Kent Academic Repository

Full text document (pdf)

Citation for published version

Zhang, Jian and Oftadeh, Elaheh (2016) Multivariate Variable Selection through Use of Null-Beamforming Principle Variable Analysis. TBD . (Submitted)

DOI

Link to record in KAR

https://kar.kent.ac.uk/64042/

Document Version

Updated Version

Copyright & reuse
Content in the Kent Academic Repository is made available for research purposes. Unless otherwise stated all content is protected by copyright and in the absence of an open licence (eg Creative Commons), permissions for further reuse of content should be sought from the publisher, author or other copyright holder.

Versions of research
The version in the Kent Academic Repository may differ from the final published version.
Users are advised to check **http://kar.kent.ac.uk** for the status of the paper. **Users should always cite the published version of record.**

Enquiries
For any further enquiries regarding the licence status of this document, please contact:
researchsupport@kent.ac.uk

If you believe this document infringes copyright then please contact the KAR admin team with the take-down information provided at **http://kar.kent.ac.uk/contact.html**

KAR Kent Academic Repository

University of Kent

Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

MULTIVARIATE VARIABLE SELECTION BY MEANS OF NULL-BEAMFORMING

By Jian Zhang[†] and Elaheh Oftadeh[†,*]

University of Kent[†]

This article extends the idea of principal component analysis to multivariate variable selection for multivariate regression models, where regression coefficients of multiple responses on each predictor are treated as values derived from a random variable. The proposed method, called principal variable analysis, aims at select predictor variables with relatively higher coefficient variations. The basic premise behind the proposal is to scan through a predictor variable space with a series of forward filters named null-beamformers; each is tailored to a particular region in the space and resistant to interference effects originating from other regions. The new approach attempts to explore the maximum amount of variation in the data with a small number of principal variables. Applying the proposal to simulated data and real cancer drug data, we show that it substantially outperforms the existing methods in terms of sensitivity and specificity. An asymptotic theory on selection consistency is established under some regularity conditions.

1. Introduction. Multivariate regression analysis is a mainstay of statistical methodologies used in science. It concerns predicting or explaining causal relationships of multiple related responses on a common set of predictors. A classical example for demonstrating this usage is about the analysis of Rohwer's experiment, which aimed to predict children' aptitude/achievement by the scores they received in paired-associate tests. In this analysis, as multiple response variables three correlated kids' aptitude measurements were regressed to the scores obtained in five tests. Friendly (2007) showed that such an analysis enhanced the original univariate analysis by taking the advantage of correlation structures between the response variables.

The advance of high-throughput technology in science has generated high dimensional data automatically with unprecedented pace. The high dimensional data hold great promises for detecting sparse predictors, which may

*The research of Elaheh Oftadeh is supported by the 50th anniversary PhD scholarship from the University of Kent.

MSC 2010 subject classifications: Primary 62J07, 62H25; secondary 62P10

Keywords and phrases: High dimensional and multivariate regression models, principal variable analysis, variable selection, null-beamforming

not be possible with small data. For example, in cancer research, various cancer genomic data have been generated in recent years. The research aims to understand biological processes, especially processes that relate to cancer occurrence, and to identify biomarkers (a set of genes or DNA variants) for cancer drug development. The particular question of interest in this paper is about whether and how the responses of cancer cell lines to various drug treatments can be predicted from gene activities in the cells. The data under investigation contain the measurements of median inhibition concentrations of drugs called IC50s in 586 cancer cell lines and expression levels of 13321 genes (Garnett et al., 2012). According to cancer encyclopedia, IC50 is a concentration of drug that reduces a biochemical activity such as cell multiplication to 50 percent of its normal value in the absence of the inhibitor. However, when the number of predictor variables is nearly as large as, or larger than the number of observations, the ordinary least squares criterion will not provide a satisfactory solution to the question. To cope with the difficulty, we make a sparsity assumption on the model that there are only a small number of true predictors useful for predicting response variables among many candidates. Under this assumption, a remedy for the shortcomings of least squares is to modify the sum of squared errors criterion by using penalties based on the magnitudes of regression coefficients. When the penalty is increasing, estimates are zeroed out, and a subset model is then identified and estimated. Such a remedy is particularly of interest when the dimension p is large and candidate predictors are thought to contain many redundant or irrelevant variables (George, 2000). The variable selection procedure LASSO (Tibshirani, 1996) followed this remedy. Over the past two decades, much progress has been made along this direction (Fan and Li, 2001; Zou and Hastie, 2005; among others). Although the recent research on variable selection mainly focuses on a univariate response setting, a few multivariate methods have been developed (e.g., Peng et al.,2009; Rothman et al., 2010; Chen and Huang, 2012; Sofer et al., 2014; Li et al., 2015).

Despite of the above progress, a few issues remain to be addressed. First, most of these methods have been developed for independent samples. There are various applications which have dependent samples. For instance, the drug sensitivity values, IC50s, of cell lines can be dependent as cell lines exhibit genetic relatedness when they are associated with the same types of cancers (Garnett et.al, 2012). In multiple genome-wide association studies, individual genotypes in a subject group are correlated (Zhou and Stephens, 2014). In neuroimaging, measurements from different sensors outside a brain are dependent as they are generated from the same neuronal sources inside the brain (Van Veen et al., 1997). In finance, returns of different stocks

are correlated due to the so-called cross-sectional dependence (Froot, 1989). Secondly, the existing methods mentioned above are mainly for estimating a multivariate fixed effect regression model, where given its design matrix, the response covariance structure is determined only by error terms. However, in a multivariate random effect regression model, given its design matrix, the response covariance depends not only on error terms but also on random coefficients. Finally, most of the existing methods are not computationally scalable to the analysis of large-scale data. This prohibits their applications to big data.

Here, we address these issues by generalizing the idea of principal component analysis to multivariate variable selection as follows. First, we develop a novel method called principal variable analysis (PVA) to identify major predictors that account for the maximum amount of variation in the data. In the PVA, unlike the existing methods, we gauge the contribution of each predictor to multiple responses by a variation index of the corresponding regression coefficients. We select variables of relatively high variation index. Such a treatment provides a principled way of combining information across multiple responses. Let $\mathbf{y}_j$ and $\mathbf{x}_k$ be column vectors containing observations on the jth response variable and the kth predictor respectively. In the PVA, we estimate the variation index called predictive power for the kth predictor by null-beamforming, i.e., minimizing the sample variance of the projected data points $\mathbf{w}^T\mathbf{y}_j, 1 \leq j \leq J$ with respect to the weighting vector $\mathbf{w}$, subject to the constraint $\mathbf{w}^T\mathbf{x}_k = 1$ and to the nulling of significant predictors identified in the previous steps. The null-beamforming takes advantage of sample dependence by means of the data projection (i.e., the sample linear combination) and reduce interferences with other predictors by using the above variance minimization. The predictive power measures the amount of information in a predictor for predicting multiple responses. The higher the power, the more information about the responses the predictor contains. Note that the projected data at each predictor may have varying background noises. To adjust for this, we consider the signal-to-noise ratio (SNR) for each predictor. The SNR values create a SNR map for the predictors. The predictors can then be ranked and selected by thresholding the map. This produces a list of highly ranked predictors called principal variables along with their estimated regression coefficients. Based on these selected predictors, a decomposition of the response covariance matrix can be made. In this sense, the PVA is viewed as a generalized principal component analysis. As the PVA can be implemented through parallel computing, it is scalable to large-scale data. Secondly, following Fan and Lv (2008) and Wang (2009), we establish an asymptotic theory underpinning the PVA, where a selection

sparsistency property holds. Finally, we conduct a wide range of simulation studies to evaluate the performance of the PVA compared to the existing variable selection methods. The results demonstrate that in terms of sensitivity and specificity the PVA can substantially outperform the existing methods such as the multivariate group LASSO, the multivariate elastic-net, the multivariate LASSO, the multivariate sparse group LASSO and among others. We also apply our method to the cancer data mentioned above, identifying a novel gene network for predicting median inhibition concentrations of drugs in cancer cell lines. Using the information extracted from the Human Protein Atlas Portal at http://www.proteinatlas.org/cancer, we show that most of the identified genes have significantly high protein staining levels at least in one or more than one of common cancers.

The remaining of the paper is organized as follows. The details of the proposed methodology and algorithm are provided in Section 2. An asymptotic theory on the proposed procedure is developed in Section 3. The simulation studies and a real data application are presented in Section 4. The discussion and conclusion are made in Section 5. The technical details and proofs can be found in the Online Supplementary Material. Throughout the paper, we denote by $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ the largest and smallest eigenvalues of a square matrix respectively. For any matrix $\mathbf{F}_n$, we define the spectral norm $||\mathbf{F}_n||$ as $\lambda_{\max}^{1/2}(\mathbf{F}_n^T\mathbf{F}_n)$. For a sequence of real numbers $\{u_n\}$, we say $\mathbf{F}_n = O(u_n)$ if $||\mathbf{F}_n||/|u_n|$ is bounded from above and $\mathbf{F}_n = o(u_n)$ if $||\mathbf{F}_n||/|u_n|$ tends to zero as n tends to infinity.

2. Methodology. Suppose that we have observations on the (centralized) responses $y_1, y_2, ..., y_J$ and on the same set of (centralized) predictors with design values $x_1, x_2, ..., x_p$, where each response is linked with the predictors through a regression model:

$$y_j = \beta_{1j}x_1 + \cdots + \beta_{pj}x_p + \varepsilon_j, 1 \le j \le J, \tag{2.1}$$

where β_{kj} is the (k, j)th unknown (random) regression coefficient and the error term ε_j has zero mean and unknown variance $0 < \sigma_{\varepsilon_j}^2 < \infty$. We let $\mathbf{Y} = \mathbf{Y}_{n\times J} = (y_{ij})_{n\times J} = (\mathbf{y}_1\mathbf{y}_2\cdots\mathbf{y}_J)$ and $\mathbf{X} = \mathbf{X}_{n\times p} = (x_{ik})_{n\times p} = (\mathbf{x}_1\cdots\mathbf{x}_p)$, where $\mathbf{y}_j$ and $\mathbf{x}_k$ are the column vectors of n observations on the jth response variable and the kth predictor respectively. Given observations $(\mathbf{Y}, \mathbf{X})$, we want to identify these predictors of significant regression coefficients. For this purpose, we reformulate the model (2.1) in the following matrix form:

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon}, \tag{2.2}$$

where $\mathbf{B} = \mathbf{B}_{p\times J} = (\boldsymbol{\beta}_1\boldsymbol{\beta}_2\cdots\boldsymbol{\beta}_J)$ and $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_{n\times J} = (\boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_2\cdots\boldsymbol{\varepsilon}_J)$ with $\boldsymbol{\beta}_j$ and $\boldsymbol{\varepsilon}_j$, respectively, containing the regression coefficients and the error terms related to the jth response variable. Note that $\mathbf{B}$ is random and $\mathbf{X}$ is fixed. It can be shown that when $p < n$, the least square solution, $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ gives the same coefficients as fitting univariate multiple regression models to $(\mathbf{y}_j, \mathbf{X}), 1 \le j \le J$ separately. Note that when we treat $\beta_{kj}, 1 \le j \le J$ as correlated random coefficients, the response variables y_j's will be dependent on each other. On the other hand, the least square principle has been designed for independent samples and not for dependent samples. Therefore, the least square solution may not be efficient when observations on the response variables y_j's are dependent. To tackle the problem, we define a predictive power for each predictor based on the projected response data below. The predictive power takes advantage of correlation structures among the response variables as well as the sample dependence. Using the predictive power, we are able to rank and select predictors.

