Kent Academic Repository

Full text document (pdf)

Citation for published version

Ding, Hui and Zhang, Jian and Zhang, Riquan (2018) Nonparametric Variable Screening for Multivariate Additive Models. TBD . (Submitted)

DOI

Link to record in KAR

https://kar.kent.ac.uk/66718/

Document Version

Pre-print

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

Nonparametric Variable Screening for Multivariate Additive Models

Hui Ding, Jian Zhang and Riquan Zhang

East China Normal University, University of Kent and East China Normal University

April 12, 2018

Abstract

We propose a novel approach to nonparametric variable screening for sparse multivariate additive models with random effects, which includes two stages. In Stage 1, each nonparametric component is approximated by a linear combination of spline basis functions. Under this approximation, the above screening problem can be treated as selecting block-matrices of regression coefficients for a multivariate regression model. In Stage 2, a series of filtering operations are conducted by projections of the multiple response observations into the covariate space; each filter is tailored to a particular covariate and resistant to interferences originating from other covariates and from background noises. The filtering is further improved by sequentially nulling significant covariates detected in the previous steps. An asymptotic theory on the selection consistency has been established under some regularity conditions. By simulations, the proposed procedure is shown to outperform the existing procedures in terms of sensitivity and specificity over a wide range of scenarios. We apply the proposed approach to the integrative analysis of the anti-cancer drug data, identifying a few biomarkers that potentially influence the concentration of drugs in cancer cell lines.

Some key words: Multivariate additive models, high-dimensional multivariate data, nonparametric variable screening and beamforming.

Short title: Nonparametric Variable Screening

*Address for correspondence: Professor Jian Zhang, School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7FS, United Kingdom. E-mail: jz79@kent.ac.uk.

1 Introduction

This paper is about nonparametric variable screening for the multivariate additive model with random effects, where the response data, $\mathbf{y}_j \in \mathbb{R}^n$, $1 \le j \le J$, depend on the same set of covariates, $\mathbf{x}_k \in \mathbb{R}^n, 1 \le k \le p$, via the equations

$$\mathbf{y}_j = \boldsymbol{\mu} + \mathbf{f}_{1j}(\mathbf{x}_1) + \cdots + \mathbf{f}_{pj}(\mathbf{x}_p) + \boldsymbol{\varepsilon}_j, \quad j = 1, 2, ..., J. \tag{1.1}$$

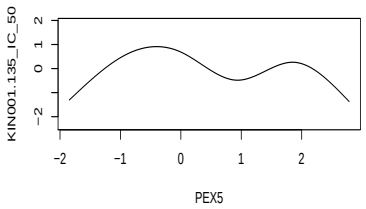
In the above model, n is the sample size, J is the number of response variables, p is the number of components (or covariates), $\boldsymbol{\mu}$ is an n-vector of fixed-effect intercepts, $\mathbf{f}_{kj}(\mathbf{x}_k) \hat{=} (f_{kj}(x_{1k}), ..., f_{kj}(x_{nk}))^T$ is a nonparametric covariate-specific random effect vector at $\mathbf{x}_k \hat{=} (x_{1k}, ..., x_{nk})^T$, $\boldsymbol{\varepsilon}_j = (\epsilon_{1j}, ..., \epsilon_{nj})^T$ is a vector of error terms, and given $\mathbf{X}$, $\boldsymbol{\varepsilon}_j$'s are conditionally independent of $\mathbf{f}_{kj}(\mathbf{x}_k)$'s. We use covariate-specific random effects to explore sample dependence as well as to pool information across multiple response variables. To make the model identifiable, for $J > 1$, we impose the constraint that conditional on $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$, for all (k, j), $E[\mathbf{f}_{kj}(\mathbf{x}_k)|\mathbf{X}] = 0$, $E[\boldsymbol{\varepsilon}_j|\mathbf{X}] = 0$, $\text{cov}(\boldsymbol{\varepsilon}_j|\mathbf{X}) = \sigma^2 I_n$ with $0 < \sigma^2 < \infty$. We assume that the random-effects $f_{kj}(\cdot)$'s are sparse in the sense that only for a few covariates, the variability, $(nJ)^{-1} \sum_{i=1}^n \sum_{j=1}^J (f_{kj}(x_{ik}) - \bar{f}_k(x_{ik}))^2$, is asymptotically positive as n and J tend to infinity, where $\bar{f}_k(x_{ik}) = \sum_{j=1}^J f_{kj}(x_{ik})/J$. The above model is different from the conventional multivariate additive models (Yee and Wild, 1996; Lin and Zhang, 1999; Rigby and Stasinopoulos, 2005) in that we allow for random-effects in nonparametric components as well as dependance in the sample. Note that the above model reduces to a multivariate linear model when $f_{kj}(\cdot), 1 \le k \le p, 1 \le j \le J$ are linear. The purpose of this paper is to provide a general screening method to identify these sparse components when p is larger than both n and J.

In the above model, the so-called vector back-fitting algorithm can be used to reduce the potential difficulty associated with high-dimensional nonparametric estimation (Yee, 2015). The estimation problem, however, becomes challenging when there are many more components (or covariates) than both the sample size and the number of response variables. For example, in cancer research, people aim 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 is about whether and how the drug sensitivity of cancer cell lines can be predicted from gene activities in the cells. The response data consist of the measurements of median inhibition concentrations, IC50s, of 131 drugs in 42 cancer cell lines while the covariate data contain expression levels of 13321 genes (Garnett et al., 2012). According to cancer encyclopedia,

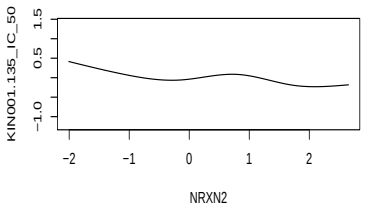
IC50 is a concentration of drug that reduces a biochemical activity such as cell multiplication to 50 percent of its normal value in the absence of the inhibitor. Note that the IC50 values observed in cell lines can be dependent as the cell lines are grouping according to cancer types. Note also that the effect of a biomarker on the concentration of drug can be nonlinear. See Figure 1 for the evidence. When the number of components is larger than the sample size, the conventional least squares criterion may not provide a satisfactory solution to the above problem. This gives rise to the above ill-posed problem of selecting a few important components (or covariates) from a large number of candidates in the model (1.1).

Extensive research has been conducted on variable selection for univariate additive models (see, e.g., Stone, 1985; Zhang et al., 2004; Lin and Zhang, 2006; Koltchinskii and Yuan 2010; Meier et al., 2009; Ravikumar et al., 2009; Huang et al., 2010; Fan et al., 2011). Zhang et al. (2004) and Lin and Zhang (2006) investigated the use of a Lasso-type penalty-based procedure in smoothing spline ANOVA with a fixed number of covariates. Meier et al. (2009) addressed the problem when the numbers of zero and nonzero components both exceed the sample size. They established a sure screening property for their proposal. Ravikumar et al. (2009) established a theory of selection consistency under the assumption that the eigenvalues of a "design matrix" are bounded away from both zero and infinity, where the "design matrix" was formed from the basis functions of the components. Huang et al. (2010) showed a selection consistency property for their procedure under a much weaker condition than those of Ravikumar et al. (2009). Fan et al. (2011) proposed an independence sure screening procedure for sparse ultra-high-dimensional univariate additive models.

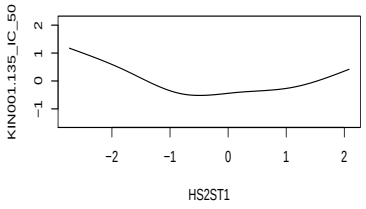
Despite of the above progress, the challenging problem of variable selection for multivariate additive models with random effects remains open in the literature, due to the lack of an appropriate framework for exploring variance components in response variables, particularly, when the sample is dependent. In this paper, we propose a novel approach for variable screening in the model (1.1), which include two stages. In Stage 1, we approximate each nonparametric component by a linear combination of spline basis functions. Under this approximation, the above variable selection problem reduces to that of selecting block-matrices of regression coefficients in a multivariate regression model. In Stage 2, we conduct a series of filtering operations (called beamforming in Zhang and Liu (2015)) on projections of $\mathbf{y}_j, 1 \leq j \leq J$ into each covariate subspace; each is tailored to a particular covariate and resistant to interferences originating from other covariates and from noises. The filtering is further improved by sequentially nulling significant covariates detected in



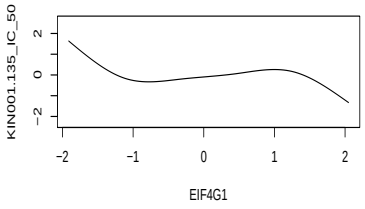
(a)



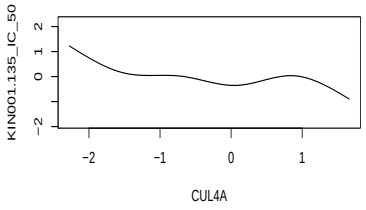
(b)



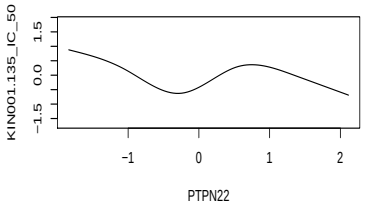
(c)



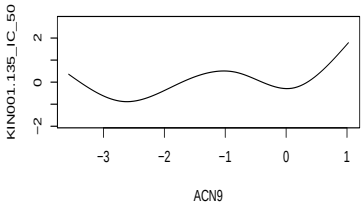
(d)



(e)



(f)



(g)

Figure 1: Estimated nonparametric sensitivity functions of selected genes, PEXS, NRXN2, HS2ST1, EIF4G1, CUL4A, PIPN22 and ACN9, to the anti-cancer drug KIN001-135.

the previous steps. The above filtering is based on the covariate power (i.e., variability), which is estimated by minimizing the trace of the sample covariance matrix of the projected data points $\mathrm{W}^T\mathbf{y}_j, 1 \le j \le J$ with respect to a weighting matrix W. The minimization is subject to the constraint that $\mathrm{W}^T\Psi(\mathbf{x}_k) =$ an identity matrix and to the nulling of significant covariates identified in the previous steps, where $\Psi(\mathbf{x}_k)$ is the $n \times \kappa_n$ "design matrix" produced by the values of the κ_n B-spline basis functions at the kth covariate. The higher the power, the more information about responses the covariate contains. Note that the projected data at each covariate 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 standardized power map of covariates. The covariates can then be ranked and selected by thresholding the map. A list of highly ranked covariates called functional principal variables are produced along with their estimated regression coefficients. Based on these selected covariates, a variance-component decomposition of the response covariance matrix can be made.

