Journal of Multivariate Analysis 192 (2022) 105069

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Nonparametric variable screening for multivariate additive models

Hui Ding [a], Jian Zhang [b,*], Riquan Zhang [c]

[a] School of Economics, Nanjing University of Finance and Economics, 3 WenYuan Road, Nanjing, 210023, PR China
[b] School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7FS, United Kingdom
[c] School of Statistics, East China Normal University, 3663 North Zhongshan Road, Shanghai, 200062, PR China

ARTICLE INFO

Article history:
Received 7 January 2021
Received in revised form 15 June 2022
Accepted 15 June 2022
Available online 23 June 2022

AMS 2020 subject classifications:
primary 62R10
secondary 62J05

Keywords:
High-dimensional multivariate data
Multivariate additive models
Nonparametric variable screening
Beamforming

ABSTRACT

In this paper we develop a novel procedure of variable screening for a multivariate additive random-effects model, based on B-spline function approximations. With these approximations, the so-called signal-to-noise ratio (SNR) can be defined to inform the importance of each covariate in the model. Then, SNR-based forward filtering is conducted on covariates by using iterative projections of the multiple response data into the space of covariates. The proposed procedure is easy to use and allows the user to pool non-linear information across heterogeneous subjects through random-effects variables. We establish an asymptotic theory on the selection consistency under some regularity conditions. By simulations, we show that the procedure has a superior performance over some existing methods in terms of sensitivity and specificity. We also apply the procedure to anti-cancer drug data, revealing a set of biomarkers that potentially influence concentrations of anti-cancer drugs in cancer cell lines.

1. Introduction

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

$$\mathbf{y}_j = \boldsymbol{\mu} + f_{1j}(\mathbf{x}_1) + \cdots + f_{pj}(\mathbf{x}_p) + \boldsymbol{\varepsilon}_j, \quad j \in \{1, \ldots, J\}. \tag{1}$$

In the above model, $\boldsymbol{\mu}$ is a vector of fixed-effects, component $f_{kj}(\mathbf{x}_k) = (f_{kj}(x_{1k}), ..., f_{kj}(x_{nk}))^\top$ with $\mathbf{x}_k = (x_{1k}, \ldots, x_{nk})^\top$ is a vector of random-effects of the kth covariate on the jth subject, $\boldsymbol{\varepsilon}_j = (\epsilon_{1j}, \ldots, \epsilon_{nj})^\top$ is a vector of error terms related to the jth subject, and given $\mathbf{X}$, $\boldsymbol{\varepsilon}_j$'s are conditionally independent of $f_{kj}(\mathbf{x}_k)$'s. Here, random-effects $f_{kj}(\mathbf{x}_k)$ are random functions used to account for unobserved heterogeneity and dependence across the subjects. The "multivariate" is for the dimension of $\mathbf{y}_j$. To make the model identifiable, for $J > 1$, we impose the constraint that conditional on $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, for all (k, j), $E[f_{kj}(\mathbf{x}_k)|\mathbf{X}] = 0$, $E[\boldsymbol{\varepsilon}_j|\mathbf{X}] = 0$, $\mathrm{cov}(\boldsymbol{\varepsilon}_j|\mathbf{X}) = \sigma^2 I_n$ with $0 < \sigma^2 < \infty$. We assume that random-effects are sparse in the sense that only for a few k's, where the random-effects matrix $\mathbf{f}_k(\mathbf{x}_k) = (f_{k1}(\mathbf{x}_k), \ldots, f_{kJ}(\mathbf{x}_k))$ has an asymptotically positive variability, $(nJ)^{-1}\sum_{i=1}^n \sum_{j=1}^J (f_{kj}(x_{ik}) - \bar{f}_k(x_{ik}))^2$, across the subjects as n and J tend to infinity, where $\bar{f}_k(x_{ik}) = \sum_{j=1}^J f_{kj}(x_{ik})/J$. Note that model (1) will reduce to a linear random-effects model if $f_{kj}(\cdot)$, $1 \le k \le p$, $1 \le j \le J$ are linear. See Laird and Ware [8] and Lin and Zhang [10]. Unlike the conventional multivariate additive models in Lin and Zhang [11], Rigby and

* Corresponding author.
E-mail address: J.Zhang-79@kent.ac.uk (J. Zhang).

https://doi.org/10.1016/j.jmva.2022.105069

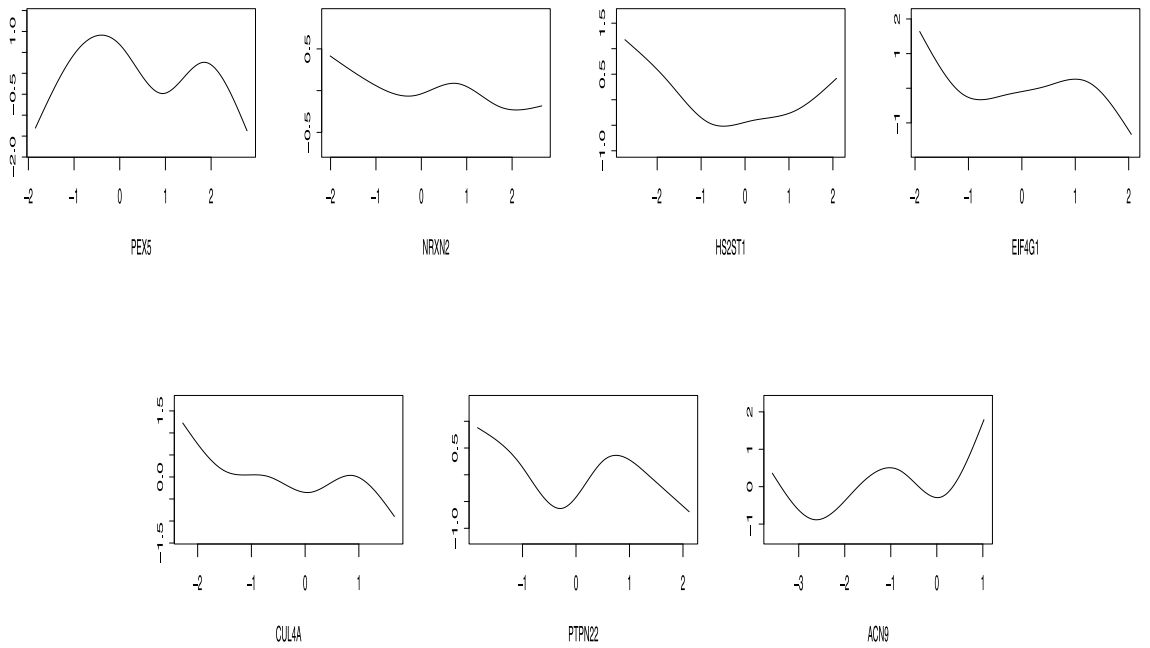

Fig. 1. Estimated nonparametric effects of selected genes on the anti-cancer drug KIN001-135, which were measured by median inhibition concentration (IC50) in a cancer cell line. In each plot, *x*-axis shows the expression level of a selected gene while *y*-axis gives the corresponding value of IC50. From the left to the right and from the top to the bottom, the plots correspond to the selected genes PEXS, NRXN2, HS2ST1, EIF4G1, CUL4 A, PIPN22 and ACN9 respectively.

Stasinopoulos [14], Yee [17] and Yee and Wild [18], the new model allows for subject-specific nonparametric random-effects as well as dependence in the response data, without any joint density assumptions on errors. In practice, the number of covariates can be larger than both n and J. For example, in cancer research, people aim to identify biomarkers (a set of genes or DNA variants) for cancer drug development. The response data considered in this paper 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. See Garnett, et al. [4]. Note that in Section 4.2, we found that the concentration of a drug depended nonlinearly on expression levels of genes as shown in Fig. 1. The purpose of this paper is to provide a general screening method to identify these sparse covariates when p is larger than both n and J.

Extensive research has been conducted on variable selection for univariate additive models (Fan et al. [3], Huang et al. [6], Koltchinskii and Yuan [7], Lin and Zhang [11], Meier et al. [12], Ravikumar et al. [13], Stone [16], Zhang et al. [22], and references therein). Despite of the above progress, the challenging problem of variable selection for the multivariate additive random-effects model remains open in the literature as for multivariate additive models, modeling each univariate additive model marginally and separately is not efficient in exploring dependence across the subjects. In this paper, we propose a novel approach for addressing the above problem. We first approximate each nonparametric component by a linear combination of spline basis functions. With these approximations, the above variable selection problem reduces to that of selecting significant block-matrices of regression coefficients in a multivariate random-effects regression model. Then, we conduct a series of SNR-based filtering operations (Zhang, Liu and Green [20], Zhang and Liu [19]) through projections of $\mathbf{y}_j, 1 \leq j \leq J$ into each covariate space; 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 important covariates detected in the previous steps. The above filtering is based on the so-called covariate power (i.e., variability), which is estimated by minimizing the trace of the sample covariance matrix of the projected data points $\mathrm{W}^\top \mathbf{y}_j$, $1 \leq j \leq J$ with respect to a weighting matrix W. For the kth covariate, the minimization is subject to the constraint that $\mathrm{W}^\top \Psi(\mathbf{x}_k)$ is equal to an identity matrix and subject to that important covariates identified in the previous steps have been nulled, where $\Psi(\mathbf{x}_k)$ is the $n \times \kappa_n$ "design matrix" derived from 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 covariate specific background noises. To adjust for this, we consider the signal-to-noise ratio (SNR) for each covariate. The covariates can then be ranked and selected by thresholding these SNR values. A list of highly ranked covariates called functional principal variables are produced along with the estimated random regression coefficient functions. Based on these selected covariates, a variance-component decomposition of the response covariance matrix can be made. We show that the proposed procedure is asymptotically consistent under some regularity conditions. We demonstrate that the proposed procedure can outperform some existing screening methods in terms of sensitivity and specificity by simulation studies and a real data application.

The rest of the article is organized as follows. In Section 2, we develop the SNR-based screening procedure for multivariate additive models. In Section 3, we establish theoretical properties of the proposed procedure. In Section 4, we conduct simulation studies and a real data analysis. We conclude with a discussion in Section 5. The lemmas and technical proofs are relegated to the Appendix and some numerical details are put in the 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 $\|\mathrm{A}_n\|$ as $\lambda_{\max}^{1/2}(\mathrm{A}_n^\top \mathrm{A}_n)$. 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. Proposed methodology

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

$$\mathbf{Y} = \boldsymbol{\mu}\mathbf{1}_J^\top + \mathbf{f}_1(\mathbf{x}_1) + \cdots + \mathbf{f}_p(\mathbf{x}_p) + \mathrm{E}. \tag{2}$$

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,

$$\Pr\left(\max_{1\le k\le p, 1\le j\le J} \sup_{z, z+\delta_* \in [a,b]} |f_{kj}^{(r)}(z+\delta_*) - f_{kj}^{(r)}(z)| \le c_*|\delta_*|^\alpha\right) = 1$$

for some non-negative integer r is a non-negative integer and universal constants $0 < \alpha \le 1$ and c_*.