2.1. *Predictive power and SNR.* The concept of predictive power, defined as the variance of a signal, is borrowed from the research field of signal processing, where sensor observations $\mathbf{y}_j, 1 \le j \le J$ are often assumed weakly stationary with covariance matrix $\mathbf{C} = \mathrm{cov}(\mathbf{y}_j)$ (van Veen et al., 1997). In genetics, the above concept describes the pleiotropic genetic effect of a single gene on multiple phenotypic traits, where multivariate linear models have been developed to connect genetic variant data to multiple quantitative traits (Chiu et al., 2017). In the multivariate regression setting, we view regression coefficients of multivariate response on a predictor as values generated from a random variable. Similar to principal component analysis, the importance of a predictor variable is measured by variations in its regression coefficients. The larger the variability of these regression coefficients $(\beta_{kj})_{1\le j\le J}$ at the kth predictor, the higher degree of uncertainty in response variables is accounted for by the kth predictor. In practice, however the regression coefficients $(\beta_{kj})_{1\le j\le J}$ (therefore its power index $\sum_{j=1}^{J}(\beta_{kj} - \bar{\beta}_k)^2/J$ with $\bar{\beta}_k = \sum_{j=1}^{J}\beta_{kj}/J$) are unknown. We estimate $(\beta_{kj})_{1\le j\le J}$ by projecting response data into the coefficient space of the kth predictor along the direction $\mathbf{w}$ that can minimize interferences with the other predictors and with the background noise. That is, for the kth predictor variable, we estimate its regression coefficients by the projected data $\mathbf{w}_k^T\mathbf{Y}$ along the direction $\mathbf{w}_k = \mathbf{C}^{-1}\mathbf{x}_k/\mathbf{x}_k^T\mathbf{C}^{-1}\mathbf{x}_k$, in which $\mathrm{var}(\mathbf{w}^T\mathbf{y}_j)$ attains the minimum, subject to $\mathbf{w}^T\mathbf{x}_k = 1$ (Zhang and Liu, 2015). This gives an estimator $\mathbf{w}_k^T\mathbf{Y} = \mathbf{x}_k^T\mathbf{C}^{-1}\mathbf{Y}/\mathbf{x}_k^T\mathbf{C}^{-1}\mathbf{x}_k$ for $(\beta_{kj})_{1\le j\le J}$. If let $\mathbf{C} = \text{constant} \times I_n$, where I_n is an $n \times n$ identity matrix (i.e., ig-

noring correlations in the sample and the regression coefficients), then the above estimator reduces to a marginal multivariate least square estimator. In the so-called beamforming, we let $\mathbf{C}$ be adaptive to the data, namely $\mathbf{C} = \text{cov}(\mathbf{y}_j)$, in order to explore correlation structures in the sample and regression coefficients. To explain why the above beamforming approach can provide an interference-minimized estimator of the underlying power, we assume that error term $\boldsymbol{\varepsilon}_j$ is independent of the p-dimensional regression coefficient $\boldsymbol{\beta}_j$ and with $\text{cov}(\boldsymbol{\varepsilon}_j) = \Lambda$ and $\text{cov}(\boldsymbol{\beta}_j) = \Sigma = (\sigma_{i_1 j_1})_{p\times p}$. Then, we have $\mathbf{C} = \mathbf{X}\Sigma\mathbf{X}^T + \Lambda$. Note that under the constraint $\mathbf{w}^T\mathbf{x}_k = 1$, we have

$$\begin{aligned}\mathbf{w}^T\mathbf{C}\mathbf{w} &= \text{var}(\mathbf{w}^T\mathbf{y}_j) = \sigma_{kk} + \left(\sum_{i_1\neq k, j_1\neq k} \sigma_{i_1 j_1}\mathbf{w}^T\mathbf{x}_{i_1}\mathbf{x}_{j_1}^T\mathbf{w} + \mathbf{w}^T\Lambda\mathbf{w}\right)\\ &\hat{=} \text{ power of the } k\text{th predictor} + \mathbf{w}\text{-dependent interference},\end{aligned}$$

which yields

$$\begin{aligned}\min\{\mathbf{w}^T\mathbf{C}\mathbf{w} : \mathbf{w}^T\mathbf{x}_k = 1\} &= \text{power of the } k\text{th predictor}\\ &\quad + \min\{\mathbf{w}\text{-dependent interference} : \mathbf{w}^T\mathbf{x}_k = 1\}.\end{aligned}$$

This implies that the constraint $\mathbf{w}^T\mathbf{x}_k = 1$ is a linear filter which allows the power σ_{kk} of the kth predictor to pass through it, whereas interferences with other predictors and with the background noise are reduced via the minimization. So, $\min\{\mathbf{w}^T\mathbf{C}\mathbf{w} : \mathbf{w}^T\mathbf{x}_k = 1\}$ is an interference-minimized estimator for the theoretical power σ_{kk}. A simple calculation shows that the above estimated power, called the predictive power of the kth predictor, can be expressed as

$$\gamma_k = \min\{\text{var}(\mathbf{w}^T\mathbf{y}_j) : \mathbf{w}^T\mathbf{x}_k = 1\} = (\mathbf{x}_k^T\mathbf{C}^{-1}\mathbf{x}_k)^{-1}.$$

When observations on response variables are white noises of noise level σ^2, the predictive power of the kth predictor reduces to $\sigma^2\mathbf{w}_k^T\mathbf{w}_k$. So we define the SNR at the kth predictor by $\gamma_k(\sigma^2\mathbf{w}_k^T\mathbf{w}_k)^{-1}$.

Analogously, for a subset of predictors $\nu = \{k_1, k_2, ..., k_m\}$, their joint predictive power (called the predictive power matrix) can be defined by $\gamma_\nu = \left(\mathbf{x}_\nu^T\mathbf{C}^{-1}\mathbf{x}_\nu\right)^{-1}$, where the data matrix $\mathbf{x}_\nu = (\mathbf{x}_{k_1}, ..., \mathbf{x}_{k_m})$ consists of the observations on the predictors in ν and the columns in $\mathbf{x}_\nu$ are assumed linearly independent. Abusing the above notation, we let $\mathbf{w}$ and $\mathbf{w}_\nu$ denote $n \times m$ matrices below. Then, we can also define the SNR of predictor set ν as $\text{SNR}_\nu = \text{tr}\left(\gamma_\nu(\sigma^2\mathbf{w}_\nu^T\mathbf{w}_\nu)^{-1}\right)$. Using the Lagrange multiplier, we can show that γ_ν is the covariance matrix of the projected data $\mathbf{w}_\nu^T\mathbf{Y}$ along interference-minimized directions $\mathbf{w}_\nu = \mathbf{C}^{-1}\mathbf{x}_\nu(\mathbf{x}_\nu^T\mathbf{C}^{-1}\mathbf{x}_\nu)^{-1}$, in the sense

that $\text{tr}(\gamma_\nu) = \min\{\text{tr}(\text{cov}(\mathbf{w}^T\mathbf{Y})) : \mathbf{w}^T\mathbf{x}_\nu = I_m\}$, where $\text{tr}(\cdot)$ is the trace operator and I_m is an $m \times m$ identity matrix. Note that $\mathbf{w}^T\mathbf{x}_\nu = I_m$ define m linear filters which null each other. The projection of $\mathbf{Y}$ along interference-minimized directions gives an estimator, $(\mathbf{x}_\nu^T\mathbf{C}^{-1}\mathbf{x}_\nu)^{-1}\mathbf{x}_\nu^T\mathbf{C}^{-1}\mathbf{Y}$, for random coefficient matrix $\mathbf{B}$. The above estimator will reduce to a marginal least square estimator if let $\mathbf{C} = \text{constant} \times I_n$. As $\mathbf{C}$ is often not diagonal, a better estimator of $\mathbf{B}$ can be obtained by estimating $\mathbf{C}$ from the data.

Predictors can be correlated. For example, in the cancer genomic data, genes as predictors can be highly correlated if they are located in the same pathway. Consequently, the predictive power of a predictor can be biased by interferences with other predictors. To address this problem, we null the previously identified predictors by adding more constraints on the linear filter in each step. Let ω and ν be two non-overlapped subsets of the predictors with sizes m_1 and m respectively. To define a ω-nulled predictive power matrix of ν, adding null constraints $\mathbf{w}^T\mathbf{x}_\omega = 0_{m\times m_1}$ into the linear filters $\mathbf{w}^T\mathbf{x}_\nu = I_m$, we consider the following optimization problem:

$$\min \text{tr}(\mathbf{w}^T\mathbf{C}\mathbf{w}), \text{ subject to } \mathbf{w}^T\mathbf{x}_\nu = I_m, \quad \mathbf{w}^T\mathbf{x}_\omega = \mathbf{0}_{m\times m_1}.$$

Using the Lagrange multiplier again, we obtain the optimal weighting matrix

$$\mathbf{w}_{\nu|\omega} = \mathbf{C}^{-1}\mathbf{x}_{\nu\cup\omega}\left(\mathbf{x}_{\nu\cup\omega}^T\mathbf{C}^{-1}\mathbf{x}_{\nu\cup\omega}\right)^{-1}\phi_{\nu|\omega},$$

where $\phi_{\nu|\omega} = (I_m, \mathbf{0})^T$ with $\mathbf{0}$ being the $m \times m_1$ matrix of 0's. The nulled predictive power matrix $\gamma(\nu|\omega)$ is then defined as $\mathbf{w}_{\nu|\omega}^T\mathbf{C}\mathbf{w}_{\nu|\omega}$, the covariance matrix of the projected data along $\mathbf{w}_{\nu|\omega}$. The nulled SNR $\text{SNR}_{\nu|\omega}$ is defined as $\text{tr}\left(\gamma_{\nu|\omega}(\sigma^2\mathbf{w}_{\nu|\omega}^T\mathbf{w}_{\nu|\omega})^{-1}\right)$. It can be shown that $\gamma_{\nu|\omega}$ is equal to the upper corner $m \times m$ block matrix of $\left(\mathbf{x}_{\nu\cup\omega}^T\mathbf{C}^{-1}\mathbf{x}_{\nu\cup\omega}\right)^{-1}$.

2.2. *Estimation of response covariance matrix.* Note that the concept of predictive power used above is based on an estimator of the response covariance matrix, for example, the sample covariance matrix $\hat{\mathbf{C}} = \sum_{j=1}^J \mathbf{y}_j\mathbf{y}_j^T/J - \bar{\mathbf{y}}\bar{\mathbf{y}}^T = (\hat{c}_{ij})$, where $\bar{\mathbf{y}} = \sum_{j=1}^J \mathbf{y}_j/J = (\bar{y}_1, ..., \bar{y}_n)^T$ and $\hat{c}_{ij} = \sum_{t=1}^J (y_{it} - \bar{y}_i)(y_{jt} - \bar{y}_j)/J$. As the sample covariance matrix is inconsistent with the true one when the dimension n is larger than J, Bickel and Levina (2008) amended it by thresholding its entries: $\hat{\mathbf{C}}_h = \hat{\mathbf{C}}(\tau_{nJ}) = (\hat{c}_{ij}I(|\hat{c}_{ij}| > h\tau_{nJ}))$, where $I(\cdot)$ is the indicator and $\tau_{nJ} = \sqrt{\log(n)/J}$ with the tuning constant $h \geq 0$. Under certain mixing conditions, Zhang and Liu (2015) showed that the thresholded sample covariance matrix was consistent with the true one. For a finite sample, the thresholded covariance matrix may still be degenerate when the dimension J is close to or smaller than the sample size n.

So, following Ledoit and Wolf (2004), we further shrink the thresholded covariance estimator to a diagonal matrix as follows:

$$\hat{\mathbf{C}}_{hs} = \frac{b_n^2}{d_n^2}\hat{\mu}_n I_n + \frac{d_n^2 - b_n^2}{d_n^2}\hat{\mathbf{C}}_h,$$

where

$$\begin{aligned}
\bar{b}_n^2 &= \frac{1}{J^2}\sum_{k=1}^{J}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}((y_{ik}-\bar{y}_i)(y_{kj}-\bar{y}_j)-\hat{c}_{ij})^2 I(|\hat{c}_{ij}|>h\tau_{nJ}),\\
\hat{\mu}_n &= <\hat{\mathbf{C}}_h, I_n>, \quad d_n^2 =<\hat{\mathbf{C}}_h-\hat{\mu}_n I_n, \hat{\mathbf{C}}_h-\hat{\mu}_n I_n>, \quad b_n^2=\min\{\bar{b}_n^2, d_n^2\},
\end{aligned}$$

and $<\mathbf{D}_1,\mathbf{D}_2>=\mathrm{tr}(\mathbf{D}_1\mathbf{D}_2^T)/n$ for any $n\times n$ matrices $\mathbf{D}_1$ and $\mathbf{D}_2$. Having defined $\hat{\mathbf{C}}_{hs}$, we estimate the power matrices γ_ν and $\gamma_{\nu|\omega}$ by $\hat{\gamma}_\nu = (\mathbf{x}_\nu\hat{\mathbf{C}}_{hs}\mathbf{x}_\nu)^{-1}$ and $\hat{\gamma}_{\nu|\omega} = \phi_{\nu|\omega}^T(\mathbf{x}_{\nu\cup\omega}^T\hat{\mathbf{C}}_{hs}\mathbf{x}_{\nu\cup\omega})^{-1}\phi_{\nu|\omega}$ respectively. Similarly, the ω-nulled SNR can be estimated by

$$\hat{\mathrm{SNR}}_{\nu|\omega} \propto \mathrm{tr}\left(\hat{\gamma}_{\nu|\omega}(\hat{\mathbf{w}}_{\nu|\omega}^T\hat{\mathbf{w}}_{\nu|\omega})^{-1}\right), \tag{2.3}$$

where $\hat{\mathbf{w}}_{\nu|\omega} = \hat{\mathbf{C}}_{hs}^{-1}\mathbf{x}_{\nu\cup\omega}\left(\mathbf{x}_{\nu\cup\omega}^T\hat{\mathbf{C}}_{hs}^{-1}\mathbf{x}_{\nu\cup\omega}\right)^{-1}\phi_{\nu|\omega}$.

2.3. *Principal variable analysis.* We are now ready to describe the PVA for multivariate variable selection. Although we focus on the SNR-based PVA below, the power-based PVA can also be defined similarly.

Initialization: To start with, find $1\le k_1\le p$ at which the SNR attains the maximum. Set $\omega_0=\emptyset$ and $\omega_1=\{k_1\}$.

Forward nulling: In the iteration m, $m\ge 2$, let ω_{m-1} denote the set of predictors selected in the first $m-1$ iterations. For any predictor k not in ω_{m-1}, using the formula (2.3), we calculate the nulled SNR, $\hat{\mathrm{SNR}}_{\{k\}|\omega_{m-1}}$, as well as an estimated optimal projection direction $\hat{\mathbf{w}}$. We then find $k_m\notin\omega_{m-1}$ in which $\hat{\mathrm{SNR}}_{\{k\}|\omega_{m-1}}$ attains the maximum.

