1. Introduction

During the past two decades, the dramatic improvement in data collection and acquisition technologies has enabled scientists to collect a great amount of high-dimensional data, for which the dimension $p$ can be much larger than the sample size $n$ (a.k.a. small-$n$-large-$p$). The current research on high-dimensional data mainly focuses on variable selection and graphical modeling. The former aims to find a consistent estimate for high-dimensional regression under a sparsity constraint. The existing methods include Lasso (Tibshirani 1996), SCAD (Fan and Li 2001), MCP (Zhang 2010), and rLasso (Song and Liang 2015a), among others. The latter aims to learn conditional independence relationships for dimension reduction is general and potentially can be extended to other high-dimensional or big data problems as well. Supplementary materials for this article are available online.

The desparsified Lasso method was proposed in van de Geer et al. (2014), which is also essentially the same as the one developed in Zhang and Zhang (2014) and Javanmard and Montanari (2014). For the high-dimensional linear regression $Y = X\beta + \epsilon$, where $\epsilon$ are zero-mean Gaussian random errors, desparsified Lasso defines a bias-corrected estimator

$$\hat{\beta}_{bc} = \hat{\beta}_{Lasso} + \hat{\Theta}X^T(y - X\hat{\beta}_{Lasso})/n,$$

where $\hat{\beta}_{Lasso}$ is the original Lasso estimator, and $\hat{\Theta}$ is an approximater to the inverse of $\hat{\Sigma} = X^TX/n$. From (1), one can obtain

$$\sqrt{n}(\hat{\beta}_{bc} - \beta) = \hat{\Theta}X^T\epsilon/\sqrt{n} + \sqrt{n}(I_p - \hat{\Theta}^\top)(\hat{\beta}_{Lasso} - \beta)$$

$$:= \hat{\Theta}X^T\epsilon/\sqrt{n} + \Delta_n,$$

where $I_p$ denotes the $p \times p$ identity matrix, and $\Delta_n$ is the error term. With an appropriate estimator $\hat{\Theta}$, for example, the one obtained by nodewise regression (Meinshausen and Bühlmann 2006), it can be shown that $||\Delta_n||_\infty = o_p(1)$ and thus $\sqrt{n}(\hat{\beta}_{bc} - \beta)$ shares the same asymptotic distribution with $\Theta X^T\epsilon/\sqrt{n}$. Further, to calculate confidence intervals for $\beta$, one needs to approximate the distribution of $\hat{\Theta}X^T\epsilon/\sqrt{n}$. For example, Javanmard and Montanari (2014) approximated it by $N(0, \hat{\sigma}^2\hat{\Theta}\Sigma\hat{\Theta}^T)$, where $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2$; and Zhang and Cheng (2017) approximated it using multiplier bootstrap.

The multi-sample-splitting method was proposed and analyzed in Meinshausen, Meier, and Bühlmann (2009), which works in the following procedure: Splitting the samples into $n$-large- $p$, the confidence interval and $p$-values for the $1, X_2, \ldots, X_p$, the proposed method is coined as Markov neighborhood regression (MNR). The proposed method is tested on high-dimensional linear, logistic, and Cox regression. The numerical results indicate that the proposed method significantly outperforms the existing ones. Based on the MNR, a method of learning causal structures for high-dimensional linear models is proposed and applied to identification of drug sensitive genes and cancer driver genes. The idea of using conditional independence relations for dimension reduction is general and potentially can be extended to other high-dimensional or big data problems as well. Supplementary materials for this article are available online.
two subsets equally, using the first half of samples for variable selection and using the second half of samples for calculating p-values based on the selected variables; repeating this process for many times; and aggregating the p-values for statistical inference. The confidence intervals can be constructed based on their duality with p-values. The idea about sample-splitting and subsequent statistical inference has also been implicitly contained in Wasserman and Roeder (2009). The multi-sample-splitting method is very general and can be applied to many different types of models. The ridge projection method was studied in Bühlmann (2013), which can be viewed as a direct extension of the low-dimensional ridge regression to the high-dimensional case. The bias of the ridge estimator has been extended of the low-dimensional ridge regression to the high-dimensional case. The bias of the ridge estimator has been assessed and approximated, and thus the method can be used for statistical inference.

The other methods include residual-type bootstrapping (Chatterjee and Lahiri 2013; Liu and Yu 2013), covariance test (Lockhart et al. 2014), and group-bound (Meinshausen 2015). A problem with the residual-type bootstrapping method is the super-efficiency phenomenon; that is, a confidence interval of the (1, 1)th element of \( X_1 \) is a set of features indexed by \( A \subset \mathbb{V} \), and use \( X_\mathbb{A} = (X_k : k \in \mathbb{A}) \) to denote a set of features indexed by \( A \subset \mathbb{V} \), and use \( X_\mathbb{P} \) to denote the joint probability distribution of \( X_\mathbb{V} \). For a triplet \( I, J, U \subset \mathbb{V} \), we use \( X_I \perp X_J | X_U \) to denote that \( X_I \) is conditionally independent of \( X_J \) given \( X_U \). A path of length \( l > 0 \) from a vertex \( v_0 \) to another vertex \( v_l \) is a sequence \( v_0, v_1, \ldots, v_l \) of distinct vertices such that \( v_{k+1} \perp v_k | v_{1:k} \) for \( k = 1, 2, \ldots, l \). The subset \( \mathbb{U} \subset \mathbb{V} \) is said to separate \( I \subset \mathbb{V} \) from \( J \subset \mathbb{V} \) if for every \( i \in I \) and \( j \in J \), all paths from \( i \) to \( j \) have at least one vertex in \( \mathbb{U} \). \( P_\mathbb{V} \) is said to satisfy the Markov property with respect to \( \mathbb{G} \) if for every triplet of disjoint sets \( I, J, U \subset \mathbb{V} \), it holds that \( X_I \perp X_J | X_U \) whenever \( U \) separates \( I \) and \( J \) in \( \mathbb{G} \). Let \( \tilde{\xi} = \{ k : \tilde{e}_k = 1 \} \) denote the neighboring set of \( X_I \) in \( \mathbb{G} \). Following from the Markov property of the Gaussian graphical model (GGM), we have \( X_I \perp X_J | X_{\tilde{\mathbb{S}}_I} \) for any \( i \in \mathbb{V} \setminus \tilde{\xi}_j \), as \( \tilde{\xi}_j \) forms a separator between \( X_I \) and \( X_J \). For convenience, we call \( \tilde{\xi}_j \) the minimum Markov neighborhood of \( X_i \) in \( \mathbb{G} \), and call any superset \( \tilde{\xi}_j \supset \xi_j \) a Markov neighborhood of \( X_i \) in \( \mathbb{G} \). The minimum Markov neighborhood is also termed as Markov blanket in Bayesian networks or general Markov networks.

To motivate the proposed method, we first look at a simple mathematical fact based on the well-known property of GGMs (see, e.g., Lauritzen 1996):

\[
X_i \perp X_j | X_{\mathbb{V} \setminus \{i,j\}} \iff \theta_{ij} = 0,
\]

where \( \theta_{ij} \) denotes the \((i, j)\)th entry of \( \Theta \). Without loss of generality, we let \( S_1 = \{ 2, \ldots, d \} \) denote a Markov neighborhood of \( X_1 \), let \( \Sigma_d \) denote the covariance matrix of \( \{ X_I \}_1 \times \mathbb{S}_1 \), and partition \( \Theta \) as

\[
\Theta = \begin{bmatrix}
\Theta_d & \Theta_{d:p-d} \\
\Theta_{p-d:d} & \Theta_{p-d}
\end{bmatrix},
\]

where the first row of \( \Theta_{d:p-d} \) and the first column of \( \Theta_{p-d:d} \) are exactly zero, as \( X_I \perp X_{\mathbb{V} \setminus \{i,j\}} | X_{\tilde{\mathbb{S}}_I} \). Inverting \( \Theta \), we have

\[
\Sigma_d^{-1} = \left[ \Theta_d - \Theta_{d:p-d} \Theta_{p-d:d}^{-1} \Theta_{p-d:d} \right]^{-1},
\]

which is equal to the top \( d \times d \)-submatrix of \( \Theta^{-1} \). Therefore,

\[
\Sigma_d^{-1} = \Theta_d - \Theta_{d:p-d} \Theta_{p-d:d}^{-1} \Theta_{p-d:d}.
\]

Since the first row of \( \Theta_{d:p-d} \) and the first column of \( \Theta_{p-d:d} \) are exactly zero, the \((1, 1)\)th element of \( \Theta_{d:p-d} \Theta_{p-d:d}^{-1} \Theta_{p-d:d} \)
Y = β₀ + X_Dj β_Dj + ε, \quad (7)

where the prime on β₀'s for k ≠ 1 indicates that those regression coefficients might be modified by the subset regression. Since Sₙ forms a Markov neighborhood of X₁ in the Markov network formed by all features, we call (6) an MNR, which breaks the high-dimensional inference problem into a series of low-dimensional inference problems. Based on this mathematical fact, we propose the following algorithm:

**Algorithm 1 (Markov neighborhood regression).**

(a) (Variable selection) Conduct variable selection for the model (3) to get a consistent estimate of Sₙ. Denote the estimate by ̂Sₙ.

(b) (Markov blanket estimation) Construct a GGM for X and obtain a consistent estimate of the Markov blanket for each variable. Denote the estimates by ̂ξ_j for j = 1, 2, . . . , p.

(c) (Subset regression) For each variable X_j, j = 1, . . . , p, let D_j = {j} ∪ ̂ξ_j ∪ ̂Sₙ and run an ordinary least square (OLS) regression with the features given by X_Dj, that is,

Y = β₀ + X_Dj β_Dj + ε, \quad (7)

where ε ~ N(0, σ²Iₚ) and Iₚ is an n × n-identity matrix. Conduct inference for β_j, including the estimate, confidence interval and p-value, based on the output of (7).

The Markov neighborhood corresponding to the subset regression (6) is {2, 3, . . . , d} ⊇ ̂ξ_j ∪ ̂Sₙ. In general, ̂ξ_j ∪ ̂Sₙ can be any a subset of {1, 2, . . . , p} depending on the ordering of features in (3). Algorithm 1 can be implemented in many different ways. For example, a variety of high-dimensional variable selection algorithms can be used for Step (a), for example, Lasso (Tibshirani 1996), SCAD (Fan and Li 2001), MCP (Zhang 2010), and rLasso (Song and Liang 2015a), which are all able to produce a consistent estimate for Sₙ under appropriate conditions. Similarly, a variety of graphical model learning algorithms can be used for Step (b), for example, graphical Lasso (Yuan and Lin 2007; Friedman, Hastie, and Tibshirani 2008), nodewise regression (Meinshausen and Bühlmann 2006), and ψ-learning (Liang, Song, and Qiu 2015), which all produce a consistent estimate for the underlying GGM under appropriate conditions. To justify Algorithm 1, we introduce the following low-dimensional linear regression and the mathematical relation shown around Equations (4) and (5). Refer to the supplementary materials for the detail of the proof.