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 $c_1 n^{-\nu} \le \min_{0\le k\le N} |z_k - z_{k+1}| \le \max_{0\le k\le N} |z_k - z_{k+1}| \le c_2 n^{-\nu}$ with $0 \le \nu < 0.5$, 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, $k = 0, \ldots, \kappa_n$ with $\kappa_n = N + 2m - 1$. Following de Boor [2], 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 functions, $\tilde{f}_{kj}(x) = \sum_{d=1}^{\kappa_n} \beta_{kjd}\psi_d(x)$ such that

$$\Pr\left(\max_{1\le k\le p, 1\le j\le J} \sup_{a\le x\le b} |f_{kj}(x) - \tilde{f}_{kj}(x)| \le c_3\kappa_n^{-r_0}\right) = 1, \tag{3}$$

for $r_0 = r + \alpha$ and a universal constant c_3. Using these approximations, we can reformulate Eq. (2) as follows:

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

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

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

2.1. SNR indices for covariates

Let $\bar{\mathbf{Y}} = (\sum_{j=1}^J \mathbf{y}_j/J)\mathbf{1}_J^\top$ and $\hat{\mathrm{C}} = (\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})^\top/J$. To define a power index for the kth covariate, we project the data $\mathbf{Y}$ into the covariate space generated by $\mathbf{x}_k$ with an $n \times \kappa_n$ direction matrix W, namely, $\mathrm{W}^\top\mathbf{Y}$, subject to $\mathrm{W}^\top\Psi(\mathbf{x}_k) = I_{\kappa_n}$. 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 $\mathrm{W}^\top\mathbf{Y}$, $\mathrm{tr}(\mathrm{W}^\top(\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})^\top\mathrm{W})/J$, is minimized, subject to $\mathrm{W}^\top\Psi(x_k) = \mathrm{I}_{\kappa_n}$. This gives an optimal solution $\hat{\mathrm{W}} = \hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_k)(\Psi(\mathbf{x}_k)^\top\hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_k))^{-1}$ with the variability

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

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

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

which shows the signal strength of the kth covariate relative to the white noise.

Note that covariates can be correlated to each other. To reduce such an effect on covariate ranking, we adopt a nulled SNR as follows. Let ω and ν be two non-overlapped subsets of covariates with sizes m_1 and m respectively. Merging $\Psi(\mathbf{x}_\nu)$ with $\Psi(\mathbf{x}_\omega)$ defines matrix $\Psi(\mathbf{x}_{\nu\cup\omega}) = (\Psi(\mathbf{x}_\nu), \Psi(\mathbf{x}_\omega))$. Similarly, for $\nu = \{k_1, \ldots, k_m\}$, we define $\Psi(\mathbf{x}_\nu) = \big(\Psi(\mathbf{x}_{k_1}), \ldots, \Psi(\mathbf{x}_{k_m})\big)$ by merging $\Psi(\mathbf{x}_k)$, $k \in \nu$. To null the effects of covariates in ω, we choose W in which the trace of the sample covariance matrix of the projected data $\mathrm{W}^\top\mathbf{Y}$, $\mathrm{tr}(\mathrm{W}^\top(\mathbf{Y}-\bar{\mathbf{Y}})(\mathbf{Y}-\bar{\mathbf{Y}})^\top\mathrm{W})/J$, is minimized, subject to $\mathrm{W}^\top\Psi(\mathbf{x}_\nu) = \mathbf{1}_m^\top \otimes \mathrm{I}_{\kappa_n}$ and $\mathrm{W}^\top\Psi(\mathbf{x}_\omega) = \mathbf{0}_{m_1}^\top \otimes \mathrm{I}_{\kappa_n}$, 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

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

and the SNR of ν after nulling ω,

$$\begin{aligned}\mathrm{SNR}_{\nu|\omega} = \mathrm{tr}\Bigg\{&\left(\mathbf{e}_{\nu|\omega}^\top\left(\Psi(\mathbf{x}_{\nu\cup\omega})^\top\hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_{\nu\cup\omega})\right)^{-1}\mathbf{e}_{\nu|\omega}\right)\left(\mathbf{e}_{\nu|\omega}^\top\left(\Psi(\mathbf{x}_{\nu\cup\omega})^\top\hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_{\nu\cup\omega})\right)^{-1}\right.\\ &\left.\times\left(\Psi(\mathbf{x}_{\nu\cup\omega})^\top\hat{\mathrm{C}}^{-2}\Psi(\mathbf{x}_{\nu\cup\omega})\right)\left(\Psi(\mathbf{x}_{\nu\cup\omega})^\top\hat{\mathrm{C}}^{-1}\Psi(\mathbf{x}_{\nu\cup\omega})\right)^{-1}\mathbf{e}_{\nu|\omega}\right)^{-1}\Bigg\},\end{aligned} \tag{4}$$

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

2.2. Nulled-beamforming procedure

Based on the SNR indices, we define a nulled-beamforming procedure called functional Principal Variable Analysis (fPVA) for identifying important covariates as follows.

Initialization: Find k_1 at which the SNR_{k_1} attains the maximum. Set $\omega_1 = \{k_1\}$ as the current selected covariate set. We call the complement of ω_1, ω_1^c the current remaining covariate set.

Sequentially screening: In the mth iteration with $m \geq 2$, 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 (4), we calculate $\mathrm{SNR}_{k|\omega_{m-1}}$. If these nulled SNR values do not satisfy the stopping rule below, find $k_m \notin \omega_{m-1}$ in which the nulled index $\mathrm{SNR}_{k_m|\omega_{m-1}}$ attains the maximum. We update the current selected covariate set ω_{m-1}, ω_{m-1}^c and $\Psi(\mathbf{x}_{\omega_{m-1}})$ by letting $\omega_m = \{k_m\} \cup \omega_{m-1}$, $\omega_m^c = \omega_{m-1}^c \setminus \{k_m\}$ and $\Psi(\mathbf{x}_{\omega_m}) = (\Psi(\mathbf{x}_{k_m}), \Psi(\mathbf{x}_{\omega_{m-1}}))$. Otherwise, we terminate the iteration and let $\omega_m = \omega_{m-1}$, $\omega_m^c = \omega_{m-1}^c$ 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 of the remaining covariates to identify an elbow point which partitions the current remaining covariate set 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 whether the upper set is also uninformative, we calculate the mean μ_l and standard deviation σ_l for the lower subset. The upper set is declared uninformative if the maximum nulled SNR value, SNR_{max}, in the upper set falls into the following confidence interval, $[\mu_l - c_0\sigma_l, \mu_l + c_0\sigma_l]$, where c_0 is a tuning constant. If the upper subset is uninformative, then the current remaining covariate set is also uninformative and the iteration will be terminated. We set the default value $c_0 = 3.5$ for the above tuning constant at the asymptotic confidence level of 99.95%. The sensitivity of this choice will be further investigated in the simulation studies. See Section A in the Supplementary Material.

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 $\hat{\mathrm{C}}$. When $n > J$, $\hat{\mathrm{C}}$ is degenerate, leading to an ill-posed definition of the power. To address this issue, we consider a thresholded estimator introduced by Bickel and Levina [1]:

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

where $I(\cdot)$ is the indicator and $\tau_{nJ} = \sqrt{\log(n)/J}$ with $h \geq 0$ being a constant (for example, $h = 0.01|\mathrm{tr}(\hat{C})/n|$). The thresholded covariance matrix estimators may not be positive definite when the dimension J is close to or smaller than n. To remedy the problem, we shrink the above thresholded covariance estimator to a diagonal matrix by using the method of Ledoit and Wolf [9] as follows:

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

where

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

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

3. Asymptotic analysis

Our analysis follows a standard framework for high-dimensional asymptotics, where the dimension p grows at a rate much bigger than both n and J as n and J tend to infinity. We need the following assumption on the stationarity of the response data and the sparsity of the model.

C2: There exists a permutation π on index j so that for each n, time series $\mathbf{y}_{\pi(j)}$, $1 \le j \le J$ is strictly stationary with conditional marginal covariance matrix C given $\mathbf{X}$. We assume that there are only q_n non-zero components in the model (1).

Without loss of generality, we let the first q_n components are non-zeros: $f_{kj}(x) \neq 0$, $1 \le k \le q_n$, but $f_{kj}(x) \equiv 0$, $q_n + 1 \le k \le p$. Let $\nu_0 = \{1, 2, \ldots, q_n\}$, $|\nu_0| = q_n$ and $f^{\oplus}_{\nu_0 j}(\mathbf{x}_{\nu_0}) = f_{1j}(\mathbf{x}_1) + \cdots + f_{q_n j}(\mathbf{x}_{q_n})$. Under Condition C2, $f^{\oplus}_{\nu_0 j}(\mathbf{x}_{\nu_0})$, $1 \le j \le J$ are strictly stationary. Let $\Delta_{ij} = \sum_{k=1}^{q_n}\left(f(x_{ik}) - \sum_{d=1}^{\kappa_n}\beta_{kjd}\psi_d(x_{ik})\right)$, $\Delta_j = (\Delta_{1j}, \ldots, \Delta_{nj})^\top$, $\mathbf{b}_{kj} = (\beta_{kj1}, \ldots, \beta_{kj\kappa_n})^\top$, $\mathrm{B}_{\nu_0 j} = (\mathbf{b}_{1j}^\top, \ldots, \mathbf{b}_{q_n j}^\top)^\top$ and $\Psi(\mathbf{x}_{\nu_0}) = (\Psi(\mathbf{x}_1), \ldots, \Psi(\mathbf{x}_{q_n}))$. Let $\Sigma \hat{=} \mathrm{cov}(\mathrm{B}_{(1:p)j}) = (\Sigma_{k_1k_2})_{p\times p}$ be a $(p\kappa_n \times p\kappa_n)$ block matrix with block $\Sigma_{k_1k_2} = \mathrm{cov}(\mathbf{b}_{k_1 j}, \mathbf{b}_{k_2 j})$. Similarly, let $\mathrm{cov}(\mathrm{B}_{\nu_0 j})$ denote the $(|\nu_0|\kappa_n \times |\nu_0|\kappa_n)$ block matrix $(\Sigma_{k_1k_2})_{k1,k2\in\nu_0}$.

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

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

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

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

$$\mathbf{a}^\top\Psi(\mathbf{x}_{\nu_0})E[\mathrm{B}_{\nu_0 j}|\mathbf{X}]E[\mathrm{B}_{\nu_0 j}^\top|\mathbf{X}]\Psi(\mathbf{x}_{\nu_0})^\top\mathbf{a} = \mathbf{a}^\top E[\Delta_j|\mathbf{X}]E[\Delta_j|\mathbf{X}]^\top\mathbf{a} \le c_3^2 n q_n^2 \kappa_n^{-2r_0}.$$

It also follows from the inequality (3) 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 $\mathbf{a} \in \mathbb{R}^n$, $\|\mathbf{a}\|_2 = 1$,

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

Consequently, there exists a constant c_4 such that

$$\mathrm{C} \le \Psi(\mathbf{x}_{\nu_0})\mathrm{cov}(\mathrm{B}_{\nu_0 j}|\mathbf{X})\Psi(\mathbf{x}_{\nu_0})^\top + (\sigma^2 + c_4 n q_n^2\kappa_n^{-r_0})\mathrm{I}_n. \tag{5}$$

To regularize the coherence structure of the design matrices, we impose the following condition on the covariates.

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 $\{1, \ldots, p\}$. Let E_ν be a selection matrix, made of p block matrices sitting next to each other, where for $j \in \{1, \ldots, p_1\}$, 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 $\Psi(\mathbf{x}_\nu) = \Psi(\mathbf{X})\mathrm{E}_\nu^\top$. Let $\mathrm{A}_\nu = \mathrm{C} - \Psi(\mathbf{x}_\nu)\mathrm{E}_\nu^\top\Sigma\mathrm{E}_\nu\Psi(\mathbf{x}_\nu)^\top$, which is the residual variance–covariance 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 proportional to an identity matrix as stated in the following condition:

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.

Let $\delta = (1 - K_1/K_2)^{1/2}$ in the following proposition which states that under Conditions C1–C3, as $n \to \infty$, the estimated power at covariate set ν_0 converges to its underlying power, the trace of the covariance matrix of regression coefficients of ν_0.

