Statistica Sinica **27** (2017), 983-996
doi:https://doi.org/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: A partial correlation-based variable selection method was proposed for normal linear regression models by Bühlmann, Kalisch and Maathuis (2010) as an alternative to regularization methods for variable selection. This paper addresses issues related to (a) whether the method is sensitive to the normality assumption, and (b) whether the method is valid when the dimension of predictor increases at an exponential rate in the sample size. To address (a), we study the method for elliptical linear regression models. Our finding indicates that the original proposal can lead to inferior performance when the marginal kurtosis of predictor is not close to that of normal distribution, and simulation results confirm this. To ensure the superior performance of the 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 results address the issue (b) directly. As a by-product, the sure screening property of the first step of TPC is obtained. Numerical examples illustrate that the TPC is comparable to the commonly-used regularization methods for variable selection.

Key words 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 such works.

As an alternative method to penalized least squares for variable selection, Bühlmann, Kalisch and Maathuis (2010) proposed a variable selection procedure, named PC-simple algorithm, which ranked the partial correlations (PC) between the predictors and the response. 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 comparable to penalized least squares variable selection approaches. With two schemes for variable selection in high-dimensional linear models, one has elevated their confidence in the selected predictors when they are chosen by both techniques.

We study two issues related to the PC-simple algorithm. The first is that the procedure proposed in Bühlmann, Kalisch and Maathuis (2010) relies on a normality assumption on the joint distribution of response and predictors, while partial faithfulness does not require this. The second is that the results established in Bühlmann, Kalisch and Maathuis (2010) require that the dimension of the predictor vector increases at a polynomial rate in the sample size. We study whether the results are valid with dimensionality increasing at an exponential rate in the sample size.

In studying the normality assumption, we consider that the response and the predictors in a linear regression model jointly follow an elliptical distribution (Fang, Kotz and Ng (1990)). The elliptical distribution family contains a much broader class of distributions than the normal distribution family, such as mixtures of normal distributions, the multivariate t-distribution, the multi-uniform distribution on unit sphere, and the 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)). 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)). The elliptical distribution family has a variety of applications. For instance, it has been used for modeling finance data (Mcneil, Frey and Embrechts (2005)) to accommodate tail dependence and the phenomenon of simultaneous extremes, which are not allowed by the multivariate normal (Schmidt (2002)).

In 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 whose marginal kurtosis is larger (smaller) than that of the normal. 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, 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 with 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 so as to address the issue of dimensionality increasing at exponential rate in the sample size.

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

This paper is organized as follows. In Section 2, we propose the TPC for the elliptical linear models, and establish its asymptotic properties. Numerical studies are conducted in Section 3. A brief conclusion is given in Section 4. The proofs are in the supplemental materials.

2. Thresholded Partial Correlation (TPC) Approach

2.1. Elliptical linear model and its partial correlation estimation

Consider the linear model

$$y = \mathbf{x}^T\boldsymbol{\beta} + \epsilon, \tag{2.1}$$

where y is the response variable, $\mathbf{x} = (x_1, \cdots, x_p)^T$ is the covariate vector, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the coefficient vector, and ϵ is the random error with $\mathrm{E}(\epsilon) = 0$, $\mathrm{var}(\epsilon) = \sigma^2$. Throughout, it is assumed without loss of generality that $E(\mathbf{x}) = 0$

and $E(y) = 0$ so that there is no intercept in (2.1). In practice, it is common that $\mathbf{x}$ and y are marginally standardized before performing variable selection. We suppose that $(\mathbf{x}_1^T, y_1)$, $\cdots$, $(\mathbf{x}_n^T, y_n)$ are independent and identically distributed (iid) random samples from an elliptical distribution $\mathrm{EC}_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ that has the characteristic function $\exp(i\mathbf{t}^T\boldsymbol{\mu})\phi(\mathbf{t}^T\Sigma\mathbf{t})$ for some characteristic generator $\phi(\cdot)$ (Fang, Kotz and Ng (1990)).

