Published in final edited form as:

Stat Sin. 2017 July ; 27(3): 983–996. doi:10.5705/ss.202015.0473.

Variable Selection via Partial Correlation

Runze Li, Jingyuan Liu, and Lejia Lou

Pennsylvania State University, Xiamen University and Ernst & Young

Abstract

Partial correlation based variable selection method was proposed for normal linear regression models by Bühlmann, Kalisch and Maathuis (2010) as a comparable alternative method to regularization methods for variable selection. This paper addresses two important issues related to partial correlation based variable selection method: (a) whether this method is sensitive to normality assumption, and (b) whether this method is valid when the dimension of predictor increases in an exponential rate of the sample size. To address issue (a), we systematically study this method for elliptical linear regression models. Our finding indicates that the original proposal may lead to inferior performance when the marginal kurtosis of predictor is not close to that of normal distribution. Our simulation results further confirm this finding. To ensure the superior performance of partial correlation based variable selection procedure, we propose a thresholded partial correlation (TPC) approach to select significant variables in linear regression models. We establish the selection consistency of the TPC in the presence of ultrahigh dimensional predictors. Since the TPC procedure includes the original proposal as a special case, our theoretical results address the issue (b) directly. As a by-product, the sure screening property of the first step of TPC was obtained. The numerical examples also illustrate that the TPC is competitively comparable to the commonly-used regularization methods for variable selection.

keywords and phrases

Elliptical distribution; model selection consistency; partial correlation; partial faithfulness; sure screening property; ultrahigh dimensional linear model; variable selection

1. Introduction

Variable selection via penalized least squares has been extensively studied during the last two decades. Popular penalized least squares variable selection procedures include LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), adaptive LASSO (Zou, 2006) and among others. See Fan and Lv (2010) for a selective overview on this topic and references therein for more works on variable selection via penalized least squares.

Runze Li, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802. rzli@psu.edu

Jingyuan Liu, Department of Statistics in School of Economics, Wang Yanan Institute for Studies in Economics and Fujian Key Laboratory of Statistical Science, Xiamen University, Xiamen 361005, China. jingyuan@xmu.edu.cn

Lejia Lou, Ernst & Young, 5 Times Square, New York, NY 10036, lejia.lou@ey.com

As a powerful alternative method to penalized least squares for variable selection, Bühlmann, Kalisch and Maathuis (2010) proposed a variable selection procedure by ranking the partial correlations (PC) between the predictors and the response, named PC-simple algorithm. The definition of partial correlations is given in section 2. The authors provided a stepwise algorithm for the linear regression models with partial faithfulness - where for each predictor, if its partial correlation with the response given a certain subset of other predictors is 0, then the partial correlation given all the other predictors is also 0. The PC-simple algorithm possesses model selection consistency for such linear models, thus is competitively comparable to the penalized least squares variable selection approaches. Therefore, scientists may have two distinct schemes for variable selection in high-dimensional linear models, which raise their confidence in the selected predictors if they are chosen by both techniques.

This work aims to study two important issues related to the PC-simple algorithm. The first issue is that the procedure proposed in Bühlmann, Kalisch and Maathuis (2010) relies on normality assumption on the joint distribution of the response and the predictors, although partial faithfulness does not require normality assumption. Thus, it is of great interest to study the impact of normality assumption on the variable selection procedure developed in Bühlmann, Kalisch and Maathuis (2010). The second issue is that the theoretical results established in Bühlmann, Kalisch and Maathuis (2010) requires that the dimension of the predictor vector increases in a polynomial rate of the sample size. It is also of great interest to study whether the theoretical results are valid with dimensionality increasing at exponential rate of the sample size.

To study the issue related to normality assumption, we consider the elliptical linear model (i.e, the response and the predictors in a linear regression model jointly follow an elliptical distribution), which has been systematically accounted in Fang, Kotz and Ng (1990). The elliptical distribution family contains a much broader class of distributions than the normal distribution family, such as mixture of normal distributions, multivariate t-distribution, multi-uniform distribution on unit sphere, Pearson Type II distribution, among others. It has been used as a tool to study the robustness of normality in the literature of multivariate nonparametric tests (Mottonen, Oja and Tienari, 1997; Oja and Randles, 2004; Chen, Wiesel and Hero, 2011; Soloveychik and Wiesel, 2015; Wang, Peng and Li, 2015). The elliptical linear regressions have been proposed in Osiewalski (1991); Osiewalski and Steel (1993); Arellano-Valle, del Pino and Iglesias (2006); Fan and Lv (2008); Liang and Li (2009); Vidal and Arellano-Valle (2010), and have received more and more attentions in the recent literature (Arellano-Valle, del Pino and Iglesias, 2006; Fan and Lv, 2008; Liang and Li, 2009; Vidal and Arellano-Valle, 2010). Furthermore, the elliptical distribution family has a variety of applications. For instance, it becomes crucial for modeling finance data (Mcneil Frey and Embrechts, 2005) due to its potential to accommodate tail dependence (the phenomenon of simultaneous extremes), which is highly useful in quantitative finance but is not allowed by the multivariate normal distribution (Schmidt, 2002).

By exploring the limiting distribution of the sample partial correlation of elliptical distribution, which is of its own significance, we find that the PC-simple algorithm tends to over-fit(under-fit) the models under those elliptical distributions whose marginal kurtosis is

larger (smaller) than the marginal kurtosis of normal distribution. To ensure the superior performance of partial correlation based variable selection procedure for the elliptical distribution family, we propose a thresholded partial correlation (TPC) approach to select significant variables in linear regression models. In the same spirit of the PC-simple algorithm, the TPC is a stepwise method for variable selection and is constructed by comparing each sample correlation and sample partial correlation with a given threshold corresponding to a given significant level. The TPC approach relies on the limiting distribution of the sample partial correlation, and coincides to the PC-simple algorithm for the normal linear models. This enables us to study the asymptotic property of the PC-sample algorithm under a broader framework in order to address the issue related to dimensionality increasing at exponential rate.

We systematically study the sampling properties of the TPC. We first derive the concentration inequality of the partial correlations without model assumption when the dimensionality of the covariates increases with the same size at an exponential rate. This enables us to conduct the TPC for the ultrahigh dimensional linear models. We further establish the theoretical properties of the TPC. This allows us to broaden the usage of this variable selection scheme. We also develop the sure screening property of the first-step TPC in the terminology of Fan and Lv (2008). Note that the first step of the TPC has the same spirit as the marginal screening based on the Pearson correlation (Fan and Lv, 2008). Thus, as a by-product, we obtain the sure screening property of the marginal screening procedure based on the Pearson correlation under different assumptions from Fan and Lv (2008).

This paper is organized as follows. In section 2, we propose the TPC for the elliptical linear models, and further establish its asymptotic properties. Numerical studies are conducted in section 3. A brief conclusion is given in section 4, and all the technical proofs are allocated in the Appendix.

2. Thresholded partial correlation (TPC) approach

In this section, we introduce the linear model with elliptical response and predictors, and bring in the motivation of the TPC approach by studying the limiting distribution of the partial correlation estimation for elliptical distributions. Then the TPC approach is proposed based on the limiting distribution, and the corresponding theoretical properties are discussed.

2.1 Elliptical linear model and its partial correlation estimation

Consider a linear model

$$y = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon, \quad (2.1)$$

where y is the response variable, $\mathbf{x} = (x_1, \dots, x_p)^T$ is the covariate vector, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is the coefficient vector, and ε is the random error with $E(\varepsilon) = 0$, $\text{var}(\varepsilon) = \sigma^2$. Throughout this paper, it is assumed without loss of generality that $E(\mathbf{x}) = 0$ and $E(y) = 0$ so that there is no intercept in model (2.1). In practice, it is common that $\mathbf{x}$ and y are marginally standardized

before performing variable selection. Furthermore, $(\mathbf{x}^T, y)$ is assumed to follow an elliptical distribution, and suppose that $(\mathbf{x}_1^T, y_1), \dots, (\mathbf{x}_n^T, y_n)$ are independent and identically distributed (iid) random samples from an elliptical distribution $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$, which has the characteristic function $\exp(i\mathbf{t}^T \boldsymbol{\mu}) \phi(\mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t})$ for some characteristic generator $\phi(\cdot)$ (Fang, Kotz and Ng, 1990).

