymbol{\beta}}_{-j}$,
    which is the LASSO estimate obtained on all data except the $j$th fold.
Value $c_c$ is determined as the last position before the curve falls sharply:
    $\hat{c}_c = \min_{c_c}\Big\{c_c > 0 : \frac{\Delta\log(\sum_{j=1}^k \|\boldsymbol{r}_c(\hat{c}_a, c_c, \hat{\boldsymbol{\beta}}_{-j})\|_2)}{\Delta\log(c_c)} \leq -10^{-5}\Big\},$   (5.7)
    where $\boldsymbol{r}_c(\hat{c}_a, c_c, \hat{\boldsymbol{\beta}}_{-j})$ denotes $\boldsymbol{r}_c$ evaluated on the $j$th fold of data at $\hat{c}_a$ given in
    (5.6), $c_c$ and the LASSO estimate $\hat{\boldsymbol{\beta}}_{-j}$ on all data except the fold $j$.
```

5.3 Computational Complexity

The dominant computational cost in MOCE is at calculating the inverse of $\hat{\Sigma}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}$ with the computational complexity being of order $O(n(p-\tilde{s})^2)$ under the operation of the Sherman-Morrison formula. In the case where LASSO uses the popular coordinate descent algorithm, the associated computational complexity is of order $O(2np)$ (Friedman et al., 2010), pertaining to iterations of all p variables under a fixed tuning parameter. Debiasing methods (van de Geer et al., 2014; Zhang and Zhang, 2014) ought to run p LASSO regressions for the node-wise LASSO in order to obtain a sparse estimate of the precision matrix. Therefore, with fixed p tuning parameters, the computational complexity of the existing methods is of order $O(2np^2)$. This comparison suggests that MOCE has significantly lower computational burden than the existing node-wise LASSO.

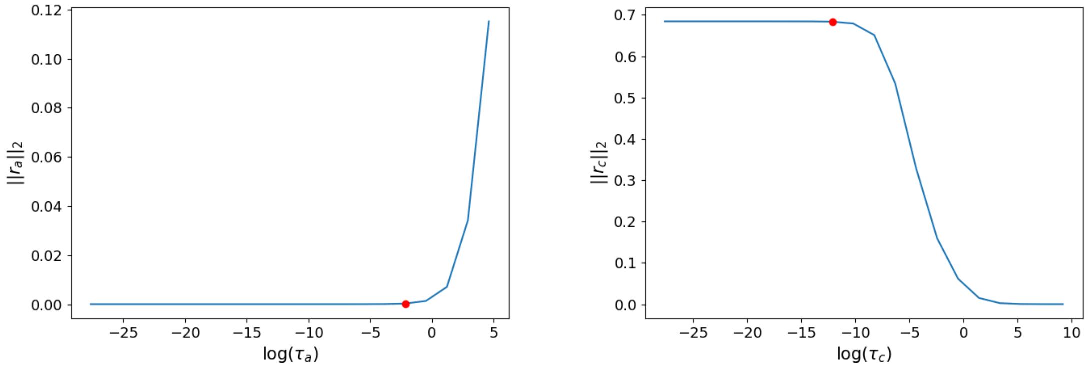

Figure 2: Plots of $\|\boldsymbol{r}_a\|_2$ versus $\log(\tau_a)$ (left) and $\|\boldsymbol{r}_c\|_2$ versus $\log(\tau_c)$ (right) in a simulation setting with $p = 400$, $n = 300$, $\alpha = 0.7$, where the red dots are two chosen values.

6. Simulation Studies

In this section, we use simulation experiments to evaluate the performance of MOCE. In particular, we compare MOCE with the popular LDP method proposed by Zhang and Zhang (2014) for their performances on inference.

6.1 Setup

We simulate 500 data according to the following linear model:

$$y = X\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} = (\epsilon_i, \ldots, \epsilon_n)^T, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n,$$

where $\sigma = 0.5$, and the s_0 signal parameters in set $\mathcal{A}$ are generated from the uniform distribution $U(0.1, 0.5)$, and the rest of $p - s_0$ parameters in $\mathcal{A}^c$ are all set at 0. Each row of the design matrix X is simulated by a p-variate normal distribution $N(\mathbf{0}, \sigma^2 R(\alpha))$, where $R(\alpha)$ is a first-order autoregressive (i.e. AR-1) correlation matrix with correlation parameter $\alpha \in \{0.5, 0.7\}$. Each of the p columns in X is normalized to satisfy ℓ_2-norm 1.

We run 500 rounds of simulations to draw summary statistics in the evaluation. Three summary metrics are used to evaluate inferential performance on individual parameters from the signal set $\mathcal{A}$ and the null signal set $\mathcal{A}^c$, separately. They include bias (Bias), coverage probability (CP), and asymptotic standard error (ASE):

$$\text{Bias}_{\mathcal{A}} = \frac{1}{s_0}\sum_{j\in\mathcal{A}}(\mathbb{E}\hat{\beta}_j - \beta_j^*), \quad \text{Bias}_{\mathcal{A}^c} = \frac{1}{p - s_0}\sum_{j\in\mathcal{A}^c}(\mathbb{E}\hat{\beta}_j - \beta_j^*),$$

$$\text{ASE}_{\mathcal{A}} = \frac{1}{s_0}\sum_{j\in\mathcal{A}}\sqrt{\mathbb{V}\text{ar}(\hat{\beta}_j)}, \quad \text{ASE}_{\mathcal{A}^c} = \frac{1}{p - s_0}\sum_{j\in\mathcal{A}^c}\sqrt{\mathbb{V}\text{ar}(\hat{\beta}_j)},$$

$$\text{CP}_{\mathcal{A}}(\eta) = \frac{1}{s_0}\sum_{j\in\mathcal{A}}\mathbb{1}\{\beta_j^* \in \text{CI}_j(\eta)\}, \quad \text{CP}_{\mathcal{A}^c}(\eta) = \frac{1}{p - s_0}\sum_{j\in\mathcal{A}^c}\mathbb{1}\{0 \in \text{CI}_j(\eta)\},$$

where $\mathbb{E}\hat{\beta}_j$ is the empirical expectation of $\hat{\beta}_j$, $\mathbb{V}\mathrm{ar}(\hat{\beta}_j)$ is the averaged asymptotic variance of $\hat{\beta}_j$, and $\mathrm{CI}_j(\eta)$ denotes the confidence interval for β_j^* derived from $\mathbb{V}\mathrm{ar}(\hat{\beta}_j)$ under the confidence level $1-\eta$, where $\eta \in (0,1)$, over 500 replicates. Average lengths of confidence intervals are also reported.