Bühlmann, Kalisch and Maathuis (2010) proposed a variable selection method, PC-simple algrithm, based on the parital correlations for (2.1) with normal response and predictors. We study the limiting distributions for correlations and partial correlations under the elliptical assumption. Denote by $\rho(y, x_j)$ and $\hat{\rho}(y, x_j)$ the population and the sample correlations between y and x_j, respectively. Then, in Theorem 5.1.6 of Muirhead (1982), the asymptotic distribution of $\hat{\rho}(y, x_j)$ is

$$\sqrt{n}\,\{\hat{\rho}(y, x_j) - \rho(y, x_j)\} \to N\Big(0, (1+\kappa)\{1-\rho^2(y, x_j)\}^2\Big), \tag{2.2}$$

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

For an index set $\mathcal{S} \subseteq \{1, 2, \cdots, p\}$, let $\mathcal{S}^c = \{1 \le j \le p : j \notin \mathcal{S}\}$, $|\mathcal{S}|$ be its cardinality, and $x_{\mathcal{S}} = \{x_j : j \in \mathcal{S}\}$ be a subset of covariates with index set $\mathcal{S}$. Denote the truly active index set by $\mathcal{A} = \{1 \le j \le p : \beta_j \ne 0\}$, with the corresponding cardinality $d_0 = |\mathcal{A}|$.

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. The corresponding sample partial correlation between y and x_j given $x_{\mathcal{S}}$ is denoted by $\hat{\rho}(y, x_j|x_{\mathcal{S}})$.*

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

$$\sqrt{n}\,\{\hat{\rho}(y, x_j|x_{\mathcal{S}})\} - \rho(y, x_j|x_{\mathcal{S}})\} \to N\Big(0, (1+\kappa)\{1-\rho^2(y, x_j|x_{\mathcal{S}})\}^2\Big). \tag{2.3}$$

To our best knowledge, this result is new; and its proof is given in the supplemental materials. Let $\emptyset$ be the empty set, and $\hat{\rho}(y, x_j|x_\emptyset)$ and $\rho(y, x_j|x_\emptyset)$ be $\hat{\rho}(y, x_j)$ 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.

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:

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

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

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

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 $\hat{\rho}(y, x_j|x_\mathcal{S})$ directly.

2.2. A variable selection algorithm

Based on the partial faithfulness condition, one has for all $j \in \{1, \ldots, p\}$ (Bühlmann, Kalisch and Maathuis (2010)),

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

Thus, 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, we propose to identify active predictors by iteratively testing the series of hypotheses

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

where $\hat{m}_{reach} = \min\{m : |\hat{\mathcal{A}}^{[m]}| \leq m\}$, and $\hat{\mathcal{A}}^{[m]}$ is the chosen model index set in the mth step with cardinality $|\hat{\mathcal{A}}^{[m]}|$. Based on (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 $\hat{\kappa}$ a consistent estimate of κ, and $\Phi^{-1}(\cdot)$ the inverse of the cumulative distribution of the standard normal. 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 the calculation of residuals. Therefore, an equivalent form of the rejection region with small sample correction is

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

where

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

In this, κ is estimated by its sample counterpart

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

where $\bar{x}_j$ is the sample mean of the j-the element of $\mathbf{x}$ and x_{ij} is the j-th element of $\mathbf{x}_i$. The sample partial correlations can be computed recursively: For any $k \in \mathcal{S}$,