Bühlmann, Kalisch and Maathuis (2010) proposed a variable selection method, PC-simple algorithm, based on the partial correlation learning for the linear model (2.1) with normal response and predictors. To extend this method to the elliptical distribution, we first study the limiting distributions for the correlation and partial correlation under the elliptical assumption. Denote by $\rho(y, x_j)$ and $\hat{\rho}(y, x_j)$ the population and the sample correlation between y and x_j , respectively. Then as shown in Theorem 5.1.6 of Muirhead (1982), the asymptotic distribution of $\hat{\rho}(y, x_j)$ is

$$\sqrt{n} \{ \hat{\rho}(y, x_j) - \rho(y, x_j) \} \rightarrow N \left(0, (1+\kappa) \{ 1 - \rho^2(y, x_j) \}^2 \right), \quad (2.2)$$

where $\kappa = \frac{\phi''(0)}{(\phi'(0))^2} - 1$ with $\phi'(0)$ and $\phi''(0)$ being the first and second derivatives of ϕ at 0. κ is the marginal kurtosis of the elliptical distribution of $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ and equals 0 for a normal distribution $N_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

For an index set $\mathcal{S} \subseteq \{1, 2, \dots, p\}$, we define $\mathcal{S}^c$ to be $\mathcal{S}^c = \{1 \leq j \leq p : j \notin \mathcal{S}\}$, $|\mathcal{S}|$ to be its cardinality, and $x_{\mathcal{S}} = \{x_j : j \in \mathcal{S}\}$ to be a subset of covariates with index set $\mathcal{S}$. Denote the truly active index set $\mathcal{A} = \{1 \leq j \leq p : \beta_j \neq 0\}$ and the corresponding cardinality $d_0 = |\mathcal{A}|$. Based on $x_{\mathcal{S}}$, the definition of the partial correlation is given below.

Definition 1—(Partial Correlation) The partial correlation between x_j and y given a set of controlling variables $x_{\mathcal{S}}$, denoted by $\rho(y, x_j | x_{\mathcal{S}})$, is defined as the correlation between the residuals $r_{x_j, x_{\mathcal{S}}}$ and $r_{y, x_{\mathcal{S}}}$ from the linear regression of x_j on $x_{\mathcal{S}}$ and that of y on $x_{\mathcal{S}}$, respectively. And the corresponding sample partial correlation between y and x_j given $x_{\mathcal{S}}$ is denoted by $\hat{\rho}(y, x_j | x_{\mathcal{S}})$.

In the next theorem, we study the asymptotic distribution of the sample partial correlation when the sample was drawn from an elliptical distribution, which provides the foundation of TPC variable selection procedure.

Theorem 1—Suppose that $(\mathbf{x}_1^T, y_1), \dots, (\mathbf{x}_n^T, y_n)$ are iid random samples from an elliptical distribution $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ with all finite fourth moments. For any $j = 1, \dots, p$, and $\mathcal{S} \subseteq \{j\}^c$ with cardinality $|\mathcal{S}| = o(\sqrt{n})$, if there exists a positive constant δ_0 , the smallest eigenvalue of the covariance matrix of $x_{\mathcal{S}}$ is greater than δ_0 , then

$$\sqrt{n} \{ \hat{\rho}(y, x_j | x_{\mathcal{S}}) - \rho(y, x_j | x_{\mathcal{S}}) \} \rightarrow N \left(0, (1+\kappa) \{ 1 - \rho^2(y, x_j | x_{\mathcal{S}}) \}^2 \right). \quad (2.3)$$

Theorem 1 seems to be a natural extension on partial correlation from normal distribution to elliptical distribution, but to our best knowledge, this results is new, and its proof is given in Appendix A. Let $\emptyset$ be the empty set, and $\hat{\rho}(y, x_j/x_{\emptyset})$ and $\rho(y, x_j/x_{\emptyset})$ stand for $\hat{\rho}(y, x_j)$ and $\rho(y, x_j)$, respectively. Then (2.3) is also valid for $\mathcal{S} = \emptyset$ by (2.2). The limiting distributions of sample correlation and partial correlation given in (2.2) and (2.3) provides insights into the impact of normality assumption on the PC-simple algorithm through the marginal kurtosis under ellipticity assumption. This enables us to modify the PC-simple algorithm by taking into account the marginal kurtosis to ensure its superior performance.

In addition, since the limiting distribution of $\hat{\rho}(y, x_j/x_{\mathcal{S}})$ in (2.3) involves $\rho(y, x_j/x_{\mathcal{S}})$ in the asymptotic variance, we consider the Fisher Z-transformation of $\hat{\rho}(y, x_j/x_{\mathcal{S}})$, whose limiting distribution no longer depends on $\rho(y, x_j/x_{\mathcal{S}})$. Specifically, let $\hat{Z}(y, x_j/x_{\mathcal{S}})$ and $Z(y, x_j/x_{\mathcal{S}})$ be the Fisher Z-transformation of $\hat{\rho}(y, x_j/x_{\mathcal{S}})$ and $\rho(y, x_j/x_{\mathcal{S}})$, respectively. That is,

$$\hat{Z}(y, x_j/x_{\mathcal{S}}) = \frac{1}{2} \log \left\{ \frac{1 + \hat{\rho}(y, x_j/x_{\mathcal{S}})}{1 - \hat{\rho}(y, x_j/x_{\mathcal{S}})} \right\}, \quad Z(y, x_j/x_{\mathcal{S}}) = \frac{1}{2} \log \left\{ \frac{1 + \rho(y, x_j/x_{\mathcal{S}})}{1 - \rho(y, x_j/x_{\mathcal{S}})} \right\}. \quad (2.4)$$

Then, it follows by the delta method and Theorem 1 that

$$\sqrt{n} \left\{ \hat{Z}(y, x_j/x_{\mathcal{S}}) - Z(y, x_j/x_{\mathcal{S}}) \right\} \rightarrow N(0, 1 + \kappa). \quad (2.5)$$

The asymptotic distribution of $\hat{Z}(y, x_j/x_{\mathcal{S}})$ no longer depends on $\rho(y, x_j/x_{\mathcal{S}})$, thus it is easier to derive the selection threshold for $\hat{Z}(y, x_j/x_{\mathcal{S}})$ rather than for $\rho(y, x_j/x_{\mathcal{S}})$ directly.

2.2 A variable selection algorithm

Based on the partial faithfulness condition, the following holds for all $j \in \{1, \dots, p\}$ (Bühlmann, Kalisch and Maathuis, 2010):

$$\rho(y, x_j/x_{\mathcal{S}}) \neq 0 \text{ for all } \mathcal{S} \subseteq \{j\}^c \text{ if and only if } \beta_j \neq 0.$$

That is, x_j is important (or $\beta_j \neq 0$) if and only if the partial correlations between y and x_j given all subsets $\mathcal{S}$ contained in $\{j\}^c$ are not zero. Extending the PC-simple algorithm proposed by Bühlmann, Kalisch and Maathuis (2010), we propose to identify active predictors by iteratively testing the series of hypotheses

$$H_0: \rho(y, x_j/x_{\mathcal{S}}) = 0 \text{ for } |\mathcal{S}| = 0, 1, \dots, \hat{m}_{reach},$$

where $\hat{m}_{reach} = \min\{m: |\hat{\mathcal{A}}^{[m]}| = m\}$, and $\hat{\mathcal{A}}^{[m]}$ is the chosen model index set in the m th step with cardinality $|\hat{\mathcal{A}}^{[m]}|$. Based on the limiting distribution (2.5), the rejection region at level

α is $|\hat{Z}(y, x_j | x_{\mathcal{S}})| > \sqrt{1+\hat{\kappa}} \Phi^{-1}(1-\alpha/2) / \sqrt{n}$ with $\hat{\kappa}$ being a consistent estimate of κ , where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution of the standard normal distribution.