**Lemma 1.** Let ̂ξ_j ⊇ ̂ξ_j denote any Markov neighborhood of feature x_j, let ̂Sₙ ⊇ Sₙ denote any reduced feature space, and let D_j = {j} ∪ ̂ξ_j ∪ ̂Sₙ. Consider the subset regression (7). Let ̂β_Dj denote the OLS estimator of β_Dj from the subset regression, and let ̂β_j denote the element of ̂β_Dj corresponding to the variable X_j. If |D_j| = o(n¹/₂), as n → ∞, the following results hold:

(i) \( \sqrt{n}(β_j - ̂β_j) \sim N(0, σ²θ_j), \) where θ_j is the (j, j)th entry of the precision matrix Θ.

(ii) \( \sqrt{n}\frac{β_j - ̂β_j}{\sqrt{σ_j²}} \sim N(0, 1), \) where \( ̂σ_j² = (y - x_Dj ̂β_Dj)^T(y - x_Dj ̂β_Dj)/(n - d - 1), \) θ_j is the (j, j)th entry of the matrix \( \frac{1}{n} \sum_{i=1}^{n} (x_Dj(x_Dj)^T)^{-1}. \)

**Remark 1.** For the case that n is finite, we have (n - |D_j| - 1) \( ̂σ_j²/σ_j² \sim χ²({n - |D_j| - 1}), \) independent of ̂β_Dj, by the standard theory of OLS estimation. Therefore, we can use t(n - |D_j| - 1) to approximate the distribution of \( \sqrt{n}\frac{β_j - ̂β_j}{\sqrt{σ_j²}}, \) that is, the estimate, p-value and confidence interval of β_j can be calculated from (7) as in conventional low-dimensional multiple linear regression.

**Lemma 1** implies that Algorithm 1 will be valid as long as the following conditions hold:

\( ̂ξ_j ⊇ ̂ξ_j, \forall j ∈ {1, 2, . . . , p}, \) \quad (8)

\( ̂Sₙ ⊇ Sₙ, \) \quad (9)

\( |D_j| = o(√n). \) \quad (10)

Condition (9) is the so-called screening property, which is known to be satisfied by many high-dimensional variable selection algorithms, such as SCAD (Fan and Li 2001), MCP (Zhang 2010), and adaptive Lasso (Zou 2006). Lasso also satisfies this condition if the design matrix satisfies the compatibility condition (van de Geer and Bühlmann 2009), and the beta-min condition holds. See Dezeure et al. (2015) for more discussions on this issue. Given the sure screening property of the above variable selection procedure, if the node-wise regression algorithm (Meinshausen and Bühlmann 2006) is applied to learn the GGM in Step (b) of Algorithm 1, then the condition (8) can be satisfied. In fact, as along as the GGM construction algorithm is consistent, the condition (8) will be asymptotically satisfied. Further, the condition (10) can be easily satisfied by a slight twist of the sparsity conditions used in the variable selection and GGM estimation algorithms.

As an example, we give in the Appendix a set of technical conditions (A0)–(A9) under which the conditions (8)–(10) can be asymptotically satisfied, provided that the SCAD algorithm is used for variable selection and the ψ-learning algorithm is used for GGM estimation. Based on these technical conditions, the validity of Algorithm 1 is justified in Theorem 1, whose proof is straightforward based on Slutsky’s theorem and some existing theoretical results. Refer to the supplementary materials for the detail. If different algorithms are used in Algorithm 1, then the conditions used in Theorem 1 should be changed accordingly. We note that many conditions we imposed in proving the theorem are purely technical and only serve to provide theoretical understanding of the proposed method. We have no intent to make the conditions the weakest possible.

**Theorem 1 (Validity of Algorithm 1).** If the conditions (A0)–(A9) hold, the SCAD algorithm is used for variable selection in Step (a), and the ψ-learning algorithm is used for GGM construction in Step (b), then for each j ∈ {1, 2, . . . , pₙ}, we
have $\sqrt{n}(\hat{\beta}_j - \beta_j) \sim N(0, 1)$ as $n \to \infty$, where $\hat{\beta}_j$ denotes the estimate of $\beta_j$ obtained from the subset regression, $\hat{\sigma}_j^2 = (y - x_{D_j} \hat{\beta}_{D_j})^T (y - x_{D_j} \hat{\beta}_{D_j})/(n - d - 1)$, $\hat{\sigma}_j$ is the $(j, j)$th entry of the matrix $\left[ \frac{1}{n} \sum_{i=1}^n x_{D_j}^{(i)^T} x_{D_j}^{(i)^T} \right]^{-1}$, and $x_{D_j}^{(i)}$ denotes the $i$th row of $X_{D_j}$.

**Remark 2.** Following Remark 1, we can conduct inference for $\beta_j$ based on the output of the subset regression (7) as in conventional low-dimensional multiple linear regression.

As implied by Theorem 1, variable selection for regression (3) can be converted as a multiple hypothesis testing problem for simultaneously testing the hypotheses

$$H_{0,j} : \beta_j = 0 \iff H_{1,j} : \beta_j \neq 0, \quad j = 1, 2, \ldots, p,$$

based on the $p$-values obtained from the subset regressions. The consistency of this test-based method follows from Theorem 2 of Liang, Song, and Qiu (2015) as discussed in Section 4.1. Compared to the regularization methods, for example, Lasso, MCP, SCAD, and adaptive Lasso (Zou 2006), a significant advantage of this method is that it controls the false discovery rate of selected features in an explicit way. In addition, since the screening property generally holds for these regularization methods, see Dezeure et al. (2015) for discussions on this issue, the new method might result in a lower false discovery rate as shown in Table 2. On the other hand, since the $p$-value measures the contribution of a feature to the regression conditioned on all other $p - 1$ features, MNR might not work well when strong collinearity exists among certain true and false features. This case has been excluded by Conditions A2 and A4, where the fixed upper bounds on correlations and $\psi$-partial correlations place some additional restrictions on the design.

The essential conditions required by MNR are only sparsity; that is, the true regression model is sparse and the conditional independence structure among the features is sparse such that (8)–(10) hold when appropriate algorithms are applied. Similar conditions have been assumed by some existing methods. For example, desparsified-Lasso requires the true model to be of size $O(\sqrt{n}/\log p)$ (see, e.g., Dezeure et al. 2015, Fact 2), which is a little more restrictive than $O(n^{1/2})$ required by MNR; desparsified-Lasso also requires the precision matrix $\Theta$ to be row-sparse at a level of $O(n/\log p)$, which is comparable with the Markov blanket size $O(\sqrt{n})$ required by MNR when $\log(p) = n^\delta$ for some $\delta \approx 1/2$. The multi-sample-splitting and the ridge projection methods require the screening property (9) only, which seems weaker than the conditions required by MNR and desparsified-Lasso. However, as shown later by numerical examples, they both essentially fail even for the simple linear regression case. The use of conditional independence relations seems important for high-dimensional inference.

For MNR, since the essential conditions are (8)–(10), a variable screening-based algorithm will also work under appropriate conditions. Based on this observation, we propose Algorithm S1, which together with some numerical results are presented in the supplementary materials. Compared to Algorithm 1, Algorithm S1 can be substantially faster but the resulting confidence intervals can be a little wider; that is, Algorithm S1 is an accuracy/efficiency trade-off version of Algorithm 1.

Finally, we note that there are indeed scenarios that conditions (8)–(10) are violated. For example, if all the features are equally correlated or there are a few features whose Markov blanket is of size $O(\sqrt{n})$ or larger, then the condition (8) will be violated, as the algorithm always restricts the Markov blanket to be of size $o(\sqrt{n})$ or smaller. Similarly, if the true model is of size $O(\sqrt{n})$ or larger, then condition (9) will be violated. These conditions can also be violated by the algorithms used for Markov blanket estimation or variable selection, particularly when the sample size is small. The screening property is itself a large sample property. Our numerical experience shows that the MNR method is pretty robust to violations of the conditions (8)–(10). This will be demonstrated in Section 3.4.

### 2.2. Generalized Linear Models

The MNR method can be easily extended to the generalized linear models (GLMs) whose density function is given in the canonical form

$$f(y|x, \beta) = \exp(\vartheta y - b(\vartheta) + c(y)),$$

where $b(\cdot)$ is continuously differentiable, and $\vartheta$ is the natural parameter relating $y$ to the features $x$ via a linear function $\vartheta = \beta_0 + x_1 \beta_1 + \cdots + x_p \beta_p$. This class of GLMs includes Poisson regression, logistic regression, and linear regression (with known variance). Note that for Cox proportional hazards models, the parameters can be estimated by maximizing the partial likelihood function (Cox 1975), based on which the Cox regression can be converted to a Poisson regression. See, for example, McCullagh and Nelder (1989, chap. 13) for the detail. This conversion is important, which enables the use of the MNR method for high-dimensional survival data analysis.

To justify this extension, we establish the following lemma, where we assume that the features follow a multivariate normal distribution $N(0, \Sigma)$ and each has been standardized to have a mean 0 and variance 1. The proof follows the same logic as that of Lemma 1, but the precision matrix involved in Lemma 1 is replaced by the inverse of the Fisher information matrix of the GLM. Refer to the supplementary materials for the detail.

**Lemma 2.** Let $\hat{\xi}_j \supseteq \xi_j$ denote any Markov neighborhood of feature $x_j$, let $\hat{S}_j \supseteq S_j$ denote any reduced feature space, and let $D_j = |j| \cup \hat{\xi}_j \cup \hat{S}_j$. Consider a subset GLM with the features $X_{D_j}$, let $\hat{\beta}_{D_j}$ denote the MLE of $\beta_{D_j}$, and let $\hat{\beta}_j$ denote the component of $\hat{\beta}_{D_j}$ corresponding to feature $X_j$. If $|D_j| = o(n^{1/2})$, then, as $n \to \infty$, the following results hold:

(i) $\sqrt{n}(\hat{\beta}_j - \beta_j) \sim N(0, k_{jj})$, where $k_{jj}$ denotes the $(j, j)$th entry of the inverse of the Fisher information matrix $K = F^{-1} = [E(b''(\vartheta) \vartheta \vartheta^T)]^{-1}$, and $\beta$ denotes the true regression coefficients.