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

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
        $$\hat{\mathcal{A}}^{[1]} = \{j = 1, \cdots, p : |\hat{\rho}(y, x_j)| > T(\alpha, n, \hat{\kappa}, 0)\}.$$
Step 2: Based on $\hat{\mathcal{A}}^{[m-1]}$, construct the $m$th step estimated active set by
        $$\hat{\mathcal{A}}^{[m]} = \{j\in\hat{\mathcal{A}}^{[m-1]} : |\hat{\rho}(y, x_j|x_{\mathcal{S}})|>T(\alpha, n, \hat{\kappa}, |\mathcal{S}|), \forall\mathcal{S}\subseteq\hat{\mathcal{A}}^{[m-1]}\backslash\{j\}, |\mathcal{S}| = m-1\}.$$
Step 3: Repeat Step 2 until $m = \hat{m}_{reach}$.
```

Algorithm 1 results in the sequence of estimated active sets

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

Since $\kappa = 0$ for the normal, the TPC is the PC-simple algorithm under a normality assumption, and Theorem 1 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 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.

(D1) The joint distribution of $(\mathbf{x}^T, y)$ satisfies partial faithfulness.

(D2) $(\mathbf{x}^T, y)$ follows $\mathrm{EC}_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ with $\boldsymbol{\Sigma} > 0$, and there exists $s_0 > 0$, such that for all $0 < s < s_0$,

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

(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
$\inf \{|\rho(y, x_j|x_{\mathcal{S}})| : j = 1, \cdots, \mathrm{p}, \mathcal{S} \subseteq \{j\}^c, |\mathcal{S}| \le d_0, \rho(y, x_j|x_{\mathcal{S}}) \ne 0\} \ge c_n$.

(D5) The partial correlations $\rho(y, x_j|x_{\mathcal{S}})$ and $\rho(x_j, x_k|x_{\mathcal{S}})$ satisfy:
i). $\sup \{|\rho(y, x_j|x_{\mathcal{S}})| : 1 \le j \le p, \mathcal{S} \subseteq \{j\}^c, |\mathcal{S}| \le d_0\} \le \tau < 1$,
ii). $\sup \{|\rho(x_j, x_k|x_{\mathcal{S}})| : 1 \le j \ne k \le p, \mathcal{S} \subseteq \{j, k\}^c, |\mathcal{S}| \le d_0\} \le \tau < 1$.

Condition (D1) guarantees the validity of the TPC method as a variable selection criterion. Condition (D2) is crucial when deriving the asymptotic distribution of the sample partial correlation, and the sub-exponential tail probability ensures that the difference between the population and sample partial correlations degenerates at an exponential rate. Many elliptical distributions satisfy the sub-exponential tail probability, such as multivariate normal and Pearson Type II distributions (Fang, Kotz and Ng (1990)). Although (D2) is widely used as a sufficient condition to facilitate the proof, it may not be the weakest condition guaranteeing 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 on 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 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 that excludes perfect collinearity between the covariates.

Since the TPC depends on the significance level $\alpha = \alpha_n$, we write the final chosen model as $\hat{\mathcal{A}}_n(\alpha_n)$.

Theorem 2. *Consider the linear model in* (2.1)*. Under* ($D1$)-($D5$)*, there exists a sequence* $\alpha_n \to 0$ *and a positive constant* C*, such that if* d_0 *is fixed, then for* $p = o(\exp(n^{\xi}))$*,* $0 < \xi < 1/5$*,*

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

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

The proof is given in the supplemental materials. The result implies that TPC enjoys the model selection consistency property when dimensionality increases at an exponential rate in the sample size. Following Bühlmann, Kalisch and Maathuis (2010), a possible choice of the theoretical significance level α_n is $\alpha_n = 2\{1 - \Phi(c_n\sqrt{n/(1+\kappa)}/2)\}$.

Bühlmann, Kalisch and Maathuis (2010) utilized the tail probability of the normal 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.

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). We establish the sure screening property (Fan and Lv (2008)) of this first step of TPC under a different set of assumptions. We need the following conditions on the population marginal correlations:

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

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

Theorem 3. *Consider the linear model in* (2.1) *and assume that* ($D1$)-($D3$)*,* ($E4$) *and* ($E5$) *hold. For* $p = O(\exp(n^{\xi}))$*, where* $0 < \xi < 1$*, there exists a sequence* $\alpha_n \to 0$ *such that* $P\{\hat{\mathcal{A}}_n^{[1]} \supseteq \mathcal{A}\} \geq 1 - O\{\exp(-n^{\nu}/C^*)\}$*, where* C^* *is a positive constant and* $\xi < \nu < 1/5$*.*

The proof is given in the supplemental materials.

3. Numerical Studies

3.1. Simulation studies

In our simulation study, data were generated according to (2.1) with $\beta_1 = 3$, $\beta_2 = 1.5$, $\beta_5 = 2$, and $\beta_j = 0$ if $j \neq 1, 2, 5$. We took $p = 200, 500$, and $2,000$; and sample size $n = 200$. The joint distribution of $\mathbf{x}$ and ϵ was taken to be $0.9N(0, \Sigma) + 0.1N(0, 9\Sigma)$, where Σ is the $(p+1) \times (p+1)$ matrix with (i, j)th entry $\rho^{|i-j|}$. We took $\rho = 0$, 0.3, and 0.8 to correspond to uncorrelated, moderately correlated, and strongly correlated. The estimated kurtosis of the mixture normal is around 1.5, indicating a heavy-tailed situation. For each case, we conducted 1,000 simulations.