In practice, the factor $\sqrt{n}$ in the rejection region is replaced by $\sqrt{n-1-|\mathcal{S}|}$ due to the loss of degrees of freedom used in calculation of residuals. Therefore, an equivalent form of the rejection region with small sample correction is

$$|\hat{\rho}(y, x_j | x_{\mathcal{S}})| > T(\alpha, n, \hat{\kappa}, |\mathcal{S}|), \quad (2.6)$$

where

$$T(\alpha, n, \hat{\kappa}, |\mathcal{S}|) = \frac{\exp \left\{ 2 \sqrt{1+\hat{\kappa}} \Phi^{-1}(1-\alpha/2) / \sqrt{n-1-|\mathcal{S}|} \right\} - 1}{\exp \left\{ 2 \sqrt{1+\hat{\kappa}} \Phi^{-1}(1-\alpha/2) / \sqrt{n-1-|\mathcal{S}|} \right\} + 1} \quad (2.7)$$

with κ being estimated by its sample counterpart:

$$\hat{\kappa} = \frac{1}{p} \sum_{j=1}^p \left\{ \frac{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^4}{3 \left\{ \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \right\}^2} - 1 \right\}, \quad (2.8)$$

where $\bar{x}_j$ is the sample mean of the j -th element of $\mathbf{x}$ and x_{ij} is the j -th element of $\mathbf{x}_i$. In practice, the sample partial correlations can be computed recursively:

For any $k \in \mathcal{S}$,

$$\hat{\rho}(y, x_j | x_{\mathcal{S}}) = \frac{\hat{\rho}(y, x_j | x_{\mathcal{S} \setminus \{k\}}) - \hat{\rho}(y, x_k | x_{\mathcal{S} \setminus \{k\}}) \hat{\rho}(x_j, x_k | x_{\mathcal{S} \setminus \{k\}})}{[\{1 - \hat{\rho}(y, x_k | x_{\mathcal{S} \setminus \{k\}})^2\} \{1 - \hat{\rho}(x_j, x_k | x_{\mathcal{S} \setminus \{k\}})^2\}]^{1/2}}. \quad (2.9)$$

We summarize the TPC variable selection by the following algorithm.

Algorithm 1

Algorithm for TPC variable selection.

Step 1: Set $m = 1$ and $\mathcal{S} = \emptyset$, obtain the marginally estimated active set by

$$\hat{\mathcal{A}}^{[1]} = \{j = 1, \dots, p: |\hat{\rho}(y, x_j)| > T(\alpha, n, \hat{\kappa}, 0)\}.$$

Step 2: Based on $\hat{\mathcal{A}}^{[m-1]}$, construct the m th step estimated active set by

$$\hat{\mathcal{A}}^{[m]} = \{j \in \hat{\mathcal{A}}^{[m-1]} : |\hat{\rho}(y, x_j | x_{\mathcal{S}})| > T(\alpha, n, \hat{\kappa}, |\mathcal{S}|), \forall \mathcal{S} \subseteq \hat{\mathcal{A}}^{[m-1]} \setminus \{j\}, |\mathcal{S}| = m-1\}.$$

Step 3: Repeat Step 2 until $m = \hat{m}_{reach}$

Algorithm 1 results in a sequence of estimated active sets

$$\hat{\mathcal{A}}^{[1]} \supseteq \hat{\mathcal{A}}^{[2]} \supseteq \dots \hat{\mathcal{A}}^{[m]} \supseteq \dots \supseteq \hat{\mathcal{A}}^{[\hat{m}_{reach}]}$$

Since $\kappa = 0$ for normal distributions, the TPC is indeed the PC-simple algorithm under normality assumption. Thus, Theorem 1 clearly shows that the PC-simple algorithm tends to over-fit (under-fit) the models under those distributions where the kurtosis is larger (smaller) than the normal kurtosis 0. Following Bühlmann, Kalisch and Maathuis (2010), we further apply the ordinary least squares approach to estimate the coefficients of predictors in $\hat{\mathcal{A}}^{[\hat{m}_{reach}]}$ after running Algorithm 1.

2.3 Theoretical properties

We impose the following regularity conditions to establish the asymptotic theory of the TPC. These regularity conditions may not be the weakest ones.

- (D1) The joint distribution of $(\mathbf{x}^T, y)$ satisfies partial faithfulness (Bühlmann, Kalisch and Maathuis, 2010).
- (D2) $(\mathbf{x}^T, y)$ follows $EC_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ with $\boldsymbol{\Sigma} > 0$. Furthermore, there exists $s_0 > 0$, such that for all $0 < s < s_0$,

$$E\{\exp(sy^2)\} < \infty, \max_{1 \leq j \leq p} E\{\exp(sx_j y)\} < \infty, \\ \text{and } \max_{1 \leq j, k \leq p} E\{\exp(sx_j x_k)\} < \infty.$$

- (D3) There exists $\delta > -1$, such that the kurtosis satisfies $\kappa > \delta > -1$.
- (D4) For some $c_n = O(n^{-d})$, $0 < d < 1/2$, the partial correlations $\rho(y, x_j / x_{\mathcal{S}})$ satisfy

$$\inf \{|\rho(y, x_j | x_{\mathcal{S}})| : j=1, \dots, p, \mathcal{S} \subseteq \{j\}^c, |\mathcal{S}| \leq d_0, \rho(y, x_j | x_{\mathcal{S}}) \neq 0\} \geq c_n.$$

- (D5) The partial correlations $\rho(y, x_j / x_{\mathcal{S}})$ and $\rho(x_j, x_k / x_{\mathcal{S}})$ satisfy:

- i. $\sup \{|\rho(y, x_j / x_{\mathcal{S}})| : 1 \leq j \leq p, \mathcal{S} \subseteq \{j\}^c, |\mathcal{S}| \leq d_0\} = \tau < 1,$
- ii. $\sup \{|\rho(x_j, x_k / x_{\mathcal{S}})| : 1 \leq j < k \leq p, \mathcal{S} \subseteq \{j, k\}^c, |\mathcal{S}| \leq d_0\} = \tau < 1.$

Condition (D1) guarantees the validity of the TPC method as a variable selection criterion. The assumption on elliptical distribution in (D2) is crucial when deriving the asymptotic distribution of the sample partial correlation, and the sub-exponential tail probability ensures the difference between the population and sample partial correlations to degenerate with an

exponential rate. Many elliptical distributions satisfy the sub-exponential tail probability, such as multivariate normal distribution and Pearson Type II distribution (Fang, Kotz and Ng, 1990). In addition, although (D2) is widely used in literature as a sufficient condition to facilitate the theoretical proof, it may not be the weakest condition to guarantee the validity of the TPC. (D3) puts a mild condition on the kurtosis, and is used to control Type I and II errors. The lower bound of partial correlations in (D4) is used to control Type II errors for the tests. This condition has the same spirit as that of the penalty-based methods which requires the non-zero coefficients to be bounded away from 0. The upper bound of partial correlations in the condition i) of (D5) is used to control Type I error, and the condition ii) of (D5) imposes a fixed upper bound on the population partial correlations between the covariates, which excludes the perfect collinearity between the covariates.

Based on the above regularity conditions, we obtain the following consistency property. First we consider the model selection consistency of the final estimated active set by TPC. Since the TPC depends on the significance level $\alpha = \alpha_n$, we rewrite the final chosen model to be $\hat{\mathcal{A}}_n(\alpha_n)$.

Theorem 2—Consider linear model (2.1). Under Conditions (D1)–(D5), there exists a sequence $\alpha_n \rightarrow 0$ and a positive constant C , such that if d_0 is fixed, then for $p = \alpha(\exp(n^\xi))$, $0 < \xi < 1/5$, the estimated active set can be identified with the following rate

$$P\{\hat{\mathcal{A}}_n(\alpha_n) = \mathcal{A}\} \geq 1 - O\{\exp(-n^\nu/C)\}, \quad (2.10)$$

where $\xi < \nu < 1/5$; and if $d_0 = O(n^b)$, $0 < b < 1/5$, then for $p = \alpha(\exp(n^\xi))$, $0 < \xi < 1/5 - b$, (2.10) still holds, with $\xi + b < \nu < 1/5$.

The proof of this theorem is given in Appendix B. Theorem 2 implies that the TPC method including the original PC-simple algorithm enjoys the model selection consistency property when dimensionality increases at exponential rate of the sample size. Following Bühlmann, Kalisch and Maathuis (2010), one possible choice of the theoretical significance level α_n is $\alpha_n = 2\{1 - \Phi(c_n \sqrt{n/(1+\kappa)/2})\}$.

Note that Bühlmann, Kalisch and Maathuis (2010) utilized the tail probability of normal distribution to control the upper bound of probabilities of Types I and II errors. Thus, they have to assume that the model dimension grows at the polynomial rate of the sample size. We take a different approach from Bühlmann, Kalisch and Maathuis (2010) to establishing the model selection consistency in Theorem 2. We first derive the concentration inequality of the partial correlations as in Step 1 of the proof of Theorem 2. In this step, we do not require assumption of ellipticity. With the concentration inequality, we allow the dimensionality of the covariates increases with the same size at an exponential rate. This enables us to conduct the TPC for ultrahigh dimensional linear models.