(ii) $\sqrt{n}(\hat{\beta}_j - \beta_j)/\sqrt{k_{jj}} \sim N(0, 1)$, where $\hat{k}_{jj}$ denotes the $(j, j)$th entry of the inverse of the observed information matrix.
WefirsttesttheconceptofMNRusingalarge-

Figure 1. The 95% confidence intervals of $\beta_1, \ldots, \beta_p$ produced by the MNR (solid line) and OLS (dashed line) methods for a dataset with $n = 2000$ and $p = 50$.

$$J_n(\hat{\beta}_{D_1}) = -\sum_{i=1}^{n} H_{\beta_{D_1}} (\log f(y_i|\beta_{D_1}, x_i) )/n \text{ and } H_{\beta_{D_1}} (\cdot)$$
denotes the Hessian matrix evaluated at the MLE $\hat{\beta}_{D_1}$.

Lemma 2 implies that the estimate, $p$-value and confidence interval of $\beta_j$ can be calculated from the subset GLM as in conventional low-dimensional GLMs. For GLMs, variable selection can be done using the SCAD, MCP, or Lasso algorithm, and variable screening can be done using the sure independence screening algorithm developed in Fan and Song (2010). By Theorem 5 of Fan and Song (2010), we can bound the size of $\hat{S}_n$ by $O(n^{1/2})$ for a small constant $\varepsilon > 0$ with a slight modification of the technical conditions therein. Therefore, the theorems parallel to Theorem 1 and Theorem S1 (in the supplementary materials) can be proved for GLMs. For simplicity, they are omitted in the article.

2.3. Joint Inference

The MNR method described above deals with only one coefficient $\beta_j$ in each subset regression. In fact, it can be easily extended to conduct joint inference for several coefficients. Let $A \subset V$ denote a set of features for which the joint inference for the corresponding coefficients is desired. Define $\xi_A = \bigcup_{j \in A} \xi_j$ as the union of the Markov blankets of the features in $A$. Let $M = A \cup \xi_A \cup \hat{S}_n$. Then a subset regression can be conducted with the features included in $M$. For high-dimensional linear regression, if $|M| = O(n^{1/2})$, then, similar to Theorem 1, we can show $\sqrt{n}(\hat{\beta}_A - \beta_A) \sim N(0, \sigma^2 \Theta_{AA})$, where $\Theta_{AA}$ denotes the submatrix of the precision matrix $\Theta$ constructed by its $A$ rows and $A$ columns. Similarly, for high-dimensional GLMs, we can show $\sqrt{n}(\hat{\beta}_A - \beta_A) \sim N(0, K_{AA})$, where $K_{AA}$ denotes the submatrix of $K = [E(x'x^T \beta)xx^T]^{-1}$ constructed by its $A$ rows and $A$ columns.

3. Simulation Studies

3.1. A Conceptual Experiment

We first test the concept of MNR using a large-$n$-small-$p$ example; that is, whether the confidence intervals generated by MNR coincide with those generated by the OLS method as the sample size $n$ becomes large. We generated a dataset from the model (3) with $n = 2000$ and $p = 50$, where $\sigma^2$ was set to 1, the covariates $X$ were generated from a zero-mean multivariate Gaussian distribution with a Toeplitz covariance matrix given by $\Sigma_{ij} = 0.9|i-j|$ for $i, j = 1, \ldots, p$, and the true regression coefficients $(\beta_0, \beta_1, \beta_2, \ldots, \beta_5) = (1, 0.2, 0.4, -0.3, -0.5, 1.0)$ and $\beta_6 = \cdots = \beta_p = 0$.

Figure 1 compares the 95% confidence intervals of $\beta_1, \ldots, \beta_p$ produced by MNR and OLS with the simulated dataset. For MNR, the nodewise regression algorithm (with SIS-Lasso performed for each node) was employed for Markov blanket estimation, and SIS-SCAD was employed for variable selection. The SIS-Lasso refers to a variable selection procedure implemented in the package SIS (Saldana and Feng 2018), where the sure independence screening (SIS) algorithm (Fan and Lv 2008) was first applied for variable screening and then the Lasso algorithm was applied to select variables from those survived from the screening procedure. The SIS-SCAD and SIS-MCP can be interpreted in the same way. As expected from Theorem 1, OLS and MNR produced almost identical confidence intervals for each regression coefficient. In this simulation, we set $n$ excessively large, which ensures the convergence of the sample covariance matrix to the true covariance matrix $\Theta^{-1}$. In fact, MNR can work well with a smaller value of $n$ as illustrated by the following small-$n$-large-$p$ examples.

3.2. An Illustrative Example

To illustrate the performance of MNR, we generated 100 independent datasets from the regression (3), where $n = 200$, $p = 500$, $\sigma^2 = 1$, the features were generated from a zero-mean multivariate Gaussian distribution with a Toeplitz covariance matrix given by $\Sigma_{ij} = 0.9|i-j|$ for $i, j = 1, \ldots, p$, and the true regression coefficients were given by $(\beta_0, \beta_1, \beta_2, \ldots, \beta_5) = (1, 2, 4, -3, -5, 10)$ and $\beta_6 = \cdots = \beta_p = 0$. We note that the same covariance matrix has been used in van de Geer et al. (2014) to illustrate the performance of desparsified-Lasso. For convenience, we call this model a Toeplitz-covariance linear regression model.
Table 1. Coverage rates and widths of the 95% confidence intervals produced by MNR with Algorithm 1 for the Toeplitz-covariance linear regression model, where "signal" and "noise" denote nonzero and zero regression coefficients, respectively.

| Measure | Desparsified-Lasso | Ridge | Multi-Split | MNR |
|---------|-------------------|-------|-------------|-----|
| Coverage | Signal | 0.384(0.049) | 0.576(0.049) | 0.202(0.040) | 0.956(0.021) |
|         | Noise  | 0.965(0.018) | 0.990(0.010) | 1.000(6.4e-4) | 0.950(0.022) |
| Width   | Signal | 0.673(0.005) | 1.086(0.010) | 2.711(0.097) | 0.822(0.011) |
|         | Noise  | 0.691(0.005) | 1.143(0.008) | 2.790(0.103) | 0.869(0.007) |

NOTE: For "signal," the reported mean value and standard deviation (in the parentheses) are defined in (13) and (14), respectively. For "noise," they are defined similarly.

Table 2. Variable selection for the Toeplitz-covariance linear regression with the MNR, SIS-SCAD, SIS-MCP, and SIS-Lasso methods.

| Measure | q = 0.0001 | q = 0.001 | q = 0.01 | SIS-SCAD | SIS-MCP | SIS-Lasso |
|---------|-------------|-----------|-----------|----------|---------|----------|
| FSR     | 0           | 0.004     | 0.022     | 0.127    | 0.175   | 0.819    |
| NSR     | 0           | 0         | 0         | 0        | 0       | 0        |

3.2.1. Illustration of MNR

Algorithm 1 was run for this example as in Section 3.1, that is, applying SIS-SCAD for variable selection and the node wise regression algorithm for Markov blanket estimation. Table 1 summarizes the coverage rates and widths of the 95% confidence intervals produced by MNR for each regression coefficient. For the non-zero regression coefficients (denoted by "signal"), the mean coverage rate and mean width of the confidence intervals are defined, respectively, by

\[ \tilde{P}_{\text{cover}} = \frac{100}{\sum_{j=1}^{100} \sum_{i \in S_s} \hat{p}_i^{(j)}/(100 \cdot |S_s|)} \]

\[ \tilde{w}_{\text{CI}} = \frac{100}{\sum_{j=1}^{100} \sum_{i \in S_s} \hat{w}_i^{(j)}/(100 \cdot |S_s|)} \]

and their respective standard deviations are defined by

\[ \sigma(\tilde{p}_{\text{cover}}) = \sqrt{\text{var}(\hat{p}_i^{(j)} : i \in S_s, j = 1, 2, \ldots, 100)}/100, \]

\[ \sigma(\tilde{w}_{\text{CI}}) = \sqrt{\text{var}(\hat{w}_i^{(j)} : i \in S_s, j = 1, 2, \ldots, 100)}/100, \]

where \( \hat{p}_i^{(j)} \) denotes the width of the 95% confidence interval of \( \beta_i \) constructed with the \( j \)th dataset, \( \hat{p}_i^{(j)} \in \{0, 1\} \) indicates the coverage of \( \beta_i \) by the confidence interval, and \( \text{var}\{\cdot\} \) denotes the variance. By dividing \( 100 \) in (14), the standard deviation represents the variability of the mean value (averaged over 100 independent datasets) for a single regression coefficient. For the zero regression coefficients (denoted by "noise"), the mean coverage rate, the mean width, and their standard deviations can be defined similarly.

For comparison, we applied the desparsified Lasso, ridge projection, and multi-split methods to this example. These methods have been implemented in the R package hdi (Meier et al. 2016). The comparison indicates that MNR significantly outperforms the existing methods: for both the nonzero and zero regression coefficients, the mean coverage rates produced by MNR are much closer to their nominal level. The reason why desparsified-Lasso suffers from coverage deficiency for nonzero regression coefficients will be explained in Section 3.3.1.

As discussed previously, MNR converts the problem of variable selection as a multiple hypothesis testing problem. To illustrate the potential of MNR in variable selection, we converted the \( p \)-values produced by the subset regressions to \( z \)-scores using the inverse probability integral transformation

\[ Z_i^{(j)} = \Phi^{-1}(1 - q_i^{(j)}), \quad i = 1, 2, \ldots, p, \quad j = 1, 2, \ldots, 100, \]

where \( q_i^{(j)} \) denotes the \( p \)-value calculated via subset regression for feature \( i \) with dataset \( j \), and \( \Phi(\cdot) \) denotes the CDF of the standard Gaussian distribution. Figure S1 (in the supplementary materials) shows the histogram of the \( z \)-scores, which indicates that the true and false features can be well separated by the \( z \)-scores. The empirical Bayesian method developed by Liang and Zhang (2008) was applied to each of the 100 datasets for simultaneously testing the hypotheses (11). At a FDR level of \( q = 0.0001 \), which is measured by the \( q \)-value of Storey (2002), the method led to exact identifications of the true and false features for all 100 datasets, that is, both the false selection rate (FSR) and negative selection rate (NSR) are 0. More results were shown in Table 2. Here, the FSR and NSR are defined by

\[ \text{FSR} = \sum_{j=1}^{100} \frac{|\hat{S}_j \setminus S_s|}{|\hat{S}_j|}, \quad \text{NSR} = \sum_{j=1}^{100} \frac{|S_s \setminus \hat{S}_j|}{|S_s|}, \]

where \( S_s \) is the set of true features, and \( \hat{S}_j \) is the set of selected features for dataset \( j \). For comparison, SIS-SCAD, SIS-MCP, and SIS-Lasso were applied to these datasets for performing variable selection under their default settings in the package SIS. Table 2 shows that MNR can significantly outperform the existing methods in high-dimensional variable selection. As mentioned previously, compared to the existing methods, a significant advantage of the MNR-based variable selection method is that it controls the FDR of selected features.