To apply MOCE, we begin with the LASSO estimate $\hat{\boldsymbol{\beta}}_\lambda$ that is calculated by the R package `glmnet` with the tuning parameter λ selected by a 10-fold cross validation, where the variance parameter σ^2 is estimated by $\hat{\sigma}^2 = \frac{1}{n-\hat{s}}\|\mathbf{y} - X\hat{\boldsymbol{\beta}}_\lambda\|_2^2$, where $\hat{s}$ is the number of nonzero entries in the LASSO estimate $\hat{\boldsymbol{\beta}}_\lambda$. It is shown in Reid et al. (2016) that this estimator $\hat{\sigma}^2$ is robust against changes in signal sparsity and strength. Starting with the LASSO selected model $\hat{\mathcal{A}}$, we construct the expanded model $\tilde{\mathcal{A}}$ via Algorithm 1 with the target size $\tilde{s} = \hat{s} + 0.05p$. The two ridge parameters τ_a in $\boldsymbol{\tau}_a = \tau_a I$ and τ_c in $\boldsymbol{\tau}_c = \tau_c I$ are chosen with the utility of Algorithm 2. Here we set $\eta = 0.05$ to allow 5% of LASSO estimated null parameters enter the expanded model. To calculate the competing LDP estimator proposed by Zhang and Zhang (2014), denoted by $\hat{\boldsymbol{\beta}}_{LDP}$, we use the existing R package `hdi` with the initial estimate obtained from the scaled LASSO.

6.2 Inference on Individual Parameters

We compare MOCE and LDP for their performances on inference for 1-dimensional parameters in the following scenarios of combinations: $n \in \{300, 500\}$, $p \in \{400, 600, 1000\}$, $s_0 \in \{5, 15\}$ and $\alpha \in \{0.5, 0.7\}$. Tables 1, 2, 3 and 4 report Bias, ASE, coverage probability (CP) at significance level 0.01 (CP99) and 0.05 (CP95), and length of confidence interval (LEN) over 500 rounds of simulations, each table for one of the four combinations of s_0 and α.

The four tables clearly show that MOCE outperforms LDP as MOCE's estimation biases are much smaller and coverage probabilities are much closer to the nominal levels regardless of correlation parameter α. Also, the coverage probabilities get closer to the nominal levels when the sample size n increases. Such a performance improvement by MOCE is due to the fact that MOCE uses different lengths of confidence intervals to cover nonzero and zero parameters. It is noted that when p is much larger than n MOCE has larger variances for signal parameters in $\mathcal{A}$ than those for null signal parameters in $\mathcal{A}^c$, confirming the theoretical result stated in Corollary 4.1. On the contrary, in the LDP method the estimated variances for both signal and null signal parameters are very similar. According to van de Geer et al. (2014), LDP tends to optimize the global coverage of all parameters with the aim of achieving the overall shortest confidence intervals for all parameters, where differences between signals and null signals are not recognized and accounted for in the inference. Reflecting to this strategy of optimality, the LDP method typically produces standard errors for all parameters in the same order of magnitude, and consequently the resulting standard errors for signal parameters are often underestimated, leading to an undercoverage for signal parameters. This phenomenon is also reported in Zhang and Cheng (2017). In regard to length of confidence interval, with no surprise, 99% confidence interval is longer than 95% confidence interval under either MOCE or LDP, and when sample size n increases, both 95% and 99% confidence intervals become shorter. On the other hand, when sample size n is fixed but the number of parameters p increases, there seems no clear patterns due possibly to the fact that the size of expanded model varies under different tuning schemes.

These simulation results show that our MOCE may occasionally have unsatisfactory performances. In general, we observe that in the cases of $p > 2n$, such as the case of $n = 300$ and $p = 1000$, our MOCE for zero parameters tends to have overcoverage and thus likely to be superefficient. Superefficiency for null signals is not necessarily detrimental; in the oracle case, the coverage of null signals would be 100%. However, in the Neyman-Pearson school of inference, overcoverage is regarded as being problematic. It is interesting to note that in these cases, our MOCE still works well for inference on nonzero parameters. If inference for null signals is the primary interest of data analyses (which is rather unlikely in practice), as a rule of thumb we recommend first reducing the model size from p to $2n$ or smaller by a certain screening procedure before applying MOCE for inference. Statistical inference for null signals deserves some serious philosophical debates for the suitability of Neyman-Pearson's statistical inference paradigm.

			MOCE				LDP			
n	p	Set	CP95(LEN)	CP99(LEN)	Bias	ASE	CP95(LEN)	CP99(LEN)	Bias	ASE
300	400	$\mathcal{A}$	0.944(0.153)	0.993(0.201)	0.0000	0.039	0.893(0.136)	0.966(0.179)	0.0193	0.034
	400	$\mathcal{A}^c$	0.951(0.248)	0.988(0.327)	0.0001	0.063	0.951(0.138)	0.990(0.181)	-0.0006	0.035
	600	$\mathcal{A}$	0.944(0.154)	0.988(0.203)	0.0004	0.039	0.872(0.135)	0.958(0.178)	0.0234	0.034
	600	$\mathcal{A}^c$	0.957(0.103)	0.990(0.136)	0.0000	0.026	0.953(0.138)	0.991(0.181)	0.0002	0.035
	1000	$\mathcal{A}$	0.935(0.164)	0.985(0.215)	0.0002	0.041	0.883(0.138)	0.968(0.182)	0.0173	0.035
	1000	$\mathcal{A}^c$	0.965(0.050)	0.993(0.065)	0.0000	0.012	0.953(0.138)	0.991(0.181)	-0.0001	0.035
500	400	$\mathcal{A}$	0.948(0.114)	0.986(0.149)	-0.0004	0.029	0.909(0.105)	0.978(0.139)	0.0119	0.027
	400	$\mathcal{A}^c$	0.946(0.219)	0.985(0.288)	0.0000	0.055	0.952(0.107)	0.990(0.141)	-0.0001	0.027
	600	$\mathcal{A}$	0.948(0.114)	0.989(0.149)	-0.0001	0.029	0.910(0.104)	0.978(0.137)	0.0133	0.026
	600	$\mathcal{A}^c$	0.949(0.282)	0.986(0.370)	-0.0001	0.072	0.951(0.107)	0.990(0.141)	0.0002	0.027
	1000	$\mathcal{A}$	0.950(0.119)	0.990(0.156)	0.0003	0.030	0.909(0.106)	0.970(0.140)	0.0114	0.027
	1000	$\mathcal{A}^c$	0.956(0.078)	0.989(0.102)	0.0000	0.019	0.951(0.107)	0.990(0.141)	0.0000	0.027

Table 1: In the scenario of $s_0 = 5$ and $\alpha = 0.5$, summary statistics of Bias, ASE, coverage probability (CP95 and CP99) and length of the confidence interval (LEN) for inference in individual parameters based on MOCE and LDP over 500 rounds of simulations.