In addition, notice that the estimated active set from the first step of the TPC, denoted by $\hat{\mathcal{A}}_n^{[1]}(\alpha_n)$, can be viewed as a feature screening procedure, and is essentially equivalent to

the sure independence screening procedure proposed by Fan and Lv (2008). Thus, we next establish the sure screening property (Fan and Lv, 2008) of this first step of TPC under a different set of assumptions. We impose the following conditions on the population marginal correlations:

$$(E4) \quad \inf \{ |\rho_n(y, x_j)| : j = 1, \dots, p, \rho_n(y, x_j) \neq 0 \} = c_n, \text{ where } c_n = O(n^{-d}), \text{ and } 0 < d < 1/2.$$

$$(E5) \quad \sup \{ |\rho_n(y, x_j)| : j = 1, \dots, p_n \} = \tau < 1.$$

Theorem 3—Consider linear model (2.1) and assume (D1)–(D3), (E4) and (E5). For $p = O(\exp(n^\xi))$, where $0 < \xi < 1$, there exists a sequence $\alpha_n \rightarrow 0$ such that

$$P\{\hat{\mathcal{A}}_n^{[1]} \supseteq \mathcal{A}\} \geq 1 - O\{\exp(-n^\nu / C^*)\}, \text{ where } C^* \text{ is a positive constant and } \xi < \nu < 1/5.$$

The proof of this theorem is given in Appendix B. This theorem confirms the sure screening property of the marginal screening procedure based on the Pearson correlation under a different set of regularity conditions from Fan and Lv (2008).

3. Numerical Studies

In this section, we assess the finite sample performance of the TPC methods and compare it with LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001) and PC-simple algorithm through Monte Carlo simulation studies. We also illustrate the application of the TPC by a rat eye expression dataset example.

3.1 Simulation studies

In our simulation study, data were generated from linear model (2.1) with $\beta_1 = 3$, $\beta_2 = 1.5$, $\beta_5 = 2$ and $\beta_j = 0$ if $j = 1, 2, 5$. In our simulation, we consider $p = 200, 500$ and 2000 . The sample size is taken to be $n = 200$. Moreover, the joint distribution of $\mathbf{x}$ and ε are taken to be $0.9N(0, \Sigma) + 0.1N(0, 9\Sigma)$, which is an elliptical distribution, and the normal distribution $N(0, \Sigma)$, where Σ is the $(p+1) \times (p+1)$ matrix with (i, j) th entry $\rho^{|i-j|}$. We consider $\rho = 0, 0.3$, and 0.8 which corresponds to “uncorrelated”, “moderately correlated”, and “strongly correlated” among $\mathbf{x}$, respectively. Note that the estimated kurtosis of the mixture normal is around 1.5, deviating from 0 for normal distribution to a large extent. Thus, the mixture of normal distributions in this study is used to illustrate a heavy-tailed situation. For each case, we conduct 1000 simulations.

In our simulation, we compare the finite sample performance of LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001), the PC-simple algorithm (Bühlmann, Kalisch and Maathuis, 2010) and the TPC. Furthermore, the following criteria are used for evaluating the performance of variable selection procedures.

1. Model error: $E_{\mathbf{x}}[\{\mathbf{x}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^2] = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \text{cov}(\mathbf{x})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.
2. True positive number (TPN) which is defined to be the average of the number of the predictors with nonzero coefficients being successfully detected in 1000 simulation.

3. False positive number (FPN) which is defined to be the average of the number of predictors with zero coefficients being erroneously selected into the model.
4. Underfit percentage (UF), which is defined to be the percentage of underfit models which fails to identify at least one important predictor in the 1000 simulations.
5. Correct-fit percentage (CF), which is defined to be the percentage of correctly-fitted models that exactly select the truly important predictors in the 1000 simulations.
6. Overfit percentage (OF), which is defined to be the percentage of overfit models that identify all the important predictors, but include at least one unimportant predictor in the 1000 simulations.

Table 3.1 depicts the simulation results for the elliptical distribution and clearly shows that the TPC performs significantly better than LASSO, SCAD and the PC-simple algorithm in most situations, regardless of the low or high model dimensionality. Specifically, LASSO constantly over-fits the model under every scenario, as indicated by literature. The models selected by SCAD are also more conservative in this case compared with those by the PC-based methods, thus the correct-fit rate is much lower while the over-fit rate is high. Furthermore, since the PC-simple relies on normality, it fails to capture the correct model with a high percentage when this elliptical distribution is considered, especially when x -variables are independent, the correct-fit rate of the PC-simple are only 25% and 7% for $p = 500$ and 2000, respectively. The PC-simple tends to overfit this model because the marginal kurtosis is around 1.5, larger than that of normal distribution, 0. For instance, when $p = 500$, the over-fit rate (OF) of PC-simple are 0.58, 0.65 and 0.22 for $\rho = 0, 0.3, 0.8$, respectively. While the OF of TPC are 0.04, 0.07 and 0.00. The TPC increases the probability of recovering the true model to a large degree, by correcting the threshold under the ellipticity assumption. Furthermore, the time for 1000 simulations with $p = 2000$ is reported in Table S1 in the supplemental material to save space. In terms of the computational cost, TPC converges much faster than the PC-simple algorithm and the SCAD, and is comparable to the LARS algorithm for LASSO.

The results for normal distribution are presented in Table S2 in the supplement material. Recall that in theory, the TPC should be equivalent to the PC-simple algorithm in this case, thus their performances are quite similar. The median model errors are comparable for all the methods except for LASSO, which yields much larger models than necessary. Overall, both LASSO and SCAD tend to provide more conservative models, and over-fit the model, compared with the partial-correlation-based methods for variable selection.

The elliptical distribution assumption is used only for deriving the asymptotic distribution of the partial correlations in Theorem 1, thus it inspired us to propose the test statistic for the series of hypotheses $H_0 : \rho(y, x_j | x_{\mathcal{S}}) = 0$. However, the model selection consistency of the TPC does not require the elliptical distribution for the response and the predictors. Therefore, the performance of TPC is expected to be relatively robust against the elliptical assumption. To illustrate this, we may consider a simulation example in which the discrete predictors are involved. Specifically, the x 's with even-subscript are generated in the same

fashion as before, while the x 's with odd-subscript take discrete values 0, 1, 2 with probabilities 0.25, 0.5 and 0.25, respectively. The result is reported in Table S3 in the supplemental material to save space. We can see from Table S3 that the TPC outperforms other methods, especially in terms of the correct-fit rate.

3.2 An application

In this section, we demonstrate the proposed methodology by an empirical analysis of microarray data set, which was studied by Scheetz et al. (2006) and Huang et al. (2008). This dataset contains 120 12-week-old male rats, and for each rat, 3000 sufficiently expressed gene probes with enough variation are studied. The purpose of the analysis is to identify the probes that are most relevant to the response – the expression level of probe TRIM32, which are recently proved to cause Bardet-Biedl syndrome (Chiang, 2006).

We apply the SCAD, LASSO, PC-simple algorithm and TPC to this data set with one outlier deleted. Table 3.2 provides the information of the chosen gene probes by different methods. As LASSO yields a much larger model leading to the difficulty of interpretation, to save space, we only report the six probes selected by SCAD, the PC-simple algorithm and TPC, and indicate whether they are included in the 20 chosen probes by LASSO. We calculate the adjusted R^2 for each model and prediction error (PE) by leave-one-out cross-validation (LOOCV) method for each model. From Table 3.2, we can see that the models selected by SCAD, LASSO and TPC have very similar performance in terms of adjusted R^2 and predictor error. The TPC method improves the PC-simple algorithm by including two probes x_5 and x_6 . These two probes lead to about 9% predictor error reduction from the model selected by PC-simple to the model selected by TPC. Note that the probe 1389584 at (x_1) and 1383996 at (x_2) are selected by all the four approaches, and also identified by Huang et al. (2008). Therefore, they are worth more comprehensively biological research. The rest of the results from TPC is more consistent with Huang et al. (2008) than the other methods.