3.2.2. Illustration of Joint Inference With MNR

To illustrate the use of MNR for joint inference, we constructed Bonferroni joint confidence intervals based on the subset regression for each of the following sets of parameters: \( (\beta_1, \beta_2), (\beta_3, \beta_4, \beta_5), (\beta_6, \beta_7), (\beta_8, \beta_{10}), \) and \( (\beta_{20}, \beta_{200}, \beta_{400}) \), which have covered the cases of combinations of non-zero coefficients, combinations of zero and nonzero coefficients, and combinations of zero coefficients. For each set of parameters, as described in Section 2.3, the subset regression was constructed by unifying the Markov neighborhoods of the corresponding...
constructed as in Section 3.1 using nodewise regression for roni method. The Markov neighborhood of each feature was set of parameters were constructed using the standard Bonfer-oni method. The algorithm consists of two screening stages. The first stage is correlation screening, which, via a multiple hypothesis test for correlation coefficients, determines for each feature a conditioning set used for calculating the \( \psi \)-partial correlation coefficient. The second stage is \( \psi \)-partial correlation screening, which, via a multiple hypothesis test for \( \psi \)-partial correlation coefficients, determines the GGM. Corresponding to the two stages, the algorithm consists of two tuning parameters, \( \alpha_1 \) and \( \alpha_2 \), which specify the significance levels of the two multiple hypothesis tests, respectively. In all applications of the \( \psi \)-learning algorithm in this article, we set \( \alpha_1 = 0.1 \) and \( \alpha_2 = 0.05 \) as suggested by Liang, Song, and Qiu (2015). In general, \( \alpha_1 \) should be slightly large to avoid potential loss of important features in the conditioning set of each feature. The nodewise regression algorithm has also been applied to this example for Markov blanket estimation, and the results are similar.

For comparison, the desparsified Lasso and ridge projection methods were applied to this example. Both methods have been implemented in the R package \texttt{hdi} (Meier et al. 2016). The multi-split method is also available in \texttt{hdi}, but it often suffered from a convergence issue in applications to this example and thus not included for comparison. Table 4 shows that MNR significantly outperforms the existing methods: The coverage rates produced by MNR are almost identical to their nominal levels for both zero and nonzero regression coefficients; while the coverage rates produced by the other methods are far from their nominal levels, especially for the nonzero regression coefficients.

For the nonzero regression coefficients, the confidence intervals produced by desparsified-Lasso have about the same widths as those by MNR, but the former have much lower coverage rates. The coverage deficiency of desparsified-Lasso are due to at least two reasons: (i) the bias-corrected estimator \( \hat{\beta}_{bc} \) is still biased; and (ii) the required sparsity condition is violated. The bias of \( \hat{\beta}_{bc} \) can be easily seen from the derivation procedure of \( \hat{\beta}_{bc} \), which is due to Zhang and Zhang (2014). Let \( Z_j \) denote the

### 3.3. Simulation Studies With More Regression Models

#### 3.3.1. Linear Regression

We simulated 100 independent datasets from the linear regression (3) where \( n = 200, p = 500, \sigma^2 = 1 \), the features were generated from a zero-mean Gaussian distribution with the precision matrix \( \Sigma^{-1} = \Theta = (\theta_{ij}) \) given by

\[
\theta_{ij} = \begin{cases} 
0.5, & \text{if } |j - i| = 1, i = 2, \ldots, (p - 1), \\
0.25, & \text{if } |j - i| = 2, i = 3, \ldots, (p - 2), \\
1, & \text{if } j = i, i = 1, \ldots, p, \\
0, & \text{otherwise},
\end{cases}
\]  

and the regression coefficients were given by \( (\beta_0, \beta_1, \beta_2, \ldots, \beta_5) = (1, 2, 2.5, 3, 3.5, 4) \) and \( \beta_6 = \cdots = \beta_p = 0 \).

Since the precision matrix has an autoregressive (AR) structure, for convenience, we call this model an AR(2)-precision linear regression model.

Algorithm 1 was first applied to this example with the numerical results summarized in Table 4. The \( \psi \)-learning algorithm has been implemented in the R-package \texttt{equiUSA} (Jia et al. 2018). It provides an equivalent measure of the partial correlation coefficient, the so-called \( \psi \)-partial correlation coefficient, for estimating GGMs under the small-n-large-\( p \) scenario.

### Table 3. Coverage rates of 95% joint confidence intervals produced by MNR for the set of parameters: \((\beta_1, \beta_2), (\beta_3, \beta_4, \beta_5), (\beta_1, \beta_6), (\beta_7, \beta_{10}), \) and \((\beta_{20}, \beta_{200}, \beta_{400})\), where the number in the parentheses represents the standard deviation of the joint coverage rate averaged over 100 independent datasets.

| Parameters | \((\beta_1, \beta_2)\) | \((\beta_3, \beta_4, \beta_5)\) | \((\beta_1, \beta_6)\) | \((\beta_7, \beta_{10})\) | \((\beta_{20}, \beta_{200}, \beta_{400})\) |
|------------|----------------|----------------|----------------|----------------|----------------|
| Joint coverage rate | 0.97(0.017) | 0.95(0.022) | 0.93(0.026) | 0.97(0.017) | 0.93(0.026) |

### Table 4. Coverage rates and widths of the 95% confidence intervals produced by MNR, desparsified-Lasso, and ridge projection for the AR(2)-precision linear, logistic, and Cox regression.

| Response | Measure | Desparsified-Lasso | Ridge | MNR |
|----------|---------|--------------------|-------|-----|
| Gaussian | Coverage | Signal | 0.230(0.0421) | 0.3340(0.0447) | 0.9500(0.0218) |
|          |         | Noise | 0.9640(0.0186) | 0.9920(0.0088) | 0.9503(0.0217) |
|          | Width   | Signal | 0.2810(0.0027) | 0.4481(0.0043) | 0.2806(0.0022) |
|          |         | Noise | 0.2723(0.0024) | 0.4335(0.0036) | 0.2814(0.0024) |
| Binary   | Coverage | Signal | 0.0040(0.0063) | 0(0) | 0.9320(0.0252) |
|          |         | Noise | 0.9953(0.0068) | 1.0(4.5e-4) | 0.9373(0.0242) |
|          | Width   | Signal | 0.6424(0.0101) | 1.0775(0.0110) | 1.9473(0.0529) |
|          |         | Noise | 0.5782(0.0081) | 1.0100(0.0095) | 0.9799(0.0132) |
| Survival | Coverage | Signal | – – | 0.9140(0.0281) | – – |
|          |         | Noise | – – | 0.9354(0.0246) | – – |
|          | Width   | Signal | – – | 0.3356(0.0018) | – – |
|          |         | Noise | – – | 0.2683(0.0017) | – – |

NOTE: Refer to the caption of Table 1 for the notation.
residual of the regression $X_j$ versus all other features $X[-j]$, and let $P_{jk} = X_k^T Z_j / X_j^T Z_j$. Then the following identity holds
\[ Y'Z_j / X_j^T Z_j = \beta_j + \sum_{k \neq j} P_{jk} \hat{\beta}_k + \epsilon' Z_j / X_j^T Z_j, \] (17)
where $Y$ and $\epsilon$ are as defined in (3). Plugging the Lasso estimator $\hat{\beta}_{\text{Lasso}}$ (of the regression $Y$ versus $X$) into (17) leads to the bias-corrected estimator
\[ \hat{\beta}_{bc,j} = Y'Z_j / X_j^T Z_j - \sum_{k \neq j} P_{jk} \hat{\beta}_{\text{Lasso},k} \]
\[ = \hat{\beta}_{\text{Lasso},j} + Z_j'(Y - X \hat{\beta}_{\text{Lasso}}) / Z_j' X_j, \quad j = 1, 2, \ldots, p, \] (18)
which is essentially the same with the estimator given in (1). Here $\hat{\beta}_{bc,j}$ and $\hat{\beta}_{\text{Lasso},j}$ denote the $j$th component of $\hat{\beta}_{bc}$ and $\hat{\beta}_{\text{Lasso}}$, respectively. The $\hat{\beta}_{bc}$ can have the bias of $\hat{\beta}_{\text{Lasso}}$ much corrected. However, as implied by (17), $\hat{\beta}_{bc}$ is still generally biased because the Lasso estimator $\hat{\beta}_{\text{Lasso}}$ is generally biased. Such a biased estimator shifts the center of the confidence interval and thus leads to the coverage deficiency problem. For the error term $\Delta_n$ defined in (2), Dezeure et al. (2015) proved that it is negligible if the sparsity condition $|S_n| = o(\sqrt{n}/\log(p))$ holds, the precision matrix is row-sparse at a level of $o(n/\log(p))$, and some other regularity conditions on the design matrix hold. Among these conditions, the model sparsity condition $|S_n| = o(\sqrt{n}/\log(p))$ is a little restrictive and can be easily violated. For example, for a problem with $|S_n| = 5$ and $p = 500$, the sample size $n$ should be at least a few thousands to satisfy the condition $s_0 \ll \sqrt{n}/\log(p)$. As the result, the error term $\Delta_n$ might not be negligible, which can also cause the coverage deficiency issue. Since $||\hat{\beta}_{\text{Lasso}} - \beta||_1 = Op(|S_n|\sqrt{\log(p)/n})$ (Dezeure et al. 2015), violation of the sparsity condition also worsens the bias of $\hat{\beta}_{bc}$. We note that the model sparsity condition $|S_n| = o(\sqrt{n})$ required by MNR is much weaker than $|S_n| = o(\sqrt{n}/\log(p))$ under the small-$n$-large-$p$ scenario.

In our numerical experience, the coverage deficiency of desparsified-Lasso is mainly due to the bias of $\hat{\beta}_{bc}$. We illustrate this issue using two simulation studies. The first one is given as follows and the other one is given in Section 3.4. In Table 5, we reported the values of $\hat{\beta}_{bc,j}$'s, $j = 1, 2, \ldots, 8$, for the AR(2)-precision linear regression. It is easy to see that the desparsified-Lasso estimate is severely biased for the nonzero coefficients $\beta_1, \ldots, \beta_5$, which significantly shifts the centers of the resulting confidence intervals and thus leads to the coverage deficiency problem. Note that $\hat{\beta}_{bc,6}$ is also biased due to the strong correlation between $X_6$ and $X_5$. For comparison, we included in Table 5 the MNR estimates of these coefficients, which are unbiased for both zero and nonzero coefficients.

### 3.3.2. Logistic Regression

We simulated 100 datasets for a logistic regression. For each dataset, we set $n = 300$, $p = 500$, $(\beta_0, \beta_1, \ldots, \beta_5) = (1, 2, 5, 3, 3.5, 4)$, $\beta_6 = \cdots = \beta_p = 0$, and generated the covariates from a zero-mean multivariate Gaussian distribution with the precision matrix given by (16). For convenience, we call this model an AR(2) precision logistic regression model. Each dataset consisted of 150 case samples and 150 control samples. To alleviate the convergence issues suffered by the GLM estimation procedure glm in R, we set $n$ slightly large for this example.