Another advantage of MOCE in comparison to LDP concerns computational efficiency. Table 5 reports the average computation time. It is evident that MOCE is significantly faster than LDP in all six scenarios considered in the simulation experiments. This numerical evidence confirms the theoretical computational complexity discussed in Section 5.3. In practice, the node-wise LASSO spends extra computational costs for calculating solution paths with a large number of varying tuning parameters which, with no doubt, will dramatically increase LDP's computation time.

6.3 Simultaneous Test for a Group of Parameters

The second simulation study illustrates the performance of Bai and Saranadasa (1996)'s test W_{bs} defined in (4.14) for a group of parameters. In this study, $s_0 = 5$, $n = 300$, $\alpha \in \{0.5, 0.7\}$ and $p \in \{200, 400, 600\}$. Other settings are the same as ones used in the first

			MOCE				LDP			
n	p	Set	CP95(LEN)	CP99(LEN)	Bias	ASE	CP95(LEN)	CP99(LEN)	Bias	ASE
300	400	$\mathcal{A}$	0.942(0.170)	0.986(0.223)	0.0004	0.043	0.901(0.140)	0.971(0.184)	0.0136	0.035
	400	$\mathcal{A}^c$	0.952(0.281)	0.989(0.369)	0.0001	0.071	0.957(0.141)	0.992(0.186)	0.0001	0.036
	600	$\mathcal{A}$	0.946(0.174)	0.987(0.229)	-0.0000	0.044	0.894(0.141)	0.967(0.186)	0.0159	0.036
	600	$\mathcal{A}^c$	0.966(0.112)	0.992(0.147)	0.0001	0.028	0.957(0.141)	0.992(0.186)	0.0001	0.036
	1000	$\mathcal{A}$	0.946(0.184)	0.988(0.242)	0.0000	0.047	0.900(0.139)	0.969(0.183)	0.0174	0.035
	1000	$\mathcal{A}^c$	0.979(0.054)	0.995(0.071)	-0.0000	0.013	0.957(0.141)	0.992(0.186)	0.0000	0.036
500	400	$\mathcal{A}$	0.948(0.122)	0.987(0.161)	-0.0000	0.031	0.910(0.108)	0.976(0.142)	0.0111	0.027
	400	$\mathcal{A}^c$	0.938(0.216)	0.985(0.284)	0.0001	0.055	0.952(0.108)	0.991(0.142)	0.0004	0.027
	600	$\mathcal{A}$	0.946(0.125)	0.992(0.164)	-0.0002	0.031	0.906(0.108)	0.973(0.142)	0.0131	0.027
	600	$\mathcal{A}^c$	0.948(0.315)	0.987(0.414)	0.0000	0.080	0.955(0.109)	0.991(0.143)	0.0004	0.027
	1000	$\mathcal{A}$	0.947(0.127)	0.988(0.167)	0.0002	0.032	0.903(0.107)	0.967(0.141)	0.0116	0.027
	1000	$\mathcal{A}^c$	0.961(0.082)	0.990(0.108)	-0.0000	0.021	0.955(0.109)	0.991(0.143)	0.0000	0.027

Table 2: In the scenario of $s_0 = 15$ and $\alpha = 0.5$, summary statistics of Bias, ASE, coverage probability (CP95 and CP99) and length of the confidence interval (LEN) for inference in individual parameters based on MOCE and LDP over 500 rounds of simulations.

			MOCE				LDP			
n	p	Set	CP95(LEN)	CP99(LEN)	Bias	ASE	CP95(LEN)	CP99(LEN)	Bias	ASE
300	400	$\mathcal{A}$	0.939(0.198)	0.985(0.260)	-0.0004	0.050	0.886(0.167)	0.964(0.220)	0.0225	0.042
	400	$\mathcal{A}^c$	0.947(0.323)	0.987(0.425)	0.0000	0.082	0.951(0.170)	0.990(0.224)	0.0005	0.043
	600	$\mathcal{A}$	0.926(0.208)	0.981(0.274)	0.0001	0.053	0.882(0.164)	0.959(0.215)	0.0237	0.041
	600	$\mathcal{A}^c$	0.955(0.128)	0.989(0.169)	-0.0000	0.032	0.953(0.168)	0.991(0.222)	0.0003	0.043
	1000	$\mathcal{A}$	0.944(0.202)	0.988(0.265)	-0.0005	0.051	0.875(0.161)	0.953(0.211)	0.0238	0.041
	1000	$\mathcal{A}^c$	0.962(0.058)	0.992(0.076)	-0.0000	0.014	0.952(0.167)	0.991(0.220)	-0.0003	0.042
500	400	$\mathcal{A}$	0.944(0.149)	0.989(0.195)	0.0001	0.038	0.899(0.131)	0.972(0.172)	0.0129	0.033
	400	$\mathcal{A}^c$	0.946(0.288)	0.986(0.379)	0.0001	0.073	0.951(0.135)	0.990(0.177)	0.0003	0.034
	600	$\mathcal{A}$	0.948(0.155)	0.986(0.204)	-0.0002	0.039	0.897(0.131)	0.970(0.172)	0.0157	0.033
	600	$\mathcal{A}^c$	0.947(0.367)	0.986(0.483)	-0.0001	0.093	0.951(0.134)	0.990(0.176)	0.0002	0.034
	1000	$\mathcal{A}$	0.946(0.144)	0.990(0.190)	0.0001	0.036	0.898(0.130)	0.960(0.171)	0.0147	0.033
	1000	$\mathcal{A}^c$	0.953(0.096)	0.989(0.127)	0.0000	0.024	0.951(0.133)	0.990(0.175)	-0.0000	0.034

Table 3: In the scenario of $s_0 = 5$ and $\alpha = 0.7$, summary statistics of Bias, ASE, coverage probability (CP95 and CP99) and length of the confidence interval (LEN) for inference in individual parameters based on MOCE and LDP over 500 rounds of simulations.