We further conduct some exploratory analysis. We compare the model with 20 probes selected by LASSO with the model with the six probes listed in Table 3.2 (denoted by M6 in the table) by the likelihood ratio test (LRT). The p-value of the corresponding LRT is 0.058. This implies that the model with the six probes fits the data well enough. The corresponding estimates and standard errors of regression coefficients are listed in the second last column in Table 3.2. The adjusted R^2 and the predictor error calculated by the LOOCV method has much improvement over the model selected by the SCAD, LASSO, PC-simple and TPC methods. For example, the predictor error has about 10% reduction. The coefficients of x_4 and x_5 seem not to be significant at level 0.05. We refit the data to model with only four probes x_1 , x_2 , x_3 and x_6 , and their estimates and standard error are reported in the last column of Table 3.2. The adjusted R^2 and predictor error of this model is very close to the one with six probes. The empirical analysis of this example implies that two comparable schemes for variable selection (i.e., regularization methods such as the SCAD and the LASSO and partial correlation based method such as PC-simple and TPC) can be used to improve each others. For example, the regularization method would miss probe x_6 , while the TPC would miss probe x_3 . Scientists may raise their confidence in the selected probes x_1 and x_2 since they are chosen by both techniques.

4. Conclusion

In this paper, we proposed the variable selection procedure via the thresholded partial correlation (TPC) and established its model selection consistency and sure screening property in the presence of ultrahigh-dimensional predictors. Our simulation and empirical analysis of a real-life data example illustrate that the TPC may serve as a potential alternative to the commonly used regularization methods for high or ultrahigh dimensional regression models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Li's research was supported by National Institute on Drug Abuse (NIDA) grants P50-DA10075 and P50 DA036107. Liu's research was supported by National Natural Science Foundation of China (NNSFC) grant 11401497 and the Fundamental Research Funds for the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry. All authors equally contribute to this paper, and the authors are listed in the alphabetic order.

References

- Arellano-Valle RB, del Pino F, Iglesias P. Bayesian inference in spherical linear models: robustness and conjugate analysis. *Journal of Multivariate Analysis*. 2006; 97:179–197.
- Bühlmann P, Kalisch M, Maathuis M. Variable selection in high-dimensional linear models: partially faithful distributions and the PC24 simple algorithm. *Biometrika*. 2010; 97:261–278.
- Chen Y, Wiesel A, Hero AO. Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*. 2011; 59:4097–4107.
- Chiang AP, Beck JS, Yen H-J, Tayeh MK, Scheetz TE, Swiderski R, Nishimura D, Braun TA, Kim K-Y, Huang J, Elbedour K, Carmi R, Slusarski DC, Casavant TL, Stone EM, Sheffield VC. Homozygosity Mapping with SNP Arrays Identifies a Novel Gene for Bardet-Biedl Syndrome (BBS10). *Proceeding of the National Academy of Sciences*. 2006; 103:6287–6292.
- Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*. 2008; 70:849–911.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
- Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*. 2010; 20:101–148. [PubMed: 21572976]
- Fang, KT., Kotz, S., Ng, KW. *Symmetric Multivariate and Related Distributions*. Chapman and Hall; New York, NY: 1990.
- Huang J, Ma SG, Zhang CH. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*. 2008; 18:1603–1618.
- Liang H, Li R. Variable selection for partially linear models with measurement errors. *Journal of American Statistical Association*. 2009; 104:234–248.
- Mcneil, AJ., Frey, R., Embrechts, P. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press; Princeton, NJ: 2005.
- Mottonen J, Oja H, Tienari J. On the efficiency of multivariate spatial sign and rank tests. *Annals of Statistics*. 1997; 25:542–552.
- Muirhead, RJ. *Aspects of Multivariate Statistical Theory*. Wiley; New York: 1982.
- Oja H, Randles RH. Multivariate nonparametric tests. *Statistical Sciences*. 2004; 19:598–605.

- Osiewalski J. A note on Bayesian inference in a regression model with elliptical errors. *Journal of Econometrics*. 1991; 48:183–193.
- Osiewalski J, Steel MFJ. Robust bayesian inference in elliptical regression models. *Journal of Econometrics*. 1993; 57:345–363.
- Scheetz TE, Kim K-YA, Swiderski RE, Philp1 AR, Braun TA, Knudtson KL, Dorrance AM, DiBona GF, Huang J, Casavant TL, Sheffield VC, Stone EM. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceeding of the National Academy of Sciences*. 2006; 103:14429–14434.
- Schmidt R. Tail Dependence for Elliptically Contoured Distributions. *Mathematical Methods of Operations Research*. 2002; 55:301–327.
- Soloveychik I, Wiesel A. Performance analysis of Tyler's covariance estimator. *IEEE Transactions on Signal Processing*. 2015; 63:418–426.
- Tibshirani R. Regression shrinkage and selection via LASSO. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
- Vidal I, Arellano-Valle RB. Bayesian inference for dependent elliptical measurement error models. *Journal of Multivariate Analysis*. 2010; 101:2587–2597.
- Wang L, Peng B, Li R. A high-dimensional nonparametric multivariate test for mean vector. *Journal of American Statistical Association*. 2015 Accepted.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.

Appendix

Appendix A: Proof of Theorem 1

Let $\mathbf{u}_i = (y_i, \mathbf{x}_i^T)$, $i = 1, \dots, n$ and denote $q = p+1$. Thus, $\mathbf{u}_1, \dots, \mathbf{u}_n$ be an independent and identically distributed random sample from $EC_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$. To study the asymptotic behaviors of partial correlation of elliptical distribution, we consider the following general partitions of $\mathbf{u}_i$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$\mathbf{u}_i = \begin{pmatrix} \mathbf{u}_{1i} \\ \mathbf{u}_{2i} \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{u}_{1i}$ and $\boldsymbol{\mu}_1$ are q_1 -dimensional, while $\mathbf{u}_{2i}$ and $\boldsymbol{\mu}_2$ are q_2 -dimensional, $\boldsymbol{\Sigma}_{11}$ is a $q_1 \times q_1$ matrix, and $\boldsymbol{\Sigma}_{22}$ is a $q_2 \times q_2$ matrix. Here $q = q_1 + q_2$. Let $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_n)^T$ and denote

$$\mathbf{A} = \frac{1}{n} \sum_{i=1}^n (\mathbf{u}_i - \bar{\mathbf{u}})(\mathbf{u}_i - \bar{\mathbf{u}})^T = \frac{1}{n} \mathbf{U}^T \left(\mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) \mathbf{U}$$

Partition $\mathbf{A}$ in the same way as $\boldsymbol{\Sigma}$. Let a_{kl2} is the (k, l) -element of

$\mathbf{A}_{11.2} \triangleq \mathbf{A}_{11} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{A}_{21}$. Then the sample partial correlation of u_{ik} and u_{il} given $\mathbf{u}_{2i}$, $\hat{\rho}(u_{ik}, u_{il} | \mathbf{u}_{2i})$, indeed equals $a_{kl2} / \sqrt{a_{kk,2} a_{ll,2}}$.

To derive the asymptotic distribution of $\mathbf{A}_{11.2}$, let

$$C = \begin{pmatrix} I & -\Sigma_{12} \\ 0 & I \end{pmatrix},$$

where I stands for the identity matrix, and $\mathbf{v}_i = C(\mathbf{u}_i - \boldsymbol{\mu})$. Using Theorem 2.16 of Fang, Kotz and Ng (1990), it follows that

$$\mathbf{v}_i \sim EC_q(0, \begin{pmatrix} \Sigma_{11.2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}, \phi) \quad (\text{A.1})$$

where $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Let $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_n)^T$. By definition of $\mathbf{v}_i$, $\mathbf{V} = (\mathbf{U} - \mathbf{1}_n \boldsymbol{\mu}^T)C^T$, where $\mathbf{1}_n$ is an $n \times 1$ vector with all elements being 1. Define

$$\mathbf{B} = \frac{1}{n} \sum_{i=1}^n (\mathbf{v}_i - \bar{\mathbf{v}})(\mathbf{v}_i - \bar{\mathbf{v}})^T = \frac{1}{n} \mathbf{V}^T \left(I - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) \mathbf{V} = C \mathbf{A} C^T.$$

Partition $\mathbf{B}$ in the same way as $\mathbf{A}$, then $\mathbf{B}_{11} = \mathbf{A}_{11} - \Sigma_{12}\mathbf{A}_{21} - \mathbf{A}_{12}\Sigma_{21} + \Sigma_{12}\mathbf{A}_{22}\Sigma_{21}$, $\mathbf{B}_{12} = \mathbf{A}_{12} - \Sigma_{12}\mathbf{A}_{22}$, $\mathbf{B}_{21} = \mathbf{A}_{21} - \mathbf{A}_{22}\Sigma_{21}$ and $\mathbf{B}_{22} = \mathbf{A}_{22}$. By direct calculation, it follows that $\mathbf{B}_{11.2} \triangleq \mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21} = \mathbf{A}_{11.2}$. This enables us to derive the asymptotic distribution of $\mathbf{A}_{11.2}$ through $\mathbf{B}_{11.2}$.