The rest of the article is organized as follows. In Section 2 we introduce the beamforming-based screening procedure for multivariate additive models. We develop the theoretical properties of the proposed procedure in Section 3. Our Monte Carlo simulations and a real data analysis in Section 4 demonstrate the effectiveness of the proposed method in terms of sensitivity and specificity. We conclude with a discussion in Section 5, and relegate the proofs to an 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 A_n, we define the spectral norm $||\mathrm{A}_n||$ by $\lambda^{1/2}_{\max}(\mathrm{A}_n^T\mathrm{A}_n)$ and the Frobenius norm $||A_n||_F$ by $\sqrt{\mathrm{tr}(A_nA_n^T)/n}$. For a sequence of real numbers $\{u_n\}$, we say $\mathrm{A}_n = O(u_n)$ if $||\mathrm{A}_n||/|u_n|$ is bounded from above and $\mathrm{A}_n = o(u_n)$ if $||\mathrm{A}_n||/|u_n|$ tends to zero as n tends to infinity. For two symmetric matrix A and B, we mean by $\mathrm{A} \le \mathrm{B}$ that $\mathrm{B} - \mathrm{A}$ is non-negative definite. Let I_q be the $q \times q$ identity matrix and $|T|$ the cardinality of a set T.

2 Methodology

Letting $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_J)$, $\mathbf{1}_J$ be a J-vector of 1's, $\mathrm{F}_k(\mathbf{x}_k) = (\mathbf{f}_{k1}(\mathbf{x}_k), ..., \mathbf{f}_{kJ}(\mathbf{x}_k))$ and $\mathrm{E} = (\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_J)$, we write the model (1.1) in the matrix equation

$$\mathbf{Y} = \boldsymbol{\mu}\mathbf{1}_J^T + \mathrm{F}_1(\mathbf{x}_1) + \cdots + \mathrm{F}_p(\mathbf{x}_p) + \mathrm{E}. \tag{2.1}$$

We assume the following condition for the additive components $f_{kj}(\cdot)$:

(C1) The additive component functions have a bounded support $[a, b]$ and satisfy the Lipschitz inequality,

$$|f_{kj}^{(r)}(z+\delta) - f_{kj}^{(r)}(z)| \leq c_0|\delta|^{\alpha}$$

for all z and $z+\delta \in [a, b]$, where r is a non-negative integer and $0 < \alpha \leq 1$.

To introduce a normalized B-spline approximation to each component function, we let $a = z_0 < z_1 < \cdots < z_{N+1} = b$ be a partition of the interval $[a, b]$, where $c_1 n^{-\nu} \leq \min_{0\leq k\leq N} |z_k - z_{k+1}| \leq \max_{0\leq k\leq N} |z_k - z_{k+1}| \leq c_2 n^{-\nu}$ with $0 \leq \nu < 0.5$, c_1 and c_2 are constants. We repeat both the lower and upper boundary knots z_0 and z_{N+1}, $m-1$ times and re-index them as $z_k, k = 0, ..., \kappa_n$ with $\kappa_n = N + 2m - 1$. Following de Boor (1978), we define a normalized B-spline basis $\{\psi_k\}_{k=1}^{\kappa_n}$ for the functions satisfying Condition (C1). Then, for each (k, j), under Condition (C1), we can find a linear combination of the normalized B-spline basis function, $\tilde{f}_{kj}(x) = \sum_{d=1}^{\kappa_n} \beta_{kjd}\psi_d(x)$ such that

$$\max_{a\leq x\leq b} |f_{kj}(x) - \tilde{f}_{kj}(x)| \leq c_3 \kappa_n^{-r_0}, \tag{2.2}$$

where $r_0 = r + \alpha$, c_3 is a constant independent of (k, j). See de Boor (1978). Using these approximations, we can reformulate the equation (2.1) as follows:

$$\mathbf{Y} = \boldsymbol{\mu}\mathbf{1}_J^T + \Psi(\mathbf{x}_k)\mathrm{B}_k + \mathrm{E}_k^*,$$

where $\mathrm{E}_k^* = \mathrm{E} + \Delta_k + \sum_{t\neq k} \mathrm{F}_t(\mathbf{x}_t)$, $\Delta_k = \mathrm{F}_k(\mathbf{x}_k) - \Psi(\mathbf{x}_k)\mathrm{B}_k$ with

$$\Psi(\mathbf{x}_k) = \begin{pmatrix} \psi_1(x_{1k}) & \cdots & \psi_{\kappa_n}(x_{1k}) \\ \vdots & \vdots & \vdots \\ \psi_1(x_{nk}) & \cdots & \psi_{\kappa_n}(x_{nk}) \end{pmatrix}, \quad \mathrm{B}_k = \begin{pmatrix} \beta_{k11} & \cdots & \beta_{kJ1} \\ \vdots & \vdots & \vdots \\ \beta_{k1\kappa_n} & \cdots & \beta_{kJ\kappa_n} \end{pmatrix}.$$

Note that if we assume that $\mathrm{F}_k(\mathbf{x}_k), 1 \leq k \leq p$ are linear functions, then model (2.1) reduces to a multivariate linear model. As in practice these functions are often nonlinear, the proposed nonparametric model is much more flexible.

2.1 SNR indices for covariates

Let $\bar{\mathbf{Y}} = (\sum_{j=1}^{J} \mathbf{y}_j/J)\mathbf{1}_J^T$ and $\hat{\mathrm{C}} = (\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})^T/J$. We project the data $\mathbf{Y}$ into the kth covariate space with an $n \times \kappa_n$ direction matrix W, namely, $\mathrm{W}^T\mathbf{Y}$, subject to $\mathrm{W}^T\Psi(\mathbf{x}_k) = I_{\kappa_n}$. The above constraint is a filter which allows the information related to B_k to pass through. To minimize the interference from other covariates and noise, we choose W in which the trace of the sample covariance matrix of the projected data $\mathrm{W}^T\mathbf{Y}$, $\mathrm{tr}(\mathrm{W}^T(\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})^T\mathrm{W})/J$, is minimized,

subject to $\mathrm{W}^T\Psi(x_k) = \mathrm{I}_{\kappa_n}$. This gives an optimal solution $\hat{\mathrm{W}} = \hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_k)(\Psi(\mathbf{x}_k)^T\hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_k))^{-1}$ with the variability

$$\mathrm{tr}(\hat{\mathrm{W}}^T\hat{\mathrm{C}}\hat{\mathrm{W}}) = \mathrm{tr}\left(\left(\Psi(\mathbf{x}_k)^T\hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_k)\right)^{-1}\right).$$

We define the above trace as the power of the kth covariate denoted by $\hat{\gamma}_k$ which gauges the amount of uncertainty in the data $\mathbf{Y}$ that can be explained by the kth covariate. If we project a white noise data set into the kth covariate by using the above weighting matrix $\hat{\mathrm{W}}$, the corresponding sample covariance matrix is approximately equal to $\hat{\mathrm{W}}\hat{\mathrm{W}}^T$. This motivates us to define the following signal-to-noise ratio (SNR) index

$$\mathrm{SNR}_k \hat{=} \mathrm{tr}\left(\hat{\mathrm{W}}^T\hat{\mathrm{C}}\hat{\mathrm{W}}\left(\hat{\mathrm{W}}^T\hat{\mathrm{W}}\right)^{-1}\right) = \mathrm{tr}\left(\Psi(\mathbf{x}_k)^T\hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_k)\left(\Psi(\mathbf{x}_k)^T\hat{\mathrm{C}}^{-2}\Psi(\mathbf{x}_k)\right)^{-1}\right),$$

which shows the signal strength of the kth covariate relative to the white noise. Using SNR_k, we can reduce the selection bias introduced by inhomogeneity of the design matrix.

Note that covariates can be correlated to each other and that we are unable to remove such an effect completely by using the above beamforming technique. To further reduce such an effect, we define a nulled SNR as follows. Let ω and ν be two non-overlapped subsets of the predictors with sizes m_1 and m respectively. Let $\Psi(\mathbf{x}_{\nu\cup\omega}) = (\Psi(\mathbf{x}_\nu), \Psi(\mathbf{x}_\omega))$. Similarly, we define $\Psi(\mathbf{x}_\nu) = (\Psi(\mathbf{x}_k) : k \in \nu)$, a matrix formed by $\Psi(\mathbf{x}_k)$, $k \in \nu$. To null the effects of covariates in ω, we choose W in which the trace of the sample covariance matrix of the projected data $\mathrm{W}^T\mathbf{Y}$, $\mathrm{tr}(\mathrm{W}^T(\mathbf{Y}-\bar{\mathbf{Y}})(\mathbf{Y}-\bar{\mathbf{Y}})^T\mathrm{W})/J$, is minimized, subject to $\mathrm{W}^T\Psi(\mathbf{x}_\nu) = \mathbf{1}_m^T \otimes \mathrm{I}_{\kappa_n}$ and $\mathrm{W}^T\Psi(\mathbf{x}_\omega) = \mathbf{0}_{m_1}^T \otimes \mathrm{I}_{\kappa_n}$, where $\mathbf{1}_m$ is an m-vector of 1's, $\mathbf{0}_{m_1}$ is an m_1-vector of 0's and $\otimes$ is the Kronecker product. This gives rise to the following nulled power $\gamma_{\nu|\omega}$,

$$\hat{\gamma}_{\nu|\omega} = \mathrm{tr}\left(\mathbf{e}_{\nu|\omega}^T(\Psi(\mathbf{x}_{\nu\cup\omega})^T\hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_{\nu\cup\omega}))^{-1}\mathbf{e}_{\nu|\omega}\right)$$

and the SNR of ν after nulling ω,

$$\begin{aligned}\mathrm{SNR}_{\nu|\omega} \;=\; & \mathrm{tr}\left\{\left(\mathbf{e}_{\nu|\omega}^T\left(\Psi(\mathbf{x}_{\nu\cup\omega})^T\hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_{\nu\cup\omega})\right)^{-1}\mathbf{e}_{\nu|\omega}\right)\left(\mathbf{e}_{\nu|\omega}^T\left(\Psi(\mathbf{x}_{\nu\cup\omega})^T\hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_{\nu\cup\omega})\right)^{-1}\right.\right. \\ & \left.\left.\times\left(\Psi(\mathbf{x}_{\nu\cup\omega})^T\hat{\mathrm{C}}^{-2}\Psi(\mathbf{x}_{\nu\cup\omega})\right)\left(\Psi(\mathbf{x}_{\nu\cup\omega})^T\hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_{\nu\cup\omega})\right)^{-1}\mathbf{e}_{\nu|\omega}\right)^{-1}\right\}, \qquad (2.3)\end{aligned}$$

where $\mathbf{e}_{\nu|\omega} = (\mathbf{1}_m^T \otimes I_{\kappa_n}, \mathbf{0}_{m_1}^T \otimes \mathrm{I}_{\kappa_n})^T$ is a $(m+m_1)\kappa_n \times \kappa_n$ block matrix.