Proposition 1. *Under Conditions C1–C3, for any ν, if $E_\nu^\top \Sigma E_\nu$ and A_ν are invertible, then the power $\gamma_\nu = tr\{E_\nu^\top \Sigma E_\nu + \left(\Psi(\mathbf{x}_\nu)^\top A_\nu^{-1}\Psi(\mathbf{x}_\nu)\right)^{-1}\}$. In particular, for the index set, ν_0, of non-zero components, $\gamma_{\nu_0} = tr(E_{\nu_0}^\top \Sigma E_{\nu_0}) + O\left(\delta^{-q_n/2} q_n^2 \kappa_n^{2-r_0}\right)$. If κ_n takes the optimal rate $n^{1/(2r_0+1)}$ with $r_0 > 2$, $q_n \le \eta_0 \log(n)$ and $0 < \eta_0 < 2(r_0-2)/\{(2r_0+1)\log(\delta^{-1})\}$, then $\gamma_{\nu_0} = tr(E_{\nu_0}^\top \Sigma E_{\nu_0}) + O\left(n^{-\delta_0}\right)$, where $\delta_0 = (r_0-2)/(2r_0+1) - 0.5\eta_0 \log(\delta^{-1})$.*

Proof. For a general ν, invoking the Woodbury matrix identity, we have

$$\mathrm{C}^{-1} = \left(\mathrm{A}_\nu + \Psi(\mathbf{x}_\nu)\mathrm{E}_\nu^\top \Sigma \mathrm{E}_\nu \Psi(\mathbf{x}_\nu)^\top\right)^{-1} = \mathrm{A}_\nu^{-1} - \mathrm{A}_\nu^{-1}\Psi(\mathbf{x}_\nu)\left\{\left(\mathrm{E}_\nu^\top \Sigma \mathrm{E}_\nu\right)^{-1} + \Psi(\mathbf{x}_\nu)^\top \mathrm{A}_\nu^{-1}\Psi(\mathbf{x}_\nu)\right\}^{-1}\Psi(\mathbf{x}_\nu)^\top \mathrm{A}_\nu^{-1}.$$

Then

$$\begin{aligned}\Psi(\mathbf{x}_\nu)^\top \mathrm{C}^{-1}\Psi(\mathbf{x}_\nu) &= \Psi(\mathbf{x}_\nu)^\top \mathrm{A}_\nu^{-1}\Psi(\mathbf{x}_\nu) - \Psi(\mathbf{x}_\nu)^\top \mathrm{A}_\nu^{-1}\Psi(\mathbf{x}_\nu)\left\{\left(\mathrm{E}_\nu^\top \Sigma \mathrm{E}_\nu\right)^{-1} + \Psi(\mathbf{x}_\nu)^\top \mathrm{A}_\nu^{-1}\Psi(\mathbf{x}_\nu)\right\}^{-1}\Psi(\mathbf{x}_\nu)^\top \mathrm{A}_\nu^{-1}\Psi(\mathbf{x}_\nu)\\ &= \left\{\mathrm{E}_\nu^\top \Sigma \mathrm{E}_\nu + \left(\Psi(\mathbf{x}_\nu)^\top \mathrm{A}_\nu^{-1}\Psi(\mathbf{x}_\nu)\right)^{-1}\right\}^{-1}.\end{aligned}$$

Thus,

$$\gamma_\nu = \mathrm{tr}\left((\Psi(\mathbf{x}_\nu)^\top \mathrm{C}^{-1}\Psi(\mathbf{x}_\nu))^{-1}\right) = \mathrm{tr}\left\{\mathrm{E}_\nu^\top \Sigma \mathrm{E}_\nu + \left(\Psi(\mathbf{x}_\nu)^\top \mathrm{A}_\nu^{-1}\Psi(\mathbf{x}_\nu)\right)^{-1}\right\}. \tag{6}$$

In particular, when $\nu = \nu_0$, it follows from Eqs. (5) and (6) that

$$\mathrm{A}_{\nu_0} = \mathrm{C} - \Psi(\mathbf{x}_{\nu_0})\mathrm{cov}(\mathrm{B}_{\nu_0 j})\Psi(\mathbf{x}_{\nu_0})^\top \le (\sigma^2 + c_4 n q_n^2 \kappa_n^{-r_0})\mathrm{I}_n.$$

$$|\gamma_{\nu_0} - \mathrm{tr}\left(\mathrm{E}_{\nu_0}^\top \Sigma \mathrm{E}_{\nu_0}\right)| \le \kappa_n \left(\lambda_{\min}(\Psi(\mathbf{x}_{\nu_0})^\top \Psi(\mathbf{x}_{\nu_0}))/n\right)^{-1}(\sigma^2/n + c_4 q_n^2 \kappa_n^{-r_0}). \tag{7}$$

Under Condition C3, combining Lemma 1 of Stone [16] with Lemma 3 of Huang et al. [6], we show that there exists a constant $c_5 > 0$ such that

$$\lambda_{\min}\left(\Psi(\mathbf{x}_{\nu_0})^\top \Psi(\mathbf{x}_{\nu_0})/n\right) \ge c_5 \delta^{(q_n-1)/2}\kappa_n^{-1}. \tag{8}$$

This together with Eq. (7) yields that

$$|\gamma_{\nu_0} - \mathrm{tr}(\mathrm{E}_{\nu_0}^\top \Sigma \mathrm{E}_{\nu_0})| \le c_5^{-1}\kappa_n \delta^{-(q_n-1)/2}\kappa_n(\sigma^2/n + c_4 q_n^2 \kappa_n^{-r_0}) = O(\delta^{-q_n/2} q_n^2 \kappa_n^{2-r_0}).$$

The remaining proof is straightforward. The proof is completed. □

To present Theorem 1, we first introduce some notations. For any subset of covariates, ν, we define the following coherence matrices for $\Psi(\mathbf{x}_\nu)$ and $\Psi(\mathbf{x}_{\nu_0})$: $\mathrm{R}_{\nu\nu} = \Psi(\mathbf{x}_\nu)^\top \mathrm{A}_{\nu_0}^{-1}\Psi(\mathbf{x}_\nu)/n$, $\mathrm{R}_{\nu\nu_0} = \Psi(\mathbf{x}_\nu)^\top \mathrm{A}_{\nu_0}^{-1}\Psi(\mathbf{x}_{\nu_0})/n$ and $\mathrm{R}_{\nu_0\nu_0} = \Psi(\mathbf{x}_{\nu_0})^\top \mathrm{A}_{\nu_0}^{-1}\Psi(\mathbf{x}_{\nu_0})/n$. 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 $\mathrm{E}_{\nu|\nu_0}$ be 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 $\nu \subseteq \{1, \ldots, p\} \setminus \nu_0$, $\lambda_{\max}(\mathrm{R}_{\nu\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-2}\mathrm{R}_{\nu_0\nu}) = 0(1)$, $\lambda_{\max}(\mathrm{F}_\nu^{-1/2}\mathrm{R}_{\nu\nu}\mathrm{F}_\nu^{-1/2}) = 0(1)$, $n\lambda_{\min}(\mathrm{F}_\nu) \to \infty$, where $\mathrm{F}_\nu = \mathrm{R}_{\nu\nu} - \mathrm{R}_{\nu\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu}$.

To verify Condition C5, we refer readers to Zhang and Oftadeh [21]. Under the assumptions that $\Psi(\mathbf{x}_\nu)$ is a random matrix satisfying some moment conditions and that the growth of the total number of covariates, p, is not too fast compared to the number of measurements per response, n. The details are as follows. Let $\kappa_n = O(n^{1/(2r_0+1)})$ and $\Omega_n = \{\nu : \nu \subseteq \{1, \ldots, p\}, \kappa_n|\nu| \le [rn]\}$, where $[rn]$ stands for the integer part of rn. Assume that $\Psi(\mathbf{x}_\nu)$ has a concentration property, i.e., for some constant c_1, any $u > 0$ and $\nu \subseteq \{1, \ldots, p\}$, $|\nu|\kappa_n \le rn$, $0 < r < 1$,

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

Then, it follows from Zhang and Oftadeh [21] that for a small positive constant α_0,

$$\Pr\left(\max_{\nu\in\Omega_n} \lambda_{\max}(\mathrm{R}_{\nu\nu}) > n^{\alpha_0} \text{or } \min_{\nu\in\Omega_n} \lambda_{\min}(\mathrm{R}_{\nu\nu}) < n^{-\alpha_0}\right) \to 0$$

as $\log(p) \le O(n^{\alpha_0})$, n and p tend to infinity. This implies that Condition C5 holds with probability tending to one when p is not bigger than $O(\exp(n^{\alpha_0}))$. Note that $\|\mathrm{R}_{\nu\nu_0}\| \le \|\mathrm{R}_{\nu\nu}\|^{1/2}\|\mathrm{R}_{\nu_0\nu_0}\|^{1/2}$. Choosing a small α_0 such that $\log(\delta) + 1 + 1/(2r_0+1) - \alpha_0 \ge 0$ and invoking inequality (8) and $q_n = O(\log(n))$, we have

$$\lambda_{\max}(\mathrm{R}_{\nu\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-2}\mathrm{R}_{\nu_0\nu}) \le \lambda_{\min}(\mathrm{R}_{\nu_0\nu_0})^{-2}\|\mathrm{R}_{\nu\nu_0}\mathrm{R}_{\nu_0\nu})\| = O(n^{-2(\log(\delta)+1+1/(2r_0+1)-\alpha_0)})$$

which holds with an overwhelming probability.

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. Note that Condition $\delta^{-q_n/2}q_n^2\kappa_n^{1-r_0} = o(1)$ is satisfied if let $q_n = O(\log(n))$ and κ_n the optimal rate $n^{1/(2r_0+1)}$. A guideline for the choice of κ_n is provided in the Section of Numerical Studies below.

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

(i) For any $\nu \subseteq \nu_0$,

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

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

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

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

where

$$D_{\nu n} = F_\nu^{-1}R_{\nu\nu_0}R_{\nu_0\nu_0}^{-1/2}\left\{R_{\nu_0\nu 0}^{-1/2}\left(E_{\nu_0}^\top \Sigma E_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1/2}n^{-1} + I_{|\nu_0|\kappa_n}\right\}^{-1} R_{\nu_0\nu_0}^{-1/2}\left(E_{\nu_0}^\top \Sigma E_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu}F_\nu^{-1}(1+o(1))$$
$$\le F_\nu^{-1}R_{\nu\nu_0}R_{\nu_0\nu_0}^{-1}\left(E_{\nu_0}^\top \Sigma E_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu}F_\nu^{-1}(1+o(1)).$$

Proof. Note that $\mathbf{Q} = \mathrm{A}_{\nu_0}^{-1/2}\Psi(\mathbf{x}_{\nu_0})(\Psi(\mathbf{x}_{\nu_0})^\top \mathrm{A}_{\nu_0}^{-1}\Psi(\mathbf{x}_{\nu_0}))^{-1}\Psi(\mathbf{x}_{\nu_0})^\top \mathrm{A}_{\nu_0}^{-1/2}$ is idempotent, i.e., $\mathbf{Q}^2 = \mathbf{Q}$. Its eigenvalues are either 1 and 0. This implies $\mathbf{Q} \le \mathrm{I}_n$. Therefore,

$$\mathrm{R}_{\nu\nu} - \mathrm{R}_{\nu\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu} = \Psi(\mathbf{x}_\nu)^\top \mathrm{A}_{\nu_0}^{-1/2}(\mathrm{I}_n - \mathbf{Q})\mathrm{A}_{\nu_0}^{-1/2}\Psi(\mathbf{x}_\nu)/n \ge \mathbf{0},$$

i.e., $\mathrm{R}_{\nu\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu} \le \mathrm{R}_{\nu\nu}$. Moreover, using (5), Lemma 1 of Stone [16] and Lemma 3 of Huang et al. [6], we have that

$$\mathrm{R}_{\nu_0\nu_0}^{-1} \le \left(\Psi(\mathbf{x}_{\nu_0})^\top\Psi(\mathbf{x}_{\nu_0})/n\right)^{-1}(\sigma^2 + c_4nq_n^2\kappa_n^{-r_0}) \le c_5^{-1}\delta^{-(q_n-1)/2}\kappa_n(\sigma^2 + c_4nq_n^2\kappa_n^{-r_0})\mathrm{I}_{|\nu_0|\kappa_n}. \tag{9}$$

Therefore,

$$n^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1/2}\left(\mathrm{E}_{\nu_0}^\top \Sigma \mathrm{E}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1/2} \le 0(1)\lambda_{\text{mino}}^{-1}\delta^{-q_n/2}q_n^2\kappa_n^{1-r_0}\mathrm{I}_{|\nu_0|\kappa_n} = o(1). \tag{10}$$