Updating and stopping criteria: After a number of iterations, the nulled SNR values will start leveling off, which indicates that the remaining predictors have no predictive power for the response. This motivates us to set the following stopping criteria in the mth iteration: Make a scree plot of the nulled SNR values and identify an elbow point. The elbow point partitions the remaining predictors into two subsets, namely upper set and lower set. The lower set, containing those predictors with SNR values lower than the elbow point, is uninformative about the responses. To test the hypothesis that the upper set is uninformative, we calculate the mean μ_l

and standard deviation θ_l for the lower subset. The hypothesis is accepted if the maximum nulled SNR value, $\hat{\text{SNR}}_{\max} = \max\{\hat{\text{SNR}}_{k|\omega_{m-1}} : k \notin \omega_{m-1}\}$, of the upper set falls into the following confidence interval, $|\hat{\text{SNR}}_{\max} - \mu_l| \le c_0\theta_l$, where c_0 is a tuning constant. We set the default value $c_0 = 5$. Applying the central limit theorem to the SNR values in the lower set, the above interval can be shown to have the asymptotic confidence level of $1 - 5.73 \times 10^{-7}$ after multiple testing adjustment. The iteration will be terminated when the upper subset is uninformative. Otherwise, we update ω_{m-1} and $\mathbf{x}_{\omega_{m-1}}$ by letting $\omega_m = \{k_m\} \cup \omega_{m-1}$ and $\mathbf{x}_{\omega_m} = (\mathbf{x}_{k_m}, \mathbf{x}_{\omega_{m-1}})$, and the iteration will continue. Note that our simulations (not shown here) did indicate that the performance of PVA was not very sensitive to the choice of c_0 when it took values between 3 and 5.

2.4. *Predictive network.* Statistical connectivity patterns in the selected predictors are a hallmark feature for connecting pleiotropic traits such as drug inhibitory concentrations to genetic variants in genetics and for studying functional networks in neuroscience (Chiu et al., 2017; Park and Kriston, 2013). Here, to quantify such patterns, we compute the regression coefficient-based Pearson correlation coefficient for each pair of the selected predictors. The details are as follows. Suppose that q predictors are selected by the PVA. Based on the multivariate least squares, we obtain $\hat{\mathbf{B}}_0$, an estimator of the $q \times J$ regression coefficient matrix for these predictors. For any pair of rows (i, j) in $\hat{\mathbf{B}}_0$, we calculate Fisher's z-transformation of their correlation coefficient r_{ij} $z_{ij} = 0.5 \ln\left((1 - r_{ij})/(1 + r_{ij})\right)$. For rows $i < j$, we want to test whether z_{ij} (i.e., r_{ij}) is significantly away from 0. There are $q(q-1)/2$ such tests in total. Note that if the underlying correlation coefficient is zero, then $z_{ij} \approx N(0, 1/(J-3))$ in distribution. After Bonferroni correction to multiple testing, we can claim that z_{ij} is significantly away from zero if $\sqrt{J-3}|z_{ij}| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the critical value of $N(0,1)$ at the level $\alpha/2 = 0.01/q(q-1)$. For example, for our cancer data in Section 4 below, we obtained $q = 37, J = 131$ and therefore $z_{\alpha/2} = 4.33$. We are now ready to construct a predictive network with q nodes, each stands for a selected predictor (a row in $\hat{\mathbf{B}}_0$). We assign an edge to link nodes i and j if z_{ij} is significantly away from zero.

3. Theory. In this section, we develop an asymptotic theory for explaining why the PVA can separate what are called active predictors from non-active ones. Its proofs are deferred to the Appendix B, the Online Supplementary Material. Here, a predictor is said to be active if it has a positive power. As before, assume that $\mathbf{B}$ and $\boldsymbol{\varepsilon}$ in the model (2.2) are independent and that the covariance matrices of $\mathbf{y}_j$, $\boldsymbol{\beta}_j$ and $\boldsymbol{\varepsilon}_j$, denoted by $\mathbf{C} = (c_{ij})_{n\times n}$,

$\Sigma = (\sigma_{ij})_{p\times p}$ and Λ respectively, are independent of index j. Then, we have $\mathbf{C} = \mathbf{X}\Sigma\mathbf{X}^T + \Lambda$. Assume that Λ is positively definite. For ease of presentation, we consider the special case, where $\Lambda = \sigma^2 I_n$ and $\mathbf{x}_k^T\mathbf{x}_k = n, 1 \le k \le n$. If $\Lambda \neq \sigma^2 I_n$, we can change $\mathbf{Y}$ and $\mathbf{X}$ by the transformations $\Lambda^{-1/2}\mathbf{Y}$ and $\Lambda^{-1/2}\mathbf{X}$ (under which the predictive power is invariant), followed by rescaling $\Lambda^{-1/2}\mathbf{X}$ and $\mathbf{B}$ (see Zhang and Liu, 2015). Then a general theory can be derived from the special case. We denote the full predictor set by $[1:p] = \{1, 2, \cdots, p\}$ corresponding to $\mathbf{x}_1, \cdots, \mathbf{x}_p$, and the true predictor set by ν_0. Let $\nu = \{k_1, \cdots, k_{p_1}\}$ denote any subset of $[1:p]$ with size $|\nu|$. The $(k_1, \cdots, k_{p_1})$th columns of $\mathbf{X}$ forms a data matrix $\mathbf{x}_\nu$ for the predictor set ν. If let $\mathbf{e}_\nu$ be a $p \times p_1$ selection matrix in which for $1 \le j \le p_1$, its (k_j, j)th entry takes value of 1 and the other entries take values of 0, then we can write $\mathbf{x}_\nu = \mathbf{X}\mathbf{e}_\nu$. Let σ_k^2 denote σ_{kk} in Σ, which shows the underlying power at the kth predictor. We begin with an ideal setting where $\mathbf{C}$ is known. This includes the case of $J = \infty$ in which we can estimate $\mathbf{C}$ exactly.

3.1. *PVA with known* $\mathbf{C}$. To establish lower bounds for the SNRs, we need the condition below.

(C0). There exists a permutation on $\mathbf{y}_j, 1 \le j \le J$ so that the resulted sequence is strictly stationary with marginal covariance matrix $\mathbf{C}$ and that $(\mathbf{y}_j, \mathbf{X})$ follows model (2.2). The error term $\boldsymbol{\varepsilon}_j$ and the p-dimensional regression coefficient $\boldsymbol{\beta}_j$ are independent of each other.

Note that under Condition $(C0)$, $\mathbf{y}_j$'s (therefore $\boldsymbol{\varepsilon}_j$'s) can be mutually dependent on each other.

PROPOSITION 3.1. *Under Condition (C0),* $\text{SNR}_\nu \ge 1$ *holds for any* $\nu \subseteq [1:p]$ *of the size* $|\nu| \le n$ *and* $\text{SNR}_{\nu|\omega} \ge 1$ *holds for any* $\nu, \omega \subseteq [1:p]$ *of the size* $|\nu \cup \omega| \le n$. *The lower bound is attained when all predictors in* $[1:p]$ *are not active.*

The above proposition shows that the SNR-based map has a sharp lower bound of 1 when $\Lambda = \sigma^2 I_n$. However, when $\Lambda \neq \sigma^2 I_n$, we apply the proposition to $(\Lambda^{-1/2}\mathbf{Y}, \Lambda^{-1/2}\mathbf{X})$ to obtain a Λ-dependent lower bound for the SNR values.

To investigate the asymptotic properties of the power-based and the SNR-based maps, let $A_{\nu_0} = \mathbf{C} - \mathbf{x}_{\nu_0}\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\mathbf{x}_{\nu_0}^T$, the remainder of $\mathbf{C}$ after subtracting the term $\mathbf{x}_{\nu_0}\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\mathbf{x}_{\nu_0}^T$. In the next proposition, we shows that the power at ν_0 can be written as the underlying power plus the interferences with the predictors not in ν_0 and with the white noise. These interferences can be negligible if predictors outside ν_0 have zero powers.

PROPOSITION 3.2. *If both* $\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}$ *and* A_{ν_0} *are invertible and* $\mathbf{x}_{\nu_0}$ *has a full column rank, then the predictive power matrix*

$$\gamma_{\nu_0} = \mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0} + \left(\mathbf{x}_{\nu_0}^T\mathbf{A}_{\nu_0}^{-1}\mathbf{x}_{\nu_0}\right)^{-1}.$$

If $\sigma_k^2 = 0$, $k \notin \nu_0$ *and* $\lambda_{min}(\mathbf{x}_{\nu_0}^T\mathbf{x}_{\nu_0}/n)$ *is bounded below from zero as the sample size* n *tends to infinity, then* $\gamma_{\nu_0} = \mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0} + O(n^{-1})$.

The above proposition shows a local consistency of the predictive power with the underlying power $\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}$ at ν_0. To establish a global consistency of the whole power map, we need a few more notations below. For any subsets ν and ν_0 of predictors, if $\mathbf{A}_{\nu_0}$ is invertible, then we define the coherence (i.e., collinearity) matrices between $\mathbf{x}_\nu$ and $\mathbf{x}_{\nu_0}$:

$$\mathbf{R}_{\nu\nu} = \mathbf{x}_\nu^T A_{\nu_0}^{-1}\mathbf{x}_\nu/n, \quad \mathbf{R}_{\nu\nu_0} = \mathbf{x}_\nu^T A_{\nu_0}^{-1}\mathbf{x}_{\nu_0}/n, \quad \mathbf{R}_{\nu_0\nu_0} = \mathbf{x}_{\nu_0}^T A_{\nu_0}^{-1}\mathbf{x}_{\nu_0}/n.$$

Suppose that for $\nu_0 = \{k_1, ..., k_{p_0}\}$ and for any $\nu \subseteq \nu_0$, we can find $j_{[1:m]} = \{j_1, ..., j_m\} \subseteq \{1, ..., p_0\}$ such that $\nu = \{k_j : j \in j_{[1:m]}\}$. Let $\mathbf{e}_{\nu\lhd\nu_0}$ be a $|\nu_0| \times |\nu|$ indicator matrix with the (j_l, l)th entry equal to 1, $1 \le l \le |\nu|$ and with other entries equal to zeros. Using $\mathbf{e}_{\nu\lhd\nu_0}$, we select sub-columns from $\mathbf{x}_{\nu_0}$ to form $\mathbf{x}_\nu$, namely $\mathbf{x}_\nu = \mathbf{x}_{\nu_0}\mathbf{e}_{\nu\lhd\nu_0}$. To identify active predictors, we impose the following regularity conditions on the covariance structures of response variables and predictors, where $\mathbf{X}$ is treated as deterministic. If we treat $\mathbf{X}$ as a random design matrix, some parallel conditions can be assumed through replacing $O(\cdot)$ by $O_p(\cdot)$ in the following conditions.

(C1). There are a constant $0 < r \le 1$ and a set of active predictors ν_0 of size $|\nu_0| \le rn$ such that $\mathbf{x}_{\nu_0}$ is of full column rank and that $\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}$ and $\mathbf{A}_{\nu_0}$ are invertible.

(C2). For ν_0 and r in Condition (C1), as n tends to infinity, there is a constant $0 \le \alpha_0 < 1$ such that uniformly for any set $\nu \subseteq [1:p]$ with $|\nu| \le rn$, $\mathbf{R}_{\nu\nu} = O(n^{\alpha_0})$ and $\mathbf{R}_{\nu\nu}^{-1} = O(n^{\alpha_0})$.

(C3). For ν_0 and r in Condition (C1), as n tends to infinity, uniformly for any $\nu \subseteq [1:p] \setminus \nu_0$ with the size $|\nu| \le rn$, $(\mathbf{R}_{\nu\nu} - \mathbf{R}_{\nu\nu_0}\mathbf{R}_{\nu_0\nu_0}^{-1}\mathbf{R}_{\nu_0\nu})^{-1} = O(n^{\alpha_0})$.

(C4). For ν_0 and r in Condition (C1), as n tends to infinity, uniformly for any $\nu \subseteq [1:p] \setminus \nu_0$ with the size $|\nu| \le rn$, $\mathbf{x}_{\nu_0}^T\mathbf{A}_{\nu_0}^{-2}\mathbf{x}_\nu = \zeta_0\mathbf{x}_{\nu_0}^T\mathbf{A}_{\nu_0}^{-1}\mathbf{x}_\nu + O(1)$, where ζ_0 and $O(1)$ are independent of ν.

REMARK 3.1. *Condition (C1) says that there are no redundant predictors in* ν_0. *Condition (C2) implies that* $||\mathbf{R}_{\nu\nu_0}|| \le ||\mathbf{R}_{\nu\nu}||^{1/2}||\mathbf{R}_{\nu_0\nu_0}||^{1/2} = O(n^{\alpha_0})$. *Note that for* $\nu \subseteq [1:p] \setminus \nu_0$, $\mathbf{R}_{\nu\nu}^{-1}/n = (\mathbf{x}_\nu^T\mathbf{A}_{\nu_0}^{-1}\mathbf{x}_\nu)^{-1}$ *is the residual power of* ν *after selecting* ν_0. *So, Condition (C2) says that the residual power*

of ν is of order $n^{-1+\alpha_0} = o(1)$, which is negligible. Similarly, Condition (C3) says that ν_0-adjusted residual power of ν is also negligible.