2.2 Null-beamforming procedure

Based on the SNR indices, we are ready to define a nulled-beamforming procedure for identifying principal covariates as follows:

Initialization: Find k_1 at which the SNR_{k_1} attains the maximum. Set $\omega_1 = \{k_1\}$.

Variable screening: In the iteration $m, m \geq 2$, let ω_{m-1} denote the set of the identified covariates in the first $m-1$ iterations. For any covariate k not in ω_{m-1}, using the formula (2.3), we find $k_m \notin \omega_{m-1}$ in which $\text{SNR}_{k_m|\omega_{m-1}}$ attains the maximum. If the rule shown later are satisfied, we update ω_{m-1} and $\Psi(\mathbf{x}_{\omega_{m-1}})$ by letting $\omega_m = \{k_m\} \cup \omega_{m-1}$ and $\Psi(\mathbf{x}_{\omega_m}) = (\Psi(\mathbf{x}_{k_m}), \Psi(\mathbf{x}_{\omega_{m-1}}))$. Otherwise, we let $\omega_m = \omega_{m-1}$ and $\Psi(\mathbf{x}_{\omega_m}) = \Psi(\mathbf{x}_{\omega_{m-1}})$.

Stopping rule: After a number of iterations, the nulled SNR values will start leveling off, which indicates that the remaining covariates have no predictive power for the response. This motivates us to set the following stopping rule in each iteration: Make a scree plot of the nulled SNR values and identify an elbow point. The elbow point partitions the remaining covariates into two subsets, namely upper set and lower set. The lower set, containing those covariates 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 σ_l for the lower subset. The hypothesis is accepted if the maximum nulled SNR value, SNR_{max}, in the upper set falls into the following confidence interval,

$$|\text{SNR}_{max} - \mu_l| \leq c_0 \sigma_l, \tag{2.4}$$

where c_0 is a tuning constant. The iteration will be terminated when the upper subset is uninformative. Otherwise, we add the covariate of the maximum nulled SNR value into the current selected covariate ω and the iteration will continue. We set the default value $c_0 = 3.5$ for above tuning constant at the confidence level of 99.7%.

The proposed procedure is called functional Principal Variable Analysis (fPVA). An analogous procedure for the multivariate linear model is called linear PVA.

2.3 Estimation of response covariance matrix

The above defined power is based on the response covariance matrix which is often estimated by the sample covariance matrix

$$\hat{\text{C}} = (\hat{c}_{ij}) \hat{=} (\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})^T / J$$

with $\bar{\mathbf{Y}} = (\sum_{j=1}^{J} \mathbf{y}_j / J)\mathbf{1}_J^T$. When $n > J$, $\hat{\text{C}}$ is degenerate, leading an ill-posed definition of the power. To address this issue, we consider a thresholded estimator introduced by Bickel and Levina (2005):

$$\hat{\text{C}}_h = \hat{\text{C}}(\tau_{nJ}) = (\hat{c}_{ij} I(|\hat{c}_{ij}| > h\tau_{nJ})),$$

where $I(\cdot)$ is the indicator and $\tau_{nJ} = \sqrt{\log(n)/J}$ with $h \geq 0$ being a constant (for example, $h = 0.001|\mathrm{tr}(\hat{C})/n|$). The thresholded covariance matrix estimators may not be positive definite when the dimension J is close to or smaller than the sample size n. To remedy the problem, we shrinkage the above thresholded covariance estimator to a diagonal matrix by using the method of Ledoit and Wolf (2004) as follows:

$$\hat{\mathrm{C}}_{hs} = \frac{b_n^2}{d_n^2}\hat{\mu}_n \mathrm{I}_n + \frac{d_n^2 - b_n^2}{d_n^2}\hat{\mathrm{C}}_h, \tag{2.5}$$

where

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

with, for any $n \times n$ matrices D_1 and D_2, $< \mathrm{D}_1, \mathrm{D}_2 >= \mathrm{tr}(\mathrm{D}_1\mathrm{D}_2^T)/n$.

3 Asymptotic analysis

To give an insight into the proposed method, we make a further assumption on the model as follows.

(C2) There exists a permutation on $\mathbf{y}_j, 1 \leq j \leq J$ so that the resulted sequence is strictly stationary with marginal covariance matrix C. We assume that there are only q_n non-zero components in the model (1.1).

Without loss of generality, we let the first q_n components are non-zeros: $f_{kj}(x) \neq 0, 1 \leq k \leq q_n$, but $f_{kj}(x) \equiv 0, q_n + 1 \leq k \leq p$. Let $\nu_0 = \{1, 2, ..., q_n\}$, $|\nu_0| = q_n$ and $\mathbf{f}_{\nu_0 j}^{\oplus}(\mathbf{x}_{\nu_0}) = \mathbf{f}_{1j}(\mathbf{x}_1) + \cdots + \mathbf{f}_{q_n j}(\mathbf{x}_{q_n})$. Let $\Delta_{ij} = \sum_{k=1}^{q_n}(f(x_{ik}) - \sum_{d=1}^{\kappa_n}\beta_{kjd}\psi_d(x_{ik}))$, $\Delta_j = (\Delta_{1j}, ..., \Delta_{nj})^T$, $\mathbf{b}_{kj} = (\beta_{kj1}, \cdots, \beta_{kj\kappa_n})^T$, $\mathrm{B}_{\nu_0 j} = (\mathbf{b}_{1j}^T, ..., \mathbf{b}_{q_n j}^T)^T$ and $\Psi(\mathbf{x}_{\nu_0}) = (\Psi(\mathbf{x}_1), ..., \Psi(\mathbf{x}_{q_n}))$. Let $\Sigma \hat{=} \mathrm{cov}(\mathrm{B}_{(1:p)j})$ be a $(p\kappa_n \times p\kappa_n)$ block matrix $(\Sigma_{k_1 k_2})_{p\times p}$ with block $\Sigma_{k_1 k_2} = \mathrm{cov}(\mathbf{b}_{k_1 j}, \mathbf{b}_{k_2 j})$. Similarly, we define $\mathrm{cov}(\mathrm{B}_{\nu_0 j})$ as the $(|\nu_0|\kappa_n \times |\nu_0|\kappa_n)$ block matrix $(\Sigma_{k_1 k_2})_{k1,k2\in\nu_0}$.

3.1 With known C

When $J = \infty$, we can estimate C exactly. So, we first consider an ideal setting where C is known. Let γ and SNR denote the corresponding power and signal-to-noise-ratio index respectively. Then,

for the fixed $\mathbf{X}$, we have

$$\begin{aligned}
\mathrm{C} &= E(\mathbf{Y}_j - \boldsymbol{\mu})(\mathbf{Y}_j - \boldsymbol{\mu})^T \\
&= E\left(\mathbf{f}^{\oplus}_{\nu_0 j}(\mathbf{x}_{\nu_0}) + \boldsymbol{\varepsilon}_j\right)\left(\mathbf{f}^{\oplus}_{\nu_0 j}(\mathbf{x}_{\nu_0}) + \boldsymbol{\varepsilon}_j\right)^T \\
&= E\left(\Psi(\mathbf{x}_{\nu_0})\mathrm{B}_{\nu_0 j} + \Delta_j + \boldsymbol{\varepsilon}_j\right)(\Psi(\mathbf{x}_{\nu_0})\mathrm{B}_{\nu_0 j} + \Delta_j + \boldsymbol{\varepsilon}_j)^T \\
&= \Psi(\mathbf{x}_{\nu_0})\mathrm{cov}(\mathrm{B}_{\nu_0 j})\Psi(\mathbf{x}_{\nu_0})^T + \Psi(\mathbf{x}_{\nu_0})E[\mathrm{B}_{\nu_0 j}]E[\mathrm{B}^T_{\nu_0 j}]\Psi(\mathbf{x}_{\nu_0})^T + E[\Delta_j\Delta_j^T] + E[\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}_j^T] \\
&\quad + \Psi(\mathbf{x}_{\nu_0})E[\mathrm{B}_{\nu_0 j}\Delta_j^T] + E[\Delta_j\mathrm{B}^T_{\nu_0 j}]\Psi(\mathbf{x}_{\nu_0})^T + \Psi(\mathbf{x}_{\nu_0})E[\mathrm{B}_{\nu_0 j}\boldsymbol{\varepsilon}_j^T] + E[\boldsymbol{\varepsilon}_j\mathrm{B}^T_{\nu_0 j}]\Psi(\mathbf{x}_{\nu_0})^T \\
&\quad + E[\Delta_j\boldsymbol{\varepsilon}_j^T] + E[\boldsymbol{\varepsilon}_j\Delta_j^T] \\
&= \Psi(\mathbf{x}_{\nu_0})\mathrm{cov}(\mathrm{B}_{\nu_0 j})\Psi(\mathbf{x}_{\nu_0})^T + \Psi(\mathbf{x}_{\nu_0})E[\mathrm{B}_{\nu_0 j}]E[\mathrm{B}^T_{\nu_0 j}]\Psi(\mathbf{x}_{\nu_0})^T + E[\Delta_j\Delta_j^T] + E[\boldsymbol{\varepsilon}_j\boldsymbol{\varepsilon}_j^T] \\
&\quad + \Psi(\mathbf{x}_{\nu_0})E[\mathrm{B}_{\nu_0 j}\Delta_j^T] + E[\Delta_j\mathrm{B}^T_{\nu_0 j}]\Psi(\mathbf{x}_{\nu_0})^T.
\end{aligned}$$

The last equality follows from the assumption that $\mathbf{f}^{\oplus}_{\nu_0 j}(\mathbf{x}_{\nu_0})$ (therefore $\Psi(\mathbf{x}_{\nu_0})\mathrm{B}_{\nu_0 j}$) is independent of $\boldsymbol{\varepsilon}_j$.