(i) For any $\nu \subseteq \nu_0$, it follows from Lemma 1 and the inequality (9) that

$$\Psi(\mathbf{x}_\nu)^\top \mathrm{C}^{-1}\Psi(\mathbf{x}_\nu) = \mathrm{E}_{\nu|\nu_0}^\top\left(\mathrm{E}_{\nu_0}^\top\Sigma\mathrm{E}_{\nu_0}\right)^{-1}\mathrm{E}_{\nu|\nu_0} - (T_1 - T_2),$$

where

$$T_1 = \mathrm{E}_{\nu|\nu_0}^\top\left(\mathrm{E}_{\nu_0}^\top\Sigma\mathrm{E}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\mathrm{E}_{\nu_0}^\top\Sigma\mathrm{E}_{\nu_0}\right)^{-1}\mathrm{E}_{\nu|\nu_0}n^{-1} \le \lambda_{\text{mino}}^{-2}\mathrm{I}_{|\nu|\kappa_n}\delta^{-q_n/2}q_n^2\kappa_n^{1-r_0}O(1),$$
$$T_2 = \mathrm{E}_{\nu|\nu_0}^\top\left(\mathrm{E}_{\nu_0}^\top\Sigma\mathrm{E}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\mathrm{E}_{\nu_0}^\top\Sigma\mathrm{E}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\mathrm{E}_{\nu_0}^\top\Sigma\mathrm{E}_{\nu_0}\right)^{-1}\mathrm{E}_{\nu|\nu_0}n^{-2}$$
$$\times(1+0(1)\lambda_{\text{mino}}^{-1}\delta^{-q_n/2}q_n^2\kappa_n^{1-r_0}) \le \lambda_{\text{mino}}^{-3}\mathrm{I}_{|\nu|\kappa_n}\delta^{-q_n}q_n^4\kappa_n^{2(1-r_0)}O(1). \tag{11}$$

Note that $\Sigma_{\nu|\nu_0}^{-1} \ge \lambda_{\text{maxo}}^{-1}\mathrm{I}_{|\nu|\kappa_n}$. This together with the equations in (10) and (11) implies that

$$\Sigma_{\nu|\nu_0}T_1\Sigma_{\nu|\nu_0}T_1\Sigma_{\nu|\nu_0} \le \lambda_{\text{maxo}}\Sigma_{\nu|\nu_0}T_1^2\Sigma_{\nu|\nu_0} \le \lambda_{\text{maxo}}\lambda_{\text{mino}}^{-4}\delta^{-q_n}q_n^4\kappa_n^{2(1-r_0)}O(1)\Sigma_{\nu|\nu_0}^2$$
$$\Sigma_{\nu|\nu_0}T_2\Sigma_{\nu|\nu_0} \le \lambda_{\text{maxo}}^2\lambda_{\text{mino}}^{-3}\delta^{-q_n}q_n^4\kappa_n^{2(1-r_0)}O(1).$$

Using the Woodbury matrix identity and the above inequalities, we obtain

$$\left(\Psi(\mathbf{x}_\nu)^\top\mathrm{C}^{-1}\Psi(\mathbf{x}_\nu)\right)^{-1} = \Sigma_{\nu|\nu_0}^{1/2}\left(I_{|\nu|\kappa_n} - \Sigma_{\nu|\nu_0}^{1/2}(T_1-T_2)\Sigma_{\nu|\nu_0}^{1/2}\right)^{-1}\Sigma_{\nu|\nu_0}^{1/2}$$
$$= \Sigma_{\nu|\nu_0} + \Sigma_{\nu|\nu_0}T_1\Sigma_{\nu|\nu_0} + \lambda_{\text{maxo}}^2\lambda_{\text{mino}}^{-3}(1+\lambda_{\text{maxo}}\lambda_{\text{mino}}^{-1})\delta^{-q_n}q_n^4\kappa_n^{2(1-r_0)}O(1)\mathrm{I}_{|\nu|\kappa_n}.$$

This implies that

$$\gamma_\nu = \mathrm{tr}(\Sigma_{\nu|\nu_0}) + \mathrm{tr}(\Sigma_{\nu|\nu_0}T_1\Sigma_{\nu|\nu_0}) + \lambda_{\text{maxo}}^2\lambda_{\text{mino}}^{-3}(1+\lambda_{\text{maxo}}\lambda_{\text{mino}}^{-1})\delta^{-q_n}q_n^5\kappa_n^{2(1-r_0)+1}O(1).$$

(ii) For any $\nu \subseteq \{1,\ldots,p\}\backslash\nu_0$, it follows from Lemma 1 that

$$\Psi(\mathbf{x}_\nu)^\top \mathrm{C}^{-1}\Psi(\mathbf{x}_\nu) = n\mathrm{F}_\nu + \mathrm{R}_{\nu\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1/2}\left\{\mathrm{R}_{\nu_0\nu 0}^{-1/2}\left(\mathrm{E}_{\nu_0}^\top \Sigma \mathrm{E}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1/2}n^{-1} + \mathrm{I}_{|\nu_0|\kappa_n}\right\}^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1/2}\left(\mathrm{E}_{\nu_0}^\top \Sigma \mathrm{E}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu}.$$

The desired result then follows from the Woodbury matrix identity and the Taylor expansion. The proof is completed. □

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

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

(i) For $a \in \{1,\ldots,p\}\setminus\nu_0$, $a \notin \nu_2$, $\nu_1 \subseteq \nu_0$ and $\nu_2 \subseteq \{1,\ldots,p\}\setminus\nu_0$, $SNR_{a|\nu_1\cup\nu_2} = \frac{\kappa_n}{\zeta_0\sigma^2}(1+o(1))$.

(ii) For $a \in \nu_0$, $a \notin \nu_1$, $\nu_1 \subseteq \nu_0$ and $\nu_2 \subseteq \{1,\ldots,p\}\setminus\nu_0$,

$$SNR_{a|\nu_1\cup\nu_2} = \frac{n(1+o(1))}{\sigma^2\zeta_0}tr\left\{\left(E_{\{a\}|\nu_0}^\top \Sigma_{\nu_0\backslash\nu_1}^{-1}\Phi_0\Sigma_{\nu_0\backslash\nu_1}^{-1}E_{\{a\}|\nu_0}\right)^{-1}E_{\{a\}|\nu_0}^\top \Sigma_{\nu_0\backslash\nu_1}^{-1}E_{\{a\}|\nu_0}\right\} + \frac{1+o(1)}{\sigma^2\zeta_0}tr\left(\Phi_2\right),$$

where

$$\begin{aligned}
\Sigma_{\nu_1|\nu_0} &= \left(E_{\nu_1|\nu_0}^\top\left(E_{\nu_0}^\top\Sigma E_{\nu_0}\right)^{-1}E_{\nu_1|\nu_0}\right)^{-1}, \quad \Sigma_{\nu_0\backslash\nu_1}^{-1} = \left(E_{\nu_0}^\top\Sigma E_{\nu_0}\right)^{-1/2}P_{\nu_0\backslash\nu_1}\left(E_{\nu_0}^\top\Sigma E_{\nu_0}\right)^{-1/2},\\
P_{\nu_0\backslash\nu_1} &= I_{|\nu_0|\kappa_n} - \left(E_{\nu_0}^\top\Sigma E_{\nu_0}\right)^{-1/2}E_{\nu_1|\nu_0}\Sigma_{\nu_1|\nu_0}E_{\nu_1|\nu_0}^\top\left(E_{\nu_0}^\top\Sigma E_{\nu_0}\right)^{-1/2},\\
F_{\nu_2} &= R_{\nu_2\nu_2} - R_{\nu_2\nu_0}R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu_2}, \quad \Phi_0 = R_{\nu_0\nu_0}^{-1} + R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu_2}F_{\nu_2}^{-1}R_{\nu_2\nu_0}R_{\nu_0\nu_0}^{-1},\\
\Phi_1 &= \left(I_{|\nu_0|\kappa_n} - E_{\nu_1|\nu_0}\Sigma_{\nu_1|\nu_0}E_{\nu_1|\nu_0}^\top\left(E_{\nu_0}^\top\Sigma E_{\nu_0}\right)^{-1}\right)\Phi_0\left(I_{|\nu_0|\kappa_n} - E_{\nu_1|\nu_0}\Sigma_{\nu_1|\nu_0}E_{\nu_1|\nu_0}^\top\left(E_{\nu_0}^\top\Sigma E_{\nu_0}\right)^{-1}\right)^\top.\\
\Phi_2 &= E_{\{a\}|\nu_0}^\top\left(E_{\nu_0}^\top\Sigma E_{\nu_0}\right)^{-1}\Phi_1\left(E_{\nu_0}^\top\Sigma E_{\nu_0}\right)^{-1}E_{\{a\}|\nu_0}\left(E_{\{a\}|\nu_0}^\top\Sigma_{\nu_0\backslash\nu_1}^{-1}\Phi_0\Sigma_{\nu_0\backslash\nu_1}^{-1}E_{\{a\}|\nu_0}\right)^{-1}\left(E_{\{a\}|\nu_0}^\top\Sigma_{\nu_0\backslash\nu_1}^{-1}E_{\{a\}|\nu_0}\right)^2.
\end{aligned}$$

The above theorem indicates that the nulled-SNR index contrast between active and non-active covariates tends to infinity if

$$\frac{n}{\kappa_n}\mathrm{tr}\left\{\left(\mathrm{E}_{\{a\}|\nu_0}^\top\Sigma_{\nu_0\backslash\nu_1}^{-1}\Phi_0\Sigma_{\nu_0\backslash\nu_1}^{-1}\mathrm{E}_{\{a\}|\nu_0}\right)^{-1}\mathrm{E}_{\{a\}|\nu_0}^\top\Sigma_{\nu_0\backslash\nu_1}^{-1}\mathrm{E}_{\{a\}|\nu_0}\right\} \to \infty \tag{12}$$

holds uniformly for any $a \in \nu_0$, $a \notin \nu_1$, $\nu_1 \subseteq \nu_0$ and $\nu_2 \subseteq \{1,\ldots,p\}\setminus\nu_0$, as n tends to infinity. Therefore, under the above condition, the covariate set selected by the nulled-SNR is consistent with the true one when n tends to infinity.

Proof. To prove the part (i), let $\nu = \nu_1 \cup \{a\} \cup \nu_2$ and abusing the notation, let $\mathbf{e}_{a|\nu_1\cup\nu_2}^\top = (\mathbf{1}_{|\nu_1|}^\top \otimes \underline{0}_{\kappa_n}, \mathbf{1}_{|\nu_2|}^\top \otimes \underline{0}_{\kappa_n}, \mathrm{I}_{\kappa_n})$ and $\mathbf{e}_{a|\nu_2}^\top = (\mathbf{1}_{|\nu_2|}^\top \otimes \underline{0}_{\kappa_n}^\top, \mathrm{I}_{\kappa_n})$. Let α_n be an arbitrary constant of order $0(1)\delta^{-q_n/2}q_n^2\kappa_n^{1-r_0}$. Then, we have

$$\mathbf{w}_{a|\nu_1\cup\nu_2}^\top\mathbf{w}_{a|\nu_1\cup\nu_2} = \mathbf{e}_{a|\nu_1\cup\nu_2}^\top\gamma_\nu\Psi(\mathbf{x}_\nu)^\top\mathrm{C}^{-2}\Psi(\mathbf{x}_\nu)\gamma_\nu\mathbf{e}_{a|\nu_1\cup\nu_2}. \tag{13}$$

Note that $\Psi(\mathbf{x}_\nu)^\top\mathrm{C}^{-1}\Psi(\mathbf{x}_\nu)$ can be naturally partitioned as follows:

$$\Psi(\mathbf{x}_\nu)^\top\mathrm{C}^{-1}\Psi(\mathbf{x}_\nu) = \begin{pmatrix} \Psi(\mathbf{x}_{\nu_1})^\top\mathrm{C}^{-1}\Psi(\mathbf{x}_{\nu_1}) & \Psi(\mathbf{x}_{\nu_1})^\top\mathrm{C}^{-1}\Psi(\mathbf{x}_{\nu_2\cup\{a\}})\\ \Psi(\mathbf{x}_{\nu_2\cup\{a\}})^\top\mathrm{C}^{-1}\Psi(\mathbf{x}_{\nu_1}) & \Psi(\mathbf{x}_{\nu_2\cup\{a\}})^\top\mathrm{C}^{-1}\Psi(\mathbf{x}_{\nu_2\cup\{a\}})\end{pmatrix},$$