In our simulation, we compared the finite sample performance of LASSO (Tibshirani (1996)) and SCAD The following criteria were used to evaluate the performance of variable selection procedures.

1. Model error: $E_{\mathbf{x}}[\{\mathbf{x}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}^2] = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathrm{cov}(\mathbf{x})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.
2. True positive number (TPN), the average number of predictors with nonzero coefficients successfully detected in 1,000 simulation.
3. False positive number (FPN), the average number of predictors with zero coefficients being erroneously selected into the model.
4. Underfit percentage (UF), the percentage of models that fail to identify at least one important predictor in the 1,000 simulations.
5. Correct-fit percentage (CF), the percentage of models that exactly select the truly important predictors in the 1,000 simulations.
6. Overfit percentage (OF), the percentage of models that identify all the important predictors, but include at least one unimportant predictor in the 1,000 simulations.

Table 1 depicts the simulation results for the elliptical distribution. It 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. Thus, LASSO constantly over-fits the model under every scenario. The models selected by SCAD are more conservative than those selected by the PC-based methods. Since the PC-simple relies on normality, it fails to capture the correct model with a high percentage; especially, when the x-variables are independent, the correct-fit rates of the PC-simple are only 25% and 7% for $p = 500$

Table 1. Simulation results for Example 1: Elliptical distribution.

p	ρ	Method	MedME(Devi)	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
200	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
200	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
500	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
500	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
2,000	0	SCAD	0.051 (0.032)	3.00	10.13	0.00	0.40	0.60
		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
2,000	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
		TPC	0.039 (0.026)	2.97	0.18	0.03	0.83	0.14
2,000	0.8	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
		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 1,000 simulations.

and $2,000$, respectively. Thus, when $p = 500$, the over-fit rates (OF) of PC-simple are 0.58, 0.65, and 0.22 for $\rho = 0$, 0.3, 0.8, respectively, the OF of TPC are 0.04, 0.07, and 0.00. The times for 1,000 simulations with $p = 2,000$ are reported in Table S.1 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 in Table S.2 in the supplemental material. 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 to over-fit, compared with the partial-correlation-based methods for variable selection.

The model selection consistency of the TPC does not require the elliptical distribution for the response and the predictors. To illustrate this, we consider a simulation example in which discrete predictors are involved. Specifically, the x's with even-subscript are generated in the same fashion as before, where the x's with odd-subscript take discrete values $0, 1, 2$ with probabilities $0.25, 0.5$ and 0.25, respectively. The results are in Table S.3 in the supplemental material. From Table S.3 that the TPC outperforms other methods, especially in terms of the correct-fit rate.

3.2. An application

We demonstrate the proposed methodology by an empirical analysis of the microarray data set that was studied by Scheetz et al. (2006) and Huang, Ma and Zhang (2008). This dataset contains 120 12-week-old male rats, and, for each rat, 3,000 sufficiently expressed gene probes with enough variation were studied. The purpose of the analysis is to identify the probes that are most relevant to the response – the expression level of probe TRIM32, recently proved to cause Bardet-Biedl syndrome (Chiang et al. (2006)).

We applied the SCAD, LASSO, PC-simple algorithm, and TPC to this data set with one outlier deleted. Table 2 provides the information on the chosen gene probes by different methods. As LASSO yields a much larger model leading to the difficulty of interpretation, 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 calculated the adjusted R^2 for each model and prediction error (PE) by the leave-one-out cross-validation (LOOCV) method for each model. From Table 2, 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 the probes x_5 and x_6. These probes lead to about 9% predictor error reduction from the model selected by PC-simple to the model selected by TPC. The probe 1389584_at (x_1) and 1383996_at (x_2) are selected by all four approaches, and also identified by

Table 2. Results for real data example.

Selected Probes	SCAD	LASSO	PC- simple	TPC	M6 Est(& SE)	M4 Est(& SE)