Algorithm 1 was run for the datasets, where the SIS-MCP learning algorithm was employed for variable selection and the $\psi$-learning algorithm was employed for Markov blanket estimation. The nodewise regression algorithm was also applied for Markov blanket estimation, the results were similar. The numerical results were summarized in Table 4, which indicates that MNR significantly outperforms the other methods. Desparsified-Lasso and ridge projection essentially fail for this example.

### 3.3.3. Cox Regression

For Cox regression, which is also known as Cox proportional-hazards model, we let $\lambda(t)$ denote the hazard rate at time $t$ and let $\lambda_0(t)$ denote the baseline hazard rate. The Cox model can then be expressed as
\[ \lambda(t) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p). \] (19)

In the simulation, we set $(\beta_1, \ldots, \beta_5) = (1, 1, 1, 1, 1)$, $\beta_6 = \cdots = \beta_p = 0$, the baseline hazard rate $\lambda_0(t) = \lambda_0 = 0.1$, and the censoring hazard rate $\lambda_c = 1$; generated the event time from the Weibull distribution with the shape parameter $= 1$ and the scale parameter $= \lambda_0 \exp(-\sum_{i=1}^p X_i \beta_i)$; generated the censoring time from the Weibull distribution with the shape parameter $= 1$ and the scale parameter $= \lambda_c$; set the observed survival time as the minimum of the event time and the censoring time for each subject; and generated the features $X_1, \ldots, X_p$ from a zero-mean multivariate normal distribution with the precision matrix given by (16). For convenience, we call this model an AR(2)-precision Cox regression model. We simulated 100 datasets from this model with $n = 300$ and $p = 500$.

Algorithm 1 was run for the datasets, where the SIS-Lasso algorithm was used for variable selection and the $\psi$-learning

| Method | Measure | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ |
|--------|---------|----------|----------|----------|----------|----------|----------|----------|----------|
| -      | True    | 2        | 2.5      | 3        | 3.5      | 4        | 0        | 0        | 0        |
| Desparsified | $\hat{\beta}_{bc}$ | 1.841    | 2.274    | 2.698    | 3.270    | 3.849    | -0.051   | -0.007   | 0.016    |
|         | SD      | (0.008)  | (0.009)  | (0.009)  | (0.007)  | (0.007)  | (0.006)  | (0.007)  | (0.007)  |
| MNR    | $\hat{\beta}_{MNR}$ | 1.997    | 2.503    | 2.994    | 3.498    | 4.001    | 0.014    | 0.004    | 0.004    |
|         | SD      | (0.006)  | (0.008)  | (0.008)  | (0.007)  | (0.006)  | (0.007)  | (0.008)  | (0.008)  |
algorithm was used for Markov blanket estimation. The numerical results were summarized in Table 4. The nodewise regression algorithm was also applied for Markov blanket estimation, the results were similar. In the MNR results, we can observe some slight bias, which mainly comes from the model selection error and the estimation error of the Markov blanks. Our numerical experience shows that the nominal level can be reached by MNR with the correct model and correct Markov neighborhoods or when n becomes large.

In addition to coverage rates, Table 4 reports mean widths of the confidence intervals resulted from different methods. For linear regression, the confidence intervals by MNR are narrower than those by ridge projection, and of about the same width as those by desparsified-Lasso. However, as analyzed previously, desparsified-Lasso often suffers from the coverage deficiency issue. For logistic and Cox regression, the comparison is not meaningful, as the other methods either fail or are not available.

To explore the potential of MNR in variable selection, we have calculated z-scores in (15) based on the p-values generated by MNR for the datasets simulated above. Figures S2–S4 show the histograms of the z-scores, which indicate that the true and false features can always be well separated by z-scores for all these datasets. This is an attractive feature of MNR and its use for feature selection will be further explored in Section 4.

### 3.4. Robustness of Markov Neighborhood Regression

This section studies the robustness of MNR to violations of the conditions (8)–(10). This issue has been partially studied in Section 2.2 of the supplementary materials, where the condition (9) is violated when the size of \( \hat{S}_j \) is restricted to 3. Recall that for the Toeplitz-covariance regression, we have \(|S_1| = 5, |\xi_j| = 2\) for \( j = 2, 3, \ldots, p - 1 \), and \(|\xi_j| = 1\) for \( j = 1 \) and \( p \). Therefore, setting \(|S_1| = 3\) leads to some true features missed in each subset regression. As shown in Table S1 (in the supplementary materials), this results in wider confidence intervals for both zero and nonzero regression coefficients, although the coverage rates are not much affected.

In what follows, we consider one linear regression example where the features were generated from a zero-mean multivariate Gaussian distribution with the covariance matrix given by

\[
\Sigma_{ij} = 0.8, \quad \text{for all } i \neq j, \quad \Sigma_{ii} = 1 \quad \text{for all } i.
\]

We set \( p = 500, n = 300, (\beta_0, \beta_1, \ldots, \beta_{10}) = (1, 2, 2.5, 3, 3.5, 4, 5, 6, 7, -8, -9), \) and \( \beta_{11} = \cdots = \beta_p = 0 \), and generated 100 independent datasets in total. For convenience, we will call this model an equi-correlation linear regression model. The same model has been used in van de Geer et al. (2014) to illustrate the performance of desparsified-Lasso, but with different sample sizes and regression coefficients. For this example, it is easy to see that for each feature \( x_j \), the Markov blanket \( \xi_j \) consists of all other \( p - 1 \) features. That is, the condition (8) is violated, as we always restrict the Markov blanket to be much smaller than \( p \).

Algorithm 1 was first applied to this example, where SIS-MCP was used for variable selection, and nodewise regression (with SIS-Lasso performed for each node) was used for Markov blanket estimation. All the algorithms were run under their default setting in the R package SIS. For comparison, desparsified-Lasso and ridge projection methods were also applied to this example. Both methods were run under their default settings in the R package hdi. The numerical results were summarized in Table 6, which indicates that MNR is pretty robust to the misspecification of the Markov blanket for this example. In terms of mean coverage rates and widths, MNR produced most accurate confidence intervals compared to the desparsified-Lasso and ridge projection methods.

Compared to the results reported in Table 4 for the AR(2)-precision linear regression, desparsified-Lasso works much better for this example. For the AR(2)-precision linear regression, Table 5 shows that \( \hat{\beta}_{bc} \) is severely biased and thus the method suffers from coverage deficiency for nonzero coefficients. To have this issue further explored, we reported in Table 7 \( \hat{\beta}_{bc} \) and \( \hat{\beta}_{MNR} \) for the nonzero coefficients \( \beta_1, \beta_2, \ldots, \beta_{10} \). The comparison with the true value shows that \( \hat{\beta}_{bc} \) is nearly unbiased for \( \beta_1, \ldots, \beta_{10} \), although it is systematically smaller than the true value in magnitudes. As the result, desparsified-Lasso produced a good coverage rate for the nonzero coefficients of this example. MNR continuously works well; \( \beta_{MNR} \) is unbiased and accurate for this example.

In summary, MNR is robust to misspecification of the Markov neighborhood. It will perform reasonably well as long as for each subset regression, the Markov neighborhood has covered the major contributors of the subset regression, which include the most significant features to the original regression as well as the most correlated features to the target feature of the subset regression.

### 3.5. Computational Complexity

The MNR method consists of three steps, variable selection, Markov blanket estimation, and subset regression. Its computational complexity is typically dominated by the algorithm used for Markov blanket estimation.

For Algorithm 1, if the Lasso algorithm is employed for variable selection, then, by Meinshausen (2007), the compu-
4. Causal Structure Discovery for High-Dimensional Regression

The causal relationship for a pair or more variables refers to a persistent association which is expected to exist in all situations without being affected by the values of other variables. Due to its attractive feature, which does not only allow better explanations for past events but also enables better predictions for the future, causal discovery has been an essential task in many disciplines. Since, for high-dimensional problems, it is difficult and expensive to identify causal relationships through intervention experiments, passively observed data has thus become an important source to be searched for causal relationships. The challenge of causal discovery from observational data lies in the fact that statistical associations detected from observational data are not necessarily causal.

In statistics, the causal relationship or persistent association can be determined using conditional independence tests. For a large set of variables, a pair of variables are considered to have no direct causal relationship if a subset of the remaining variables can be found such that conditioning on this subset of variables, the two variables are independent. Based on conditional independence tests, Spirtes, Glymour, and Scheines (2000) proposed the famous PC algorithm for learning the structure of causal Bayesian networks. Later, Bühlmann, Kalisch, and Maathuis (2010) extended the PC algorithm to high-dimensional variable selection. The extension is called the PC-simple algorithm which can be used to search for the causal structure around the response variable. Note that the causal structure includes all the possible direct causes and effects of the response variable, that is, all the parents and children in the terminology of directed acyclic graphs (DAGs). For certain problems, we may be able to determine in logic which are for parents and which are for children, although PC-simple cannot tell. An alternative algorithm that can be used for local causal discovery is the HITON-PC algorithm (Aliferis et al. 2010), which is also an extension of the PC algorithm. The major issue with the PC-simple and

| Method      |  $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ | $\beta_9$ | $\beta_{10}$ |
|-------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|-------------|
| Desparsified| $\hat{\beta}_{D}$<br>S.D | 1.96<br>(0.02) | 2.43<br>(0.02) | 2.98<br>(0.02) | 3.47<br>(0.02) | 3.96<br>(0.02) | 4.97<br>(0.02) | 5.95<br>(0.02) | 6.95<br>(0.02) | 7.87<br>(0.02) | 8.84<br>(0.02) |
| MNR         | $\hat{\beta}_{MNR}$<br>S.D | 2.01<br>(0.01) | 2.49<br>(0.01) | 3.00<br>(0.02) | 3.50<br>(0.01) | 3.99<br>(0.02) | 5.02<br>(0.02) | 5.98<br>(0.02) | 6.99<br>(0.02) | 8.01<br>(0.01) | 9.01<br>(0.02) |

Finally, we note that MNR can be substantially accelerated via parallel computing, for which both the Markov blanket estimation and subset regression steps can be done in an embarrassingly parallel way. As described in Section 3.3.1, the $\psi$-learning algorithm consists of two screening stages, for which both the correlation coefficients and $\psi$-partial correlation coefficients can be calculated in parallel. Refer to Liang, Song, and Qiu (2015) for more discussions on this issue. If the nodewise regression algorithm is used for Markov blanket estimation, its parallel implementation is obvious.
HITON-PC algorithms have their time complexity. For both algorithms, in the worst scenario, that is, when for each of the $p$ features all conditional independence tests of order from 1 to $p-1$ are conducted, the total number of conditional tests is $O(p^2p)$. Even under the sparsity constraint, the total number of conditional tests can still be of a high order polynomial of $p$. See Bühlmann, Kalisch, and Maathuis (2010) for more discussions on this issue.