It follows from the inequality (2.2) that $|\Delta_{ij}| \le c_3 q_n \kappa_n^{-r_0}$. For any $\mathbf{a} \in \mathbb{R}^n$, $||\mathbf{a}||_2 = 1$,

$$\mathbf{a}^T\Delta_j\Delta_j^T\mathbf{a} = \left(\sum_{i=1}^n a_i\Delta_{ij}\right)^2 \le ||\mathbf{a}||_2^2||\Delta_j||_2^2 \le c_3^2 n q_n^2 \kappa_n^{-2r_0}.$$

$$\begin{aligned}
\mathbf{a}^T\Psi(\mathbf{x}_{\nu_0})E[\mathrm{B}_{\nu_0 j}]E[\mathrm{B}^T_{\nu_0 j}]\Psi(\mathbf{x}_{\nu_0})^T\mathbf{a} &= \mathbf{a}^T E[\mathbf{f}^{\oplus}_{\nu_0 j}(\mathbf{x}_{\nu_0}) - \Delta_j]E[\mathbf{f}^{\oplus}_{\nu_0 j}(\mathbf{x}_{\nu_0}) - \Delta_j]^T\mathbf{a} \\
&= \mathbf{a}^T E[\Delta_j]E[\Delta_j]^T\mathbf{a} \le c_3^2 n q_n^2 \kappa_n^{-2r_0}.
\end{aligned}$$

It also follows from the inequality (2.2) that $|\sum_{k=1}^{q_n}\sum_{d=1}^{\kappa_n}\beta_{kjd}\psi_d(x_{ik})| \le \sum_{k=1}^{q_n}(|f_{kj}(x_{ik})| + c_3\kappa_n^{-r_0}) = O(q_n)$. Therefore, for any $\mathbf{a} \in \mathbb{R}^n$, $||\mathbf{a}||_2 = 1$,

$$|\mathbf{a}^T\Psi(\mathbf{x}_{\nu_0})\mathrm{B}_{\nu_0 j}\Delta_j^T\mathbf{a}| \le ||\Psi(\mathbf{x}_{\nu_0})\mathrm{B}_{\nu_0 j}||_2||\Delta_j||_2 \le O(q_n)O(n^{1/2})c_3 n^{1/2} q_n\kappa_n^{-r_0} = O(q_n^2)n\kappa_n^{-r_0}.$$

Consequently, we have

$$\begin{aligned}
\mathrm{C} &= \Psi(\mathbf{x}_{\nu_0})\mathrm{cov}(\mathrm{B}_{\nu_0 j})\Psi(\mathbf{x}_{\nu_0})^T + \sigma^2\mathrm{I}_n + 2\times O(nq_n^2\kappa_n^{-2r_0}) + 2\times O(nq_n^2\kappa_n^{-r_0}) \\
&= \Psi(\mathbf{x}_{\nu_0})\mathrm{cov}(\mathrm{B}_{\nu_0 j})\Psi(\mathbf{x}_{\nu_0})^T + \sigma^2\mathrm{I}_n + O(nq_n^2\kappa_n^{-r_0}). \\
\mathrm{C} &\le \Psi(\mathbf{x}_{\nu_0})\mathrm{cov}(\mathrm{B}_{\nu_0 j})\Psi(\mathbf{x}_{\nu_0})^T + (\sigma^2 + c_4 nq_n^2\kappa_n^{-r_0})\mathrm{I}_n.
\end{aligned}$$

To regularize the coherence structure of the design matrices, we impose the following condition on the covariates, which was used Huang et al. (2010).

(C3) There exist some positive constants K_1 and K_2 such that the marginal density function of the jth covariate, $g_j(\cdot)$, satisfies $0 < K_1 \le g_j(x) \le K_2 < \infty$ for all $1 \le j \le p$ and $x \in [a, b]$.

Let $\nu = \{k_1, \ldots, k_{p_1}\}$ denote an arbitrary subset of $1:p$. Let E_ν be a selection matrix, made of p block matrices sitting next to each other, where for $j = 1, ..., p_1$, its k_jth sub-block matrix takes the value of the $\kappa_n \times \kappa_n$ identity matrix and the remaining sub-block matrices take the value of the $\kappa_n \times \kappa_n$ zero matrix. Then, the B-spline basis for the covariates indexed by ν can be written as $\Psi(\mathbf{x}_\nu) = \Psi(\mathbf{X})\mathrm{E}_\nu^T$. Let $\mathrm{A}_\nu = \mathrm{C} - \Psi(\mathbf{x}_\nu)\mathrm{E}_\nu^T \Sigma \mathrm{E}_\nu \Psi(\mathbf{x}_\nu)^T$, which is the remaining variance-component after removing those belonging to covariate set ν. Note that if $\Psi(\mathbf{x}_{\nu_0})\mathrm{B}_{\nu_0 j}$ can approximate $\mathbf{f}^{\oplus}_{\nu_0 j}(\mathbf{x}_{\nu_0})$ perfectly, then $\sigma^{-2}\mathrm{A}_{\nu_0}$ is an identity matrix. In general, $\sigma^{-2}\mathrm{A}_{\nu_0}$ is assumed to be asymptotically dominated by a diagonal matrix in the following sense:

(C4) For some positive constant ζ_0, $\sigma^{-2}\mathrm{A}_{\nu_0} = \zeta_0 \mathrm{I}_{|\nu_0|\kappa_n}(1 + o(1))$ as n tends to infinity.

The following proposition states that under Conditions (C1)$\sim$(C3), the power of ν_0 converges to its underlying power, the trace of the covariance matrix of regression coefficients at ν_0.

Proposition 3.1 *Under Conditions (C1)$\sim$(C3), if $E_\nu^T \Sigma E_\nu$ and A_ν are invertible, then the power*

$$\gamma_\nu = tr\{E_\nu^T \Sigma E_\nu + \left(\Psi(\mathbf{x}_\nu)^T A_\nu^{-1} \Psi(\mathbf{x}_\nu)\right)^{-1}\}.$$

In particular,

$$\gamma_{\nu_0} = tr(E_{\nu_0}^T \Sigma E_{\nu_0}) + O\left(\delta^{-q_n/2} q_n^2 \kappa_n^{2-r_0}\right),$$

where $\delta = (1 - K_1/K_2)^{1/2}$. If κ_n takes the optimal rate $n^{1/(2r_0+1)}$ with $r_0 > 2$ and $q_n \leq \eta_0 \log(n)$ with $0 < \eta_0 < \frac{2(r_0-2)}{(2r_0+1)\log(\delta^{-1})}$, then

$$\gamma_{\nu_0} = tr(E_{\nu_0}^T \Sigma E_{\nu_0}) + O\left(n^{-\delta_0}\right),$$

where $\delta_0 = \frac{r_0-2}{2r_0+1} - \frac{\eta_0}{2}\log(\delta^{-1})$.

To introduce Theorem 1 below, we first introduce some notations. For any subset of covariates, ν, we define the following coherence matrices for $\Psi(\mathbf{x}_\nu)$ and $\Psi(\mathbf{x}_{\nu_0})$.

$$\mathrm{R}_{\nu\nu} = \Psi(\mathbf{x}_\nu)^T \mathrm{A}_{\nu_0}^{-1} \Psi(\mathbf{x}_\nu)/n,\ \mathrm{R}_{\nu\nu_0} = \Psi(\mathbf{x}_\nu)^T \mathrm{A}_{\nu_0}^{-1} \Psi(\mathbf{x}_{\nu_0})/n,\ \mathrm{R}_{\nu_0\nu_0} = \Psi(\mathbf{x}_{\nu_0})^T \mathrm{A}_{\nu_0}^{-1} \Psi(\mathbf{x}_{\nu_0})/n.$$

For any $\nu \subseteq \nu_0$, we can find $\phi = \{j_1, \ldots, j_m\} \subseteq \{1, \ldots, |\nu_0|\}$ such that $\nu = \{k_j : j \in \phi\}$. Let $\mathrm{E}_{\nu|\nu_0}$ is a selection block matrix, made of $|\nu_0|$ sub-blocks sitting next to each other, where for $j \in \phi$, its k_jth sub-block matrix takes value of I_{κ_n} and the remaining sub-block matrices take value of the $\kappa_n \times \kappa_n$ zero matrix. We assume the following irrepresentability condition that ν_0 is separable from its outside in terms of coherence and that the coherence within ν_0 and the cross-sectional coherence between ν_0 and its outside are of the same scale order, that is,

(C5) For any $\nu \subseteq [1:p] \setminus \nu_0$, $\lambda_{\max}(\mathrm{R}_{\nu\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-2}\mathrm{R}_{\nu_0\nu}) = 0(1)$, $\lambda_{\max}(\mathrm{F}_\nu^{-1/2}\mathrm{R}_{\nu\nu}\mathrm{F}_\nu^{-1/2}) = 0(1)$, $n\lambda_{\min}(\mathrm{F}_\nu) \to \infty$, where $\mathrm{F}_\nu = \mathrm{R}_{\nu\nu} - \mathrm{R}_{\nu\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu}$.

In the following theorem, we show that the power index is consistent. This implies that the power index based screening can have a sure screening property under the ideal scenario where the response covariance matrix is known.