Following the same block dimensions as above, we partition γ_ν and $\mathbf{x}_\nu^\top\mathrm{C}^{-2}\mathbf{x}_\nu$, namely

$$\gamma_\nu = \begin{pmatrix}\mathrm{A}^{11} & \mathrm{A}^{12}\\ \mathrm{A}^{21} & \mathrm{A}^{22}\end{pmatrix}, \quad \Psi(\mathbf{x}_\nu)^\top\mathrm{C}^{-2}\Psi(\mathbf{x}_\nu) = \begin{pmatrix}\mathrm{B}_{11} & \mathrm{B}_{12}\\ \mathrm{B}_{21} & \mathrm{B}_{22}\end{pmatrix}.$$

Substituting them into Eq. (13), we have

$$\mathbf{w}_{a|\nu_1\cup\nu_2}^\top\mathbf{w}_{a|\nu_1\cup\nu_2} = \mathbf{e}_{a|\nu_2}^\top\left(\mathrm{A}^{21}\mathrm{B}_{11}\mathrm{A}^{12} + \mathrm{A}^{22}\mathrm{B}_{21}\mathrm{A}^{12} + \mathrm{A}^{21}\mathrm{B}_{12}\mathrm{A}^{22} + \mathrm{A}^{22}\mathrm{B}_{22}\mathrm{A}^{22}\right)\mathbf{e}_{a|\nu_2}.$$

Combining this with Lemmas 2 and 3 and after some algebraic manipulation, we have that under Condition C3,

$$\begin{aligned}\mathbf{w}_{a|\nu_1\cup\nu_2}^\top\mathbf{w}_{a|\nu_1\cup\nu_2} &= \zeta_0 n^{-1}\big(\mathrm{R}_{aa} - \mathrm{R}_{a\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0 a} - (\mathrm{R}_{a\nu_2} - \mathrm{R}_{a\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu_2})\\ &\quad\times(\mathrm{R}_{\nu_2\nu_2} - \mathrm{R}_{\nu_2\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu_2})^{-1}(\mathrm{R}_{\nu_2 a} - \mathrm{R}_{\nu_2\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0 a})\big)^{-1}(1+o(1)).\end{aligned}$$

This together with Lemma 2 completes the proof of the part (i).

To prove the part (ii), let $\nu = \{a\}\cup\nu_1\cup\nu_2$. Define $\mathbf{e}_{a|\nu_1\cup\nu_2}^\top = (\mathrm{I}_{\kappa_n}, \mathbf{1}_{|\nu_1|}^\top\otimes 0_{\kappa_n}, \mathbf{1}_{|\nu_2|}^\top\otimes 0_{\kappa_n})$ and $\mathbf{e}_{a|\nu_1}^\top = (\mathrm{I}_{\kappa_n}, \mathbf{1}_{|\nu_1|}^\top\otimes 0_{\kappa_n})$. Then, we have $\mathbf{w}_{a|\nu_1\cup\nu_2}^\top\mathbf{w}_{a|\nu_1\cup\nu_2} = \mathbf{e}_{a|\nu_1\cup\nu_2}^\top\gamma_\nu\Psi(\mathbf{x}_\nu)^\top\mathrm{C}^{-2}\Psi(\mathbf{x}_\nu)\gamma_\nu\mathbf{e}_{a|\nu_1\cup\nu_2}$. Under Conditions C3 and C4, the part (ii) is obtained by following the steps similar to those in the part (i). □

3.2. With estimated C

To state a consistency property when C is estimated, we need the following two conditions used by Fan et al. [3], where we treat $\mathbf{X}$ as a fixed design matrix. 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 $\max_{1\le i\le n} E|y_{i1}|^{4\eta_0} < +\infty$, where $\eta_0 > 1$ is a constant.

In the second condition, we assume that there exists a permutation π on $\{1, \ldots, J\}$ 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

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

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 $\alpha(k_0, k) \le \exp(-\tau_2(k-k_0)^{\kappa_2})$. 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 $\{\mathbf{y}_j\}_{1\le j\le J}$ can be divided into mutually independent segments $\{\mathbf{y}_j\}_{j_{k-1}\le j<j_k}$, $1 \le k \le m$.

Note that under Conditions C1–C7, we show that the optimal shrinkage covariance estimator $\hat{\mathrm{C}}_{hs}$ is consistent with the true covariance C in Appendix A. This allows us to extend Theorems 1 and 2 to the case where unknown C is estimated by $\hat{\mathrm{C}}_{hs}$. Let $m_n = \max_{1\le i\le n}\sum_{1\le j\le n} I(c_{ij} \ne 0)$. Let $s_n = nm_n$ describe the worst-case sparsity of C.

Theorem 3. *Suppose that $\delta^{-q_n/2} q_n^2 \kappa_n^{1-r_0} = o(1)$, $\lambda_{\text{mino}}^{-1} = 0(1)$ and $\tau_{nJ} s_n = o(1)$ as both n and J tend to infinity. Then, under Conditions C1–C7, we have:*

(i) For any $\nu \subseteq \nu_0$,

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

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

(ii) For any $\nu \subseteq \{1, \ldots, p\} \setminus \nu_0$,

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

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 fPVA.

Theorem 4. *Suppose that $\delta^{-q_n/2} q_n^2 \kappa_n^{1-r_0} = o(1)$, $\lambda_{\text{mino}}^{-1} = 0(1)$ and $\tau_{nJ} s_n = o(1)$ as both n and J tend to infinity. Then, under Conditions C1–C7, we have:*

(i) For $a \in [1 : p] \setminus \nu_0$, $a \notin \nu_2 \subseteq \{1, \ldots, p\} \setminus \nu_0$, $\nu_1 \subseteq \nu_0$, $SNR_{a|\nu_1\cup\nu_2} = \frac{\kappa_n}{\zeta_0\sigma^2}(1 + o(1)) + O_p(s_n\tau_{nJ})$.

(ii) For $a \in \nu_0$, $a \notin \nu_1$, $\nu_1 \subseteq \nu_0$ and $\nu_2 \subseteq \{1, \ldots, p\} \setminus \nu_0$, $SNR_{a|\nu_1\cup\nu_2}$ is equal to

$$\frac{n(1 + o(1))}{\sigma^2\zeta_0} tr\left\{\left(E_{\{a\}|\nu_0}^\top \Sigma_{\nu_0\setminus\nu_1}^{-1} \Phi_0 \Sigma_{\nu_0\setminus\nu_1}^{-1} E_{\{a\}|\nu_0}\right)^{-1} E_{\{a\}|\nu_0}^\top \Sigma_{\nu_0\setminus\nu_1}^{-1} E_{\{a\}|\nu_0}\right\} + \frac{1 + o(1)}{\sigma^2\zeta_0} tr(\Phi_2) + O_p(s_n\tau_{nJ}),$$

where $\Sigma_{\nu_0\setminus\nu_1}^{-1}$, Φ_0 and Φ_2 are defined in Theorem 2.

The above theorem implies that under some regularity conditions the nulled-SNR contrast between active and non-active covariates will tend to infinity if the limit in (12) holds. Using the same arguments as Zhang and Oftadeh [21], we obtain the selection consistency of covariates selected by the nulled-SNR in the sense that $P(\nu_0 = \hat{\nu}_0) \to 1$ as n and J tend to infinity, where $\hat{\nu}_0$ is the set of selected covariates.

Proofs of Theorems 3 and 4. They follow from Theorems 1–2 and Lemmas 4–8, which are similar to the proof of Theorem 2 in Zhang and Liu [19]. □

4. Numerical studies

In this section, we evaluate the performance of the proposed procedure fPVA on simulated and real data. In our simulations, we compare the fPVA to the linear PVA and the Multiple Independence Screening (MIS), a multivariate extension of the nonparametric variable screening procedure (Fan et al. [3]), in terms of sensitivity and specificity. See the definitions of the linear PVA and the MIS in Appendix A. Here, sensitivity and specificity are referred to 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 $\hat{T}^c$. Under some sparsity conditions, SPE can be shown quite high when p is much larger than q_n, the number of true active covariates. To fairly compare the sensitivities of these procedures, we adjust the corresponding thresholds in the fPVA, the linear PVA and the MIS so that their specificities are all around the same level. Note by Theorem 4 that the SNR values for non-active covariates are proportional to the value of κ_n. So, a too large κ_n will reduce the contrast between the active covariates and non-covariates. On the other hand, the approximation $\tilde{f}_{kj}(x)$ used in the fPVA can attain the optimal convergence rate if we let $\kappa_n = \lfloor n^{0.2} \rfloor$ and set κ_n equally spaced interior knots for normalized B-splines (Huang et al. [6], Stone [16]). To test robustness of the above choice, we consider $\kappa_n = c\lfloor n^{\alpha_0} \rfloor$ with tuning constants $c \in \{0.5, 1, 1.5\}; \alpha_0 \in \{0.1, 0.2, 0.3\}$. In the same spirit, we let $h = h_0|\text{tr}(\hat{C})/n|$ with $h_0 \in \{0.005, 0.01, 0.02\}$ in the covariance estimator $\hat{C}_{hs}$ and $c_0 \in \{3, 3.5, 4\}$ in the stopping rule. For the kth component in a multivariate additive model, we define its oracle signal-to-noise ratio (OSNR) as follows:

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

under the oracle assumption that $\{f_{ij}(\mathbf{x}_k)\}_{i\neq k}$ are known. The higher the OSNR of a component, the higher chance it will be selected.

4.1. Simulated data

We considered the following multivariate additive model:

$$y_{ij} = \sum_{k=1}^{p} f_{kj}(x_{ik}) + \epsilon_{ij},\ 1 \le i \le n,\ 1 \le j \le J, \tag{14}$$

with

$$\begin{aligned}
f_{1j}(x_{i1}) &= r_1(x_{i1} - r_2)\sin j, \quad f_{2j}(x_{i2}) = r_3\left(r_4\,(x_{i2} - r_5)^2\right)\cos j - E\left[r_3\left(r_4\,(x_{i2} - r_5)^2\right)\cos j\right],\\
f_{3j}(x_{i3}) &= r_6\frac{\sin\left(2\pi x_{i3}\sqrt{j}\right)}{r_7 - \sin\left(2\pi x_{i3}\sqrt{j}\right)} - E\left[r_6\frac{\sin\left(2\pi x_{i3}\sqrt{j}\right)}{r_7 - \sin\left(2\pi x_{i3}\sqrt{j}\right)}\right],\\
f_{4j}(x_{i4}) &= r_8\left[r_9\sin(2\pi x_{i4}) + r_{10}\cos(2\pi x_{i4}) + r_{11}\sin^2(2\pi x_{i4}) + r_{12}\cos^3(2\pi x_{i4}) + r_{13}\sin^3(2\pi x_{i4})\right]\\
&\quad - E\left\{r_8\left[r_9\sin(2\pi x_{i4}) + r_{10}\cos(2\pi x_{i4}) + r_{11}\sin^2(2\pi x_{i4}) + r_{12}\cos^3(2\pi x_{i4}) + r_{13}\sin^3(2\pi x_{i4})\right]\right\},\\
f_{kj}(x_{ik}) &= 0, \quad 5 \le k \le p,
\end{aligned}$$

where given $(\mathbf{Y}, \mathbf{X})$, r_i, $1 \le i \le 13$ were independent Normal random variables. Model (14) involved J subjects with heterogeneous and nonlinear random-effects functions for each covariate.