In what follows we describe how the causal structure around the response variable can be discovered for high-dimensional regression based on the output of MNR. The proposed algorithm has a much favorable computational complexity, which is $O(n^{3/2}p)$ in all scenarios. For Gaussian, binary and proportional-hazards response data, the MNR method can be described one by one as follows.

4.1. Gaussian Response

Assume that $Z = (Y, X)$ jointly follows a multivariate Gaussian distribution $N_{p+1}(0, \Sigma)$. To distinguish the notation from that used in previous sections, we let $G_{z} = (V_{z}, E_{z})$ denote the graph underlying the joint Gaussian distribution. Let $\zeta_j = \xi_j \cup S_j$, where $\xi_j$ denotes the Markov blanket of $X_j$ in the graph $G_{z}$. It is easy to see that $\zeta_j$ forms a separator of $Y$ and $X_j$ in the graph $G_{z}$. Then, under the faithfulness condition for the joint distribution $N_{p+1}(0, \Sigma)$, we can show as in Liang, Song, and Qiu (2015) that $\rho(Y, X_j | X_{\zeta_j}) \neq 0 \Leftrightarrow \rho(Y, X_j | X_{\xi_j}) \neq 0$, where $\rho(\cdot, \cdot)$ denotes the partial correlation coefficient. The validity of the faithfulness condition is supported by the Lebesgue measure zero argument (Meek 1995); that is, the problems that violate the faithfulness condition usually correspond to some particular parameter values that form a zero measure set in the space of all possible parameterizations. Further, by the relationship between partial correlations and regression coefficients (see, e.g., Bühlmann and van de Geer 2011, p. 436), we have $\rho(Y, X_j | X_{\zeta_j}) \neq 0 \Leftrightarrow \beta_j \neq 0$, where $\beta_j$ is the coefficient of $X_j$ in the Markov neighborhood regression $Y \sim X_j + X_{\zeta_j}$. Therefore, the test for $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ can be conducted via the MNR. Given the $p$-values of individual tests, the causal structure around the response variable $Y$ can be determined via a multiple hypothesis test.

Since the problem of causal structure discovery is to identify a small set of variables that have causal or effect relations with the response variable, a simplified version of Algorithm S1 can be used, which avoids to assess the effect of all variables on the response. In the simplified algorithm, the Markov blankets only need to be found for the variables survived from the variable screening step. The simplified algorithm can be described as follows.

Algorithm 2 (Simplified MNR for causal structure discovery).

(a) (Variable screening) Apply a sure independence screening procedure with $Y$ as the response variable and $X$ as features, to obtain a reduced feature set, $\hat{S}_s = \{1, \ldots, p\}$, with the size $|\hat{S}_s| = O(\sqrt{n}/\log(n))$.

(b) (Markov blanket estimation) For each variable $X_j \in \hat{S}_s$, apply a sure independence screening procedure to obtain a reduced neighborhood $\hat{e}_j \subseteq \{1, \ldots, p\}$ with the size $|\hat{e}_j| = O(\sqrt{n}/\log(n))$.

(c) (Subset regression) For each feature $X_j \in \hat{S}_s$, run a subset regression with the features given by $X_j \cup \hat{e}_j \cup X_{\hat{S}_s}$. Conduct inference for $\beta_j$, including the estimate, confidence interval and $p$-value, based on the output of the subset regression.

(d) (Causal structure discovery) Conduct a multiple hypothesis test to identify causal features based on the $p$-values calculated in Step (c).

The consistency of the algorithm for causal structure identification can be established under slightly modified conditions of Theorem 1. To be more precise, we only need to restate the conditions (A1)–(A4) for the joint distribution of $(Y, X)$, and then the proof directly follows Theorem 2 of Liang, Song, and Qiu (2015). It is easy to see that the computational complexity of this algorithm is $O(n^{3/2}p)$, as the computational complexity of the SIS algorithm is $O(np)$ (Fan and Lv 2008) and there are a total of $O(\sqrt{n}/\log(n))$ features for which the Markov blanket needs to be estimated. Hence, Algorithm 2 can potentially be much faster than the PC-simple and HITON-PC algorithms, especially when $p$ is large. Again, this algorithm can have many implementations. For example, the SIS algorithm (Fan and Lv 2008) can be used for both variable screening and Markov blanket estimation. The HZ-SIS algorithm (Xue and Liang 2017) can also be used for both of them.

4.2. Binary Response

For binary response data, if we assume that the features $X$ follow a Gaussian distribution, then a joint distribution of $(Y, X)$ can be defined as in Lee and Hastie (2015), for which the conditional distribution of each component of $X$ is Gaussian with a linear regression model, and the conditional distribution of $Y$ is a binomial distribution as given by a logistic distribution. Further, we can assume that the joint distribution is faithful to the mixed Graphical model formed by $(Y, X)$ (Meek 1995). As pointed out in Lee and Hastie (2015), the mixed graphical model is a pairwise Markov network and the zero regression coefficients (in the nodewise regression) correspond to the conditional independence. Therefore, Algorithm 2 is also applicable to the binary response data, for which variable screening can be done using the GLM SIS algorithm (Fan and Song 2010). Extending the algorithm to multinomial response data is straightforward. The consistency of the approach directly follows from Theorem 2 of Xu, Jia, and Liang (2019), which shows the consistency of a conditional independence test based approach for learning mixed graphical models. Following Xu, Jia, and Liang (2019), the consistency of the proposed approach can be established under appropriate conditions such as the faithfulness of the joint distribution of $(Y, X)$ with respect to the underlying mixed graphical model, the sparsity of Markov blankets, the sparsity of the true model, and some conditions on GLMs.

4.3. Proportional-Hazards

For Gaussian and binary response data, we justify Algorithm 2 for causal structure discovery by presenting $(Y, X)$ as an
undirected graph for which the causal structure around \( Y \) contains both direct causes and effects of the response variable. Unfortunately, extending this justification to survival data is hard. For survival data, the response variable is proportional hazard, which is non-Gaussian and non-multinomial and thus the joint distribution of \((Y, X)\) is difficult to define with respect to an undirected graph. However, this difficulty can be resolved by modeling \((Y, X)\) as a Bayesian network with \( Y \) being a child node only. If \( Y \) is a child of \( X_{j} \), then \( \tilde{\xi}_{j} = \xi_{j} \cup \{ p + 1 \} \cup S_{j} \) forms the Markov blanket of \( X_{j} \), where \( \xi_{j} \) is the sub-Markov blanket formed with \( X \) as implied by the PC algorithm (Spirtes, Glymour, and Scheines 2000). \( p + 1 \) is the index of \( Y \) (by defining \( X_{p+1} = Y \)), and \( S_{j} \) contains all siblings of \( X_{j} \) with respect to the common child \( Y \). By the total conditioning property shown in Pellet and Elisseeff (2008) for Bayesian networks, we have

\[
X_{j} \perp Y | X_{\tilde{\xi}_{j}\backslash\{j,p+1\}} \iff X_{j} \perp Y | X_{V\backslash\{j,p+1\}},
\]

(21)

which implies that Algorithm 2 is still valid for survival data. However, construction of Bayesian networks for non-Gaussian and non-multinomial and with missing data is beyond the scope of this article. Therefore, there will be no illustrative examples for this part. In (21), if \( \tilde{\xi}_{i} \) is replaced by some super Markov blanket \( \tilde{\xi}_{i} \supseteq \tilde{\xi}_{j} \), the equivalence still holds.

This justification is very general and can be applied to the Gaussian and multinomial response data as well. The only shortcoming is that it assumes that \( Y \) can only be a child of \( X \), while this might be too restrictive for the problems considered with the Gaussian and multinomial response data.

5. Real Data Studies

This section reports two applications of Algorithm 2, one is for identification of anti-cancer drug sensitive genes, and the other is for identification of cancer driver genes.

5.1. Identification of Drug Sensitive Genes

Disease heterogeneity is often observed in complex diseases such as cancer. For example, molecularly targeted cancer drugs are only effective for patients with tumors expressing targets (Grünewald and Hidalgo 2003; Buzdar 2009). The disease heterogeneity has directly motivated the development of precision medicine, which aims to improve patient care by tailoring optimal therapies to an individual patient according to his/her molecular profile and clinical characteristics. Identifying sensitive genes to different drugs is an important step toward the goal of precision medicine.

To illustrate the MNR method, we considered the cancer cell line encyclopedia (CCLE) dataset, which is publicly available at www.broadinstitute.org/ccle. The dataset consists of 8-point dose-response curves for 24 chemical compounds across over 400 cell lines. For different chemical compounds, the numbers of cell lines are slightly different. For each cell line, it consists of the expression values of \( p = 18,988 \) genes. We used the area under the dose-response curve, which was termed as activity area in Barretina et al. (2012), to measure the sensitivity of a drug to each cell line. Compared to other measurements, such as IC\(_{50}\) and EC\(_{50}\), the activity area could capture the efficacy and potency of the drug simultaneously. An exploratory analysis indicates that treating the activity area as the response of a linear regression with respect to the gene expression values is appropriate.

Since the purpose of this study is to identify the drug sensitive genes instead of assessing the drug effect for all genes, Algorithm 2 was applied with the HZ-SIS algorithm used for variable screening and Markov blanket estimation. In both steps, we set the neighborhood size to be 40. After getting \( p \)-values from the subset regressions, the adjusted \( p \)-values (Holm 1979) were calculated, and the genes with the adjusted \( p \)-values less than 0.05 were identified as the drug sensitive genes. For some drugs, if there are no genes identified at this significance level, we just selected one gene with the smallest \( p \)-value. The results were summarized in Table 9. For Algorithm 2, different neighborhood sizes have been tried, the results are similar.

For comparison, desparsified Lasso, ridge projection, and multi-sample-splitting were also applied to this example. As in the MNR method, for each drug, we selected the genes with the adjusted \( p \)-values less than 0.05 as significant; and if there were no genes selected at this significance level, we just reported one gene with the smallest adjusted \( p \)-value. The results were also summarized in Table 9.

Compared to the existing methods, MNR performs reasonably well for this real data example. First of all, for all drugs, desparsified Lasso is simply inapplicable due to the ultra-high dimensionality of the dataset; the package hdi aborted due to the excess of memory limit. Due to the same issue, hdi also aborted for some drugs when performing ridge regression. For multi-sample-splitting and MNR, it is easy to see that if the same gene is selected by both methods, then the 95% confidence interval produced by MNR is narrower.