Theorem 1 *Suppose that* $\delta^{-q_n/2}q_n^2\kappa_n^{1-r_0} = o(1)$ *and that* $\lambda_{mino}^{-1} = 0(1)$. *Under Conditions (C1)∼(C3), as n tends to infinity, we have*

(i) For any $\nu \subseteq \nu_0$, *we have*

$$\begin{aligned}\gamma_\nu &= tr(\Sigma_{\nu|\nu_0}) + n^{-1}tr\left(\Sigma_{\nu|\nu_0}E_{\nu|\nu_0}^T\left(E_{\nu_0}^T\Sigma E_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1}\left(E_{\nu_0}^T\Sigma E_{\nu_0}\right)^{-1}E_{\nu|\nu_0}\Sigma_{\nu|\nu_0}\right)\\ &\quad +\lambda_{maxo}^2\lambda_{mino}^{-3}(1+\lambda_{maxo}\lambda_{mino}^{-1})\delta^{-q_n}q_n^5\kappa_n^{3-2r_0}O(1),\end{aligned}$$

where $\Sigma_{\nu|\nu_0}^{-1} = E_{\nu|\nu_0}^T\left(E_{\nu_0}^T\Sigma E_{\nu_0}\right)^{-1}E_{\nu|\nu_0}$, λ_{maxo} *and* λ_{mino} *are the largest and the smallest eigenvalues of* $E_{\nu_0}^T\Sigma E_{\nu_0}$ *respectively.*

(ii) For any $\nu \subseteq \{1,\ldots,p\}\backslash\nu_0$, *if* $n\lambda_{min}(F_\nu) \to \infty$ *as* $n \to \infty$, *then*

$$\gamma_\nu = n^{-1}tr(F_\nu^{-1}) - n^{-2}tr(D_{\nu n}),$$

where

$$\begin{aligned}D_{\nu n} &= F_\nu^{-1}R_{\nu\nu_0}R_{\nu_0\nu_0}^{-1/2}\\ &\quad\times\left\{R_{\nu_0\nu 0}^{-1/2}\left(E_{\nu_0}^T\mathbf{\Sigma}E_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1/2}n^{-1} + I_{|\nu_0|\kappa_n}\right\}^{-1}\\ &\quad\times R_{\nu_0\nu_0}^{-1/2}\left(E_{\nu_0}^T\Sigma E_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu}F_\nu^{-1}(1+o(1))\\ &\leq F_\nu^{-1}R_{\nu\nu_0}R_{\nu_0\nu_0}^{-1}\left(E_{\nu_0}^T\mathbf{\Sigma}E_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu}F_\nu^{-1}(1+o(1)).\end{aligned}$$

We now state the following theorem about the discriminability of the nulled-SNR index, the extent to which active and non-active covariates can be distinguished by the nulled-SNR index.

Theorem 2 *Suppose that* $\delta^{-q_n/2}q_n^2\kappa_n^{1-r_0} = o(1)$ *and that* $\lambda_{mino}^{-1} = 0(1)$. *Then, under Conditions (C1)∼(C5), as n tends to infinity, we have:*

(i) For $a \in [1:p]\setminus\nu_0$, $a \notin \nu_2$, $\nu_1 \subseteq \nu_0$ *and* $\nu_2 \subseteq [1:p]\setminus\nu_0$,

$$SNR_{a|\nu_1\cup\nu_2} = \frac{\kappa_n}{\zeta_0\sigma^2}(1+o(1)).$$

(ii) For $a \in \nu_0$, $a \notin \nu_1$, $\nu_1 \subseteq \nu_0$ and $\nu_2 \subseteq [1:p] \setminus \nu_0$,

$$\begin{aligned} {}^{SNR}{}_{a|\nu_1 \cup \nu_2} &= \frac{n(1+o(1))}{\sigma^2 \zeta_0} tr\left\{ \left(E_{\{a\}|\nu_0}^T \Sigma_{\nu_0 \setminus \nu_1}^{-1} \Phi_0 \Sigma_{\nu_0 \setminus \nu_1}^{-1} E_{\{a\}|\nu_0} \right)^{-1} E_{\{a\}|\nu_0}^T \Sigma_{\nu_0 \setminus \nu_1}^{-1} E_{\{a\}|\nu_0} \right\} \\ &+ \frac{1+o(1)}{\sigma^2 \zeta_0} tr\Big\{ E_{\{a\}|\nu_0}^T \left(E_{\nu_0}^T \Sigma E_{\nu_0} \right)^{-1} \Phi_1 \left(E_{\nu_0}^T \Sigma E_{\nu_0} \right)^{-1} E_{\{a\}|\nu_0} \\ &\times \left(E_{\{a\}|\nu_0}^T \Sigma_{\nu_0 \setminus \nu_1}^{-1} \Phi_0 \Sigma_{\nu_0 \setminus \nu_1}^{-1} E_{\{a\}|\nu_0} \right)^{-1} \left(E_{\{a\}|\nu_0}^T \Sigma_{\nu_0 \setminus \nu_1}^{-1} E_{\{a\}|\nu_0} \right)^2 \Big\}, \end{aligned}$$

where

$$\begin{aligned} \Sigma_{\nu_1|\nu_0} &= \left(E_{\nu_1|\nu_0}^T \left(E_{\nu_0}^T \Sigma E_{\nu_0} \right)^{-1} E_{\nu_1|\nu_0} \right)^{-1}, \\ \Sigma_{\nu_0 \setminus \nu_1}^{-1} &= \left(E_{\nu_0}^T \Sigma E_{\nu_0} \right)^{-1/2} P_{\nu_0 \setminus \nu_1} \left(E_{\nu_0}^T \Sigma E_{\nu_0} \right)^{-1/2}, \\ P_{\nu_0 \setminus \nu_1} &= I_{|\nu_0|\kappa_n} - \left(E_{\nu_0}^T \Sigma E_{\nu_0} \right)^{-1/2} E_{\nu_1|\nu_0} \Sigma_{\nu_1|\nu_0} E_{\nu_1|\nu_0}^T \left(E_{\nu_0}^T \Sigma E_{\nu_0} \right)^{-1/2}, \\ F_{\nu_2} &= R_{\nu_2\nu_2} - R_{\nu_2\nu_0} R_{\nu_0\nu_0}^{-1} R_{\nu_0\nu_2}, \\ \Phi_0 &= R_{\nu_0\nu_0}^{-1} + R_{\nu_0\nu_0}^{-1} R_{\nu_0\nu_2} F_{\nu_2}^{-1} R_{\nu_2\nu_0} R_{\nu_0\nu_0}^{-1}, \\ \Phi_1 &= \left(I_{|\nu_0|\kappa_n} - E_{\nu_1|\nu_0} \Sigma_{\nu_1|\nu_0} E_{\nu_1|\nu_0}^T \left(E_{\nu_0}^T \Sigma E_{\nu_0} \right)^{-1} \right) \Phi_0 \\ &\times \left(I_{|\nu_0|\kappa_n} - E_{\nu_1|\nu_0} \Sigma_{\nu_1|\nu_0} E_{\nu_1|\nu_0}^T \left(E_{\nu_0}^T \Sigma E_{\nu_0} \right)^{-1} \right)^T. \end{aligned}$$

The above theorem indicates that if uniformly for any $a \in \nu_0$, $a \notin \nu_1$, $\nu_1 \subseteq \nu_0$ and $\nu_2 \subseteq [1:p] \setminus \nu_0$,

$$\frac{n}{\kappa_n} \mathrm{tr} \left\{ \left(\mathrm{E}_{\{a\}|\nu_0}^T \Sigma_{\nu_0 \setminus \nu_1}^{-1} \Phi_0 \Sigma_{\nu_0 \setminus \nu_1}^{-1} \mathrm{E}_{\{a\}|\nu_0} \right)^{-1} \mathrm{E}_{\{a\}|\nu_0}^T \Sigma_{\nu_0 \setminus \nu_1}^{-1} \mathrm{E}_{\{a\}|\nu_0} \right\} \to \infty$$

as n tends to infinity, then the nulled-SNR index contrast between active and non-active covariates also tends to infinity. Therefore, under the above condition, the covariate set selected by the nulled-SNR is consistent with the true one when the sample size tends to infinity.

3.2 With estimated C

To state an index consistency property for the case of unknown C, we need the following two conditions used by Fan et al. (2011). In the first one, we regularize the tail behavior of $\mathbf{y}_j$.

(C6): 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 $||C||_F$ is bounded as $n \to \infty$.

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

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

The mixing coefficient $\alpha(k)$ quantifies the degree of the dependence of the process $\{\mathbf{y}_{\pi(j)}\}$ at lag k. We assume that $\alpha(k_0, k)$ is decreasing exponentially fast as lag k is increasing, i.e.,

(C7): There exist positive constants κ_2 and τ_2 such that $\alpha(k_0, k) \leq \exp(-\tau_2(k - k_0)^{\kappa_2})$.

Note that (C6) holds if y_{ij}'s are Gaussian. And (C7) holds if there exist $1 = j_0 < j_1 < \cdots < j_m = J$ such that $\{\mathbf{y}_j\}_{1 \leq j \leq J}$ can be divided into mutually independent segments $\{\mathbf{y}_j\}_{j_{k-1} \leq j < j_k}, 1 \leq k \leq m$.

Note that under Conditions (C1)$\sim$(C7), we show that the optimal shrinkage covariance estimator $\hat{\mathrm{C}}_{hs}$ is consistent with the true covariance C in the Appendix B, the Online Supplementary Material. This allows us to extend Theorems 1$\sim$2 to the case where unknown C is estimated by $\hat{\mathrm{C}}_{hs}$. We state the following theorem.

Theorem 3 *Suppose that* $\delta^{-q_n/2} q_n^2 \kappa_n^{1-r_0} = o(1)$, $\lambda_{mino}^{-1} = 0(1)$ *and* $\tau_{nJ} n^2 = o(1)$ *as both* n *and* J *tend to infinity. Suppose that* $||C - < C, I_n > I_n||_F$ *is bounded below from zero. Then, under Conditions (C1)$\sim$(C7), we have:*

(i) For any $\nu \subseteq \nu_0$, *we have*

$$\begin{aligned} \hat{\gamma}_\nu &= tr(\Sigma_{\nu|\nu_0}) + n^{-1} tr\left(\Sigma_{\nu|\nu_0} E_{\nu|\nu_0}^T \left(E_{\nu_0}^T \Sigma E_{\nu_0}\right)^{-1} R_{\nu_0\nu_0}^{-1} \left(E_{\nu_0}^T \Sigma E_{\nu_0}\right)^{-1} E_{\nu|\nu_0} \Sigma_{\nu|\nu_0}\right) \\ &\quad + \lambda_{maxo}^2 \lambda_{mino}^{-3} (1 + \lambda_{maxo} \lambda_{mino}^{-1}) \delta^{-q_n} q_n^5 \kappa_n^{3-2r_0} O(1) + O_p(n^2 \tau_{nJ}). \end{aligned}$$

where $\Sigma_{\nu|\nu_0}^{-1}$, λ_{maxo} *and* λ_{mino} *are the same as in Theorem 1.*

(ii) For any $\nu \subseteq [1 : p] \setminus \nu_0$, *then*

$$\hat{\gamma}_\nu = n^{-1} tr(F_\nu^{-1}) + O(n^{-2} \lambda_{min}^{-1}(F_\nu)) + O_p(n^2 \tau_{nJ}).$$

The above theorem implies that $\hat{\gamma}_a$ converges to zero in probability when $a \notin \nu_0$ and to a non-zero limit when $a \in \nu_0$. This make it possible to use $\hat{\gamma}_a$ to screen for the covariates with a pre-specified threshold. The selected active set will have a sure screening property.