Define $\mathbf{W}_{11} = \sqrt{n}(\mathbf{B}_{11} - \Sigma_{11.2})$, $\mathbf{W}_{12} = \sqrt{n}\mathbf{B}_{12}$, $\mathbf{W}_{21} = \sqrt{n}\mathbf{B}_{21}$, and

$\mathbf{W}_{22} = \sqrt{n}(\mathbf{B}_{22} - \Sigma_{22})$. The assumption that all fourth moments of $\mathbf{u}_i$ are finite implies that all fourth moments of $\mathbf{v}_i$ are finite. Thus, it follows by the central limit theorem that $\mathbf{W}_{kl}$ for $k = 1, 2$ and $l = 1, 2$, has an asymptotic normal distribution with mean zero and a finite covariance matrix. Then

$$\mathbf{B}_{11.2} = \frac{1}{\sqrt{n}} \mathbf{W}_{11} + \sum_{11.2} - \frac{1}{n} \mathbf{W}_{12} \left(\Sigma_{22} + \frac{1}{\sqrt{n}} \mathbf{W}_{22} \right)^{-1} \mathbf{W}_{21}.$$

By the assumptions of Theorem 1, the largest eigenvalue of $\left(\Sigma_{22} + \frac{1}{\sqrt{n}} \mathbf{W}_{22} \right)^{-1}$ is positive and finite. Therefore it follows that if $q_2 = o(\sqrt{n})$,

$$\begin{aligned} \sqrt{n}(\mathbf{A}_{11.2} - \Sigma_{11.2}) &= \sqrt{n}(\mathbf{B}_{11.2} - \Sigma_{11.2}) = \mathbf{W}_{11} + o_P(1) \\ &= \sqrt{n}(\mathbf{B}_{11} - \Sigma_{11.2}) + o_P(1). \end{aligned}$$

This implies that $\mathbf{A}_{11.2}$ and $\mathbf{B}_{11}$ have the same asymptotic normal distribution, and hence $a_{kl.2}/\sqrt{a_{kk.2}a_{ll.2}}$ and $b_{kl}/\sqrt{b_{kk}b_{ll}}$ have the same asymptotic distribution, where b_{kl} is the (k, l) -element of $\mathbf{B}_{11}$. Further notice that $\mathbf{v}_{1i} \sim EC_{q_1}(0, \Sigma_{11.2}), \phi$ by (A.1), where $\mathbf{v}_{1i}$ consists of the first q_1 elements of $\mathbf{v}_i$. Therefore, the asymptotic normal distribution of the sample correlation coefficient $\hat{\rho}(v_{ik}, v_{il})$, which indeed equals to $b_{kl}/\sqrt{b_{kk}b_{ll}}$, can be derived from (2.2) with replacing Σ by $\Sigma_{11.2}$. Thus, Theorem 1 holds by setting $\mathbf{u}_{li} = (y_i, x_{i\mathcal{S}^c}^T)^T$ and $\mathbf{u}_{2i} = \mathbf{x}_{i\mathcal{S}}$.

Appendix B: Proof of Theorems 2 and 3

In this section, we introduce the following lemmas which are used repeatedly in the proofs of the theorems.

Lemma 1

(Hoeffding's Inequality) Assume the independent random sample $\{X_i: i = 1, \dots, n\}$ satisfies $P(X_i \in [a_i, b_i]) = 1$ for some a_i and $b_i, \forall i = 1, \dots, n$. Then, for any $\varepsilon > 0$, the sample mean $\bar{X}$ satisfies

$$P(|\bar{X} - E(\bar{X})| > \varepsilon) \leq 2\exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \quad (\text{B.1})$$

Lemma 2

Suppose X is a random variable with $E(e^{aX}) < \infty$ for some $a > 0$. Then for any $M > 0$, there exist positive constants b and c such that

$$P(|X| \geq M) \leq be^{-cM}. \quad (\text{B.2})$$

Lemma 3

Suppose $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are estimates of the finite parameters γ_1 and γ_2 based on a size- n sample, respectively. Assume there exist positive constants b_1, b_2 and ν such that for any $0 < \varepsilon < 1$,

$$P\{|\hat{\gamma}_j - \gamma_j| > \varepsilon\} \leq b_j \exp(-n^\nu / b_j), \quad j=1, 2. \quad (\text{B.3})$$

Then

$$\begin{aligned} P\{|\hat{\gamma}_1 - \hat{\gamma}_2 - (\gamma_1 - \gamma_2)| > \varepsilon\} &\leq b_3 \exp(-n^\nu/b_3), \\ P\{|\hat{\gamma}_1 \hat{\gamma}_2 - \gamma_1 \gamma_2| > \varepsilon\} &\leq b_4 \exp(-n^\nu/b_4), \end{aligned}$$

where $b_3 = b_1 + b_2$, and $b_4 = 2b_1 + b_2$. If $\gamma_2 = 0$,

$$P\left\{\left|\frac{\hat{\gamma}_1}{\hat{\gamma}_2} - \frac{\gamma_1}{\gamma_2}\right| > \varepsilon\right\} \leq b_5 \exp(-n^\nu/b_5),$$

where $b_5 = b_1 + 3b_2$. If we further assume $\gamma_2 > 0$, then

$$\begin{aligned} P\{|\sqrt{\hat{\gamma}_2} - \sqrt{\gamma_2}| > \varepsilon\} &\leq b_6 \exp(-n^\nu/b_6), \\ P\{|\log \hat{\gamma}_2 - \log \gamma_2| > \varepsilon\} &\leq b_2 \exp(-n^\nu/b_2), \end{aligned}$$

where $b_6 = 2b_2$.

The proof of Lemma 3 can be found at the supplemental material of this work.

For ease of notation, denote

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \bar{x}_j \bar{y} = \frac{1}{n} \sum_{i=1}^n x_{ij} y_i, \bar{x}_j^2 = \frac{1}{n} \sum_{i=1}^n x_{ij}^2, \text{ and } \bar{y}^2 = \frac{1}{n} \sum_{i=1}^n y_i^2. \text{ Then}$$

$$\hat{\rho}(y, x_j) = \frac{\bar{x}_j \bar{y} - \bar{x}_j \bar{y}}{\sqrt{(\bar{x}_j^2 - \bar{x}_j^2)(\bar{y}^2 - \bar{y}^2)}} \quad (\text{B.4})$$

The proof of Theorem 2

We divide the proof into three parts.

Step 1

Study the consistency of $\hat{Z}_n(y, x_j | x_{\mathcal{S}}) / \sqrt{1 + \hat{\kappa}}$. First consider $\bar{x}_j$. For any $0 < \varepsilon < 1$ and any $M > 0$,

$$\begin{aligned} P(|\bar{x}_j - \mathbb{E}x_j| > \varepsilon) &\leq P(|\bar{x}_j - \mathbb{E}x_j| > \varepsilon, \max_{1 \leq i \leq n} |x_{ij}| \leq M) + P(\max_{1 \leq i \leq n} |x_{ij}| > M) \\ &\leq P(|\bar{x}_j - \mathbb{E}x_j| > \varepsilon, \max_{1 \leq i \leq n} |x_{ij}| \leq M) + nP(|x_{ij}| > M) \\ &\leq 2\exp(-\frac{n\varepsilon^2}{2M^4}) + nC_2 \exp(-C_1 M) \end{aligned} \quad (\text{B.5})$$

for some positive constants C_1 and C_2 . The first term above is obtained by Hoeffding's inequality in Lemma 1, and the second term is by condition (D2) and Lemma 2. Take $M = O(n^{1/5})$, then for large n , (B.5) is simplified as $P(|\bar{x}_j - \mathbb{E}x_j| > \varepsilon) \leq C_3 \exp(-n^\nu/C_3)$, where $0 <$

$\nu < 1/5$ and $C_3 > 0$. In the same fashion, there exist some positive constants C_4 , C_5 , C_6 and C_7 , such that for large n ,

$$\begin{aligned} P(|\bar{y} - E y| > \varepsilon) &\leq C_4 \exp(-n^\nu / C_4), \\ P(|\bar{x}_j^2 - E(x_j^2)| > \varepsilon) &\leq C_5 \exp(-n^\nu / C_5), \\ P(|\bar{y}^2 - E(y^2)| > \varepsilon) &\leq C_6 \exp(-n^\nu / C_6), \\ P(|\bar{x}_j \bar{y} - E(x_j y)| > \varepsilon) &\leq C_7 \exp(-n^\nu / C_7). \end{aligned}$$

Therefore by (B.4) and Lemma 3,

$$P\{|\hat{\rho}(y, x_j) - \rho(y, x_j)| > \varepsilon\} \leq C_8 \exp(-n^\nu / C_8),$$

where the positive constant C_8 is determined by $C_3, \dots, C_7$.