Intercept	Yes	Yes	Yes	Yes	0.0147 (0.0465)	0.0164 (0.0467)
1389584_at(x_1)	Yes	Yes	Yes	Yes	0.3669 (0.0823)***	0.4098 (0.0710)***
1383996_at(x_2)	Yes	Yes	Yes	Yes	0.1400 (0.0595)*	0.1583 (0.0590)**
1382452_at(x_3)	Yes	Yes	/	/	0.2450 (0.0606)***	0.2279 (0.0547)***
1370429_at(x_4)	/	/	Yes	Yes	0.0464 (0.0815)	
1383110_at(x_5)	/	Yes	/	Yes	0.1543 (0.0840)	
1374106_at(x_6)	/	Yes	/	Yes	0.2203 (0.0727)**	0.2580 (0.0688)***
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.

Huang, Ma and Zhang (2008). The results from TPC are more consistent with Huang, Ma and Zhang (2008) than these of the other methods.

We further conducted some exploratory analysis. We compared the model with 20 probes selected by LASSO with the model with the six probes listed in Table 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 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 the 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 2. The adjusted R^2 and predictor error of this model is very close to the one with six probes. This empirical analysis implies that two comparable schemes for variable selection (i.e., regularization methods such as the SCAD and the LASSO, and partial correlation based methods such as PC-simple and TPC) can be used to improve each other. Thus, the regularization method would miss probe x_6, while the TPC would miss probe x_3. Confidence in the selected probes x_1 and x_2 could increase 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 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 Materials

Proofs, as well as the additional simulation results, are included in the online supplemental materials.

Acknowledgment

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 contributed to this paper, and the authors are listed in the alphabetic order.

References

Arellano-Valle, R. B. del Pino, F. and Iglesias, P. (2006). Bayesian inference in spherical linear models: robustness and conjugate analysis. *Journal of Multivariate Analysis* **97**, 179–197.

Bühlmann, P., Kalisch, M. and Maathuis, M. (2010). Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. *Biometrika* **97**, 261–278.

Chen, Y., Wiesel, A. and Hero, A. O. (2011). Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing* **59**, 4097–4107.

Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R., Nishimura, D., Braun, T. A., Kim, K.-Y., Huang, J., Elbedour, K., Carmi, R., Slusarski, D. C., Casavant, T. L., Stone, E. M. and Shefield, V. C. (2006). Homozygosity mapping with SNP arrays identifies a novel gene for Bardet- Biedl syndrome (BBS10). *Proceeding of the National Academy of Sciences* **103**, 6287–6292.

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

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

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101–148.

Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman and Hall, New York, NY.

Huang, J., Ma, S. G. and Zhang, C. H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* **18**, 1603–1618.

Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of American Statistical Association* **104**, 234–248.

Mcneil, A. J., Frey, R. and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools.* Princeton University Press, Princeton, NJ.

Mottonen, J., Oja, H. and Tienari, J. (1997). On the efficiency of multivariate spatial sign and rank tests. *Annals of Statistics* **25**, 542–552.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory.* Wiley, New York.

Oja, H. and Randles, R. H. (2004). Multivariate nonparametric tests. *Statistical Sciences* **19**, 598–605.

Osiewalski, J. (1991). A note on Bayesian inference in a regression model with elliptical errors. *Journal of Econometrics* **48**, 183–193.

Osiewalski, J. and Steel, M. F. J. (1993). Robust bayesian inference in elliptical regression models. *Journal of Econometrics* **57**, 345–363.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Shefield, V. C. and Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceeding of the National Academy of Sciences* **103**, 14429–14434.

Schmidt, R. (2002). Tail dependence for elliptically contoured distributions. *Mathematical Methods of Operations Research* **55**, 301–327.

Soloveychik, I. and Wiesel, A. (2015). Performance analysis of Tyler's covariance estimator. *IEEE Transactions on Signal Processing* **63**, 418–426.

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

Vidal, I. and Arellano-Valle, R. B. (2010). Bayesian inference for dependent elliptical measurement error models. *Journal of Multivariate Analysis* **101**, 2587–2597.

Wang, L., Peng, B. and Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *Journal of American Statistical Association.* Accepted.

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

Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802, USA.

E-mail: rzli@psu.edu

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.

E-mail: jingyuan@xmu.edu.cn

Ernst & Young, 5 Times Square, New York, NY 10036, USA.

E-mail: lejia.lou@ey.com

(Received December 2015; accepted July 2016)