We further present the following asymptotic analysis on active and non-active covariates for the SNR-based fPVA.

Theorem 4 *Suppose that* $\delta^{-q_n/2} q_n^2 \kappa_n^{1-r_0} = o(1)$, $\lambda_{mino}^{-1} = 0(1)$ *and* $\tau_{nJ} n^2 = o(1)$ *as both* n *and* J *tend to infinity. Suppose that* $||C- < C, I_n > I_n||_F$ *is bounded below from zero. Then, under Conditions (C1)∼(C7), we have:*

(i) For $a \in [1:p] \setminus \nu_0$, $a \notin \nu_2$, $\nu_1 \subseteq \nu_0$ *and* $\nu_2 \subseteq [1:p] \setminus \nu_0$,

$$SNR_{a|\nu_1 \cup \nu_2} = \frac{\kappa_n}{\zeta_0 \sigma^2}(1 + o(1)) + O_p(n^2 \tau_{nJ}).$$

(ii) For $a \in \nu_0$, $a \notin \nu_1$, $\nu_1 \subseteq \nu_0$ *and* $\nu_2 \subseteq [1:p] \setminus \nu_0$,

$$\begin{aligned} SNR_{a|\nu_1 \cup \nu_2} &= \frac{n(1+o(1))}{\sigma^2 \zeta_0} tr\left\{ \left(E_{\{a\}|\nu_0}^T \Sigma_{\nu_0 \setminus \nu_1}^{-1} \Phi_0 \Sigma_{\nu_0 \setminus \nu_1}^{-1} E_{\{a\}|\nu_0} \right)^{-1} E_{\{a\}|\nu_0}^T \Sigma_{\nu_0 \setminus \nu_1}^{-1} E_{\{a\}|\nu_0} \right\} \\ &+ \frac{1+o(1)}{\sigma^2 \zeta_0} tr\Big\{ E_{\{a\}|\nu_0}^T \left(E_{\nu_0}^T \Sigma E_{\nu_0} \right)^{-1} \Phi_1 \left(E_{\nu_0}^T \Sigma E_{\nu_0} \right)^{-1} E_{\{a\}|\nu_0} \\ &\times \left(E_{\{a\}|\nu_0}^T \Sigma_{\nu_0 \setminus \nu_1}^{-1} \Phi_0 \Sigma_{\nu_0 \setminus \nu_1}^{-1} E_{\{a\}|\nu_0} \right)^{-1} \left(E_{\{a\}|\nu_0}^T \Sigma_{\nu_0 \setminus \nu_1}^{-1} E_{\{a\}|\nu_0} \right)^2 \Big\} + O_p(n^2 \tau_{nJ}), \end{aligned}$$

where $\Sigma_{\nu_0 \setminus \nu_1}^{-1}$, Φ_0 *and* Φ_1 *are the same as in Theorem 2.*

The above theorem shows that the nulled-SNR is of order κ_n in probability when the covariate is truly not active while it is of order $n\kappa_n$ in probability when the covariate is truly active. This contrast allows us to asymptotically discriminate active covariates from non active ones. Thus, the set of selected active covariates will be consistent with the underlying one if n and J are large enough.

4 Numerical results

In this section, we evaluate the performance of the proposed procedure fPVA on simulated and real data. Following Fan et al. (2011), we approximated $\mathbf{f}_{kj}$ by a linear combination of $\kappa_n = \lfloor n^{0.2} \rfloor$ normalized B-splines, where we set $\lfloor n^{0.2} \rfloor$ equally spaced interior knots for these B-splines and $\lfloor z \rfloor$ denotes the integer part of z. In the fPVA and the linear PVA, we set $h = 0.01|\mathrm{tr}(\hat{C})/n|$ in equation (2.5) and the tuning constant (defined in inequality (2.4)) $c_0 = 4$ in the simulations and $c_0 = 3.5$ in the real data analysis. The results for other choices of c_0 will not be presented here as the results are not sensitive when $3.5 \leq c_0 \leq 4$. In our simulations, we compare the fPVA to the linear PVA of Zhang and Oftadeh (2016) and the MIS, a multivariate extension of the nonparametric variable screening procedure of Fan et al. (2011), in terms of sensitivity and specificity. Here, sensitivity and specificity are defined as the survival rates of true active covariates and of true non-active

covariates in a screening procedure, namely

$$\text{SEN} = \frac{|\hat{T} \bigcap T|}{|T|}, \quad \text{SPE} = \frac{|\hat{T}^c \bigcap T^c|}{|T^c|},$$

where T and T^c are the sets of true active covariates and of true non-active covariates respectively with estimators $\hat{T}$ and $\hat{T}^c$. See the Appendix B, the Online Supplementary Material for a detailed description of the MIS.

For the kth component in a multivariate additive model, we define its oracle signal-to-noise ratio (OSNR) as follows:

$$\text{OSNR} = \text{tr}\left((\Psi(\mathbf{x}_k)^T \text{cov}(\mathbf{f}_{kj}(\mathbf{x}_k))^{-1}\Psi(\mathbf{x}_k))^{-1}\right) / \text{tr}\left((\Psi(\mathbf{x}_k)^T \text{cov}(\boldsymbol{\varepsilon}_j)^{-1}\Psi(\mathbf{x}_k))^{-1}\right)$$

under the oracle assumption that $\{\mathbf{f}_{ij}\}_{i\neq k}$ are known. The OSNR shows the oracle signal strength of each component. The higher the OSNR of a component, the higher chance it will be selected.

4.1 Simulated data

In simulations, we investigated the behavior of the above three procedures for multivariate additive and multivariate linear models respectively.

Example 4.1. We considered the following multivariate additive model similar to Fan et al. (2011):

$$y_{ij} = \sum_{k=1}^{p} f_{kj}(x_{ik}) + \epsilon_{ij}, 1 \leq i \leq n, 1 \leq j \leq J, \tag{4.1}$$

where

$$\begin{aligned}
f_{1j}(x_{i1}) &= 2(x_{i1} - 0.157)\sin j, \quad f_{2j}(x_{i2}) = 2\left((2x_{i2}-1)^2 - 0.111\right)\cos j, \\
f_{3j}(x_{i3}) &= 2.5\frac{\sin\left(2\pi x_{i3}\sqrt{j}\right)}{2-\sin\left(2\pi x_{i3}\sqrt{j}\right)} - 2.5E\left\{\frac{\sin\left(2\pi x_{i3}\sqrt{j}\right)}{2-\sin\left(2\pi x_{i3}\sqrt{j}\right)}\right\}, \\
f_{4j}(x_{i4}) &= 3\left[0.1\sin\left(2\pi x_{i4}\right) + 0.2\cos\left(2\pi x_{i4}\right) + 0.3\sin^2\left(2\pi x_{i4}\right) + 0.4\cos^3\left(2\pi x_{i4}\right)\right. \\
&\quad \left. +0.5\sin^3\left(2\pi x_{i4}\right)\right] \\
&\quad -3E\left\{\left[0.1\sin\left(2\pi x_{i4}\right) + 0.2\cos\left(2\pi x_{i4}\right) + 0.3\sin^2\left(2\pi x_{i4}\right) + 0.4\cos^3\left(2\pi x_{i4}\right)\right.\right. \\
&\quad \left.\left. +0.5\sin^3\left(2\pi x_{i4}\right)\right]\right\}, \\
f_{kj}(x_{ik}) &= 0, \quad 5 \leq k \leq p,
\end{aligned}$$

and x_{ij}'s and ϵ_{ij}'s are random variables specified below.

The above model involves J responses with multiple nonlinear coefficient functions for each covariate. We sampled 100 data sets from the above model for each of the combinations of (t, n, J, p) with $t = 0, 1$, $n = 50$, 100 and 200, $J = 30$, 70 and 140, and $p = 500$ and 1000. Each data set was generated as follows. First, to generate covariates $x_{i1}, \ldots, x_{ip}$, we sampled $w_{i1}, ..., w_{ip}, u_i$ independently from $N(0, 1)$ and truncated them into the interval $[0, 1]$, $i = 1, ..., n$. We set $x_{ik} = (w_{ik} + tu_i)/(1 + t)$ for $k = 1, ..., p$. The simple calculation can show that the pairwise correlations between covariates are equal to $t^2/(1 + t^2)$. In particular, the covariates are independent of each other if $t = 0$. Then, we independently drew the error row vectors $(\epsilon_{ij})_{1 \le j \le J}, 1 \le i \le n$ from the multivariate normal with mean zeros and covariance matrix $(0.7^{|j_1 - j_2|})_{J \times J}$, stacking them together to form an $n \times J$ error matrix. Finally, we generated $\mathbf{y}_j, 1 \le j \le J$ by using the equation (4.1). In our simulated data, we can see that the OSNR values vary significantly across the 4 components as shown in Table 1.

[Put Table 1 here.]

We applied the fPVA, the linear PVA and the MIS to each data set respectively, obtaining a list of the SEN and SPE values for each procedure. We then calculated their averages and standard deviations over 100 replicates respectively, expressing them in percentage. The results for various combinations of (t, n, J) are presented in Table 2 when $p = 1000$ and in Table 1 in the Appendix C, the Online Supplementary Material when $p = 500$.

[Put Table 2 here.]

It is easy to see from these tables that the MIS had lower specificity (thus selected more non-active covariates) than did the fPVA. The results show that the average sensitivity and specificity values were increasing in the sample size n when J, p and t were fixed. The average sensitivity was decreasing in the pairwise correlations between covariates. This reflects that the increasing correlations between covariates could make it difficult to identify true active covariates. The average sensitivity was also decreasing in the number of covariates p. This is again not surprising because the larger the number of irrelevant covariates in the model, the harder the selection of true covariates will be. When the covariates were uncorrelated (i.e., $t = 0$), the fPVA performed slightly better than both the MIS and the linear PVA in the terms of sensitivity and specificity for most of combinations of (n, J, p). In contrast, when the covariates were correlated (i.e., $t = 1$), the fPVA substantially outperformed the MIS and the linear PVA for most of combinations of (n, J, p). This demonstrates that the fPVA was more effective in taking advantage of correlation structures in covariates. We also compared the average CPU times used to run these procedures on the simulated data in a PC.