Conditions (C1) to (C3) are the assumptions commonly used in the large sample theory for linear regression models (e.g., Wang, 2009). To verify Conditions (C1)∼(C3), we refer readers to Fan and Lv (2008) and Wang (2009) under the assumptions that $\sigma_k^2 = 0$, $k \notin \nu_0$and that $\mathbf{X}$ is assumed to be a random matrix satisfying some moment conditions and that the growth of the dimension p is not too fast compared to the sample size n. For example, following Fan and Lv (2008), we assume that $\mathbf{X}$ has a concentration property, i.e., for some constant c_1, any $u > 0$ and $\nu \subseteq [1:p]$, $|\nu| \leq rn$,

$$P\left(\lambda_{max}(\mathbf{R}_{\nu\nu}) > u \text{ or } \lambda_{min}(\mathbf{R}_{\nu\nu}) < u^{-1}\right) \leq c_1 \exp(-nu/c_1).$$

Letting $\Omega_n = \{\nu : \nu \subseteq [1:p], |\nu| \leq [rn]\}$, where $[rn]$ stands for the integer part of rn, we have

$$\begin{aligned} \max_{\nu\in\Omega_n} \lambda_{max}(\mathbf{R}_{\nu\nu}) &= \max_{\nu\in\Omega_n, |\nu|=[rn]} \lambda_{max}(\mathbf{R}_{\nu\nu}), \\ \min_{\nu\in\Omega_n} \lambda_{min}(\mathbf{R}_{\nu\nu}) &= \min_{\nu\in\Omega_n, |\nu|=[rn]} \lambda_{min}(\mathbf{R}_{\nu\nu}) \end{aligned}$$

and hence as $\log(p) \leq n^{\alpha_0}/c_1 - 1 + \log(r) + (1-1/n)\log(n)$, n and p tend to infinity,

$$\begin{aligned} &P\left(\max_{\nu\in\Omega_n} \lambda_{max}(\mathbf{R}_{\nu\nu}) > n^{\alpha_0} \text{ or } \min_{\nu\in\Omega_n} \lambda_{min}(\mathbf{R}_{\nu\nu}) < n^{-\alpha_0}\right) \\ &\leq c_1 \binom{p}{[rn]} \exp\left(-n^{1+\alpha_0}/c_1\right) \leq (pe/n)^n \exp\left(-n^{1+\alpha_0}/c_1\right) \leq c_1/n \to 0, \end{aligned}$$

This implies that Condition (C2) holds with an overwhelming probability. Analogously, Condition (C3) holds if $\mathbf{x}_\nu$ and $\mathbf{x}_{\nu_0}$ are asymptotically, uniformly noncoherent with respect to $\nu \subseteq [1:p] \setminus \nu_0$, in the sense that $\mathbf{R}_{\nu\nu_0} = o(1)$. Condition (C4) is a technical condition which holds when $\sigma_k^2 = 0$ (or sufficiently close to zero in a sense), $k \notin \nu_0$.

We now in position to state a theorem on the global sparsistency property of the power map. In the theorem, we show that for an active predictor, the predictive power has a positive limit whereas for a non-active predictor, the predictive power tends to zero. This allows us to separate active predictors from non-active ones by thresholding the power map.

THEOREM 3.1. *Suppose that there exist constants $0 \leq \alpha_1 \leq (1-3\alpha_0)/2$ and $c_2 > 0$, $c_2 n^{-\alpha_1} \leq \lambda_{min}\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right) \leq \lambda_{max}\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right) = O(1)$. Let $\Sigma_{\nu \triangleleft \nu_0} =$*

$\left(\mathbf{e}_{\nu\lhd\nu_0}^T \left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu\lhd\nu_0}\right)^{-1}$, *a partial covariance matrix of* ν *with respect to* ν_0. *Then, under Conditions (C0)∼ (C3), as* n *tends to infinity, we have:*

(i) Uniformly for any $\nu \subseteq \nu_0$ *with* $|\nu| \leq rn$,

$$\begin{aligned}\gamma_\nu &= \Sigma_{\nu\lhd\nu_0} + n^{-1}\Sigma_{\nu\lhd\nu_0}\mathbf{e}_{\nu\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{R}_{\nu_0\nu_0}^{-1}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu\lhd\nu_0}\Sigma_{\nu\lhd\nu_0} \\ &\quad +O(n^{-2+2\alpha_0+4\alpha_1}) \\ &= \Sigma_{\nu\lhd\nu_0} + O(n^{-1+\alpha_0+2\alpha_1}).\end{aligned}$$

(ii) Uniformly for any $\nu \subseteq [1:p]\setminus\nu_0$ *with* $|\nu| \leq rn$,

$$\gamma_\nu = n^{-1}\left(\mathbf{R}_{\nu\nu} - \mathbf{R}_{\nu\nu_0}\mathbf{R}_{\nu_0\nu_0}^{-1}\mathbf{R}_{\nu_0\nu}\right)^{-1} + O(n^{-2+4\alpha_0+\alpha_1}) = O(n^{-1+\alpha_0}).$$

(iii) Uniformly for any $\nu = \nu_1 \cup \nu_2$ *with* $\nu_1 \subseteq \nu_0$ *and* $\nu_2 \subseteq [1:p]\setminus\nu_0$ *and* $|\nu| \leq rn$, γ_ν *can be partitioned into*

$$\gamma_\nu = \begin{pmatrix} \gamma_\nu^{11} & \gamma_\nu^{12} \\ \gamma_\nu^{21} & \gamma_\nu^{22}\end{pmatrix}$$

with

$$\begin{aligned}\gamma_\nu^{11} &= \Sigma_{\nu_1\lhd\nu_0} + O(n^{-1+3\alpha_0+2\alpha_1}), \\ \gamma_\nu^{12} &= -n^{-1}\Sigma_{\nu_1\lhd\nu_0}\mathbf{e}_{\nu_1\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{R}_{\nu_0\nu_0}^{-1}\mathbf{R}_{\nu_0\nu_2} \\ &\quad \times\left(\mathbf{R}_{\nu_2\nu_2} - \mathbf{R}_{\nu_2\nu_0}\mathbf{R}_{\nu_0\nu_0}^{-1}\mathbf{R}_{\nu_0\nu_2}\right)^{-1} + o(n^{-1+2\alpha_0+\alpha_1}) \\ &= O(n^{-1+2\alpha_0+\alpha_1}), \\ \gamma_\nu^{21} &= -n^{-1}\left(\mathbf{R}_{\nu_2\nu_2} - \mathbf{R}_{\nu_2\nu_0}\mathbf{R}_{\nu_0\nu_0}^{-1}\mathbf{R}_{\nu_0\nu_2}\right)^{-1}\mathbf{R}_{\nu_2\nu_0}\mathbf{R}_{\nu_0\nu_0}^{-1} \\ &\quad \times\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\lhd\nu_0}\Sigma_{\nu_1\lhd\nu_0} + o(n^{-1+2\alpha_0+\alpha_1}) \\ &= O(n^{-1+2\alpha_0+\alpha_1}), \\ \gamma_\nu^{22} &= n^{-1}\left(\mathbf{R}_{\nu_2\nu_2} - \mathbf{R}_{\nu_2\nu_0}\mathbf{R}_{\nu_0\nu_0}^{-1}\mathbf{R}_{\nu_0\nu_2}\right)^{-1} + O(n^{-2+4\alpha_0+2\alpha_1}) \\ &= O(n^{-1+\alpha_0}).\end{aligned}$$

The above theorem also indicates that compared to the underlying power matrix, $\mathbf{e}_\nu^T\Sigma\mathbf{e}_\nu$, the predictive power matrix γ_ν may be not consistent if the collinearity between a pair of the predictors does not converge to zero as n tends infinity. This can be seen from the derivation of the predictive power at the predictor $k_j \in \nu_0$ below. Let $\sigma_{k_j[-k_j]}$ denote $(\sigma_{k_jk_i} : i \neq j)$, the jth row in $\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}$ excluding the jth coordinate. Let $\sigma_{[-k_j]k_j}$ denote $(\sigma_{k_ik_j} : i \neq j)$, the jth column in Σ excluding the jth coordinate. Let $\sigma_{[-k_j][-k_j]}$ denote the remaining matrix after removing the jth row and the

jth column from $\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}$. Then, the (j,j)th entry in $\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1}$ is equal to $\left(\sigma_{k_j}^2 - \sigma_{k_j[-k_j]}\sigma_{[-k_j][-k_j]}^{-1}\sigma_{[-k_j]k_j}\right)^{-1}$. The following corollary says that under Condition (C1), the predictor k_j does have positive predictive power although the power has deteriorated due to the interferences with other predictors. Therefore, if we employ the estimated predictive power to screen predictors and if $\hat{\mathbf{C}}$ is consistent with $\mathbf{C}$, then under Conditions (C0)∼(C3), the screening procedure can have a sure screening property that for an appropriately chosen threshold, all predictors in ν_0 can be detected with a probability approaching to one.

COROLLARY 3.1. *Under the conditions in Theorem 3.1, as n tends infinity, we have:*

(i) Uniformly for any $k_j \in \nu_0$, the predictive power of the predictor k_j can be expressed as

$$\gamma_{k_j} = \sigma_{k_j}^2 - \sigma_{k_j[-k_j]}\sigma_{[-k_j][-k_j]}^{-1}\sigma_{[-k_j]k_j} + O(n^{-1+\alpha_0+2\alpha_1}).$$

(ii) Uniformly for any $k \notin \nu_0$, the predictive power of the predictor k can be expressed as $\gamma_k = O(n^{-1+\alpha_0})$.

Let a be the current predictor under investigation and $\nu_1 \cup \nu_2$ be the predictors identified in the previous steps by PVA, with the size $|\nu_1 \cup \nu_2| < rn$, where $0 < r \leq 1$, $\nu_1 \subseteq \nu_0$ and $\nu_2 \subseteq [1:p] \setminus \nu_0$. For $a = k_j \in \nu_0$, let $\mathbf{e}_{a \lhd \nu_0} = \mathbf{e}_{\{a\} \lhd \nu_0}$, a $|\nu_0|$-dimensional column vector with the $j$the coordinate equal to 1 and other coordinates equal to zero. In the next theorem, we show that the global sparsistency property continues holds for the nulled predictive power and that the nulling can improve the accuracy of power estimation.

THEOREM 3.2. *Suppose that there exist constants $0 \leq \alpha_1 \leq (1-6\alpha_0)/5$ and $c_2 > 0$, $c_2 n^{-\alpha_1} \leq \lambda_{min}\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right) \leq \lambda_{max}\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right) = O(1)$. Then, under Conditions (C0)∼(C3), as n tends to infinity, we have:*

(i) Uniformly for $a \in [1:p] \setminus \nu_0$ and $a \notin \nu_1 \cup \nu_2$ with $|\nu_1 \cup \nu_2| < rn$, the $(\nu_1 \cup \nu_2)$-nulled predictive power of a admits the form $\gamma_{a|\nu_1 \cup \nu_2} = O(n^{-1+\alpha_0})$.

(ii) Uniformly for $a \in \nu_0$ and $a \notin \nu_1 \cup \nu_2$ with $|\nu_1 \cup \nu_2| < rn$, the $(\nu_1 \cup \nu_2)$-

nulled predictive power of predictor a admits the form

$$\begin{aligned}\gamma_{a|\nu_1\cup\nu_2} &= \left(\mathbf{e}_{a\lhd\nu_0}^T\Sigma_{\nu_0\setminus\nu_1}^{-1}\mathbf{e}_{a\lhd\nu_0}\right)^{-1}\\ &\quad +n^{-1}\mathbf{e}_{a\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{\Psi}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{a\lhd\nu_0}\\ &\quad +O(n^{-2+6\alpha_0+5\alpha_1}),\end{aligned}$$

where

$$\begin{aligned}\Sigma_{\nu_1\lhd\nu_0} &= \left(\mathbf{e}_{\nu_1\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\lhd\nu_0}\right)^{-1},\\ \Sigma_{\nu_0\setminus\nu_1}^{-1} &= \left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1/2}\mathbf{P}_{\nu_0\setminus\nu_1}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1/2},\\ \mathbf{P}_{\nu_0\setminus\nu_1} &= I_{|\nu_0|}-\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1/2}\mathbf{e}_{\nu_1\lhd\nu_0}\Sigma_{\nu_1\lhd\nu_0}\mathbf{e}_{\nu_1\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1/2},\\ \mathbf{F}_{\nu_2} &= \mathbf{R}_{\nu_2\nu_2}-\mathbf{R}_{\nu_2\nu_0}\mathbf{R}_{\nu_0\nu_0}^{-1}\mathbf{R}_{\nu_0\nu_2},\\ \mathbf{\Phi} &= \mathbf{R}_{\nu_0\nu_0}^{-1}+\mathbf{R}_{\nu_0\nu_0}^{-1}\mathbf{R}_{\nu_0\nu_2}\mathbf{F}_{\nu_2}^{-1}\mathbf{R}_{\nu_2\nu_0}\mathbf{R}_{\nu_0\nu_0}^{-1},\\ \mathbf{\Psi} &= \left(I_{|\nu_0|}-\mathbf{e}_{\nu_1\lhd\nu_0}\Sigma_{\nu_1\lhd\nu_0}\mathbf{e}_{\nu_1\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\right)\mathbf{\Phi}\\ &\quad \left(I_{|\nu_0|}-\mathbf{e}_{\nu_1\lhd\nu_0}\Sigma_{\nu_1\lhd\nu_0}\mathbf{e}_{\nu_1\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\right)^T.\end{aligned}$$