MNR produced promising results in selection of drug sensitive genes. For example, for both drugs Topotecan and Irinotecan, MNR selected the gene SLFN11 as the top drug sensitive gene. In the literature, Barretina et al. (2012) and Zoppoli et al. (2012) reported that SLFN11 is predictive of treatment response for Topotecan and Irinotecan. For drug 17-AAG, MNR selected NQO1 as the top gene; in the literature, Hadley and Hendricks (2014) and Barretina et al. (2012) reported NQO1 as the top predictive biomarker for 17-AAG. For drug Paclitaxel, BNN selected BCL2L1 as the top gene. In the literature, many publications, such as H. Lee et al. (2016) and Dorman et al. (2016), reported that the gene BCL2L1 is predictive of treatment response for Paclitaxel. For drug PF2341066, Lawrence and Salgia (2010) reported that HGF, which was selected by MNR as the top drug sensitive gene, is potentially responsible for the effect of PF2341066. For drug LBW242, RIPK1 is selected by MNR. Gaither et al. (2007) and Moriwaki et al. (2015) stated that RIPK1 is one of the presumed target of LBW242, which is involved in increasing death of cells. Finally, we pointed out that the genes selected by MNR have some overlaps with those selected by the multi-sample-splitting method, although for the overlapped genes the 95% confidence intervals produced by MNR tend to be narrower.

5.2. Identification of Cancer Driver Genes

We considered the Lymph dataset (Hans, Dobra, and West 2007), which consists of \( n = 148 \) samples with 100 node-negative
cases (low risk for breast cancer) and 48 node-positive cases (high risk for breast cancer) as our binary response. For each sample, there are \( p = 4512 \) genes that showed evidence of variation above the noise level for further study. This dataset has been analyzed by multiple authors, such as Hans, Dobra, and Kato, and Noguchi (2007). ATP6V1F has been reported by many authors in lymph node status studies (see, e.g., Hans, Dobra, and West 2007; Dobra 2009). For comparison, desparsified Lasso and ridge projection methods were also applied to this example. As aforementioned, the multi-sample-splitting algorithm is not yet available for logistic regression. Both desparsified Lasso and ridge projection selected only the gene RGS3 as the cancer driver gene.

A closer look at Table 10 shows that MNR outperforms desparsified Lasso and ridge projection for this example. This can be explained from two perspectives. First, MNR is the only method that identifies RGS3 as a cancer driver gene at an acceptable significance level. While the desparsified Lasso and ridge projection can only identify that RGS3 has a smaller adjusted \( p \)-value than other genes, and its adjusted \( p \)-value is greater than 0.05. Second, for the gene RGS3, the 95\% confidence interval produced by MNR is narrower than that produced by desparsified Lasso. Moreover, the 95\% confidence interval produced by ridge projection even contains 0 and is thus less significant.

### 6. Discussion

This article has proposed the MNR method for constructing confidence intervals and assessing \( p \)-values for high-dimensional regression. The MNR method has successfully broken the high-dimensional inference problem into a series of low-dimensional inference problems based on conditional independence relations among different variables. The embarrassingly parallel structure of the MNR method, where the Markov blanket, confidence interval and \( p \)-value can be calculated for each variable in parallel, enables it potentially to be run very fast on multicore computers. The MNR method has been tested on high-dimensional linear, logistic and Cox regression. The numerical results indicate that the MNR method signifi- cantly outperforms the existing ones. The MNR method has been used to learn causal structures for high-dimensional linear models with the real data examples for identification of drug sensitive genes and cancer driver genes presented.

This article has assumed that the features are Gaussian. Extension of the MNR method to non-Gaussian features is straightforward. In this case, the conditional independence

### Table 9. Comparison of drug sensitive genes selected by desparsified Lasso, ridge projection, multi-sample-splitting (multi-split), and MNR for 24 anti-cancer drugs, where * indicates that this gene was significantly selected and the number in the parentheses denotes the width of the 95\% confidence interval produced by the method.

| Drug     | Desparsified Lasso | Ridge | Multi-Split | MNR       |
|----------|--------------------|-------|-------------|-----------|
| 17-AAG   | –                  | –     | –           | NQ01*(0.138) |
| AEW541   | –                  | F3(0.076) | –           | SP10(1.176) |
| AZD0530  | –                  | PPy2(0.966) | –           | SYN3(0.705) |
| AZD6244  | –                  | OSBPL3(0.161) | –           | SPRY2*(0.084) |
| Erlotinib | –                  | LRRN1(0.102) | –           | PCDHGC3(0.684) |
| Irinotecan | –              | SLFN1(0.091) | –           | ARHGAP19*(0.134) |
| Lapatinib | –                  | WDFY4(0.509) | –           | LOC100009676(0) |
| LB2422   | –                  | RXFP3(0.86) | –           | LOB37(1.187) |
| Nutlin-3  | –                  | TCCB(0.119) | –           | ATP6V1F(1.104) |
| Paclitaxel | –             | ABCB1(0.229) | –           | ABCB1*(0.183) |
| Panobinostat | –           | C17orf105(1.104) | –          | PUM2(0.589) |
| PD-0325901 | –            | ZNF646(0.498) | –           | LYZ*(0.064) |
| PD-032991 | –                  | GRM6(0.719) | –           | LOC100506972(0.569) |
| PFZ341066 | –                  | WDFY4(0.487) | –           | SPN*(0.124) |
| PHA-665752 | –             | –               | –           | GHR(0.845) |
| PLX4720   | –                  | ADA1(0.692) | –           | INHBB(0.509) |
| RAF265    | –                  | LOC10057235(0.748) | –          | SPK(0.22) |
| Sorafenib | –                  | LTB10(0.679) | –           | SEPT11*(0.078) |
| TAE684    | –                  | –               | –           | ARID3(0.111) |
| TKI258    | –                  | –               | –           | ARID3(0.108) |
| Topotecan | –                  | –               | –           | KDRB3(0.251) |
| ZD-6474   | –                  | MID11P1(0.158) | –          | NOD1(0.363) |

**Table 10.** Comparison of the cancer driver genes selected by the MNR, desparsified Lasso, and ridge projection methods for the Lymph dataset, where * indicates that this gene was significantly selected.

| Gene     | Desparsified Lasso | Ridge | MNR       |
|----------|--------------------|-------|-----------|
| RGS3     | (1.145,5.748)     | (−0.251,2.249) | (0.859,5.178) |
| Width    | 4.603             | 2.500 | 4.319     | 5.058     |

**Algorithm 2** was applied to this dataset, where variable screening was done using the GLM SIS algorithm developed in Fan and Song (2010), and the Markov blanket estimation step was done using the HZ-SIS algorithm (Xue and Liang 2017). In both steps, we set the neighborhood size to be 5. For this dataset, MNR selected two genes, RGS3 and ATP6V1F, with adjusted \( p \)-value less than 0.05. The details were given in Table 10, which are consistent with our existing knowledge. For example, RGS3 is known to play a role in modulating the ability of motile lymphoid cells (Bowman et al. 1998), and to be upregulated in p53-mutated breast cancer tumors (Ooe, Kato, and Noguchi 2007). ATP6V1F has been reported by many authors in lymph node status studies (see, e.g., Hans, Dobra, and West 2007; Dobra 2009). For comparison, desparsified Lasso and ridge projection methods were also applied to this example. As aforementioned, the multi-sample-splitting algorithm is not yet available for logistic regression. Both desparsified Lasso and ridge projection selected only the gene RGS3 as the cancer driver gene.
relations among the features can be figured out using Bayesian networks based on the concept of Markov blanket as described in Section 4.3. The theory developed in Section 2 will still hold. The idea of using conditional independence relations for dimension reduction is general and potentially can be extended to other high-dimensional or big data problems as well.

Finally, we note that the performance of the MNR method relies on the algorithms used for variable selection and Markov blanket estimation, while each of these algorithms can rely on a nontrivial amount of tuning. A suboptimal performance of these algorithms may adversely affect the performance of the MNR method, for example, the resulting confidence intervals can be wider.

Appendix A: Conditions for Theorem 1

(A0) The dimension \( p \) is \( O(\exp(n^{0.5})) \) for some constant \( 0 \leq \delta < 1/2 \).

(A1) The distribution \( P_X \) is multivariate Gaussian, and it satisfies the Markov property and adjacency faithfulness condition with respect to the undirected underlying graph \( G \).

(A2) The correlation satisfies \( \min \{|r_{ij}|: e_{ij} = 1, i, j = 1, \ldots, p, i \neq j\} \geq c_0 n^{-\kappa} \) for some constants \( c_0 > 0 \) and \( 0 < \kappa < (1-\delta)/2 \), and \( \max \{|r_{ij}|: i, j = 1, \ldots, p, i \neq j\} \leq M_\delta < 1 \) for some constants \( 0 < M_\delta < 1 \).

(A3) There exists constant \( c_1 > 0 \) such that \( 0 < \kappa' \leq \kappa \), and \( 0 < \tau < 1-2\kappa' \) such that \( \lambda_{\max}(\Sigma) \leq c_1 n^{-\delta} \).

(A4) The \( \psi \)-partial correlation coefficients satisfy \( \inf \{|\psi_{ij}|: \psi_{ij} \neq 0, i, j = 1, \ldots, p, i \neq j, |S_{ij}| \leq q_n\} \geq c_2 n^{-d} \), where \( q_n = O(n^{\kappa+\gamma}) \), \( 0 < \gamma < \infty \), \( 0 < d < (1-\delta/2) \) are some constants, and \( S_{ij} \) denotes the conditioning set used in calculating \( \psi_{ij} \). In addition, \( \sup \{|\psi_{ij}|: i, j = 1, \ldots, p, i \neq j, |S_{ij}| \leq q_n\} \geq \kappa n^{-d} \leq M_\psi < 1 \) for some constants \( 0 < M_\psi < 1 \).

(A5) \( \max_{i=1,\ldots,p} |e_{ij}| = o(n^{1/2}) \).

(A6) There exist constants \( c_3 > 0 \) and \( c_4 \) such that \( \min \{e_{ij} \Sigma \}: |\beta_j| > c_3 n^{-\kappa} \) and \( \min \{e_{ij} \Sigma \Sigma^{1/2} \Sigma: |\Sigma^{1/2} \Sigma^{1/2} \Sigma| \} \geq c_4 \).

(A7) \( |S_i| = o(n^{1/3}) \).

(A8) Other assumptions in Theorem 2 of Fan and Peng (2004).

(A9) Other assumptions in Theorem 1 of Fan, Guo, and Hao (2012) (for the case of random design).

Supplementary Materials

Supplement description: (i) proofs of Lemma 1, Lemma 2, and Theorem 1, (ii) a variable screening-based MNR method and related theory and numerical results, and (iii) some figures illustrating the performance of MNR in variable selection.

Acknowledgments

The authors thank the editor, associate editor, and two referees for their encouraging and constructive comments which have led to significant improvement of this article.

Funding

Liang’s research was partially supported by the NIH grants R01-GM117597 and R01-GM126089.