			MOCE				LDP			
n	p	Set	CP95(LEN)	CP99(LEN)	Bias	ASE	CP95(LEN)	CP99(LEN)	Bias	ASE
300	400	$\mathcal{A}$	0.932(0.226)	0.982(0.297)	0.0002	0.057	0.888(0.167)	0.966(0.220)	0.0173	0.042
	400	$\mathcal{A}^c$	0.947(0.368)	0.985(0.483)	0.0001	0.093	0.956(0.171)	0.992(0.225)	-0.0002	0.043
	600	$\mathcal{A}$	0.911(0.222)	0.972(0.292)	0.0002	0.056	0.874(0.168)	0.954(0.222)	0.0188	0.043
	600	$\mathcal{A}^c$	0.958(0.135)	0.989(0.178)	-0.0000	0.034	0.957(0.172)	0.992(0.227)	-0.0001	0.044
	1000	$\mathcal{A}$	0.907(0.237)	0.953(0.312)	-0.0006	0.060	0.861(0.168)	0.947(0.221)	0.0249	0.042
	1000	$\mathcal{A}^c$	0.978(0.064)	0.995(0.085)	0.0000	0.016	0.956(0.169)	0.992(0.223)	-0.0002	0.043
500	400	$\mathcal{A}$	0.944(0.161)	0.986(0.212)	0.0003	0.041	0.904(0.132)	0.972(0.174)	0.0102	0.033
	400	$\mathcal{A}^c$	0.936(0.282)	0.982(0.371)	-0.0000	0.072	0.954(0.136)	0.991(0.178)	-0.0003	0.034
	600	$\mathcal{A}$	0.937(0.158)	0.986(0.208)	0.0001	0.040	0.888(0.131)	0.962(0.172)	0.0117	0.033
	600	$\mathcal{A}^c$	0.941(0.408)	0.984(0.536)	0.0000	0.104	0.953(0.135)	0.991(0.178)	0.0000	0.034
	1000	$\mathcal{A}$	0.940(0.160)	0.987(0.211)	0.0001	0.041	0.883(0.133)	0.958(0.175)	0.0157	0.034
	1000	$\mathcal{A}^c$	0.955(0.100)	0.989(0.131)	-0.0000	0.025	0.954(0.134)	0.991(0.176)	-0.0002	0.034

Table 4: In the scenario of $s_0 = 15$ and $\alpha = 0.7$, summary statistics of Bias, ASE, coverage probability (CP95 and CP99) and length of the confidence interval (LEN) for inference in individual parameters based on MOCE and LDP over 500 rounds of simulations.

		Computation Time (seconds)	
α	p	MOCE	LDP
0.5	400	15.35	38.60
	600	24.94	63.78
0.7	400	14.48	32.08
	600	24.17	54.12

Table 5: Average computation time in one simulated data set with sample size $n = 300$ for MOCE and LDP.

simulation study. We consider a hypothesis $H_0 : \boldsymbol{\beta}_{0,\mathcal{G}} = \mathbf{0}$ vs $H_a : \boldsymbol{\beta}_{0,\mathcal{G}} \neq \mathbf{0}$, where the size of $\mathcal{G}$ is set at 5, 50 and 100. We also consider varying different size of intersection $\mathcal{G} \cap \mathcal{A}$. When $|\mathcal{G} \cap \mathcal{A}| = 0$, the null hypothesis H_0 is true, otherwise the alternative hypothesis H_a is the case.

Empirical type I error rates and power are computed under the significance level 0.05 over 500 replications. We observe from Table 6 that in all the cases, the test can properly control type I error, although it appears a little conservative when p is relatively small, which may consequently affect the power. According to this table, the key message learned from this simulation study is that the test W_{bs} performs well for cases of large p, with satisfactory type I error control and desirable power to detect any violation of the null hypothesis. Noting that MOCE increases the size of expanded model along with an increase in p, we learn from a different perspective the effectiveness of the proposed the ridge-type debiasing method in high-dimensional linear models.

		$\alpha = 0.5$			$\alpha = 0.7$		
$\lvert\mathcal{G}\rvert$	$\lvert\mathcal{G} \cap \mathcal{A}\rvert$	$p = 200$	$p = 400$	$p = 600$	$p = 200$	$p = 400$	$p = 600$
5	0	0.0300	0.0300	0.0400	0.0350	0.0600	0.0450
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
50	0	0.0250	0.0250	0.0500	0.0100	0.0250	0.0300
	2	0.6350	0.4650	1.0000	0.2050	0.1750	1.0000
	3	0.8850	0.7500	1.0000	0.3850	0.2750	1.0000
100	0	0.0150	0.0350	0.0550	0.0100	0.0200	0.0250
	2	0.2350	0.1500	1.0000	0.0450	0.0550	1.0000
	3	0.3650	0.2550	1.0000	0.0900	0.0900	1.0000

Table 6: Empirical type I error and power of W_{bs} over 200 replications under AR-1 correlated predictors with correlation $\alpha = 0.5$ and $\alpha = 0.7$.

7. Discussion

We have developed a new method of contraction and expansion (MOCE) for simultaneous inference in high-dimensional linear models. The key technical challenge in the paradigm of post-model selection inferences is to quantify the model selection uncertainty in a regularized estimation procedure, which is however notoriously difficulty. The proposed step of model expansion overcomes this difficulty. Different from the existing low dimensional projection (LDP) method, MOCE takes a step of model expansion to reduce the model selection uncertainty due to the LASSO regularization so that the selection uncertainty is asymptotically ignorable in comparison to the sampling uncertainty. The model expansion is carried out by the means of the forward screening approach (Wang, 2009), and under certain regularity conditions, the resulting expanded model has been shown to be viable and of expansion consistency. Thus, MOCE provides a realistic solution to valid simultaneous post-model selection inferences. We have thoroughly discussed the issue of determining the size of expanded model and established as a series of theorems to guarantee the validity of MOCE. We showed both analytically and numerically that MOCE gives better control of type I error and higher power as well as faster computation than the existing LDP method.