Here, abusing the notation, we let $\Sigma_{\nu_0\setminus\nu_1}^{-1}$ denote the generalized inverse of $\Sigma_{\nu_0\setminus\nu_1}$. Note that $\mathbf{P}_{\nu_0\setminus\nu_1}$ is a projection matrix of the ν_1-nulled precision space spanned by predictors $\nu_0\setminus\nu_1$. Therefore, $\Sigma_{\nu_0\setminus\nu_1}$ can be viewed as an ν_1-nulled projected precision matrix for $\nu_0\setminus\nu_1$ and $\mathbf{e}_{a\lhd\nu_0}^T\Sigma_{\nu_0\setminus\nu_1}^{-1}\mathbf{e}_{a\lhd\nu_0}$ can be viewed as a weighted, ν_1-nulled precision for predictor a. It can be seen that for $a\in\nu_0$,

$$\begin{aligned}\left(\mathbf{e}_{a\lhd\nu_0}^T\Sigma_{\nu_0\setminus\nu_1}^{-1}\mathbf{e}_{a\lhd\nu_0}\right)^{-1} &= \lambda_{\min}\left(\left(\mathbf{e}_{a\lhd\nu_0}^T\Sigma_{\nu_0\setminus\nu_1}^{-1}\mathbf{e}_{a\lhd\nu_0}\right)^{-1}\right)\\ &\geq \lambda_{\min}\left(\left(\mathbf{e}_{\{a\}\cup\nu_1}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\{a\}\cup\nu_1}\right)^{-1}\right)\\ &= \left(\lambda_{\max}\left(\mathbf{e}_{\{a\}\cup\nu_1}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\{a\}\cup\nu_1}\right)\right)^{-1}\\ &\geq \left(\lambda_{\max}\left(\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\right)\right)^{-1}\\ &= \lambda_{\min}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)\geq c_2n^{-\alpha_1}.\end{aligned}$$

The last inequality above follows from the assumption on the growth rate of $\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}$. Note also that when $a=k_j$, $\left(\mathbf{e}_{a\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{a\lhd\nu_0}\right)^{-1}=\sigma_{k_j}^2-\sigma_{k_j[-k_j]}\sigma_{[-k_j][-k_j]}^{-1}\sigma_{[-k_j]k_j]}$. Therefore, it follows from the definition of $\gamma_{a|\nu_1\cup\nu_2}$,

Corollary 3.1 and Theorem 3.1 that $\gamma_a \leq \gamma_{a|\nu_1\cup\nu_2}$ and that both γ_a and $\gamma_{a|\nu_1\cup\nu_2}$ can be asymptotically less than or equal to $\sigma^2_{k_j}$ due to interferences with other predictors. Furthermore, we have a sharp result as follows.

COROLLARY 3.2. *Under conditions in Theorem 3.1, as n tends to infinity, we have:*

(i) Uniformly for $a \in [1:p] \setminus \nu_0$ *and* $|\nu_1 \cup \nu_2| < rn$, *both* γ_a *and* $\gamma_{a|\nu_1\cup\nu_2}$ *converge to zero in the rate of* $O(n^{-1+\alpha_0})$.

(ii) Uniformly for $a \in \nu_0$ *and* $|\nu_1 \cup \nu_2| < rn$, $\frac{\gamma_a}{\gamma_{a|\nu_1\cup\nu_2}} = \left(1 - f_{a|\nu_1}\right)(1 + o(1)) < 1$, *where* $g_{a\nu_1} = \mathbf{e}^T_{\nu_1\triangleleft\nu_0}\left(\mathbf{e}^T_{\nu_0}\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{a\triangleleft\nu_0}$ *and*

$$f_{a|\nu_1} = \frac{g^T_{a\nu_1}\left(\mathbf{e}^T_{\nu_1\triangleleft\nu_0}\left(\mathbf{e}^T_{\nu_0}\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0}\right)^{-1} g_{a\nu_1}}{\mathbf{e}^T_{a\triangleleft\nu_0}\left(\mathbf{e}^T_{\nu_0}\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{a\triangleleft\nu_0}}.$$

The power-based variable screening may not be efficient due to the inhomogeneous power background $\sigma^2\mathbf{w}^T_k\mathbf{w}_k$. This calls for the SNR-based variable screening. We show that active predictors can be asymptoticaaly separated from non-active ones by means of the nulled-SNR.

THEOREM 3.3. *Under the conditions in Theorem 3.2 and Condition (C4), as n tends to infinity, we have:*

(i) Uniformly for $a \in [1:p] \setminus \nu_0$ *and* $a \notin \nu_1 \cup \nu_2$ *with* $|\nu_1 \cup \nu_2| < rn$, $SNR_{a|\nu_1\cup\nu_2} = \frac{1}{\zeta_0\sigma^2} + O(n^{-2+5\alpha_0+2\alpha_1}) > 0$.

(ii) Uniformly for $a \in \nu_0$ *and* $a \notin \nu_1 \cup \nu_2$ *with* $|\nu_1 \cup \nu_2| < rn$,

$$\begin{aligned} SNR_{a|\nu_1\cup\nu_2} &= \frac{n\mathbf{e}^T_{a\triangleleft\nu_0}\Sigma^{-1}_{\nu_0\setminus\nu_1}\mathbf{e}_{a\triangleleft\nu_0}}{\sigma^2\zeta_0\mathbf{e}^T_{a\triangleleft\nu_0}\Sigma^{-1}_{\nu_0\setminus\nu_1}\mathbf{\Phi}\Sigma^{-1}_{\nu_0\setminus\nu_1}\mathbf{e}_{a\triangleleft\nu_0}(1+o(1))} \\ &+ \frac{\left(\mathbf{e}^T_{a\triangleleft\nu_0}\Sigma^{-1}_{\nu_0\setminus\nu_1}\mathbf{e}_{a\triangleleft\nu_0}\right)^2\mathbf{e}^T_{a\triangleleft\nu_0}\left(\mathbf{e}^T_{\nu_0}\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{\Psi}\left(\mathbf{e}^T_{\nu_0}\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{a\triangleleft\nu_0}}{\sigma^2\zeta_0\mathbf{e}^T_{a\triangleleft\nu_0}\Sigma^{-1}_{\nu_0\setminus\nu_1}\mathbf{\Phi}\Sigma^{-1}_{\nu_0\setminus\nu_1}\mathbf{e}_{a\triangleleft\nu_0}(1+o(1))} \\ &\to \infty, \end{aligned}$$

where $\Sigma^{-1}_{\nu_0\setminus\nu_1}$, $\mathbf{\Phi}$ *and* $\mathbf{\Psi}$ *are defined in Theorem 3.2.*

3.2. *PVA with estimated* $\mathbf{C}$. To state a sparsistency property for the case of unknown $\mathbf{C}$, we need the following two conditions used by Fan et al. (2011). In the first one, we regularize the tail behavior of $\mathbf{y}_j$.

(C5): There exist positive constants κ_1 and τ_1 such that for any $u > 0$, $1 \le j \le J$,

$$\max_{1\le i\le n} P(|y_{ij}| > u) \le \exp(1 - \tau_1 u^{\kappa_1})$$

and $\max_{1\le i\le n} E|y_{i1}|^{4\eta_0} < +\infty$, where $\eta_0 > 1$ is a constant.

In the second condition, we assume that there exists a permutation π on $\{1, ..., J\}$ so that $\mathbf{y}_{\pi(j)}, 1 \le j \le J$ are strong mixing. Let $\mathcal{F}_0^{k_0}$ and $\mathcal{F}_k^{\infty}$ denote the σ-algebras generated by $\{\mathbf{y}_{\pi(j)} : 0 \le j \le k_0\}$ and $\{\mathbf{y}_{\pi(j)} : j \ge k\}$ respectively. Define the mixing coefficient

$$\alpha(k) = \sup_{A\in\mathcal{F}_0^{k_0}, B\in\mathcal{F}_k^{\infty}} |P(A)P(B) - P(AB)|.$$

The mixing coefficient $\alpha(k)$ quantifies the degree of the dependence of the process $\{\mathbf{y}_{\pi(j)}\}$ at lag k. We assume that $\alpha(k)$ is decreasing exponentially fast as lag k is increasing, i.e.,

(C6): There exist positive constants κ_2 and τ_2 such that $\alpha(k) \le \exp(-\tau_2 k^{\kappa_2})$.

Note that (C5) holds if y_{ij}'s are Gaussian. And (C6) holds if there exist $1 = j_0 < j_1 < \cdots < j_m = J$ such that $\{\mathbf{y}_j\}_{1\le j\le J}$ can be divided into mutually independent segments $\{\mathbf{y}_j\}_{j_{k-1}\le j<j_k}, 1 \le k \le m$.

Note that in Lemma 4, the Appendix B of the Online Supplementary Material, under Conditions (C1)∼(C6), we show that the optimal shrinkage covariance estimator $\hat{\mathbf{C}}_{hs}$ is consistent with the true covariance $\mathbf{C}$. This allows us to extend Theorems 1∼3 to the case where unknown $\mathbf{C}$ is estimated by $\hat{\mathbf{C}}_{hs}$.

THEOREM 3.4. *Suppose that conditions in Theorem 3.1 and Conditions (C5)∼(C6) hold and that $\tau_{nJ}n^2 = o(1)$ as both n and J tend to infinity. Then, we have:*

(i) Uniformly for any $\nu \subseteq \nu_0$ with $|\nu| \le rn$, $\hat{\gamma}_\nu = \Sigma_{\nu\lhd\nu_0} + O_p(n^{-1+\alpha_0+2\alpha_1} + n^2\tau_{nJ})$, where $\Sigma_{\nu\lhd\nu_0} = \left(\mathbf{e}_{\nu\lhd\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\nu\lhd\nu_0}\right)^{-1}$.

(ii) Uniformly for any $\nu \subseteq [1 : p] \setminus \nu_0$ with $|\nu| \le rn$, $\hat{\gamma}_\nu = O_p(n^{-1+\alpha_0} + n^2\tau_{nJ})$.

(iii) Uniformly for any $\nu = \nu_1 \cup \nu_2$ with $\nu_1 \subseteq \nu_0$ and $\nu_2 \subseteq [1 : p] \setminus \nu_0$ of size $|\nu| \le rn$,

$$\hat{\gamma}_\nu = \begin{pmatrix} \hat{\gamma}_\nu^{11} & \hat{\gamma}_\nu^{12} \\ \hat{\gamma}_\nu^{21} & \hat{\gamma}_\nu^{22} \end{pmatrix},$$

where

$$\begin{aligned}\hat{\gamma}_{\nu}^{11} &= \Sigma_{\nu_1 \triangleleft \nu_0} + O_p(n^{-1+\alpha_0+2\alpha_1} + n^2\tau_{nJ}),\\ \hat{\gamma}_{\nu}^{12} &= O_p(n^{-1+\alpha_0+2\alpha_1} + n^2\tau_{nJ}), \quad \hat{\gamma}_{\nu}^{21} = O_p(n^{-1+\alpha_0+2\alpha_1} + n^2\tau_{nJ}),\\ \hat{\gamma}_{\nu}^{22} &= O_p(n^{-1+\alpha_0} + n^2\tau_{nJ}),\end{aligned}$$

where $\Sigma_{\nu_1 \triangleleft \nu_0} = \left(\mathbf{e}_{\nu_1|\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\nu_1|\nu_0}\right)^{-1}$.

The above theorem implies that the sparsistency property holds for the estimated predictor power $\hat{\gamma}_a$. Using $\hat{\gamma}_a$, we can screen the predictors with a pre-specified threshold, say $n^{-1+\alpha_0}\log(n)$, obtaining an estimated set of active predictors, $\hat{\nu}_d = \{1 \le a \le p : \hat{\gamma}_a > n^{-1+\alpha_0}\log(n)\}$. We can prove the following sure screening property for $\hat{\nu}_d$.

COROLLARY 3.3. *Under the conditions in Theorem 3.4, if* $\alpha_1 < \min\{(1-\alpha_0)/3, (1-3\alpha_0)/2\}$, $n^{2+\alpha_0}\tau_{nJ} = o(1)$ *and* $n^{2+\alpha_1}\tau_{nJ} = o(1)$, *then as both* n *and* J *tend to infinity,* $P(\nu_0 = \hat{\nu}_d) \to 1$.

Letting $\nu_1 \subseteq \nu_0$ and $\nu_2 \subseteq [1:p] \setminus \nu_0$, in the next theorem, we show that the sparsistency property holds for the estimated nulled-predictive powers.