We sampled 100 data sets from the above model for each of the combinations of (t, n, J, p) with $t \in \{0, 1\}$, $n \in \{50, 100, 200\}$, $J \in \{30, 70, 140\}$, and $p \in \{500, 1000\}$. Each data set was generated as follows. First, for $i \in \{1, \ldots, n\}$, to generate covariates x_{ik}, $1 \le k \le p$, we sampled u_i, w_{ik}, $1 \le k \le p$ independently from $N(0, 1)$ and truncated them into the interval $[0, 1]$. We set $x_{ik} = (w_{ik} + tu_i)/(1 + t)$, $1 \le k \le p$. 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 $(0.9^{|j_1 - j_2|})_{J\times J}$, stacking them together to form an $n \times J$ error matrix. We sampled r_1, r_3, r_4, r_7 independently from $N(2, 0.1^2)$, and $r_2, r_5, r_6, r_8, r_9, r_{10}, r_{11}, r_{12}, r_{13}$ independently from $N(0.46, 0.1^2)$, $N(1, 0.1^2)$, $N(2.5, 0.1^2)$, $N(3, 0.1^2)$, $N(0.1, 0.01^2)$, $N(0.2, 0.01^2)$, $N(0.3, 0.01^2)$, $N(0.4, 0.01^2)$, $N(0.5, 0.01^2)$, respectively. Finally, we generated $\mathbf{y}_j$, $1 \le j \le J$ by using the equation in model (14). In our simulated data, we can see from Table 1 that the OSNR values vary significantly across the 4 non-vanishing components and increase in the number of subjects J.

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, p) with the tuning constants $c = 1, \alpha_0 = 0.2, h_0 = 0.01$, are presented in Tables 2 and 3 and in Figures in Section A, the Supplementary Material.

Table 1
Oracle signal-to-noise ratio (OSNR) values of non-vanishing component k for different J in model (14).

		k			
		1	2	3	4
J	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
Comparison of methods PVA, fPVA and MIS by their average sensitivity (SEN) and specificity (SPE) for $p = 500$. The bigger the average SEN (SPE), the better the corresponding method is. The SEN (SPE) in percentage was averaged over 100 replicates with standard deviations (in parentheses). PVA: (Linear) Principal Variable Analysis. fPVA: Functional PVA. MIS: Multiple Independence Screening.

$t = 0$					$t = 1$				
n	J	Method	SEN	SPE	n	J	Method	SEN	SPE
50	30	fPVA	91.8(12.34)	99.5(0.28)	50	30	fPVA	63.3(14.85)	99.6(0.38)
		PVA	32.8(16.55)	99.2(0.48)			PVA	12.8(15.69)	99.2(0.52)
		MIS	55.8(19.10)	98.8(0.15)			MIS	50.3(9.06)	98.8(0.07)
	70	fPVA	90.3(14.61)	99.6(0.31)		70	fPVA	77.8(14.61)	99.6(0.35)
		PVA	28.0(15.19)	99.3(0.47)			PVA	17.5(19.94)	99.2(0.44)
		MIS	44.8(21.70)	98.7(0.18)			MIS	45.8(12.33)	98.8(0.10)
	140	fPVA	67.3(15.77)	99.7(0.34)		140	fPVA	66.8(15.91)	99.6(0.39)
		PVA	21.3(16.43)	99.2(0.47)			PVA	8.5(15.17)	99.2(0.49)
		MIS	34.8(22.16)	98.7(0.18)			MIS	36.3(15.23)	98.7(0.12)
100	30	fPVA	100(0.00)	99.9(0.20)	100	30	fPVA	93.8(11.98)	99.8(0.23)
		PVA	46.5(16.67)	99.4(0.39)			PVA	26.8(23.10)	99.4(0.39)
		MIS	82.5(17.23)	99.1(0.14)			MIS	55.0(10.66)	98.8(0.09)
	70	fPVA	100(0.00)	99.9(0.15)		70	fPVA	100(0.00)	99.8(0.23)
		PVA	46.5(23.58)	99.5(0.30)			PVA	31.8(26.30)	99.5(0.32)
		MIS	89.0(13.91)	99.1(0.11)			MIS	56.8(12.23)	98.8(0.10)
	140	fPVA	100(0.00)	99.9(0.19)		140	fPVA	100(0.00)	99.9(0.18)
		PVA	32.8(25.30)	99.5(0.29)			PVA	23.5(28.40)	99.6(0.23)
		MIS	77.5(18.97)	99.0(0.15)			MIS	55.0(15.08)	98.8(0.12)
200	30	fPVA	100(0.00)	99.9(0.15)	200	30	fPVA	99.5(3.52)	99.8(0.20)
		PVA	60.8(16.76)	99.6(0.33)			PVA	42.0(20.68)	99.4(0.36)
		MIS	99.5(3.52)	98.8(0.03)			MIS	72.5(17.59)	99.0(0.14)
	70	fPVA	100(0.00)	99.9(0.13)		70	fPVA	100(0.00)	99.9(0.12)
		PVA	63.0(22.89)	99.7(0.21)			PVA	41.8(27.07)	99.6(0.24)
		MIS	100(0.00)	98.8(0.00)			MIS	73.8(15.64)	99.0(0.13)
	140	fPVA	100(0.00)	99.9(0.12)		140	fPVA	100(0.00)	99.9(0.15)
		PVA	48.5(33.30)	99.7(0.22)			PVA	24.5(31.18)	99.7(0.17)
		MIS	96.5(8.72)	99.2(0.07)			MIS	80.3(15.61)	98.6(0.13)

We see from these tables and figures that the fPVA had a superior sensitivity over the PVA and the MIS when their specificities were fixed around the same level. The performance was also robust to the choices of $c \in \{0.5, 1, 1.5\}$, $\alpha_0 \in \{0.1, 0.2, 0.3\}$ and $h_0 \in \{0.005, 0.01, 0.02\}$. For uncorrelated covariates (i.e., $t = 0$), the fPVA had a superior performance over both the MIS and the linear PVA in the terms of sensitivity and specificity for combinations of (n, J, p). For example, for $(n, J, p) = (50, 70, 500)$, the sensitivity value of the fPVA increased by 46% and 62% over the MIS and the linear PVA respectively. For correlated covariates (i.e., $t = 1$), the improvements were also striking. For example, for $(n, J, p) = (50, 70, 500)$, the sensitivity value of the fPVA increased by 32% over the MIS and by 60% over the linear PVA. This demonstrates that the fPVA was more effective in exploring both correlated and uncorrelated nonlinear 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 $(n, p) = (100, 500)$ and $t = 0, 1$ in Figures in Section A, the Supplementary Material. It demonstrates that the fPVA computationally costs less than the MIS but more than the linear PVA. The results show that on average sensitivity and specificity values were increasing in n when J, p and t were fixed. The sensitivity was decreasing on average in the size of pairwise correlations between covariates. This reflects that the increasing correlations between covariates could make it difficult to identify true active covariates. The sensitivity was also decreasing on average 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.

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. [4]. 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

Table 3
Comparison of methods PVA, fPVA and MIS by their average sensitivity (SEN) and average specificity (SPE) for $p = 1000$. The bigger the average SEN (SPE), the better the corresponding method is. The SEN (SPE) in percentage was averaged over 100 replicates with standard deviations (in parentheses). PVA: (Linear) Principal Variable Analysis. fPVA: Functional PVA. MIS: Multiple Independence Screening.

$t = 0$					$t = 1$				
n	J	Method	SEN	SPE	n	J	Method	SEN	SPE
50	30	fPVA	93.3(11.71)	99.7(0.10)	100	30	fPVA	63.3(14.85)	99.6(0.31)
		PVA	36.0(16.79)	98.9(0.45)			PVA	14.8(15.93)	98.8(0.45)
		MIS	59.5(15.40)	98.7(0.06)			MIS	54.5(10.29)	98.7(0.04)
	70	fPVA	87.8(13.53)	99.7(0.10)		70	fPVA	72.3(15.28)	99.6(0.19)
		PVA	34.5(14.12)	99.0(0.42)			PVA	17.3(19.69)	98.8(0.45)
		MIS	48.8(21.13)	98.7(0.08)			MIS	51.0(12.77)	98.7(0.08)
	140	fPVA	66.8(15.10)	99.6(0.19)		140	fPVA	64.0(17.16)	99.6(0.22)
		PVA	27.5(13.53)	99.0(0.43)			PVA	8.75(15.23)	98.9(0.48)
		MIS	35.5(22.80)	98.6(0.09)			MIS	46.3(14.81)	98.7(0.06)
100	30	fPVA	100(0.00)	99.8(0.25)	100	30	fPVA	94.5(10.41)	99.7(0.30)
		PVA	49.8(18.29)	99.4(0.36)			PVA	28.3(20.92)	99.3(0.39)
		MIS	86.8(16.07)	98.8(0.06)			MIS	60.8(15.99)	98.7(0.06)
	70	fPVA	100(0.00)	99.9(0.21)		70	fPVA	100(0)	99.8(0.30)
		PVA	48.3(18.55)	99.5(0.26)			PVA	35.8(23.37)	99.6(0.27)
		MIS	87.3(15.28)	98.3(0.06)			MIS	58.3(13.32)	98.7(0.05)
	140	fPVA	100(0.00)	99.8(0.31)		140	fPVA	99.8(2.50)	99.7(0.32)
		PVA	41.3(23.13)	99.5(0.26)			PVA	26.8(25.44)	99.6(0.27)
		MIS	88.0(15.68)	98.8(0.06)			MIS	57.5(13.99)	98.7(0.06)
200	30	fPVA	100(0.00)	99.9(0.14)	200	30	fPVA	98.5(5.97)	99.8(0.17)
		PVA	61.0(13.45)	99.5(0.29)			PVA	43.3(19.74)	99.4(0.31)
		MIS	99.8(2.50)	98.4(0.01)			MIS	78.0(17.51)	98.8(0.07)
	70	fPVA	100(0.00)	99.9(0.10)		70	fPVA	100(0.00)	99.9(0.13)
		PVA	66.3(18.93)	99.7(0.18)			PVA	51.0(22.73)	99.7(0.20)
		MIS	100(0.00)	98.9(0.00)			MIS	75.5(18.11)	98.3(0.07)
	140	fPVA	100(0.00)	99.9(0.08)		140	fPVA	100(0.00)	99.9(0.12)
		PVA	58.3(28.22)	99.7(0.17)			PVA	39.0(29.37)	99.7(0.16)
		MIS	100(0.00)	98.9(0.00)			MIS	77.0(16.55)	98.3(0.07)

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) for the data, where we took genes as covariates and IC50 values of multiple drugs as the responses. The expression values of these genes form a design matrix $\mathbf{X}$. We began with standardizing the expression values for each gene and centralizing the response values across the drugs. We then applied the fPVA to the data, obtaining 7 active covariates (i.e., genes): PEX5, NRXN2, HS2ST1, EIF4GH1, CUL4 A, PIPN22 and ACN9. Finally, we fitted the multivariate additive model to the data set with covariates restricted to the above selected covariates. We used the so-called post-approximations to f_{kj}'s by linear combinations of 5 spline functions as did in Fan et al. [3]. To save space, we only present the estimated non-vanishing nonparametric components related to the drug KIN001-135 in Fig. 1. The results suggest that the IC50s of the drug KIN001-135 depended nonlinearly on these selected genes. These effects were also drug-specific random-effects.

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. [15]). We gathered such information from the Human Protein Atlas Portal at http://www.proteinatlas.org for 5 of the selected 7 genes, namely PEX5, NRXN2, EIF4GH1, CUL4 A, 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 1, Section B, the Supplementary Material. We then carried out a Bonferroni correction as well as Holm's correction (Holm [5]) for multiple testing respectively. The number of the rejected null-hypotheses for each selected gene are shown in Table 2 in Section B, the Supplementary Material. The results indicate that all the 5 selected genes had positive protein 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, whose biological roles are difficult to explain by using the above Portal.