Ridge-type shrinkage is adopted by MOCE in qualification of debiasing terms, which not only enjoys computational speed but also produces different lengths of confidence intervals for signal and null signal parameters. It is worth noting that MOCE attempts to provide an adaptive construction of confidence interval with respect to signal strength, instead of signal sparsity as proposed by Cai and Guo (2017). The optimality studied in Cai and Guo (2017) might offer an opportunity to develop a more efficient procedure for the selection of ridge $\boldsymbol{\tau}$-matrices, which is certainly an interesting future research direction. Alternative to the forward screening method for model expansion, as suggested by one of the reviewers, other methods, such as the sure screening (Fan and Lv, 2008) and a direct use of the magnitude of $\boldsymbol{\kappa}$ in the KKT condition, may be used to construct a viable expanded model.

It is certainly of great interest in the future work to relax the condition of minimal signal strength required in Assumption 6. With the utility of the forward screening procedure to establish a viable expanded model, Assumption 6 is a sufficient condition for the proposed MOCE method. In Appendix B we use a simulation study to show numerically that the assumption is not a necessary condition, as MOCE can still yield proper coverage probabilities when some of signal strengths are close to zero. A better strategy than the forward screening method, such as ABESS algorithm by Zhu et al. (2020), is worth further exploration to improve the performance of model expansion for MOCE. In addition, our extensive numerical experience has suggested that post-model selection inference based on large-sample asymptotic distributions is challenged by the fact that weak signals may fall in or fall out of a selected model in a certain random fashion. Although our model expansion strategy can greatly alleviate such a swinging behavior, it remains a technical issue in finite-sample situations. This swinging phenomenon essentially leads to a two-mode mixture distribution in a finite-sample situation, instead of a unimodal asymptotic distribution. It is worth studying the means of double bootstrap suggested by Efron (2014) to establish an inference procedure for weak signals, in which the resulting inference may depart from the reliance on asymptotic distributions.

Furthermore, one future research direction is to understand MOCE's potential connection to elastic-net (Zou and Hastie, 2005). Because both MOCE and elastic-net perform a combined regularization via ℓ_1-norm and ℓ_2-norm, there might exist a certain connection between the two approaches; unveiling such relationship may point to a new direction of future research. Another topic is related to Bootstrap tests based on MOCE estimator to reduce a test's sensitivity on the ridge tuning parameters and the expanded model. Fundamentally, it is of great interest to reexamine the Neyman-Pearson school of statistical inference in connection to the superefficiency of null parameters when such parameters are identified in the model selection analysis.

In summary, the new contributions of our MOCE useful in real-world applications include (i) MOCE allows to construct confidence interval with different lengths in light of signal and null-signal parameters, leading to satisfactorily control type I error and improved power; and (ii) MOCE enjoys fast computation and scalability under less stringent regularity conditions.

Acknowledgments

This research is supported by a National Institutes of Health grant R01ES024732 and National Science Foundation grants DMS 1811734 and DMS 2113564 to Song, and the China Fund of National Natural Science (Nos. 11901470, 11931014 and 11829101) and by the China Fundamental Research Funds for the Central Universities (Nos. JBK1806002) to Zhou. We are grateful to Dr. Cun-Hui Zhang for providing an R code for the low-dimensional projection (LDP) method. Drs. F. Wang and L. Zhou are the co-first author of this paper.

Appendix A. Technical Details

A.1 Proof of Lemma 3.1

Proof Let $S = UDU^T$ be the singular value decomposition of S, whose singular values are arranged in $D = \text{diag}\{\rho_{(1)}, \ldots, \rho_{(m)}, 0, \ldots, 0\}$ with $\rho_{(1)} \geq \cdots \geq \rho_{(m)} > 0 = \rho_{(m+1)} = \cdots = \rho_{(p)}$. Let $\boldsymbol{\tau}^{-1/2} S \boldsymbol{\tau}^{-1/2} = U_1 D U_1^T$ be the singular value decomposition of $\boldsymbol{\tau}^{-1/2} S \boldsymbol{\tau}^{-1/2}$. Denote $U = \boldsymbol{\tau}^{1/2} U_1$. Then we have $\boldsymbol{\tau} = UU^T$ and $S = UDU^T$. By some simple calculations we obtain

$$\begin{aligned}\|\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}\|_F^2 =& \text{tr}\left\{(D+I)^{-1}(U^TU)^{-1}(D+I)^{-1}U^TU\right\} \\ =& \sum_{j=1}^{p} \frac{1}{(\rho_{(j)}+1)^2} \leq \max(p-n, 0) + \frac{\min(n,p)}{(\rho_{(m)}+1)^2},\end{aligned}$$

where the second equality holds due to the equation $[(U^TU)^{-1}(D+I)^{-1}U^TU]_{jj} = \frac{1}{\rho_{(j)}+1}$. Here $[A]_{jj}$ denotes the jth diagonal element of matrix A. Likewise,