To save space, we only plot the log-CPU-times for the combinations $(n, J, p) = (100, 70, 1000)$ and $t = 0, 1$ in Figure 2(a,b). It demonstrates that the fPVA computationally cost less than the MIS but more than the linear PVA.

[Put Figure 2 here.]

In the above, we compared the fPVA to the linear PVA and the MIS when the data were generated from a nonparametric additive model. However, in practice, we did not know whether the underlying model was nonparametric or not. So, in the next example, we compared the fPVA to the linear PVA and the MIS in an unfavorable setting where the underlying model was linear in the form:

$$\mathbf{Y} = \mathbf{X}\mathrm{B} + \mathrm{E}, \tag{4.2}$$

with B being random effects.

Example 4.2. Following Zhang and Oftadeh (2016), we generated 100 data sets from the above model for each combination of (n, p, J) with $n = 50, 100, 200$, $p = 500, 1000$, $J = 30, 70, 140$, where each was produced in the following steps. We first obtained $\mathrm{B} = (u_{kj}\eta_{kj})_{p\times J}$ by sampling u_{kj} and η_{kj} independently from the uniform distribution $U(-1, 1)$ and the Bernoulli distribution $Bin(0.1)$ respectively. We then drew an i.i.d. sample of size np from $N(0, 1)$, stacking them together to form matrix $\mathbf{X}$. We further drew n independent row-vectors from a J-dimensional normal $N_J(\mathbf{0}, \Sigma_0)$, where $\Sigma_0 = (0.7^{|i-j|})_{J\times J}$. We stacked them together to form an $n \times J$ error matrix E. Finally, using the equation (4.2), we generated $\mathbf{Y}$.

For each combination of (n, p, J), we applied the fPVA, the linear PVA and the MIS to each of 100 data sets, obtaining the corresponding sensitivity and specificity. For each of the procedures and each combination of (n, p, J), we calculated the average and standard deviation of sensitivity and specificity over 100 replicates respectively.

The results presented in Table 3 demonstrate that the linear PVA outperformed the fPVA obviously when the sample size is small and the underlying model is linear. But the performance of linear PVA and fPVA were getting closer as sample size increases. This is not surprising as the linear PVA was designed for the linear model (4.2) which the data were generated from. The results also show that the MIS had lower specificity (thus selected more non-active covariates) than did the fPVA when the underlying model was linear. The MIS had higher sensitivity than did the fPVA. But for large samples (for example, $n \geq 200$) the MIS tended to perform worse than the fPVA in terms of sensitivity and specificity.

[Put Table 3 here.]

4.2 Anti-cancer drug data

We evaluated the performance of our approach on a data set, which was discussed in details by Garnett et al. (2012). The data contain $p = 13321$ gene expressions and fifty percent inhibitory concentration (IC50) values of $J = 131$ drugs across $n = 42$ cell lines. According to cancer encyclopaedia, 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. We considered a sparse multivariate additive model in (1.1) for the data, where we took genes as covariates and IC50's of multiple drugs as the response variable and let gene expressions form a design matrix $\mathbf{X}$. We first standardizing the expressions for each gene and centralizing the response variable. We then applied the fPVA to the data set, identifying 7 active covariates. Finally, we fitted the multivariate additive model to the data set with covariates restricted to the above selected covariates, where we used the post-approximations to f_{kj}'s by linear combinations of 5 nature spline functions as did in Fan et. al. (2011). To save space, we only present the estimated non-vanishing nonparametric components related to the drug KIN001-135 in Figure 1. The results suggest the relationship between the IC50s of the drug KIN001-135 and the selected active covariates were indeed nonlinear.

To highlight the medical relevance of these selected genes to the drug sensitivity, we investigated the protein staining of these selected genes in 20 common cancers as the protein products would indicate the functions of these genes (Stewart et al., 2017). We gathered such information from the Human Protein Atlas Portal at `http://www.proteinatlas.org` for 5 of the selected 7 genes: PEX5, NRXN2, EIF4GH1, CUL4A and ACN9. In these tables, we classified the protein staining levels into 4 categories: high, medium, low and not detected. We assigned the scores of 3, 2, 1 and 0 to these categories respectively. If a gene had not played a role in the sensitivity of an anti-cancer drug, we might obtain a score of zero as its protein staining at that cancer cell line would be hardly detectable. Therefore, the hypotheses of interest can be stated as follows:

$H_0 : \mu = 0$ v.s. $H_1 : \mu > 0$, where μ is the population median of the protein staining score of a gene.

For each of the 20 cancers under investigation, we performed a one-sample Wilcoxon signed-rank test on the above scores, obtaining a p-value. The p-values for these cancers are displayed in Table 4.

We then carried out a Bonferroni correction as well as Holm's correction (Holm, 1979) for

multiple testing respectively. The number of the rejected null-hypotheses for each selected gene are shown in Table 2 in the Appendix C, the Online Supplementary Material. The results indicate that all the 5 selected genes had positive staining levels in most of 20 cancers at the significance level of 0.05 after the correction. We further conducted both the Bonferroni and Holm corrections for multiple testing across all cancer-gene pairs, in which two cancer-gene pairs, (EIF4G1, Colorectal cancer) and (ACN9, Glioma cancer), survived after the correction. We also applied the MIS to the above data set, resulting in 236 active genes. However, we are unable to explain their biological roles for such a large number of genes by using the above Portal.

[Put Tables 4 here.]

To assess stability of the above analysis, we simulated $(\mathbf{Y}, \mathbf{X})$ by using the fitted multivariate additive model for each combination of (n, p, J), where $n = 42, 100, 200$, $p = 500, 1000$, $J = 131$ in the following steps. First, we set the above selected 7 genes as the true active covariates in the simulated model. We also randomly selected $p - 7$ gene covariates from the remaining 13314 genes in the above anti-cancer drug data and put them into the simulated model to form p covariates. Secondly, we calculated the $p \times p$ sample covariance matrix $\mathbf{\Omega}$ of these p genes by using the original gene expression data. Given $\mathbf{\Omega}$, we drew n random row-vectors from the p-dimension Normal $N_p(\mathbf{0}, \mathbf{\Omega})$ and stacked them row by row to form the design matrix $\mathbf{X}$. Thirdly, we computed a $J \times J$ sample covariance matrix Σ_0 by using the 131 residuals of IC50 data derived from the above real data analysis. We drew n random row-vectors from the J-dimensional Normal $N(\mathbf{0}, \Sigma_0)$ and stacked them row by row to obtain the error matrix E. Fourthly, we adopted fitted nonparametric functions $\hat{f}_{kj}(\cdot)$, $1 \leq j \leq J$ as the true component functions in the simulated model, where k ran over the 7 genes obtained in the previous data analysis and assigned zero functions to the remaining $p - 7$ components. Finally, we generated $\mathbf{Y}$ according to the model (1.1). We repeated the above procedure 100 times, obtaining 100 simulated datasets. For each combination of (n, p, J), we applied the fPVA to each of the 100 simulated data sets in order to recover the underlying active covariates, pretending they were unknown. This allowed us to estimate sensitivity and specificity values. In Table 5, we display the averages of these values over 100 replicates. The table shows that on average the fPVA could recover 4.8 out of 7 truly active covariates with the average specificity being bigger than 99% when the sample size $n = 42$ and $p = 1000$. This gave a recovering rate of 68% which was surprisingly high compared to the small sample size 42. In addition, the rate was 100% on average when the sample size $n \geq 100$. The above results suggest that the fPVA based data analysis was quite stable.

[Put Table 5 here.]

5 Discussion and Conclusion

We have proposed a novel approach to nonparametric component screening for multivariate additive models with random effects by using the B-spline approximation and the null-beamforming technique. The null-beamforming technique involves a series of spatial filters, nulled-SNR indices, each is tailored to a covariate related to a particular additive component and minimizes interferences originating from other covariates and from background noises. The null-beamforming substantially outperforms the ordinary beamforming used in Zhang and Liu (2015). In the proposed procedure, we iteratively search for the covariates at which the nulled-SNR index attains the maximum. In each iteration, the covariates identified in the previous steps have been nulled. We have conducted an asymptotic analysis on the behavior of the proposed procedure. In particular, under some regularity conditions, we have shown that the SNR-index can make a sharp contrast between active and non-active covariate. This has resulted in the selection consistency of the proposed procedure.

We have assessed the performance of the proposed procedure by use of simulated and real data. The simulations have demonstrated that our new procedure can substantially outperform the linear counterpart PVA and the marginal screening procedure MIS in terms of sensitivity and specificity in a wide range of scenarios. We have applied the proposed procedure to the integrative analysis of an anti-cancer drug data set, identifying 7 genes which might have influenced IC50 values. By use of the existing protein staining data, we have demonstrated that in most of common cancers, at least 5 of these selected genes had positive protein stainings at the significance level of 5% after some multiple testing correction. This suggests that these identified genes may have played certain roles in determining the concentrations of these drugs in cancer cell lines.

Acknowledgments

We are grateful to Professor Martin Micheales for discussions on the anti-cancer drug data. Research was completed while the second author was visiting the Institute for Mathematical Sciences, National University of Singapore on 5th∼16th, February in 2018.

Supplementary materials

The detailed proofs of the propositions, lemmas, theorems and some extra information on numerical results can be found in the Online Supplementary Material.

References

Bickel, P., and Levina, E. (2008). Covariance regularization by thresholding, *Ann. Stat.*, **36**, 25772604.

de Boor, C. (1978) A Practical Guide to Splines. Springer, New York.

Fan, J., Feng, Y., Song, R.(2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Ameri. Statist. Assoc.*, **106**, 544-557.

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

Huang, J., Horowitz, J.L. and Wei, F. (2010). Variable selection in nonparametric additive models. *Ann. Stat.*, **38**, 22822313.

Koltchinskii, V. and Yuan, M. (2010). Sparsity in multiple kernel learning. *Ann. Stat.*, **38**, 3660-3695.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Jour. Multi. Analy.*, **88**, 365411.

Li, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines, *J. R. Statist. Soc. B*, **61**, 381-400.

Lin, Y. and Zhang, H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Stat.* **34**, 22722297.

Meier, L., van de Geer, S., Bühlmann, P. (2009). High-dimensional additive modeling. *Ann. Stat.*, **37**,37793821.