THEOREM 3.5. *Suppose that the conditions in Theorem 3.2 and (C5)∼(C6) hold and that* $\tau_{nJ}n^2 = o(1)$ *as both* n *and* J *tend to infinity. Then, we have:*

(i) Uniformly for $a \in [1:p] \setminus \nu_0$, $a \notin \nu_1 \cup \nu_2$ *and* $|\nu_1 \cup \nu_2| < rn$, *the* $(\nu_1 \cup \nu_2)$*-nulled predictive power of* a *admits the form*

$$\begin{aligned}\hat{\gamma}_{a|\nu_1\cup\nu_2} &= n^{-1}\Big(\mathbf{R}_{aa} - \mathbf{R}_{a\nu_0}\mathbf{R}_{\nu_0\nu_0}^{-1}\mathbf{R}_{\nu_0 a} - \left(\mathbf{R}_{a\nu_2} - \mathbf{R}_{a\nu_0}\mathbf{R}_{\nu_0\nu_0}^{-1}\mathbf{R}_{\nu_0\nu_2}\right)\\ &\quad \times \left(\mathbf{R}_{\nu_2\nu_2} - \mathbf{R}_{\nu_2\nu_0}\mathbf{R}_{\nu_0\nu_0}^{-1}\mathbf{R}_{\nu_0\nu_2}\right)^{-1}\left(\mathbf{R}_{\nu_2 a} - \mathbf{R}_{\nu_2\nu_0}\mathbf{R}_{\nu_0\nu_0}^{-1}\mathbf{R}_{\nu_0 a}\right)\Big)^{-1}\\ &\quad + O_p(n^{-2+4\alpha_0+2\alpha_1} + n^2\tau_{nJ}).\end{aligned}$$

(ii) Uniformly for $a \in \nu_0 \setminus \nu_1$ *and* $|\nu_1 \cup \nu_2| < rn$, *the* $(\nu_1 \cup \nu_2)$*-nulled predictive power of* a *admits the form*

$$\begin{aligned}\hat{\gamma}_{a|\nu_1\cup\nu_2} &= \left(\mathbf{e}_{a\triangleleft\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\triangleleft\nu_0}\right)^{-1} + O_p(n^{-2+6\alpha_0+5\alpha_1} + n^2\tau_{nJ})\\ &\quad + n^{-1}\mathbf{e}_{a\triangleleft\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{\Psi} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{a\triangleleft\nu_0},\end{aligned}$$

where $\Sigma_{\nu_0\setminus\nu_1}^{-1}$ *and* $\mathbf{\Psi}$ *are defined in Theorem 3.2.*

The above theorem implies that uniformly for $a \in \nu_0 \setminus \nu_1$ and $|\nu_1 \cup \nu_2| < rn$, the $(\nu_1 \cup \nu_2)$-nulled predictive power of a admits the form $\gamma_{a|\nu_1\cup\nu_2} = \left(\mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0}\right)^{-1} + O(n^{-1+3\alpha_0+2\alpha_1} + n^2\tau_{nJ})$. This shows that the sparsistency property holds for the nulled-power-based PVA. We further show that the sparsistency property also holds for the SNR-based PVA.

THEOREM 3.6. *Suppose that the conditions in Theorem 3.3 and Conditions (C5)∼(C6) hold and that $\tau_{nJ}n^2 = o(1)$ as both n and J tend to infinity. Then, we have:*

(i) Uniformly for $a \in [1:p] \setminus \nu_0$, $a \notin \nu_1 \cup \nu_2$ and $|\nu_1 \cup \nu_2| < rn$, the $(\nu_1 \cup \nu_2)$-nulled predictive power of a admits the form $\hat{SNR}_{a|\nu_1\cup\nu_2} = \frac{1}{\zeta_0\sigma^2} + O_p(n^{-2+4\alpha_0+2\alpha_1} + n^2\tau_{nJ})$.

(ii) Uniformly for $a \in \nu_0 \setminus \nu_1$ and $|\nu_1 \cup \nu_2| < rn$, the $(\nu_1 \cup \nu_2)$-nulled SNR of predictor a admits the form

$$\hat{SNR}_{a|\nu_1\cup\nu_2} = \frac{n\mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0}}{\sigma^2\eta_0 \mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{\Phi} \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0}(1+o(1))} + O_p(n^2\tau_{nJ})$$
$$+ \frac{\left(\mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0}\right)^2 \mathbf{e}_{a\lhd\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{\Psi} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{a\lhd\nu_0}}{\sigma^2\zeta_0 \mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{\Phi} \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0}(1+o(1))},$$

where $\Sigma_{\nu_0\setminus\nu_1}^{-1}$, $\mathbf{\Psi}$ and $\mathbf{\Phi}$ are defined in Theorem 3.3.

Note that for $a \in \nu_0 \setminus \nu_1$ and $|\nu_1 \cup \nu_2| < rn$,

$$\begin{aligned}\lambda_{\min}\left(\mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0}\right) &\geq \lambda_{\min}\left(\mathbf{e}_{\{a\}\cup\nu_1}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\{a\}\cup\nu_1}\right) \\ &\geq \left(\lambda_{\max}\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)\right)^{-1},\end{aligned}$$

which is bounded below from zero as $\lambda_{\max}\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right) = O(1)$. It can also be shown that $\sigma^2\zeta_0 \mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{\Phi} \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0} = O(n^{3\alpha_0+2\alpha_1})$. Consequently, the leading term in Theorem 3.6 (ii) tends to infinity as $n^{1-3\alpha_0-2\alpha_1}$ tends to infinity. In contrast, for $a \notin \nu_0$, $\hat{\text{SNR}}_{a|\nu_1\cup\nu_2}$ converges to a constant as stated in Theorem 3.6 (i). Compared to Theorem 3.4, we can see that Theorem 3.6 provides a sharper contrast between active and non-active predictors.

As in Subsection 3.3, let $\omega_{\hat{m}}$ denote the set of predictors derived from the (SNR-based) PVA. We have the following selection consistency for $\omega_{\hat{m}}$.

COROLLARY 3.4. *Under the conditions in Theorem 3.6, as both n and J tend to infinity, we have the selection consistency in the sense that $P(\omega_{\hat{m}} = \nu_0) \to 1$.*

4. Numerical results. In this section, we assess the performance of the PVA in identifying active predictors using synthetic and real data. As our simulations suggest that the SNR-based PVA performs better than the power-based PVA, we consider only four versions of the SNR-based PVA below corresponding the four different estimators of $\mathbf{C}$, namely, Ledoit-Wolf's shrinkage estimator and the optimal shrinkage of thresholded estimator $\hat{\mathbf{C}}_{hs}$ with $h = 0.01, 0.005, 0.001$.

4.1. *Synthetic data.* We compare the sensitivity of the PVA to those implemented in the R-packages 'glmnet' (Friedman, Hastie, Simon, Tibshirani, version 2.1), 'lsgl' (Vincent, version 1.3.5) and 'mrce' (Rothman, version 2.1): the multivariate group LASSO (MGL), the multivariate elastic-net (MENET), the multivariate LASSO (ML), the multivariate group sparse LASSO (MGSL) and multivariate regression with covariance estimation (MRCE) when all these procedures fix their specificity values approximately at the same level as the PVA. A brief introduction to these methods can be found in the Appendix A, the Online Supplementary Material.

Specificity and sensitivity are defined as the survival rates of true active predictors and of true non-active predictors respectively in screening, namely $\text{SEN}_D = |\hat{T} \cap T|/|T|$ and $\text{SPE}_D = |\hat{T}^c \cap T^c|/|T^c|$, where T and T^c are respectively the sets of true active predictors and of true non-active predictors, $\hat{T}$ and $\hat{T}^c$ are their estimators, and the symbol $|\cdot|$ denotes the size of a set. Note that if $|\hat{T}| \leq m$ and $T \cup T^c = \hat{T} \cup \hat{T}^c = \{1, 2, ..., p\}$, then we have

$$\text{SPE}_D = \frac{|\hat{T}^c - \hat{T}^c \cap T|}{|T^c|} \geq \frac{p - m - |T|}{p - |T|}.$$

So the specificity SEN_D is close to 1 when $p \gg |T| + m$. This holds for most of our simulations, for example for $m = 42, p = 2000, |T| = 37$, we have $\text{SPE}_D \geq 0.978$.

Setting 4.1 ($\mathbf{B}$ *was uncorrelated both within rows and between rows)*: Modifying a simulation setting in Rothman et al. (2010), we obtained 50 datasets of $(\mathbf{Y}, \mathbf{X})$ from the model (2.2). Each dataset was generated in the following steps. First, we drew an i.i.d. sample of size np from the standard normal $N(0, 1)$ to form an $n \times p$ matrix $\mathbf{X}$. Secondly, we drew n independent auto-regressive row-vectors from the J-dimensional multivariate normal $N_J(0, E_0)$, where $E_0 = (0.7^{|i-j|})_{J \times J}$. We stacked these row vectors to generate an $n \times J$ error term matrix $\boldsymbol{\varepsilon}$. Thirdly, we generated $\mathbf{B} = (\beta_{kj})_{p \times J} = s_0 \mathbf{B}_0$, where s_0 was a scale factor, $\mathbf{B}_0 = (b_{kj})_{p \times J}$, $b_{kj} = \eta_{kj} u_{kj}$, with η_{kj} and u_{kj} independently sampled from the Bernoulli distribution Bin(0.1) (0.1 is the success probability) and the uniform distribution

$U(s_1, s_2)$ respectively. We considered combinations of $(n, p, J, p_0, \alpha, s_0, s_1, s_2)$ with $n = 50$, $p = 100, 1000$, $J = 20$, $p_0 = 5$, $\alpha = 0, 1$, $s_0 = 0.45$, 0.6, $(s_1, s_2) = (-1, 1), (0.5, 1)$ and $(1, 2)$. Note that $\alpha = 0$ and 1 corresponded to row-wisely uncorrelated and row-wisely correlated $\mathbf{B}$s respectively. We let $(s_1, s_2) = (-1, 1), (0.5, 1)$ and $(1, 2)$ to represent three scenarios of $\mathbf{B}$: (i) rows with non-zero entries were oscillates around (thus not well separated from) the background 0; (ii) rows with non-zero entries were uniformly bigger than (thus separated from) 0 by amounts not less than $0.5s_0$; (iii) rows with non-zero entries were uniformly bigger than (thus separated from) 0 by amounts not less than s_0. Then, we randomly selected a subset S_{p_0} of size p_0 from integers from 1 to p and for any j, set $\beta_{kj} = 0$ when $k \notin S_{p_0}$. Finally, we let $\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon}$.

Setting 4.2 ($\mathbf{B}$ *was uncorrelated within rows but correlated between rows)*: We adopted Setting 4.1 except that we multiplied the above $\mathbf{B}_0$ by a matrix factor $\mathbf{B}_f = (0.6^{|k-j|})_{p\times p}$, resulting in new $\mathbf{B} = s_0\mathbf{B}_f B_0$ with correlations between non-zero rows.

Setting 4.3 ($\mathbf{B}$ *was weakly correlated within rows)*: We generated 50 datasets of $(\mathbf{Y}, \mathbf{X})$ from the model (2.2) for each combination of (n, p, J, p_0), where $n = 42, 88, 150$ is the sample size, $p = 2000$ is the regression dimension, $J = 20, 34, 131$ is the dimension of the response variable, and $p_0 = 37, 50, 70$ is the number of true active predictors underpinning the model. Each dataset was generated in the following steps. We began with calculating a $J \times J$ sample covariance matrix Ω by using the $n \times J$ weakly correlated sub-data matrix of the imputed IC50 data. Given Ω, we randomly generated p row-vectors from a J-dimensional normal $N_J(\mathbf{0}, \Omega)$, stacking them together to form a matrix $\mathbf{B}$. We then modified entries of $\mathbf{B}$ so that the resulting matrix contained exactly p_0 non-zero rows which would be taken as p_0 active predictors later. See Section C, the Online Supplementary Material for further details. To obtain matrix $\mathbf{X}$, we let $\boldsymbol{F}_0$ be the $p \times p$ sample covariance matrix of the gene expressions in our cancer drug data mentioned in the Introduction. Given $\boldsymbol{F}_0$, we then generated n iid row vectors from a multivariate normal $N_p(\mathbf{0}, \boldsymbol{F}_0)$, stacking them together to form matrix $\mathbf{X}$. We generated the error term matrix, $\boldsymbol{\varepsilon}$, by sampling from $N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ J times as its column vectors, where $\sigma^2 = 0.1$. Finally, we obtained $\mathbf{Y}$ by setting $\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon}$.

Setting 4.4 ($\mathbf{B}$ *was strongly correlated within rows)*: Similar to Setting 4.3, we generated 50 datasets of $(\mathbf{Y}, \mathbf{X})$ from the model (2.2) for each combination of (n, p, J, p_0), where $n = 42, 88, 150$ is the sample size, $p = 2000$ is the regression dimension, $J = 20, 34, 131$ is the dimension of the response variable, and $p_0 = 37, 50, 70$ is the number of true active predictors underpinning the model. Each dataset was generated in the same steps as Setting 4.3, ex-

cept that matrix Ω was replaced by one with high correlation coefficients. See Section C, the Online Supplementary Material for further details.

*Setting 4.5 (***B** *was moderately correlated within rows)*: Similar to Setting 4.3, we generated 50 datasets of $(\mathbf{Y}, \mathbf{X})$ from the model (2.2) for each combination of (n, p, J, p_0), where $n = 20, 42$, $p = 2000$, $J = 131$, and $p_0 = 20, 37$. Here, Ω was generated from the n non-missing rows of the IC50 data while $\mathbf{X}$ was produced by use of the gene expression data corresponding to the above n non-missing rows. The error term matrix was generated by sampling from $N_n(0, \sigma^2 I_n)$ J times as before but with $\sigma^2 = 0.0645$. See Section C, the Online Supplementary Material for further details.