$$\|\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}\|_F^2 \geq \max(p-n, 0) + \frac{\min(n,p)}{(\rho_{(1)}+1)^2}.$$

By combining the above two inequalities, the first inequality with the Frobenius norm of part (ii) follows. Now we turn to the proof of the second inequality. By Theorem 4.3.1 in Horn and Johnson (2012), we know

$$\xi + \rho_{\min}^{+}(\boldsymbol{\tau}) \leq \rho_{\min}^{+}(\hat{\Sigma}_{\boldsymbol{\tau}}) \leq \rho_{\max}^{+}(\hat{\Sigma}_{\boldsymbol{\tau}}) \leq \rho_{\max}^{+}(S) + \rho_{\max}^{+}(\boldsymbol{\tau}),$$

where $\xi = 0$ if $p > n$ and $\xi = \rho_{\min}^{+}(S)$ if $p \leq n$. It follows immediately that

$$\frac{1}{\rho_{\max}^{+}(S) + \rho_{\max}^{+}(\boldsymbol{\tau})} \leq \rho_{\min}^{+}(\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}) \leq \rho_{\max}^{+}(\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}) \leq \frac{1}{\xi + \rho_{\min}^{+}(\boldsymbol{\tau})}.$$

Since $\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}$ is positive definite, the largest element of $\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}$ always occurs on its main diagonal, equal to $|\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}|_\infty = \max_{1\leq i\leq p} \mathbf{e}_i^T \hat{\Sigma}_{\boldsymbol{\tau}}^{-1} \mathbf{e}_i$, which satisfies

$$\frac{1}{\rho_{\max}^{+}(S) + \rho_{\max}^{+}(\boldsymbol{\tau})} \leq \max_{1\leq j\leq p} \mathbf{e}_j^T \hat{\Sigma}_{\boldsymbol{\tau}}^{-1} \mathbf{e}_j \leq \frac{1}{\xi + \rho_{\min}^{+}(\boldsymbol{\tau})},$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_p$ are the standard basis of Euclidean $\mathbb{R}^p$ space. Because diagonal matrix $\boldsymbol{\tau} \succ 0$ (positive-definite),

$$|\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}|_\infty \leq |\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}|_\infty |\boldsymbol{\tau}|_\infty \leq \frac{\rho^+_{\max}(\boldsymbol{\tau})}{\xi + \rho^+_{\min}(\boldsymbol{\tau})} = \begin{cases} \frac{\rho^+_{\max}(\boldsymbol{\tau})}{\rho^+_{\min}(\boldsymbol{\tau})}, & \text{if } p > n; \\ \frac{\rho^+_{\max}(\boldsymbol{\tau})}{\rho^+_{\min}(S)+\rho^+_{\min}(\boldsymbol{\tau})}, & \text{if } p \leq n, \end{cases}$$

and

$$|\hat{\Sigma}_{\boldsymbol{\tau}}^{-1}\boldsymbol{\tau}|_\infty \geq \frac{\rho^+_{\min}(\boldsymbol{\tau})}{\rho^+_{\max}(S) + \rho^+_{\max}(\boldsymbol{\tau})}.$$

Then the inequality in part (ii) for the ∞-norm follows. ■

A.2 Proof of Lemma 4.1

Proof By the expression of $\boldsymbol{r}_a$ in (4.6), it suffices to show that three terms I_{11}, I_{12} and I_{13} are all of order $o_p(1/\sqrt{n})$. Similarly, by the expression of $\boldsymbol{r}_c$ in (4.9), the order of $\boldsymbol{r}_c$ is established if both terms I_{21} and $I_{22}\boldsymbol{r}_a$ are all at the order of $o_p(1/\sqrt{n})$.

For term I_{11}, it follows from Assumptions 1-2 and (4.11) that

$$\|I_{11}\|_2 \leq \|I_{11}\|_1 \leq |\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}\boldsymbol{\tau}_a|_\infty \|\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}}\|_1 \leq O_p\Big(\rho^+_{\max}(\boldsymbol{\tau}_a)\sqrt{\frac{\log p}{n}} s_0\Big) = o_p(1/\sqrt{n}), \quad \text{(A.1)}$$

where the third inequality holds from Lemma 3.1 with $\tilde{s} < n$ and $\rho^+_{\min}(S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}})$ being bounded from below by Assumption 2.

For term I_{12}, because of $P(\{\hat{\mathcal{A}} \cup \mathcal{A}\} \subseteq \tilde{\mathcal{A}}) \to 1$, we know $\hat{\boldsymbol{\beta}}_{\mathcal{B}_{tn} \cap \hat{\mathcal{A}}} = \mathbf{0}$ with probability 1. Thus, from Assumptions 1 and 2 and (4.11), we further know $\|\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} S_{\tilde{\mathcal{A}},\mathcal{B}_{tn}\cap\hat{\mathcal{A}}}\|_2 \leq \frac{\sqrt{\lambda_{\max}(\tilde{s})\lambda_{\max}(\hat{s})}}{\lambda_{\min}(\tilde{s})+\rho^+_{\min}(\boldsymbol{\tau}_a)} = O_p(1)$, and therefore

$$\|I_{12}\|_2 = \|\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} S_{\tilde{\mathcal{A}},\mathcal{B}_{tn}\cap\hat{\mathcal{A}}} \hat{\boldsymbol{\beta}}_{\mathcal{B}_{tn}\cap\hat{\mathcal{A}}}\|_2 \leq O_p(1)\|\hat{\boldsymbol{\beta}}_{\mathcal{B}_{tn}\cap\hat{\mathcal{A}}}\|_2 = 0. \quad \text{(A.2)}$$

Similar to the proof of term I_{12}, $P(\{\hat{\mathcal{A}} \cup \mathcal{A}\} \subseteq \tilde{\mathcal{A}}) \to 1$ leads to $\mathcal{B}_{fn} = \emptyset$, $\|\boldsymbol{\beta}^*_{\mathcal{B}_{fn}}\|_2 = 0$ and $\|\hat{\boldsymbol{\beta}}_{\mathcal{B}_{fn}}\|_2 = 0$ with probability 1. Thus, from Assumptions 1 and 2 and (4.11) we obtain $\|\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}} S_{\tilde{\mathcal{A}}\mathcal{B}_{fn}}\|_2 \leq \frac{\sqrt{\lambda_{\max}(\tilde{s})\lambda_{\max}(b_{fn})}}{\lambda_{\min}(\tilde{s})+\rho^+_{\min}(\boldsymbol{\tau}_a)} = O_p(1)$ and

$$\|I_{13}\|_2 \leq O_p(1)\|\hat{\boldsymbol{\beta}}_{\mathcal{B}_{fn}} - \boldsymbol{\beta}^*_{\mathcal{B}_{fn}}\|_2 = 0. \quad \text{(A.3)}$$

Thus, (A.1), (A.2) and (A.3) lead to $\|\boldsymbol{r}_a\|_2 = o_p(1/\sqrt{n})$.