Ravikumar, P, Liu H, Lafferty J, Wasserman L. (2009). Sparse additive models. *J. R. Statist. Soc. B*, **71**, 10091030.

Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape. *J. R. Statist. Soc. C*, **54**, 507-554.

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

Stone, C.J. (1985). Additive regression and other nonparametric models. *Ann. Stat.*, **13**, 689705.

Yee, T.W. and Wild, C.J. (1996). Vector generalized additive models. *J. R. Statist. Soc. B,* **58**, 481-493.

Yee, T.W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R.* Springer, New York.

Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. and Klein, B. (2004). Variable selection and model building via likelihood basis pursuit. *Jour. Amer. Stat. Assoc.*, **99**, 659672.

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

Zhang, J., and Oftadeh, E. (2016). Multivariate variable selection by means of null-beamforming . *Kent Academic Repository (KAR)*, University of Kent.

Table 1: OSNRs of components

J	1	2	3	4
30	0.026	0.035	0.093	1.211
70	0.159	0.213	0.175	1.551
140	1.022	1.308	0.562	2.443

Table 2: The average sensitivity (SEN) and specificity (SPE) over 100 replicates with standard deviations (in parentheses) in percentage for Example 4.1 when $p = 1000$.

$t = 0$					$t = 1$				
n	J	Method	SEN	SPE	n	J	Method	SEN	SPE
50	30	fPVA	78.8(13.00)	99.8(0.18)	50	30	fPVA	53.8(8.97)	99.9(0.18)
		PVA	33.0(12.75)	99.7(0.26)			PVA	18.0(12.35)	99.7(0.24)
		MIS	60.3(15.53)	95.5(0.68)			MIS	53.5(9.41)	95.6(0.72)
	70	fPVA	63.8(16.81)	99.9(0.10)		70	fPVA	58.3(16.30)	99.9(0.11)
		PVA	27.8(7.86)	99.8(0.23)			PVA	26.5(6.94)	99.8(0.19)
		MIS	72.3(19.75)	90.2(0.10)			MIS	58.8(12.50)	90.4(1.03)
	140	fPVA	52.5(12.05)	99.9(0.09)		140	fPVA	51.5(14.57)	99.9(0.11)
		PVA	26.5(5.97)	99.8(0.17)			PVA	25.3(2.50)	99.8(0.17)
		MIS	80.0(18.80)	81.8(1.29)			MIS	66.3(18.25)	82.4(1.27)
100	30	fPVA	99.0(4.92)	99.9(0.17)	100	30	fPVA	71.3(16.81)	100(0.09)
		PVA	45.0(13.76)	99.9(0.15)			PVA	28.0(8.17)	99.8(0.19)
		MIS	86.0(14.35)	96.3(0.54)			MIS	58.8(13.47)	95.9(0.65)
	70	fPVA	100(0.00)	99.9(0.13)		70	fPVA	94.0(10.73)	100(0.07)
		PVA	35.3(12.36)	99.9(0.13)			PVA	28.5(9.41)	99.9(0.11)
		MIS	95.8(10.69)	91.8(0.74)			MIS	68.0(16.70)	91.0(0.85)
	140	fPVA	98.8(7.43)	99.9(0.11)		140	fPVA	97.5(8.33)	100(0.09)
		PVA	32.0(11.83)	100(0.07)			PVA	27.8(7.86)	99.9(0.10)
		MIS	99.0(4.92)	84.7(1.03)			MIS	78.5(17.77)	83.7(1.19)
200	30	fPVA	100(0.00)	100(0.05)	200	30	fPVA	87.5(13.06)	100(0.06)
		PVA	49.3(13.03)	99.9(0.14)			PVA	28.8(8.97)	99.9(0.18)
		MIS	100(0.00)	97.3(0.56)			MIS	83.5(15.17)	95.7(0.63)
	70	fPVA	100(0.00)	100(0.06)		70	fPVA	100(0.00)	100(0.05)
		PVA	42.8(13.43)	100(0.07)			PVA	30.5(10.41)	100(0.07)
		MIS	100(0.00)	93.6(0.62)			MIS	92.3(13.15)	91.2(0.90)
	140	fPVA	100(0.00)	99.9(0.11)		140	fPVA	100(0.00)	100(0.04)
		PVA	40.5(16.20)	99.9(0.08)			PVA	36.0(13.45)	100(0.05)
		MIS	100(0.00)	88.1(0.84)			MIS	95.0(10.66)	84.7(1.06)

Table 3: The average sensitivity and specificity over 100 replicates with standard deviations(in parentheses) in percentage for Example 4.2.

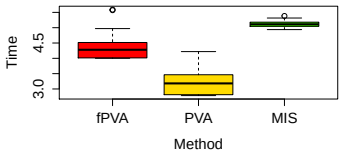
$p = 500$					$p = 1000$				
n	J	Method	SEN	SPE	n	J	Method	SEN	SPE
50	30	fPVA	28.9(15.47)	99.9(0.14)	50	30	fPVA	28.2(16.86)	99.9(0.11)
		PVA	78.1(20.85)	99.9(0.16)			PVA	79.7(18.41)	99.8(0.21)
		MIS	70.8(20.13)	91.6(1.33)			MIS	71.0(20.59)	95.6(0.58)
	70	fPVA	35.2(15.07)	100(0.08)		70	fPVA	37.6(15.64)	100(0.09)
		PVA	89.2(14.05)	99.9(0.13)			PVA	89.2(13.16)	99.9(0.13)
		MIS	93.8(9.72)	81.7(1.70)			MIS	90.0(12.87)	89.8(0.84)
	140	fPVA	35.0(14.60)	100(0.08)		140	fPVA	37.4(17.21)	100(0.09)
		PVA	98.2(5.75)	99.9(1.08)			PVA	96.6(8.07)	99.9(0.10)
		MIS	99.6(2.81)	66.2(2.12)			MIS	99.0(4.38)	80.5(1.20)
100	30	fPVA	62.5(24.86)	100(0.07)	100	30	fPVA	65.5(24.25)	100(0.05)
		PVA	89.8(14.04)	99.9(0.16)			PVA	87.5(16.25)	99.8(0.15)
		MIS	87.3(13.61)	92.5(0.99)			MIS	84.4(17.45)	96.0(0.54)
	70	fPVA	83.0(15.14)	100(0.06)		70	fPVA	83.6(14.04)	100(0.04)
		PVA	98.6(7.11)	99.9(0.13)			PVA	98.6(5.13)	99.9(0.09)
		MIS	99.0(5.22)	83.8(1.49)			MIS	98.6(5.86)	99.2(8.38)
	140	fPVA	93.8(10.13)	100(0.06)		140	fPVA	93.4(9.87)	100(0.05)
		PVA	100(0.00)	100(0.09)			PVA	99.8(2.00)	99.9(0.09)
		MIS	100(0.00)	70.2(2.08)			MIS	100(5.46)	83.0(0.96)
200	30	fPVA	87.3(14.95)	100(0.08)	200	30	fPVA	86.3(16.19)	100(0.06)
		PVA	93.8(11.13)	99.8(0.19)			PVA	94.6(10.00)	99.9(0.14)
		MIS	92.0(12.97)	93.2(1.03)			MIS	92.3(12.46)	96.6(0.54)
	70	fPVA	97.2(6.97)	100(0.06)		70	fPVA	98.2(7.02)	100(0.04)
		PVA	99.4(3.43)	99.9(0.12)			PVA	99.8(2.00)	100(0.08)
		MIS	98.8(4.77)	85.2(1.36)			MIS	99.6(2.81)	92.1(0.77)
	140	fPVA	100(0.00)	100(0.06)		140	fPVA	99.8(2.00)	100(0.03)
		PVA	100(0.00)	100(0.08)			PVA	99.8(2.00)	100(0.06)
		MIS	100(0.00)	73.0(1.78)			MIS	100(0.00)	84.7(0.85)

Table 4: P-values of the one-sample Wilcoxon signed-rank test for the significant cancer-gene pairs.

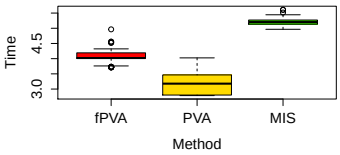
Cancer	Gene				
	PEX5	NRXN2	EIF4G1	PTPN22	ACN9
Breast	0.0009	0.0008	0.0009	0.0030	0.0017
Carcinoid	0.0745	0.0473	0.0868	0.1729	0.0445
Cervical	0.0009	0.0009	0.0010	0.0035	0.0009
Colorectal	0.0035	0.0010	0.0005	0.0060	0.0014
Endometrial	0.0006	0.0008	0.0006	0.0008	0.0009
Glioma	0.0018	0.0063	0.0030	0.5000	0.0005
Head and neck	0.0745	0.0488	0.0445	0.0745	0.0445
Liver	0.0007	0.0005	0.0025	0.0011	0.0009
Lung	0.0010	0.0009	0.0023	0.0024	0.0007
Lymphoma	0.0041	0.0010	0.0009	0.5000	0.0024
Melanoma	0.0010	0.0015	0.0010	0.0024	0.0015
Ovarian	0.0014	0.0016	0.0009	0.0009	0.0018
Pancreatic	0.0030	0.0037	0.0012	0.0023	0.0096
Prostate	0.0008	0.0007	0.0010	0.0359	0.0042
Renal	0.0007	0.0012	0.0017	0.0030	0.0022
Skin	0.0012	0.0024	0.0015	0.0014	0.0015
Stomach	0.0021	0.0007	0.0009	0.0024	0.0018
Testis	0.0012	0.0006	0.0009	0.0037	0.0005
Thyroid	0.0745	0.0445	0.0868	0.0359	0.0359
Urothelial	0.0023	0.0017	0.0012	0.0012	0.0074

Table 5: Average sensitivity and specificity of fPVA with standard deviations(in parentheses) in percentage for the stability analysis.

n	p	SEN	SPE
42	500	66.1(11.06)	99.3(0.66)
	1000	68.0(13.02)	99.4(0.56)
100	500	100(0.00)	99.3(0.74)
	1000	100(0.00)	99.2(0.52)
200	500	100(0.00)	99.1(0.44)
	1000	100(0.00)	98.8(0.31)



(a)



(b)

Figure 2: Box plots of the logarithms of CPU-times in seconds. (a) Example 4.1 with independent covariates when $(t, n, J, p) = (0, 100, 70, 1000)$. (b) Example 4.1 with dependent covariates when $(t, n, J, p) = (1, 100, 70, 1000)$.