Note that

$$\rho(y, x_j | x_{\mathcal{S}}) = \frac{\rho(y, x_j | x_{\mathcal{S} \setminus \{k\}}) - \rho(y, x_k | x_{\mathcal{S} \setminus \{k\}}) \rho(x_j, x_k | x_{\mathcal{S} \setminus \{k\}})}{[1 - \rho^2(y, x_k | x_{\mathcal{S} \setminus \{k\}})] \{1 - \rho^2(x_j, x_k | x_{\mathcal{S} \setminus \{k\}})\}}^{1/2}, \quad (\text{B.6})$$

for any $k \in \mathcal{S}$.

Under the bounded condition (D5), applying Lemma 3 to the sample version of (B.6) and the Z-transformation (2.4) recursively, we conclude that for some $C_9 > 0$ and $C_{10} > 0$,

$$\begin{aligned} P\{|\hat{\rho}(y, x_j | x_{\mathcal{S}}) - \rho(y, x_j | x_{\mathcal{S}})| > \varepsilon\} &\leq C_9 \exp(-n^\nu / C_9), \text{ and} \\ P\{|\hat{Z}(y, x_j | x_{\mathcal{S}}) - Z(y, x_j | x_{\mathcal{S}})| > \varepsilon\} &\leq C_{10} \exp(-n^\nu / C_{10}). \end{aligned}$$

Furthermore, by the same argument, the sample kurtosis is consistent to the population version with the same rate, that is, there exists $C_{11} > 0$ such that $P\{|\hat{\kappa} - \kappa| > \varepsilon\} \leq C_{11} \exp(-n^\nu / C_{11})$, and hence for some $C_{12} > 0$,

$$P\left\{\left|\frac{\hat{Z}(y, x_j | x_{\mathcal{S}})}{\sqrt{1 + \hat{\kappa}}} - \frac{Z(y, x_j | x_{\mathcal{S}})}{\sqrt{1 + \kappa}}\right| > \varepsilon\right\} \leq C_{12} \exp(-n^\nu / C_{12}).$$

Step 2

Compute $P(E_{j|\mathcal{S}}) = P\{\text{an error occurs when testing } \rho(y, x_j | x_{\mathcal{S}}) = 0\}$. Denote

$E_{j|\mathcal{S}} = E_{j|\mathcal{S}}^I \cup E_{j|\mathcal{S}}^{II}$, where $E_{j|\mathcal{S}}^I$ is the event that the type I error occurs and $E_{j|\mathcal{S}}^{II}$ is the event that the type II error occurs. Then by choosing $\alpha_n = 2\{1 - \Phi(\frac{c_n}{2} \sqrt{\frac{n}{1+\kappa}})\}$,

$$\begin{aligned}
P(E_{j|\mathcal{S}}^I) &= P \left\{ \left| \frac{\hat{Z}(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} \right| > \frac{\Phi^{-1}(1-\alpha_n/2)}{(n-|\mathcal{S}|-1)^{1/2}} \text{ when } Z(y, x_j | x_{\mathcal{S}}) = 0 \right\} \\
&\leq P \left\{ \left| \frac{\hat{Z}(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\kappa}} \right| > \frac{\Phi^{-1}(1-\alpha_n/2)}{(n-|\mathcal{S}|-1)^{1/2}} \right\} \\
&= P \left\{ \left| \frac{\hat{Z}(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\kappa}} \right| > \frac{c_n \sqrt{n}}{2} \{(n-|\mathcal{S}|-1)(1+\kappa)\}^{-1/2} \right\} \\
&\leq P \left\{ \left| \frac{\hat{Z}(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\kappa}} \right| > \frac{c_n}{2\sqrt{1+\kappa}} \right\} \\
&\leq C_{12} \exp(-n^\nu / C_{12}),
\end{aligned}$$

and

$$\begin{aligned}
P(E_{j|\mathcal{S}}^{II}) &= P \left\{ \left| \frac{\hat{Z}(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} \right| \leq \frac{\Phi^{-1}(1-\alpha_n/2)}{(n-|\mathcal{S}|-1)^{1/2}} \text{ when } Z(y, x_j | x_{\mathcal{S}}) \neq 0 \right\} \\
&\leq P \left\{ \left| \frac{Z(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\kappa}} \right| - \left| \frac{\hat{Z}(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\kappa}} \right| \leq \frac{\Phi^{-1}(1-\alpha_n/2)}{(n-|\mathcal{S}|-1)^{1/2}} \right\} \\
&\leq P \left\{ \left| \frac{\hat{Z}(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\kappa}} \right| \geq \left| \frac{Z(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\kappa}} \right| - \frac{c_n}{2\sqrt{1+\kappa}} \right\}
\end{aligned}$$

Note that $|g(u)| = \frac{1}{2} \log \{(1+u)/(1-u)\} \geq |u|$ for all $u \in (-1, 1)$, then $|Z(y, x_j | x_{\mathcal{S}})| \geq \rho_n(y, x_j | x_{\mathcal{S}}) - c_n$ under condition (D4). Thus,

$$\begin{aligned}
P(E_{j|\mathcal{S}}^{II}) &\leq P \left\{ \left| \frac{\hat{Z}(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\kappa}} \right| \geq \frac{c_n}{\sqrt{1+\kappa}} - \frac{c_n}{2\sqrt{1+\kappa}} \right\} \\
&= P \left\{ \left| \frac{\hat{Z}(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\hat{\kappa}}} - \frac{Z(y, x_j | x_{\mathcal{S}})}{\sqrt{1+\kappa}} \right| \geq \frac{c_n}{2\sqrt{1+\kappa}} \right\} \\
&\leq C_{12} \exp(-n^\nu / C_{12}).
\end{aligned}$$

Therefore, $P(E_{j|\mathcal{S}}) = P(E_{j|\mathcal{S}}^I) + P(E_{j|\mathcal{S}}^{II}) \leq 2C_{12} \exp(-n^\nu / C_{12})$.

Step 3

Study $P\{\hat{\mathcal{A}}_n(\alpha_n) = \mathcal{A}\}$. Now consider all $j = 1, \dots, p$ and all $\mathcal{S} \subseteq \{j\}^c$ subject to $|\mathcal{S}| \leq m_n$, where $m_n \leq \hat{m}_{reach}$. Define $K_j^{m_n} = \{\mathcal{S} \subseteq \{j\}^c, |\mathcal{S}| \leq m_n\}$, $j = 1, \dots, p$.

$$\begin{aligned}
P\{\hat{\mathcal{A}}_n(\alpha_n) \neq \mathcal{A}\} &= P\{\text{an error occurs for some } j \text{ and some } \mathcal{S}\} \\
&= P \left\{ \bigcup_{j=1, \dots, p; \mathcal{S} \in K_j^{m_n}} E_{j|\mathcal{S}} \right\} \leq \sum_{j=1, \dots, p; \mathcal{S} \in K_j^{m_n}} P(E_{j|\mathcal{S}}) \\
&\leq p^{m_n+1} \cdot \sup_{j=1, \dots, p; \mathcal{S} \in K_j^{m_n}} P(E_{j|\mathcal{S}}) \leq 2p^{m_n+1} C_{12} \exp(-n^\nu / C_{12}) \\
&\leq 2p^{d_0+1} C_{12} \exp(-n^\nu / C_{12}). \tag{B.7}
\end{aligned}$$

The second inequality holds since the number of possible choices of j is p and there are p^{m_n} possible choices for $\mathcal{S}$. The last inequality in (B.7) is obtained because $P(\hat{m}_{reach} = m_{reach})$

$\rightarrow 1$ and $m_{reach} \rightarrow d_0$ by the same technique as Lemma 3 in Bühlmann, Kalisch and Maathuis (2010). Thus for large n , $m_n \rightarrow \hat{m}_{reach} \rightarrow d_0$.