To assess stability of the above analysis, we conducted a parametric bootstrap analysis. We simulated $(\mathbf{Y}, \mathbf{X})$ by using the fitted multivariate additive model for each combination of (n, p, J), where $n \in \{42, 100, 200\}$, $p \in \{500, 1000\}$, $J = 131$ 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

Table 4
Average sensitivity (SEN) and specificity (SPE) (in percentage) of fPVA with standard deviations (in parentheses) for the stability analysis. Sensitivity and specificity were averaged over 100 parametric bootstrap samples drawn from the fitted multivariate additive model.

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)

into the simulated model to form p covariates. Secondly, we calculated the $p \times p$ sample covariance matrix $\boldsymbol{\Omega}$ of these p genes by using the original gene expression data. Given $\boldsymbol{\Omega}$, we drew n random row-vectors from the p-dimension Normal $N_p(\mathbf{0}, \boldsymbol{\Omega})$ 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 $\hat{f}_{kj}(\cdot)$, $1 \le j \le J$ 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). 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 4, we display the averages of these values over 100 replicates. The results show 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 relatively small n and J. In addition, the recovering rate was increasing in n and attained 100% on average when $n \ge 100$. The results suggest that the fPVA based data analysis was quite stable.

5. Discussion and conclusion

We have proposed a novel approach to nonparametric component screening for multivariate additive random-effects models by using the B-spline approximation and the null-beamforming technique. Here, we have used nonparametric variance components to model overall effects of each covariate on all subjects. The proposed procedure 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 proposed method provides a perfect forward variable selection in which the additional variable selection with penalization is not required. 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 in the sense that under certain regularity conditions, the selected active set is asymptotically equal to the underlying one. As pointed out by one of reviewers, the regularity condition C1 can be weakened as follows: for any positive η, there exists a universal positive constant c_* such that

$$\Pr\left(\max_{1\le k\le p, 1\le j\le J} \sup_{z, z+\delta_* \in [a,b]} |f_{kj}^{(r)}(z+\delta_*) - f_{kj}^{(r)}(z)| > c_* |\delta_*|^{\alpha}\right) < \eta$$

for some non-negative integer r and $0 < \alpha \le 1$.

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 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 expression levels 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.

In practice, both fixed-effects and random-effects models can be fitted to a multivariate regression dataset when the number of subjects, J, is not too large. Fixed-effects allow for structureless variation across subjects while random-effects require that unobserved heterogeneity obeys some probability distribution. However, when n, p and J are large, fixed-effects can result in a huge number of parameters ($pJ\kappa_n$ parameters in the approximate model) with group structures, where dimensionality quickly becomes an issue. Instead of estimating many individual effects, random-effects only need to estimate no more than $\min\{n, J\} \ll p$ nonparametric variance components, which greatly reduces the parameter dimension. Once the active variance components are identified, we further use the data projection to conduct a post-estimation of the trajectories of these selected random-effects on subjects. This suggests an advantage of random-effects models over fixed-effects models in multivariate nonparametric settings.

CRediT authorship contribution statement

Hui Ding: Conceptualization, Data curation, Numerical investigation, Methodology, Coding, Writing. **Jian Zhang:** Conceptualization, Theoretical analysis, Numerical investigation, Methodology, Project administration, Supervision, Writing. **Riquan Zhang:** Conceptualization, Supervision, Funding acquisition, Writing.

Acknowledgments

Research was completed while the second author was visiting the Institute for Mathematical Sciences, National University of Singapore in 2018. He would like to thank Dr. Zhigang Yao for his hospitality. The research of the first author was partially supported by the East China Normal University Fund for a Short Term Overseas Academic Visit and the National Natural Science Foundation of China (11901286). The research of the third author was partially supported by the National Natural Science Foundation of China (11971171) and 111 project, PR China (B14019). We are grateful to the Editor-in-Chief, an associate editor and reviewers for their valuable comments on the manuscript that have helped to improve the paper.

Appendix A

In this Appendix, we introduce the lemmas and define the MIS and the linear PVA, which are used in Sections 3 and 4 respectively. Lemmas 1–3 follow from the Woodbury matrix identity while Lemmas 4–8 are derived by using the concentration inequality for non i.i.d. samples (see, for example, Fan et al. [3], Zhang and Liu [19]): There exist constants $d_t, t \in \{1, \ldots, 5\}$, such that for any $u > 0$,

$$\Pr\left(\max_{1\le i,j\le n}\left|\frac{1}{J}\sum_{k=1}^{J} y_{ik}y_{kj} - c_{ij}\right| > u\right) \le n^2 J \exp\left(-\frac{(Ju)^{\kappa}}{d_1}\right) + n^2\exp\left(-\frac{(Ju)^2}{d_2(1+Jd_3)}\right) + n^2\exp\left(-\frac{(Ju)^2}{d_4 J}\exp\left(\frac{(Ju)^{\kappa(1-\kappa)}}{d_5(\log(Ju))^{\kappa}}\right)\right).$$

Lemma 1. *Suppose that $\delta^{-q_n/2}q_n^2\kappa_n^{1-r_0} = o(1)$ and that $\lambda_{\text{mino}}^{-1} = 0(1)$. Denote by α_n any constant of order $0(1)\delta^{-q_n/2}q_n^2\kappa_n^{1-r_0}$. Under Conditions C1–C3, for any subset of covariates, ν, we have*

$$\Psi(\mathbf{x}_\nu)^\top C^{-1}\Psi(\mathbf{x}_\nu)/n = F_\nu + R_{\nu\nu_0}R_{\nu_0\nu_0}^{-1}\left(E_{\nu_0}^\top \Sigma E_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu}n^{-1} - R_{\nu\nu_0}R_{\nu_0\nu_0}^{-1/2}\left(R_{\nu_0\nu_0}^{-1/2}\left(E_{\nu_0}^\top\Sigma E_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1/2}\right)^2 R_{\nu_0\nu_0}^{-1/2}R_{\nu_0\nu}n^{-2} + R_{\nu\nu_0}R_{\nu_0\nu_0}^{-1/2}\left(R_{\nu_0\nu_0}^{-1/2}\left(E_{\nu_0}^\top\Sigma E_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1/2}\right)^3 R_{\nu_0\nu_0}^{-1/2}R_{\nu_0\nu}n^{-3}(1+\alpha_n).$$

Proof. Using the Woodbury matrix identity, we have

$$\Psi(\mathbf{x}_\nu)^\top \mathrm{C}^{-1}\Psi(\mathbf{x}_\nu)/n = \Psi(\mathbf{x}_\nu)^\top\left[\mathrm{A}_{\nu_0}^{-1} - \mathrm{A}_{\nu_0}^{-1}\Psi(\mathbf{x}_{\nu_0})\left\{\left(\mathrm{E}_{\nu_0}^\top\Sigma\mathrm{E}_{\nu_0}\right)^{-1} + n\mathrm{R}_{\nu_0\nu_0}\right\}^{-1}\Psi(\mathbf{x}_{\nu_0})^\top\mathrm{A}_{\nu_0}^{-1}\right]\Psi(\mathbf{x}_\nu)/n.$$

The proof is completed by using the Taylor expansion. □

Lemma 2. *Suppose that $\delta^{-q_n/2}q_n^2\kappa_n^{1-r_0} = o(1)$ and that $\lambda_{\text{mino}}^{-1} = 0(1)$. Denote by α_n any constant of order $0(1)\delta^{-q_n/2}q_n^2\kappa_n^{1-r_0}$. Then, under Conditions C1–C3 and C5, as n tends to infinity, we have:*

(i) Uniformly for $a \in \{1, \ldots, p\} \setminus \nu_0$ and $a \notin \nu_1 \cup \nu_2$ with $|\nu_1 \cup \nu_2| < rn$, the $(\nu_1 \cup \nu_2)$-nulled predictive power $\Upsilon_{a|\nu_1\cup\nu_2}$ admits the form

$$\gamma_{a|\nu_1\cup\nu_2} = \frac{1}{n}\left\{F_a - (R_{a\nu_2} - R_{a\nu_0}R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu_2})F_{\nu_2}^{-1}(R_{\nu_2 a} - R_{\nu_2\nu_0}R_{\nu_0\nu_0}^{-1}R_{\nu_0 a})\right\}^{-1}(1+o(1)).$$

(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 power $\gamma_{a|\nu_1\cup\nu_2}$ admits the form

$$\gamma_{a|\nu_1\cup\nu_2} = \left(E_{\{a\}|\nu_0}^T\Sigma_{\nu_0\setminus\nu_1}^{-1}E_{\{a\}|\nu_0}\right)^{-1} + n^{-1}E_{\{a\}|\nu_0}^T\left(E_{\nu_0}^T\Sigma E_{\nu_0}\right)^{-1}\Phi_1\left(E_{\nu_0}^T\Sigma E_{\nu_0}\right)^{-1}E_{\{a\}|\nu_0}(1+o(1)),$$

where

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

Proof. To prove (i), let ν denote $\nu_1 \cup \nu_2 \cup \{a\}$. We partition $\Psi(\mathbf{x}_\nu)^\top \mathrm{C}^{-1} \Psi(\mathbf{x}_\nu)$ and $(\Psi(\mathbf{x}_\nu)^\top \mathrm{C}^{-1} \Psi(\mathbf{x}_\nu))^{-1}$ in the same way into the block matrices below:

$$\begin{aligned}
\Psi(\mathbf{x}_\nu)^\top \mathrm{C}^{-1} \Psi(\mathbf{x}_\nu) &= \begin{pmatrix} \Psi(\mathbf{x}_{\nu_1})^\top \mathrm{C}^{-1} \Psi(\mathbf{x}_{\nu_1}) & \Psi(\mathbf{x}_{\nu_1})^\top \mathrm{C}^{-1} \Psi(\mathbf{x}_{\nu_2\cup\{a\}}) \\ \Psi(\mathbf{x}_{\nu_2\cup\{a\}})^\top \mathrm{C}^{-1} \Psi(\mathbf{x}_{\nu_1}) & \Psi(\mathbf{x}_{\nu_2\cup\{a\}})^\top \mathrm{C}^{-1} \Psi(\mathbf{x}_{\nu_2\cup\{a\}}) \end{pmatrix},\\
(\Psi(\mathbf{x}_\nu)^\top \mathrm{C}^{-1} \Psi(\mathbf{x}_\nu))^{-1} &= \begin{pmatrix} \mathrm{A}^{11} & \mathrm{A}^{12} \\ \mathrm{A}^{21} & \mathrm{A}^{22} \end{pmatrix},
\end{aligned}$$

where $\Psi(\mathbf{x}_{\nu_1}) = \Psi(\mathbf{x}_{\nu_0})\mathrm{E}_{\nu_1|\nu_0}$. Then, by definition we have $\gamma_{a|\nu_1\cup\nu_2} = (\mathbf{1}_{|\nu_2|}^\top \otimes \mathbf{0}_{\kappa_n}^\top, \mathrm{I}_{\kappa_n}) \mathrm{A}^{22} (\mathbf{1}_{|\nu_2|}^\top \otimes \mathbf{0}_{\kappa_n}^\top, \mathrm{I}_{\kappa_n})^\top$. The proof of (i) is completed by using Lemma 1, the Taylor expansion and the fact that the smallest eigenvalue of any main block matrix is larger than that of the whole matrix. Similarly, we can complete the proof of (ii). □

Lemma 3. *Suppose that* $\delta^{-q_n/2} q_n^2 \kappa_n^{1-r_0} = o(1)$ *and that* $\lambda_{\mathrm{mino}}^{-1} = 0(1)$. *Denote by* α_n *any constant of order* $0(1)\delta^{-q_n/2} q_n^2 \kappa_n^{1-r_0}$. *Then, under Conditions C1–C4, as* $n \to \infty$, *we have:*

(i) for any $\nu \subseteq \nu_0$,

$$\Psi(\mathbf{x}_\nu)^\top C^{-2} \Psi(\mathbf{x}_\nu) = E_{\nu|\nu_0}^T \left(E_{\nu_0}^\top \Sigma E_{\nu_0}\right)^{-1} R_{\nu_0\nu_0}^{-1} \Psi(\mathbf{x}_{\nu_0})^\top A_{\nu_0}^{-2} \Psi(\mathbf{x}_{\nu_0}) R_{\nu_0\nu_0}^{-1} \left(E_{\nu_0}^\top \Sigma E_{\nu_0}\right)^{-1} E_{\nu|\nu_0} n^{-2}(1+\alpha_n).$$