For each combination of $(n, p, J, p_0, s_0, s_1, s_2)$ in Settings 4.1 and 4.2, we applied the PVA, MGL, MENET, ML, MSGL and MRCE to each of 50 datasets respectively and calculated their sensitivity values when the specificity value was fixed approximately at the same level. Note that in Settings 4.3 to 4.5, it was too time-consuming to run MRCE on a PC. In light of this, we skipped MRCE in our comparison in these settings. For the MGL, MENET, ML, MSGL and MRCE, we adjusted their penalty coefficients to achieve approximately the same specificity as that of the PVA. These sensitivity and specificity values were summarized using box-plots as shown in Figures 1 and 2. In these figures sh_0, hs1, hs2 and hs3 correspond to PVA based on the shrunk and thresholded covariance estimators with tuning constants $h = 0, 0.01, 0.005, 0.001$ respectively. And mgl, menet, mrce, ml, msgl stand for the multivariate group LASSO, the multivariate elastic-net, the multivariate regression with covariance estimation, the multivariate LASSO and the multivariate sparse group LASSO respectively.

The results indicated that the PVA substantially outperformed the MGL, MENET, ML, MSGL and MRCE in terms of sensitivity and specificity in all the scenarios under consideration. In Settings 4.1 and 4.2, the results suggested that the performances of the MGL, MENET, ML, MGSL and MRCE had deteriorated sharply when the separation between active and non-active predictors, in terms of regression coefficients, was decreasing. In contrast, the performance of the PVA was much more robust than the other procedures to interferences between active and non-active predictors. This was due to interferences being minimized through the optimization in the null-beamforming. This explained why the PVA substantially outperformed the other procedures as the separation between active and non-active predictors was decreasing. For example, for the oscillated case where $p = 1000$, $(n, J, p_0, s_0, s_1, s_2) = (50, 20, 5, 0.45, -1, 1)$, the average percentage sensitivity improvements of PVA(hs3) over the MGL, MENET, MRCE, ML and MSGL were respectively 130%, 190%, 202%, 343% and 853% when the speci-

ficity values were fixed roughly at the same level. In contrast, for the well-separated case where $p = 1000$, $(n, J, p_0, s_0, s_1, s_2) = (50, 20, 5, 0.45, 1, 2)$, the average percentage sensitivity improvements of the PVA(hs3) over the MGL, MENET, MRCE, ML and MSGL were respectively 42%, 44%, 98%, 12% and 35% when the specificity values were also fixed roughly at the same level. Only in the well-separated case, the other five procedures had competitive performances with the PVA. A similar conclusion can be made for the other settings. For example, for $p = 2000$, $(n, J, p_0) = (88, 20, 50)$, $(150, 20, 50)$, $(88, 34, 50)$, $(150, 34, 50)$ in Setting 4.4, when the specificity values were fixed roughly at the same level, compared to the MGL, on average the sensitivity values of the PVA(hs3) were increased by 74%, 97%, 136%, and 237% respectively. Compared to the MENET, on average the sensitivity values were increased by 312%, 478%, 443% and 968% respectively. In comparison to the ML, on average the sensitivity values were increased by 103%, 133%, 163% and 250% respectively. In comparison to the MSGL, on average the sensitivity values were increased by 53%, 85%, 110% and 169%. The results also suggested that the sensitivity improvements of the PVA(hs3) over the other procedures were decreasing when p_0 changed from 50 to 70, although they were still large. This was expected as the model complexity increased but the sample size did not increase. In Setting 4.3, we considered a weakly correlated regression coefficient matrix $\mathbf{B}$. With the same combinations of (n, p, J, p_0) as before, compared to highly correlated $\mathbf{B}$ setting, the improvements over the other procedures reduced but they were also substantial. This reflected a fact that the higher the correlations in columns or rows of $\mathbf{B}$, the stronger intra-correlations the response variable would receive. Therefore, more accurate variable selection would be derived from the PVA as it could explore correlation structures in the data better than the other methods. The results also indicated that the sensitivity improvements of the PVA over the other procedures were increasing in J and n. The similar result was also obtained in Setting 4.5.

We recorded the running times of performing the above procedures on each of the 50 datasets in each setting. The results, displayed in Section D, the Online Supplementary Material, showed that on average the PVA was run much faster than the ML and MSGL and was also very competitive with the MGL and MENET when we applied them to these datasets in terms of log-CPU-times in seconds.

4.2. *Cancer drug data.* Cancer drugs exert their function through binding to one or more protein targets (Wang et al., 2014). Early "one gene, one drug, one cancer" paradigm considers the role of individual genes and their

changes in drug-perturbed states, which largely ignore a target's cellular and physiological context. Meanwhile, cancer gene-centric methods largely ignore the multi-factor-driven attribute of cancer diseases at the cellular level. With the generation of rich data resources for genome-wide gene expressions and drug- and cancer-induced perturbations, data integrative approaches such as PVA try to provide systematic insights into mechanisms of drugs and cancers in a "multiple genes, multiple drugs, multiple types of cancers" paradigm.

In this section we performed PVA(hs3) on such a kind of dataset first discussed in Garnett et.al(2012). The dataset contains gene expression levels of 13321 genes and median inhibitory concentrations (IC50s) of 131 drugs across 586 cell lines. Among these cell lines, only 42 had complete records of their response to 131 drugs. Here, we considered only the 42 completed cell lines. The challenging problem of imputing remaining cell lines will be addressed in a separate work. Letting $\mathbf{X}$ be log-gene-expression levels and $\mathbf{Y}$ be IC50 values of 42 completely observed cell lines, we considered the model (2.2) for $(\mathbf{Y}, \mathbf{X})$ with the sample size $n = 42$, the number of predictors $p = 13321$ and the dimension of the response variable $J = 131$. As $p \gg n$ and $p \gg J$, the model estimation was ill-posed. To reduce the number of predictors, we performed PVA(hs3)-based variable selection on the dataset, identifying 37 active predictors (i.e., genes) for the response variable (i.e., IC50s). We then fitted a reduced multivariate regression model to the dataset by restricting the predictors to the selected, obtaining an estimated vector of the 131-dimensional regression coefficients for each selected gene. Surprisingly, although the selected genes were uncorrelated in their expression levels, they were strongly correlated when they reacted to cancer drugs as shown in the Appendix E, the Online Supplementary Material. This suggests that these genes are potentially correlated in a high function level (e.g., protein level).

Following the procedure in Subsection 2.4, we constructed a predictive network, displayed in Figure 3, for the selected genes based on their regression coefficients across 131 drugs. The network was strongly connected as there always existed a path from any node to any other node.

To reveal the potential roles of these selected genes played in cancer drug sensitivity, we investigated their protein stainings in 20 common cancers as the protein products would dictate their functions (Stewart and Wild, 2014). The tables in the Appendix F, the Online Supplementary Material provide such information gathered from the Human Protein Atlas Portal at http://www.proteinatlas.org/cancer. In these tables, as in the Portal, we classified the protein expression/staining levels into 4 categories: high, medium, low and not detected. We assigned the scores of $3, 2, 1$ and 0 to

the 4 categories respectively. If a gene did not play a role in a cancer, it would receive a score of zero as its protein staining at that cancer would be hardly detectable. We found 34 of the selected genes, which had positive staining levels on at least one of these cancers. This implied that these genes might play certain functional roles in growths of some of these cancers. In the Portal, there was no information available on the remaining 3 of the selected genes.

5. Discussion and Conclusion. In this paper, we have developed a novel approach called PVA for multivariate variable selection. Unlike the classical principal component analysis, in the PVA we project the data of the response variable along a direction in restricted eigenvector space determined by each predictor. The restricted eigenvalues called predictive powers are then used to rank predictors. The highly ranked predictors are called principal variables. By the PVA, we try to find a small number of principal variables to explain the maximum amount of variation in the data. We have established a sparsistency theory for both the power-based and the SNR-based mapping: Under certain sparsity and regularity conditions, true active predictors are asymptotically separable from non-active predictors in terms of their power or SNR values when the sample size and the dimension of the response variable tend to infinite. We have also shown that the nulled-predictor power has a higher value than a non-nulled predictor power. This has explained why the PVA can outperform the existing multivariate variable selection procedures in the literature. We have conducted a wide range of simulation studies to compare the PVA with the multivariate group LASSO, the multivariate elastic-net, the multivariate LASSO, the multivariate sparse group LASSO and the MRCE. The simulation results have shown that the PVA can substantially perform better than its competitors in all the scenarios under considerations while the PVA is scalable to the data size by iteratively calculating the power or SNR values. A limitation of the theory we have developed is that we need assumptions of stationarity and sparsity. However, the stationarity can be largely reduced if using local non-parametric regression models where only a local stationarity is required. The simulation studies in Settings 1$\sim$4 have shown that even when the response covariance matrix is not sparse or when J is much smaller than n, PVA can still have a superior performance than the existing methods.

To demonstrate the usage of the PVA in practice, we have conducted PVA on a cancer drug dataset and identified a list of principal genes and the related network to predict the drug's sensitivity to cancers in a "multiple genes, multiple drugs, multiple types of cancers" paradigm. The correlations of the

selected genes in the RNA expression levels are largely different from those in their functional levels (their contributions to the IC50 values). The results have been further validated by the protein expression levels of these genes in 20 common cancers. We should mention that we have applied the cross-validation-based multivariate group LASSO and the multivariate elastic-net to the same dataset. Unfortunately, we have ended up with a few thousand genes being selected, which were very difficult to interpret in practice.

We note that the PVA depends on the covariance matrix estimation for the response variable. In this paper, we opt for the thresholded and the shrinkage estimators as well as their hybrid version. The performance of PVA does not change much by using the alternative covariance estimators discussed in Cai and Liu (2011).

Acknowledgements. We are grateful to Professor Martin Michaelis and Dr Mark Wass from School of Bioscience, University of Kent for discussions on cancer drug studies.

SUPPLEMENTARY MATERIAL

This supplementary material provides the proofs of all propositions, corollaries and theorems in Section 3 as well as some extra numerical results and codes for Section 4.
(http://www.e-publications.org/ims/support/download/pvasupple.pdf).

References.

[1] Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding, *Ann. Stat.*, **36**, 2577-2604.

[2] Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, **106**, 672-684.

[3] Chen, L. and Huang, J. (2011). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, **107**, 1533-1545.

[4] Chiu, C.Y., Jung, J., Chen, W., Weeks, D.E., Ren, H., Boehnke, M., Amos, C.I., Liu, A., Mills, J.L., Ting Lee, M.L., Xiong, M. and Fan, R. (2017). Meta-analysis of quantitative pleiotropic traits for next-generation sequencing with multivariate functional linear models. *Eur. Jour. Hum. Genet.*, **25**, 350-359.

[5] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Jour. Ameri. Stat. Assoc.*, **96**, 1348-1360.

[6] Fan, J. , Liao, Y., and Mincheva, M. (2011) . High dimensional covariance matrix estimation in approximate factor models. *Ann. Stat.*, **39**, 3320-3356.

[7] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. B.*, **70**, 849-911.

[8] Friendly, M. (2007). HE plots for multivariate linear models. *Journal of Computational and Graphical Statistics*, **16**, 421-444.

[9] Froot, K.A. (1989). Consistent covariance matrix estimation with cross-sectional dependence and heteroskedasticity in cross-sectional financial data. *Journal of Financial and Quantitative Analysis*, **24**, 333-355.

[10] Garnett, M. J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570-575.

[11] George, E. I. (2000). The variable selection problem. *Jour. Amer. Statist. Assoc.*, **95**, 1304-1308.

[12] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Jour. Multi. Analy.*, **88**, 365-411.

[13] Li, Y., Nan, B. and Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, **71**, 354-363.

[14] Park, H.J. and Kriston, K.(2013). Structural and functional brain networks: from connections to cognition. *Science*, **342**, 1238411.

[15] Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.Y., Pollack, J.R. and Wang, P. (2010). Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer. *Ann Appl Stat.* **4**, 53-77.

[16] Rothman, A.J., Levina, E. and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Jour. Comput. Graph. Stat.*, **19**, 947-962.

[17] Sofer, T., Dicker, L. and Lin, X. (2014). Variable selection for high dimensional multivariate outcomes. *Stat. Sinica*, **24**, 1633-1654.

[18] Stewart, B. W., and Wild, C.P. (2014). World Cancer Report 2014. *International Agency for Research on Cancer. World Health Organization.* WHO Press.

[19] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B*, **58**, 267-288.

[20] Van Veen, B.D., Van Drongelen, W., Yuchtman, M. and Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering*, **44**, 867-880.

[21] Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Jour. Ameri. Stat. Assoc.*, **104**, 1512-1524.

[22] Wang, L., Wang, Y., Hu, Q. and Li, S. (2014). Systematic analysis of new drug indications by drug-gene-disease coherent subnetworks. *CPT Pharmacometric Syst. Pharmacol.*, **3**, e146.