Now we turn to the term $\boldsymbol{r}_c$. For term I_{21}, it follows from (4.11) and $P(\{\hat{\mathcal{A}} \cup \mathcal{A}\} \subseteq \tilde{\mathcal{A}}) \to 1$ that $\|\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\boldsymbol{\tau}_c\|_2 \leq 1$ and $\|\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}^c}\|_2 = 0$ with probability 1. Thus,

$$0 \leq \|I_{21}\|_2 \leq \|\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\boldsymbol{\tau}_c\|_2 \|\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}^c}\|_2 \leq 0. \quad \text{(A.4)}$$

For term $I_{22}\boldsymbol{r}_a$, under (4.11), $\|\boldsymbol{r}_a\|_2 = o_p(1/\sqrt{n})$ and Assumptions 1 and 2, we obtain $\|\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\|_2\|S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}}\|_2 \leq \frac{\sqrt{\lambda_{\max}(\tilde{s})\lambda_{\max}(p-\tilde{s})}}{\rho^+_{\min}(\boldsymbol{\tau}_c)} = O_p(1)$ and

$$\|I_{22}\boldsymbol{r}_a\|_2 \leq \|\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\|_2\|S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}}\|_2\|\boldsymbol{r}_a\|_2 \leq O_p(1)\|\boldsymbol{r}_a\|_2 = o_p(1/\sqrt{n}). \tag{A.5}$$

Therefore, (A.4) and (A.5) complete the proof for order of $\|\boldsymbol{r}_c\|_2$ being $o_p(1/\sqrt{n})$. ■

A.3 Proof of Corollary 4.1

Proof Using similar arguments in Lemma 3.1, we know the minimal variance of estimator $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a}$ satisfies

$$\min_{1\leq i\leq \tilde{s}} \mathbf{e}_i^T\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}\mathbf{e}_i \geq \rho^+_{\min}(\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}) \geq \frac{\rho^+_{\min}(S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}})}{(\rho^+_{\min}(S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}})+\tau_a)^2}.$$

It is easy to verify that

$$\sigma^2\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c} \succ \sigma^2\big[L^{-1}S(L^{-1})^T\big]_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}.$$

Consequently, we can prove the result by assessing the diagonal entries of $\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}$. The maximal variance of estimator $\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c}$ is bounded by

$$\max_{1\leq i\leq p-\tilde{s}}(\mathbf{e}_i^\perp)^T\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\mathbf{e}_i^\perp \leq \rho^+_{\max}(\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}) \leq \frac{\rho^+_{\max}(S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c})}{\tau_c^2}.$$

Therefore, assumptions for τ_a and τ_c in Theorem 4.1 imply

$$\begin{aligned}\min_{1\leq i\leq \tilde{s}} \mathbf{e}_i^T\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}\mathbf{e}_i \geq \frac{c_1}{\rho^+_{\min}(S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}})} \geq \frac{c_2}{\rho^+_{\max}(S_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}})} \geq \\ \max_{1\leq i\leq p-\tilde{s}}(\mathbf{e}_i^\perp)^T\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\mathbf{e}_i^\perp,\end{aligned}$$

where c_1 and c_2 are two positive constants. ■

A.4 Proof of Corollary 4.2

Proof Assumptions 2 and 5 and conditions for τ_a and τ_c imply that on $\tilde{\mathcal{A}}$ there exists

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}\boldsymbol{\tau}_a} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}}\|_2 &\leq \|\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}\|_2\frac{1}{n}\|X^T_{\tilde{\mathcal{A}}}\boldsymbol{\epsilon}\|_2 + \|r_a\|_2 \\ &\leq \|\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}\|_2\sqrt{\tilde{s}}\frac{1}{n}\|X^T_{\tilde{\mathcal{A}}}\boldsymbol{\epsilon}\|_\infty + o_p(1/\sqrt{n}) \\ &\leq \frac{O_p(\sqrt{\tilde{s}\log\tilde{s}/n})}{\lambda_{\min}(\tilde{s})+\tau_a} + o_p(1/\sqrt{n}) = O_p(\sqrt{\tilde{s}\log\tilde{s}/n}).\end{aligned}$$

Similarly on $\tilde{\mathcal{A}}^c$, based on the same assumptions, we obtain

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_{\tilde{\mathcal{A}}^c\boldsymbol{\tau}_c} - \boldsymbol{\beta}^*_{\tilde{\mathcal{A}}^c}\|_2 \leq & \|\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\|_2 \frac{1}{n}\|X^T_{\tilde{\mathcal{A}}^c}\boldsymbol{\epsilon} - S_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}X^T_{\tilde{\mathcal{A}}}\boldsymbol{\epsilon}\|_2 + \|r_c\|_2 \\ = & \|\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\|_2 \frac{1}{n}\|X^T_{\tilde{\mathcal{A}}^c}(I_n - \frac{1}{n}X_{\tilde{\mathcal{A}}}\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}\tilde{\mathcal{A}}}X^T_{\tilde{\mathcal{A}}})\boldsymbol{\epsilon}\|_2 + o_p(1/\sqrt{n}) \\ \leq & \|\hat{\Sigma}^{-1}_{\tilde{\mathcal{A}}^c\tilde{\mathcal{A}}^c}\|_2 \frac{\lambda_{\max}(\tilde{s})}{\lambda_{\max}(\tilde{s}) + \tau_a}\sqrt{p-\tilde{s}}\frac{1}{n}\|X^T_{\tilde{\mathcal{A}}^c}\boldsymbol{\epsilon}\|_\infty + o_p(1/\sqrt{n}) \\ \leq & \frac{1}{\tau_c}\frac{\lambda_{\max}(\tilde{s})}{\lambda_{\max}(\tilde{s}) + \tau_a}\sqrt{(p-\tilde{s})\log(p-\tilde{s})/n} + o_p(1/\sqrt{n}) \\ = & o_p(\max\{1/\sqrt{n}, \sqrt{(p-\tilde{s})\log(p-\tilde{s})/n}/\tau_c\}).\end{aligned}$$

■

A.5 Proof of Theorem 4.2

Proof Following similar arguments to the proof of Theorem 4.1, we have

$$\sqrt{n}\boldsymbol{d}^T(\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} - \boldsymbol{\beta}^*) = \frac{1}{\sqrt{n}}\boldsymbol{d}^T L^{-1}_{\boldsymbol{\tau}} X^T\boldsymbol{\epsilon} + \sqrt{n}\boldsymbol{d}^T_{\tilde{\mathcal{A}}}\boldsymbol{r}_a + \sqrt{n}\boldsymbol{d}^T_{\tilde{\mathcal{A}}^c}\boldsymbol{r}_c = \frac{1}{\sqrt{n}}\sum_{i=1}^n w_i\epsilon_i + o_p(1).$$

From Assumption 4 the Lindeberg's Condition holds because for any $\delta > 0$, as $n \to \infty$,