Moreover, recall that ν can be chosen arbitrarily in $(0, 1/5)$. Therefore, if d_0 is fixed, for $p = \alpha(\exp(n^\xi))$, $0 < \xi < 1/5$, (B.7) is simplified as $P\{\hat{\mathcal{A}}_n(\mathbf{a}_n) \neq \mathcal{A}\} = O\{\exp(-n^\nu/C_{12})\}$, provided $\xi < \nu < 1/5$. If $d_0 = O(n^b)$, $0 < b < 1/5$, for $p = \alpha(\exp(n^\xi))$, $0 < \xi < 1/5 - b$, (B.7) becomes $P\{\hat{\mathcal{A}}_n(\mathbf{a}_n) \neq \mathcal{A}\} = O\{\exp(-n^\nu/C_{12})\}$, provided that $\xi + b < \nu < 1/5$. This completes the proof of Theorem 2 with $C = C_{12}$.

Proof of Theorem 3

We only need to consider the first step of the thresholded partial correlation approach, where we have

$$P\left(\left|\frac{\hat{Z}(y, x_j)}{\sqrt{1+\hat{\kappa}}} - \frac{Z(y, x_j)}{\sqrt{1+\kappa}}\right| > \varepsilon\right) \leq C_{13}\exp(-n^\nu/C_{13})$$

for some $C_{13} > 0$. Define $E_j^{II} = \{\text{fail to include } x_j \text{ when } x_j \text{ is a true predictor}\}$, then using the same technique as the proof of Theorem 2,

$$\begin{aligned} P(E_j^{II}) &= P\left\{\left|\frac{\hat{Z}_n(y, x_j)}{\sqrt{1+\hat{\kappa}}}\right| \leq \frac{\Phi^{-1}(1-\alpha_n/2)}{(n-1)^{1/2}} \text{ when } \beta_j \neq 0\right\} \\ &\leq P\left\{\left|\frac{\hat{Z}_n(y, x_j)}{\sqrt{1+\hat{\kappa}}} - \frac{Z_n(y, x_j)}{\sqrt{1+\kappa}}\right| \geq \frac{c_n}{2\sqrt{1+\kappa}}\right\} \\ &\leq C_{13}\exp(-n^\nu/C_{13}). \end{aligned}$$

Then

$$P\{\hat{\mathcal{A}}_n^{[1]} \neq \mathcal{A}\} = P\left\{\bigcup_{j=1}^p E_j^{II}\right\} \leq \sum_{j=1}^p P(E_j^{II}) \leq pC_{13}\exp(-n^\nu/C_{13}),$$

for any $0 < \nu < 1/5$. Therefore, for $p = \alpha(\exp(n^\xi))$ and $P\{\hat{\mathcal{A}}_n^{[1]} \neq \mathcal{A}\} \leq O\{\exp(-n^\nu/C_{13})\}$, provided $\xi < \nu < 1/5$. This completes the proof of Theorem 3 with $C^* = C_{13}$.

Table 3.1

Simulation Results for Example 1: Elliptical Distribution

p	ρ	Method	MedME(Dev)	TPN	FPN	UF	CF	OF
200	0	SCAD	0.050 (0.024)	3.00	4.52	0.00	0.51	0.49
		LASSO	8.984 (0.219)	3.00	33.63	0.00	0.00	1.00
		PC-simple	0.082 (0.050)	2.92	0.82	0.08	0.41	0.51
		TPC	0.045 (0.032)	2.84	0.13	0.16	0.81	0.03
	0.3	SCAD	0.046 (0.023)	3.00	3.90	0.00	0.50	0.50
		LASSO	11.195 (0.216)	3.00	30.26	0.00	0.00	1.00
		PC-simple	0.063 (0.036)	3.00	0.46	0.00	0.58	0.42
		TPC	0.036 (0.024)	2.99	0.04	0.01	0.96	0.03
	0.8	SCAD	0.044 (0.026)	3.00	2.51	0.00	0.50	0.50
		LASSO	20.925 (0.158)	3.00	16.44	0.00	0.02	0.98
		PC-simple	0.039 (0.026)	2.94	0.17	0.06	0.83	0.11
		TPC	0.057 (0.040)	2.79	0.20	0.19	0.80	0.01
500	0	SCAD	0.041 (0.022)	3.00	5.57	0.00	0.41	0.59
		LASSO	8.960 (0.212)	3.00	45.25	0.00	0.00	1.00
		PC-simple	0.096 (0.051)	2.83	1.22	0.17	0.25	0.58
		TPC	0.043 (0.031)	2.74	0.21	0.26	0.70	0.04
	0.3	SCAD	0.043 (0.024)	3.00	7.05	0.00	0.40	0.60
		LASSO	11.172 (0.230)	3.00	38.94	0.00	0.00	1.00
		PC-simple	0.077 (0.043)	3.00	0.83	0.00	0.35	0.65
		TPC	0.030 (0.018)	2.98	0.08	0.02	0.91	0.07
	0.8	SCAD	0.042 (0.026)	3.00	4.07	0.00	0.40	0.60
		LASSO	20.879 (0.187)	3.00	20.86	0.00	0.00	1.00
		PC-simple	0.049 (0.031)	2.91	0.37	0.09	0.69	0.22
		TPC	0.044 (0.032)	2.73	0.26	0.25	0.75	0.00
		SCAD	0.051 (0.032)	3.00	10.13	0.00	0.40	0.60

d	ρ	Method	MedME(Devi)	TPN	FPN	UF	CF	OF
2000	0	LASSO	9.140 (0.179)	3.00	66.84	0.00	0.00	1.00
		PC-simple	0.112 (0.056)	2.90	1.73	0.10	0.07	0.83
		TPC	0.050 (0.037)	2.83	0.35	0.17	0.67	0.16
	0.3	SCAD	0.045 (0.028)	3.00	8.58	0.00	0.33	0.67
		LASSO	11.345 (0.189)	3.00	61.97	0.00	0.00	1.00
		PC-simple	0.105 (0.044)	2.99	1.36	0.01	0.17	0.82
2000	0.8	TPC	0.039 (0.026)	2.97	0.18	0.03	0.83	0.14
		SCAD	0.049 (0.030)	3.00	7.33	0.00	0.28	0.72
		LASSO	20.960 (0.136)	3.00	37.81	0.00	0.00	1.00
	0.8	PC-simple	0.077 (0.046)	2.96	0.59	0.04	0.48	0.48
		TPC	0.045 (0.034)	2.82	0.24	0.17	0.81	0.02

* The numbers in the parentheses are median absolute deviations over 1000 simulations.

Table 3.2

Results for Real Data Example

Selected Probes	SCAD	LASSO	PC-simple	TPC	M6 Est(& SE)	M4 Est(& SE)
Intercept	Yes	Yes	Yes	Yes	.0147 (.0465)	.0164 (.0467)
1389584 at(x_1)	Yes	Yes	Yes	Yes	.3669 (.0823)***	.4098 (.0710) ***
1383996 at(x_2)	Yes	Yes	Yes	Yes	.1400 (.0595) *	.1583 (.0590) **
1382452 at(x_3)	Yes	Yes	/	/	.2450 (.0606)***	.2279 (.0547) ***
1370429 at(x_4)	/	/	Yes	Yes	.0464 (.0815)	
1383110 at(x_5)	/	Yes	/	Yes	.1543 (.0840)	
1374106 at(x_6)	/	Yes	/	Yes	.2203 (.0727)**	.2580 (.0688) ***
15 more probe	/	Yes	/	/		
Size	4	21	4	6	7	5
Adjusted- R^2 (%)	69.37	69.55	66.64	69.10	74.60	74.38
PE	0.297	0.298	0.326	0.301	0.275	0.270

The 15 probes selected only by LASSO are omitted. ‘‘Yes’’ means the probe is selected by this method.
M6 stands for the linear model with six probes x_1 - x_6 ; M4 for the model with four probes x_1 , x_2 , x_3 and x_6 .
‘**’ stands for significant at level 0.05, ‘***’ for level 0.01, and ‘****’ for level 0.001.