[23] Zhang, J. and Liu, C. (2015). On linearly constrained minimum variance beamforming. *Journal of Machine Learning Research*, **16**, 2099-2145.

[24] Zhou, X. and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, **11**, 407-409.

[25] Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Jour. Roy. Stat. Soc. Ser. B*, **67**, 301-320.

School of Mathematics, Statistics and Actuarial Science
University of Kent, Canterbury, Kent CT2 7FU, U.K.
E-mail: jz79@kent.ac.uk E-mail: eo217@kent.ac.uk

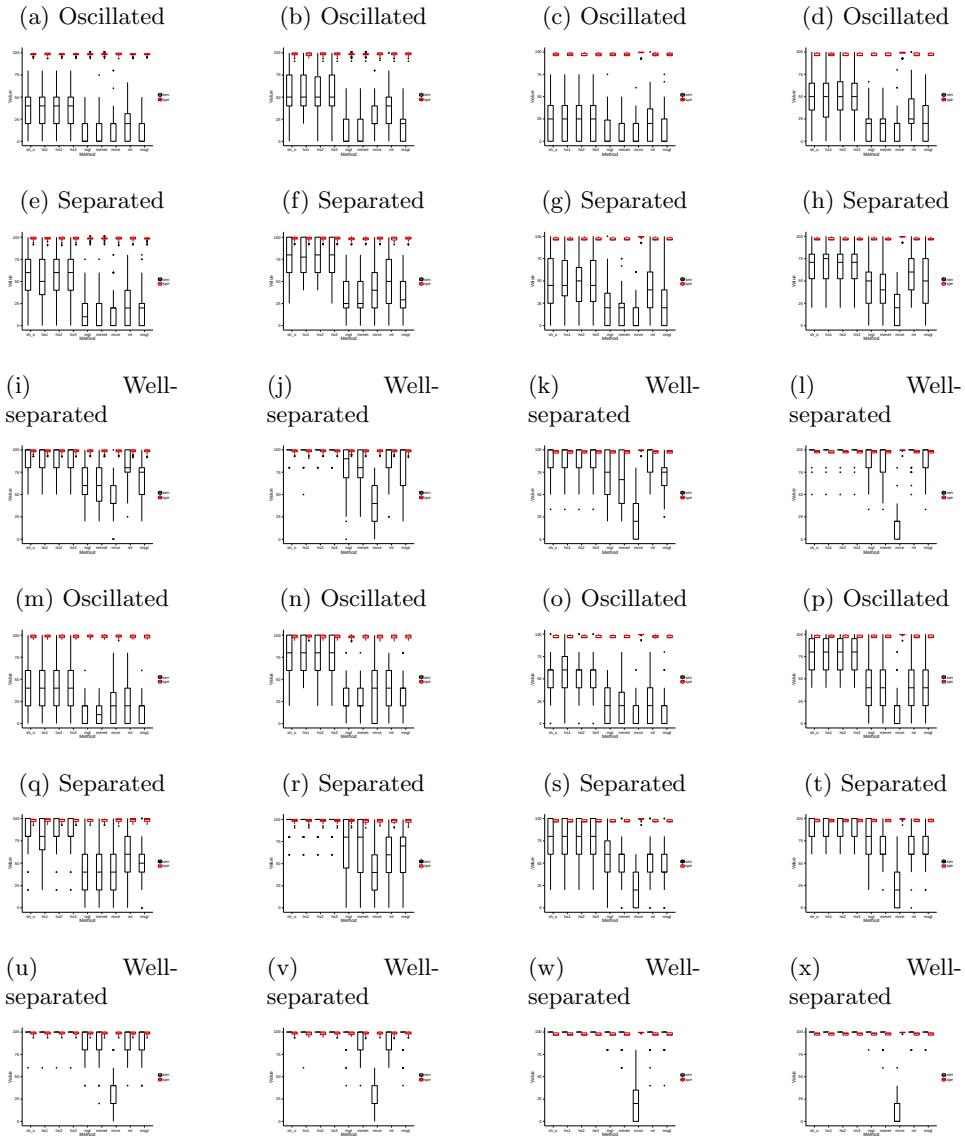
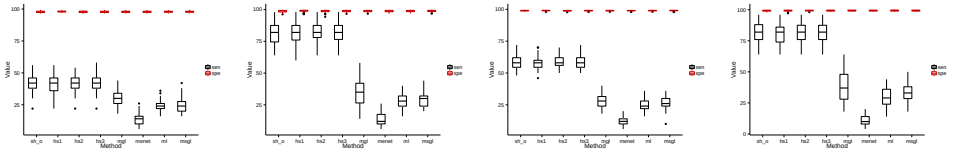
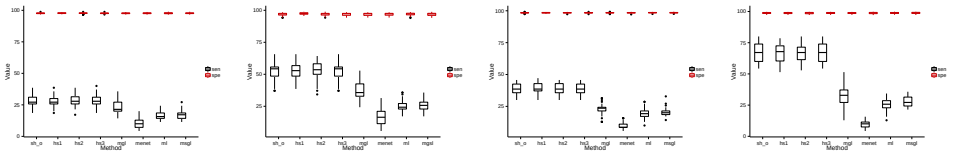


Fig 1: Box plots of sensitivity and specificity. (a)∼(l) and (m)∼(x) are for Settings 4.1 and 4.2 respectively. (a) and (m): $(n, p, J, p_0, s_0, s_1, s_2) = (50, 100, 20, 5, 0.45, -1, 1)$. (b) and (n): $(n, p, J, p_0, s_0, s_1, s_2) = (50, 100, 20, 5, 0.6, -1, 1)$. (c) and (o): $(n, p, J, p_0, s_0, s_1, s_2) = (50, 1000, 20, 5, 0.45, -1, 1)$. (d) and (p): $(n, p, J, p_0, s_0, s_1, s_2) = (50, 1000, 20, 5, 0.6, -1, 1)$. (e) and (q): $(n, p, J, p_0, s_0, s_1, s_2) = (50, 100, 20, 5, 0.45, 0.5, 1)$. (f) and (r): $(n, p, J, p_0, s_0, s_1, s_2) = (50, 100, 20, 5, 0.6, 0.5, 1)$. (g) and (s): $(n, p, J, p_0, s_0, s_1, s_2) = (50, 1000, 20, 5, 0.45, 0.5, 1)$. (h) and (t): $(n, p, J, p_0, s_0, s_1, s_2) = (50, 1000, 20, 5, 0.6, 0.5, 1)$. (i) and (u): $(n, p, J, p_0, s_0, s_1, s_2) = (50, 100, 20, 5, 0.45, 1, 2)$. (j) and (v): $(n, p, J, p_0, s_0, s_1, s_2) = (50, 100, 20, 5, 0.6, 1, 2)$. (k) and (w): $(n, p, J, p_0, s_0, s_1, s_2) = (50, 1000, 20, 5, 0.45, 1, 2)$. (l) and (x): $(n, p, J, p_0, s_0, s_1, s_2) = (50, 1000, 20, 5, 0.6, 1, 2)$. In each panel, from the left to the right, the odd columns are for sensitivity while the even columns are for specificity. In each panel, box-plots from the left to the right are for sh_0, hs1, hs2, hs3, mgl, menet, mrce, ml and msgl respectively.

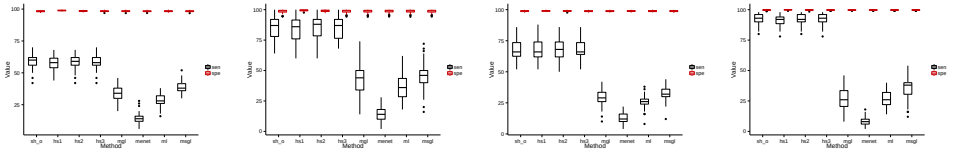
(a) LCW, $(n, J, p_0) = (88, 20, 50)$ (b) LCW, $(n, J, p_0) = (150, 20, 50)$ (c) LCW, $(n, J, p_0) = (88, 34, 50)$ (d) LCW, $(n, J, p_0) = (150, 34, 50)$



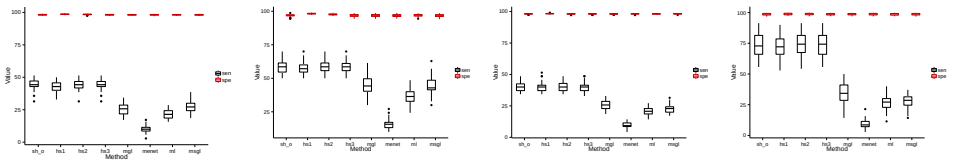
(e) LCW, $(n, J, p_0) = (88, 20, 70)$ (f) LCW, $(n, J, p_0) = (150, 20, 70)$ (g) LCW, $(n, J, p_0) = (88, 34, 70)$ (h) LCW, $(n, J, p_0) = (150, 34, 70)$



(i) HCW, $(n, J, p_0) = (88, 20, 50)$ (j) HCW, $(n, J, p_0) = (150, 20, 50)$ (k) HCW, $(n, J, p_0) = (88, 34, 50)$ (l) HCW, $(n, J, p_0) = (150, 34, 50)$



(m) HCW, $(n, J, p_0) = (88, 20, 70)$ (n) HCW, $(n, J, p_0) = (150, 20, 70)$ (o) HCW, $(n, J, p_0) = (88, 34, 70)$ (p) HCW, $(n, J, p_0) = (150, 34, 70)$



(q) MCW,$(n, J, p_0) = (42, 131, 20)$ (r) MCW, $(n, J, p_0) = (42, 131, 37)$ (s) MCW, $(n, J, p_0) = (20, 131, 20)$ (t) MCW, $(n, J, p_0) = (20, 131, 37)$

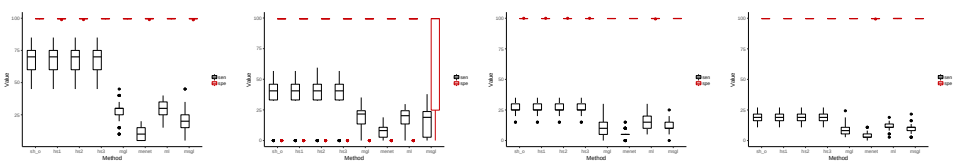


Fig 2: Box plots of sensitivity and specificity for Setting 4.3 (Low correlations within rows, short for LCW), Setting 4.4 (High correlations within rows, short for HCW) and Setting 4.5 (Moderated correlations within rows, short for MCW) when $p = 2000$ and $c_0 = 5$. Here, we adopt the same notations as in Figure 4.1.

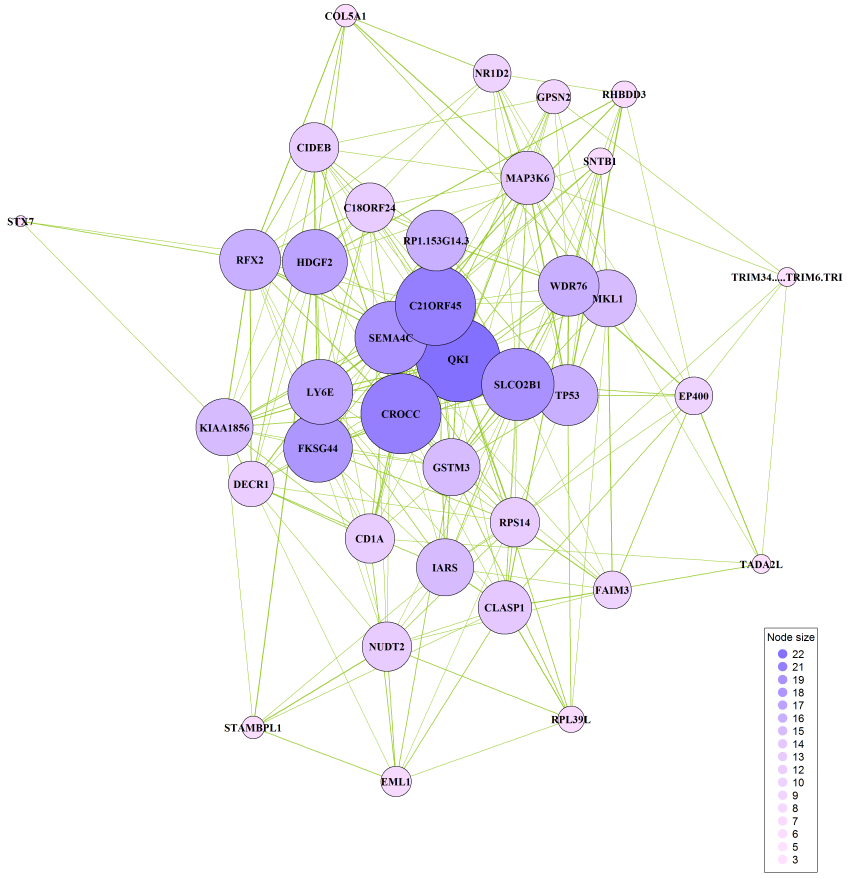


Fig 3: A network of the 37 selected genes based on their regression coefficients across 131 drugs. The size (called degree) of each node is proportional to the number of connections of that node with other nodes. The thickness of each edge represents the magnitude of the correlation coefficient between the nodes linked by this edge. The higher the correlation coefficient, the thicker the edge is. The largest degree 22 and the smallest degree 3 were attained by gene QKI and gene STX7 respectively.