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\Big\{\frac{w_i^2}{v^2}\epsilon_i^2\mathbb{1}(\Big|\frac{w_i}{v}\epsilon_i\Big| > \delta)\Big\} &\leq \sum_{i=1}^n \mathbb{E}\Big(\frac{|w_i|^r}{v^r}|\epsilon_i|^r\frac{1}{\delta^{r-2}}\Big) \\ &\leq n\max_{1\leq i\leq n}\Big(\frac{|w_i|^2}{\sum_{i=1}^n w_i^2\sigma^2}\Big)^{r/2}\frac{\max_{1\leq i\leq n}\mathbb{E}|\epsilon_i|^r}{\delta^{r-2}} \to 0.\end{aligned}$$

The Lindeberg Central Limit Theorem implies that $\frac{1}{\sqrt{n}}v^{-1}\boldsymbol{d}^T L^{-1}_{\boldsymbol{\tau}}X^T\boldsymbol{\epsilon}$ converges in distribution to $N(0,1)$. ■

A.6 Proof of Theorem 4.3

Proof Let $M_n = n\hat{\boldsymbol{\beta}}^T_{\boldsymbol{\tau}\mathcal{G}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}\mathcal{G}} - \sigma^2\text{tr}(\hat{\Sigma}_{\mathcal{G}\mathcal{G}})$. Theorem 4.1 implies that $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}\mathcal{G}} - \boldsymbol{\beta}_{0,\mathcal{G}}) \xrightarrow{d} N(0, \sigma^2\Sigma^*_{\mathcal{G}\mathcal{G}})$, which further indicates $EM_n \to 0$ given assumptions in Theorem 4.3. Furthermore, we can verify that $\text{var}(M_2) = 2\sigma^4\text{tr}\{\Sigma^*_{\mathcal{G}\mathcal{G}}\Sigma^*_{\mathcal{G}\mathcal{G}}\}$. Applying the same arguments given by Bai and Saranadasa (1996), we can show W_{bs} converges in distribution to $N(0,1)$ as $n \to \infty$. ■

A.7 Proof of Proposition 5.1

Proof If suffices to verify that Lemma 1 in Wang (2009) holds under Assumption 6. It follows from Wang's Lemma 1 that (5.3) holds. Moreover, both (5.2) and (5.4) are resulted from applying Theorem 1 in Wang (2009).

Now we prove Wang's Lemma 1. According to Theorems 4.1 and 4.2 in Kuchibhotla and Chakrabortty (2020), under Assumptions (6.a) and (6.d), we know that $\|S - \Sigma\|_\infty = O_p(\sqrt{\frac{t+2\log p}{n}})$ if $\log p = o(n(\log n)^{-2})$. Thus, the convergence rate of $\|S - \Sigma\|_\infty$ in the marginal sub-Gaussian case is of the same order as that of the Gaussian distribution. Then, Lemma 1 in Wang (2009) can be proved under Assumption 6 by following the same arguments as those given in Wang (2009).

From Assumption (6.d), $n^\xi = o(n(\log n)^{-2})$ and $\xi + 6\xi_0 + 12\xi_b < 1$, we know $0 < \xi < 1$, $\xi_0 \in (0, 1/6)$, $\xi_b \in (0, 1/12)$ and $\xi + \xi_0 < 1$. Thus, $\frac{\nu n^{\xi_0}}{n/\log p} = \nu^2 n^{\xi_0+\xi-1} = o(1)$. Because $2\xi_0 + 4\xi_b < 2/3$ and $\hat{s} = O(s_0)$, it is easy to know $n^{2\xi_0+4\xi_b} = o(n)$ and $\hat{s} = o(n^{2\xi_0+4\xi_b})$. Thus, (5.5) holds.

■

Appendix B. An Additional Simulation

Similar to the simulation settings given in Section 6.2, here we set $p = 400$, $n = 300$, and $\alpha = 0.5$. Among s_0 signal parameters, $s_0 - s_w$ parameters are simulated from $U(0.1, 0.5)$ (i.e. strong signals), and the other s_w parameters are such small signals simulated from $U(0, 0.02)$ that may be arbitrarily close to zero. Results from MOCE and LDP are summarized in the following Table 7. We see that MOCE performs well with the coverage probability close to the nominal level, while LDP remains contentious for inference on the signal set.

	MOCE				LDP			
Set	CP95	CP99	Bias	ASE	CP95	CP99	Bias	ASE
$\mathcal{A}$	0.9472	0.9868	-0.0003	0.0451	0.9044	0.9740	0.0172	0.0347
$\mathcal{A}^c$	0.9457	0.9860	0.0001	0.0626	0.9513	0.9906	-0.0007	0.0351

Table 7: In the setting of $s_0 = 5$, $s_w = 2$ and $\alpha = 0.5$, summary statistics of Bias, ASE, coverage probability (CP95 and CP99) and length of the confidence interval (LEN) for inference in individual parameters based on MOCE and LDP over 500 rounds of simulations.

References

Zhidong Bai and Hewa Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6(3):311–329, 1996.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer Science & Business Media, 2011.

T. Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

Djalil Chafaï and Konstantin Tikhomirov. On the convergence of the extremal eigenvalues of empirical covariance matrices with dependence. *Probability Theory and Related Fields*, 170(3-4):847–889, Apr 2017.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: an elementary, general approach. *Annual Review of Economics*, 7(1):649–688, 2015.

Bradley Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911, 2008.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, 2012.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2020.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Hannes Leeb and Benedikt M. Pötscher. Sparse estimators and the oracle property, or the return of hodges' estimator. *Journal of Econometrics*, 142(1):201–211, 2008.

Hanzhong Liu and Bin Yu. Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics*, 7:3124–3169, 2013.

Nicolai Meinshausen. Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(5):923–945, 2015.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

Jessica Minnier, Lu Tian, and Tianxi Cai. A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106(496): 1371–1382, 2011.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, 26(1):35–67, 2016.

Muni S. Srivastava. Multivariate theory for analyzing high dimensional data. *Journal of The Japan Statistical Society*, 37(1):53–86, 2007.

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

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Hansheng Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, 2009.

Pavel Yaskov. Controlling the least eigenvalue of a random gram matrix. *Linear Algebra and its Applications*, 504(1):108–123, 2016.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Xianyang Zhang and Guang Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768, 2017.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(90):2541–2563, 2006.

Junxian Zhu, Canhong Wen, Jin Zhu, Heping Zhang, and Xueqin Wang. A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences*, 117(52):33127–33123, 2020.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.