References

Aliferis, C., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. (2010), “Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification, Part I: Algorithms and Empirical Evaluation,” Journal of Machine Learning Research, 11, 171–234. [1209]

Barrettina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F. A., Thibault, J., Bric-Furlong, E., Raman, P., Shipwayan, A., Engels, I. H., Cheng, J., Yu, G. K., Yu, J., Aspeli, P., de Silva, M., Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palensardo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefield, T., Macconnaill, L., Winckler, W., Reich, M., Li, N., Mirkin, J. P., Gabriel, S. B., Getz, G., Ardile, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R., and Garraway, L. A. (2012), “The Cancer Cell Line Encyclopedia Enables Predictive Modeling of Anticancer Drug Sensitivity,” Nature, 483, 603–607. [1211]

Belloni, A., Chernozhukov, V., and Kato, K. (2013), “Uniform Post-Selection Inference for Least Absolute Deviation Regression and Other Z- Estimation Problems,” Biometrika, 102, 77–94. [1201]

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), “Valid Post-Selection Inference,” The Annals of Statistics, 41, 802–837. [1201]

Bühlmann, P., van de Geer, S. (2011), Statistics for High-Dimensional Data: Methods, Theory and Applications, New York: Springer. [1210]

Buzdar, A. (2009), “Role of Biologic Therapy and Chemotherapy in Hormone Receptor and HER2-Positive Breast Cancer,” The Annals of Oncology, 20, 993–999. [1211]

Chatterjee, A., and Lahiri, S. (2013), “Rates of Convergence of the Adaptive LASSO Estimators to the Oracle Distribution and Higher Order Refinements by the Bootstrap,” The Annals of Statistics, 41, 1232–1259. [1201]

Cox, D. (1975), “Partial Likelihood,” Biometrika, 62, 269–276. [1203]

Dabney, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015), “High-Dimensional Inference: Confidence Intervals, p-Values and R-Software package,” Statistical Science, 30, 533–558. [1200,1202,1203,1207]

Dobra, A. (2009), “Variable Selection and Dependency Networks for Genomewide Data,” Bioinformatics, 10, 621–639. [1212]

Dormann, S., Baranova, K., Knoll, J., Urquhart, B., Mariani, G., Carcangiu, M., and Rogan, P. (2016), “Genomic Signatures for Paclitaxel and Gemcitabine Resistance in Breast Cancer Derived by Machine Learning,” Molecular Oncology, 10, 85–100. [1211]

Fan, J., Guo, S., and Hao, N. (2012), “Variable Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression,” Journal of the Royal Statistical Society: Series B, 74, 37–65. [1213]

Fan, J., and Li, R. (2001), “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties,” Journal of the American Statistical Association, 96, 1348–1360. [1200,1202]

Fan, J., and Lv, J. (2008), “Sure Independence Screening for Ultrahigh Dimensional Feature Space” (with discussion), Journal of the Royal Statistical Society, Series B, 70, 849–911. [1204,1210]

Fan, J., and Peng, H. (2004), “Nonconcave Penalized Likelihood With a Diverging Number of Parameters,” The Annals of Statistics, 32, 928–961. [1213]

Fan, J., and Song, R. (2010), “Sure Independence Screening in Generalized Linear Model With NP-Dimensionality,” The Annals of Statistics, 38, 3567–3604. [1204,1210,1212]

Fithian, W., Sun, D., and Taylor, J. (2014), “Optimal Inference After Model Selection,” arXiv no. 1410.2597. [1201]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), “Sparse Inverse Covariance Estimation With the Graphical Lasso,” Biostatistics, 9, 432–441. [1200,1202]
Gaither, A., Porter, D., Yao, Y., Borawski, J., Yang, G., Donovan, J., Sage, D., Slisz, J., Tran, M., Straub, C., Ramsey, T., Iourgenko, V., Huang, A., Chen, Y., Schlegel, R., Labow, M., Fawell, S., Sellers, W., and Zawel, L. (2007), "A Smac Mimetic Rescue Screen Reveals Roles for Inhibitor of Apoptosis Proteins in Tumor Necrosis Factor-Alpha Signaling." Cancer Research, 67, 11493–11498. [1211]

Grünwald, V., and Hidalgo, M. (2003), "Developing Inhibitors of the Epidermal Growth Factor Receptor for Cancer Treatment," Journal of the National Cancer Institute, 95, 851–867. [1211]

Hadley, K., and Hendrickx, D. (2014), "Use of NQO1 Status as a Selective Biomarker for Oesophageal Squamous Cell Carcinomas With Greater Sensitivity to 17-AAG," BMC Cancer, 14, 1–8. [1211]

Hans, C., Dobra, A., and West, M. (2007), "Shotgun Stochastic Search for Regression With Many Candidate Predictors," Journal of the American Statistical Association, 102, 507–516. [1211,1212]

Holm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," Scandinavian Journal of Statistics, 6, 65–70. [1211]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," Journal of Machine Learning Research, 15, 2869–2909. [1200]

Jia, B., Liang, F., Shi, R., and Xu, S. (2018), "equSA: Estimate Directed and Undirected Graphical Models and Construct Networks," R Package Version 1.1.5, available at https://CRAN.R-project.org/package=equSA. [1206]

Lauritzen, S. (1996), Graphical Models, Oxford: Oxford University Press. [1201]

Lawrence, R., and Salgia, R. (2010), "MET Molecular Mechanisms and Therapies in Lung Cancer," Cell Adhesion & Migration, 4, 146–152. [1211]

Lee, H., Hanibuchi, M., Lim, S.-J., Yu, H., Kim, M., He, J., Langley, R., Lehembre, F., Regnass, U., and Fidler, I. (2016), "Treatment of Experimental Human Breast Cancer and Lung Cancer Brain Metastases in Mice by Macitentan, a Dual Antagonist of Endothelin Receptors, Combined With Paclitaxel," NeuroOncology, 18, 486–496. [1211]

Lee, J. D., and Hastie, T. J. (2015), "Learning the Structure of Mixed Graphical Models," Journal of Computational and Graphical Statistics, 24, 230–253. [1210]

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016), "Exact Post-Selection Inference, With Application to the Lasso," The Annals of Statistics, 44, 907–927. [1201]

Liang, F., Song, Q., and Qiu, P. (2015), "An Equivalent Measure of Partial Correlation Coefficients for High Dimensional Gaussian Models," Journal of the American Statistical Association, 110, 1248–1265. [1200,1202,1203,1206,1210]

Liang, F., Song, Q., and Yu, K. (2013), "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, 77, 923–945. [1201]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," The Annals of Statistics, 34, 1436–1462. [1200,1202]

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), "p-Values for High-Dimensional Regression," Journal of the American Statistical Association, 104, 1671–1681. [1200]

Moriwaki, K., Bertin, J., Gough, P., Orłowski, G., and Chan, F. (2015), "Differential Roles of RIPK1 and RIPK3 in TNF-Induced Necroptosis and Chemotherapeutic Agent-Induced Cell Death," Cell Death & Disease, 6, e1636. [1211]

Ooe, A., Kato, K., and Noguchi, S. (2007), "Possible Involvement of CCT5, RG3, and YKT6 Genes Up-Regulated in p53-Mutated Tumors in Resistance to Docetaxel in Human Breast Cancers," Breast Cancer Research and Treatment, 101, 305–315. [1212]

Pellet, J.-P., and Elisseef, A. (2008), "Using Markov Blankets for Causal Structure Learning," Journal of Machine Learning Research, 9, 1295–1342. [1211]

Saldana, D., and Feng, Y. (2018), "SIS: An R Package for Sure Independence Screening in Ultrahigh-dimensional Statistical Models," Journal of Statistical Software, 83, 1–25. [1204]

Song, Q., and Liang, F. (2015a), "High Dimensional Variable Selection With Reciprocal L1-Regularization," Journal of the American Statistical Association, 110, 1607–1620. [1200,1202]

Saldana, D., and Feng, Y. (2018), "SIS: An R Package for Sure Independence Screening in Ultrahigh-dimensional Statistical Models," Journal of Statistical Software, 83, 1–25. [1204]

Song, Q., and Liang, F. (2015b), "A Split-and-Merge Bayesian Variable Selection Approach for Ultra-high Dimensional Regression," Journal of the Royal Statistical Society, Series B, 77, 947–972. [1200]

Spiroff, P., Glymour, C., and Scheines, R. (2000), Causation, Prediction, and Search, Boston: MIT Press. [1209,1211]

Storey, J. (2002), "A Direct Approach to False Discovery Rates," Journal of the Royal Statistical Society, Series B, 64, 479–498. [1205]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the LASSO," Journal of the Royal Statistical Society, Series B, 58, 267–288. [1200,1202]

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016), "Exact Post-Selection Inference for Sequential Regression Procedures," Journal of the American Statistical Association, 111, 600–620. [1201]

van de Geer, S., and Bühlmann, P. (2009), "On the Conditions Used to Prove Oracle Results for the Lasso," Electronic Journal of Statistics, 3, 1360–1392. [1202]

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," The Annals of Statistics, 42, 1166–1202. [1200,1204,1208]

Wasserman, L., and Roeder, K. (2009), "High-Dimensional Variable Selection," The Annals of Statistics, 37, 2178–2201. [1201]

Witten, D., Friedman, J., and Simon, N. (2011), "New Insights and Faster Computations for the Graphical Lasso," Journal of Computational and Graphical Statistics, 20, 892–900. [1209]

Xu, S., Jia, B., and Liang, F. (2019), "Learning Moral Graphs in Construction of High-Dimensional Bayesian Networks for Mixed Data," Neural Computation, 31, 1183–1214. [1210]

Xue, J., and Liang, F. (2017), "A Robust Model-Free Feature Screening Method for Ultrahigh Dimensional Data," Journal of Computational and Graphical Statistics, 26, 803–818. [1210,1212]

Yang, Y. (2017), "Statistical Inference for High Dimensional Regression via Constrained Lasso," arXiv no. 1704.05098v1. [1201]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," Biometrika, 95, 19–35. [1200,1202]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," The Annals of Statistics, 38, 894–942. [1200,1202,1209]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," Journal of the Royal Statistical Society, Series B, 76, 217–242. [1200,1206]

Zhang, X., and Cheng, G. (2017), "Simultaneous Inference for High-Dimensional Linear Models," Journal of the American Statistical Association, 112, 757–768. [1200]

Zoppoli, G., Regairaz, M., Leo, E., Reinhold, W., Varma, S., Ballestrero, A., Doroshow, J., and Pommier, Y. (2012), "Putative DNA/RNA Helicase Schlaflen (SLFN11) Sensitizes Cancer Cells to DNA-Damaging Agents," Proceedings of the National Academy of Sciences of the United States of America, 109, 15030–15035. [1211]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," The Annals of Statistics, 38, 894–942. [1200,1203]