(ii) for any $\nu \subseteq \{1, \ldots, p\} \setminus \nu_0$,

$$\begin{aligned}
\Psi(\mathbf{x}_\nu)^\top C^{-2} \Psi(\mathbf{x}_\nu) = n \left(\Psi(\mathbf{x}_\nu)^T A_{\nu_0}^{-2} \Psi(\mathbf{x}_\nu)/n\right) + n(1+\alpha_n) \big\{ &-R_{\nu\nu_0} R_{\nu_0\nu_0}^{-1} \Psi(\mathbf{x}_{\nu_0})^T A_{\nu_0}^{-2} \Psi(\mathbf{x}_\nu)/n\\
&-(\Psi(\mathbf{x}_\nu)^T A_{\nu_0}^{-2} \Psi(\mathbf{x}_{\nu_0})/n) R_{\nu_0\nu_0}^{-1} R_{\nu_0\nu} + R_{\nu\nu_0} R_{\nu_0\nu_0}^{-1} (\Psi(\mathbf{x}_{\nu_0})^T A_{\nu_0}^{-2} \Psi(\mathbf{x}_{\nu_0})/n) R_{\nu_0\nu_0}^{-1} R_{\nu_0\nu} \big\}.
\end{aligned}$$

(iii) for any $\nu = \nu_1 \cup \nu_2$ *with* $\nu_1 \subseteq \nu_0$ *and* $\nu_2 \subseteq \{1, \ldots, p\} \setminus \nu_0$,

$$\Psi(\mathbf{x}_\nu)^\top C^{-2} \Psi(\mathbf{x}_\nu) = \begin{pmatrix} \Psi(\mathbf{x}_{\nu_1})^\top C^{-2} \Psi(\mathbf{x}_{\nu_1}) & \Psi(\mathbf{x}_{\nu_1})^\top C^{-2} \Psi(\mathbf{x}_{\nu_2}) \\ \Psi(\mathbf{x}_{\nu_2})^\top C^{-2} \Psi(\mathbf{x}_{\nu_1}) & \Psi(\mathbf{x}_{\nu_2})^\top C^{-2} \Psi(\mathbf{x}_{\nu_2}) \end{pmatrix}$$

where

$$\begin{aligned}
\Psi(\mathbf{x}_{\nu_1})^\top C^{-2} \Psi(\mathbf{x}_{\nu_1}) &= \frac{1}{n} E_{\nu_1|\nu_0}^T \left(E_{\nu_0}^\top \Sigma E_{\nu_0}\right)^{-1} R_{\nu_0\nu_0}^{-1} \left(\Psi(\mathbf{x}_{\nu_0})^\top A_{\nu_0}^{-2} \Psi(\mathbf{x}_{\nu_0})/n\right) R_{\nu_0\nu_0}^{-1} \left(E_{\nu_0}^\top \Sigma E_{\nu_0}\right)^{-1} E_{\nu_1|\nu_0}(1+\alpha_n),\\
\Psi(\mathbf{x}_{\nu_1})^\top C^{-2} \Psi(\mathbf{x}_{\nu_2}) &= \frac{1}{n} E_{\nu_1|\nu_0}^T \left(E_{\nu_0}^\top \Sigma E_{\nu_0}\right)^{-1} R_{\nu_0\nu_0}^{-1} \left(E_{\nu_0}^\top \Sigma E_{\nu_0}\right)^{-1} \zeta_0 R_{\nu_0\nu_0}^{-1} R_{\nu_0\nu_2}(1+o(1)),\\
\Psi(\mathbf{x}_{\nu_2})^\top C^{-2} \Psi(\mathbf{x}_{\nu_2}) &= n \left(\Psi(\mathbf{x}_{\nu_2})^\top A_{\nu_0}^{-2} \Psi(\mathbf{x}_{\nu_2})/n\right) + n(1+\alpha_n) \big\{ -R_{\nu_2\nu_0} R_{\nu_0\nu_0}^{-1} \left(\Psi(\mathbf{x}_{\nu_0})^\top A_{\nu_0}^{-2} \Psi(\mathbf{x}_{\nu_2})/n\right)\\
&\quad - \left(\Psi(\mathbf{x}_{\nu_2})^\top A_{\nu_0}^{-2} \Psi(\mathbf{x}_{\nu_0})/n\right) R_{\nu_0\nu_0}^{-1} R_{\nu_0\nu_2} + R_{\nu_2\nu_0} R_{\nu_0\nu_0}^{-1} \left(\Psi(\mathbf{x}_{\nu_0})^\top A_{\nu_0}^{-2} \Psi(\mathbf{x}_{\nu_0})/n\right) R_{\nu_0\nu_0}^{-1} R_{\nu_0\nu_2} \big\}.
\end{aligned}$$

Proof. Using the Woodbury matrix identity, we have

$$\begin{aligned}
\Psi(\mathbf{x}_{\nu_0})^\top \mathrm{C}^{-2} \Psi(\mathbf{x}_{\nu_0}) = \Psi(\mathbf{x}_{\nu_0})^\top \mathrm{A}_{\nu_0}^{-2} \Psi(\mathbf{x}_{\nu_0}) &\mathrm{R}_{\nu_0\nu_0}^{-1/2} \{\mathrm{R}_{\nu_0\nu_0}^{-1/2} \left(\mathrm{E}_{\nu_0}^\top \Sigma \mathrm{E}_{\nu_0}\right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1/2} n^{-1} + \mathrm{I}_{|\nu_0|\kappa_n}\}^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1/2} \left(\mathrm{E}_{\nu_0}^\top \Sigma \mathrm{E}_{\nu_0}\right)^{-1} n^{-1}\\
&- \mathrm{R}_{\nu_0\nu_0}^{1/2} \{\mathrm{R}_{\nu_0\nu_0}^{-1/2} \left(\mathrm{E}_{\nu_0}^\top \Sigma \mathrm{E}_{\nu_0}\right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1/2} n^{-1} + \mathrm{I}_{|\nu_0|\kappa_n}\}^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1/2} \Psi(\mathbf{x}_{\nu_0})^\top \mathrm{A}_{\nu_0}^{-2} \Psi(\mathbf{x}_{\nu_0})\\
&+ \mathrm{R}_{\nu_0\nu_0}^{1/2} \{\mathrm{R}_{\nu_0\nu_0}^{-1/2} \left(\mathrm{E}_{\nu_0}^\top \Sigma \mathrm{E}_{\nu_0}\right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1/2} n^{-1} + \mathrm{I}_{|\nu_0|\kappa_n}\}^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1/2} \Psi(\mathbf{x}_{\nu_0})^\top \mathrm{A}_{\nu_0}^{-2} \Psi(\mathbf{x}_{\nu_0})\\
&\times \{\mathrm{R}_{\nu_0\nu_0}^{-1/2} \left(\mathrm{E}_{\nu_0}^\top \Sigma \mathrm{E}_{\nu_0}\right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1/2} n^{-1} + \mathrm{I}_{|\nu_0|\kappa_n}\}^{-1} \mathrm{R}_{\nu_0\nu_0}^{1/2}.
\end{aligned}$$

The proof of (i) is completed by using the Taylor expansion and the assumption that $\lambda_{\min}(\mathrm{E}_{\nu_0}^\top \Sigma \mathrm{E}_{\nu_0})^{-1} = 0(1)$. The statements in (ii) and (iii) can be proved by a similar approach. □

Let $\kappa_3 = \max\{2(2/\kappa_1 + 1/\kappa_2) - 1, (4/3)(1/\kappa_1 + 1/\kappa_2) - 1/3, 1\}$. As before, let $\|D\|_F = \sqrt{\mathrm{tr}(DD^T)/n}$ be the size-normalized Frobenius norm and $\|D\|$ be the spectral norm of D respectively. Let $\mu_n = \mathrm{tr}(\mathrm{C})/n$. We have $\|D\|_F \le \|D\|$. The following lemmas are partially adapted from Zhang and Oftadeh [21].

Lemma 4. *Under Conditions C1–C7, if $\tau_{nJ} = o(1)$ as $n \to \infty$ and $J \to \infty$,*

$$E \max_{1\le i,j\le n} \left|\hat{c}_{ij} - c_{ij}\right| = O(\tau_{nJ}), \quad E \max_{1\le i,j\le n} \left(\hat{c}_{ij} - c_{ij}\right)^2 = O(\tau_{nJ}^2), \quad \max_{1\le i,j\le n} \left|\hat{c}_{ij} - c_{ij}\right| = O_p(\tau_{nJ}).$$

Proof. For notational simplicity, denote $Q_{ij} = \left|\frac{1}{J}\sum_{k=1}^{J} y_{ik}y_{kj} - c_{ij}\right|$. For a large sequence of constants $0 < h_n = O(1)$, invoking the inequality (15), we have

$$\begin{aligned} E \max_{1\le i,j\le n} Q_{ij} &\le h_n\tau_{nJ} + E\left[\max_{1\le i,j\le n} Q_{ij} I(\max_{1\le i,j\le n} Q_{ij} > h_n\tau_{nJ})\right] \le 2h_n\tau_{nJ} + \int_{h_n\tau_{nJ}}^{\infty} \Pr\left(\max_{1\le i,j\le n} Q_{ij} > u\right) du \\ &\le 2h_n\tau_{nJ} + \frac{n^2 d_1}{\kappa(h_nJ\tau_{nJ})^{\kappa-1}} \exp\left(\frac{(h_nJ\tau_{nJ})^{\kappa}}{d_1}\right) + \frac{n^2 d_2(1/J+d_3)}{2h_nJ\tau_{nJ}} \exp\left(-\frac{(h_n\sqrt{J}\tau_{nJ})^2}{d_2(1/J+d_3)}\right) \\ &\quad + \frac{n^2}{2h_nJ\tau_{nJ}} \exp\left(-\frac{(h_n\sqrt{J}\tau_{nJ})^2(1-o(1))}{d_4}\right) \\ &= \tau_{nJ}(2h_n + o(1)) = O(\tau_{nJ}), \end{aligned}$$

which yields $\max_{1\le i,j\le n} Q_{ij} = O_p(\tau_{nJ})$. We also have $E[\max_{1\le i,j\le n} Q_{ij}^2] \le (\tau_{nJ})^2(2h_n^2 + o(1)) = O(\tau_{nJ}^2)$. Similarly, we can show

$$E \max_{1\le i,j\le n} |\bar{y}_i\bar{y}_j| = O(\tau_{nJ}), \quad E \max_{1\le i,j\le n} |\bar{y}_i\bar{y}_j|^2 = O(\tau_{nJ}^2), \quad \max_{1\le i,j\le n} |\bar{y}_i\bar{y}_j| = O_p(\tau_{nJ}).$$

Combining these with the other equalities shown before, we complete the proof. □

In the next lemma, we show the convergence rates of the threshold estimator.

Lemma 5. *Under Conditions C1–C7, if $m_n\tau_{nJ} = o(1)$ as $n \to \infty$ and $J \to \infty$, then for $h > 0$,*

$$E\|\hat{C}_h - C\| = O(m_n\tau_{nJ}), \quad E\|\hat{C}_h - C\|^2 = O(m_n\tau_{nJ})^2, \quad \|\hat{C}_h - C\| = O_p(m_n\tau_{nJ}).$$

For $h = 0$, the above results continue to hold if m_n is replaced by n.

Proof. Without loss of generality, we assume $\pi(j) = j$, $1 \le j \le J$. For $h > 0$, we have

$$E\|\hat{C}_h - C\| \le E[\mathrm{I}] + E[\mathrm{II}] + E[\mathrm{III}] + h\tau_{nJ}m_n, \tag{15}$$

where

$$\mathrm{I} = \max_i \sum_{j=1}^{n} |\hat{c}_{ij} - c_{ij}| I(|\hat{c}_{ij}| > h\tau_{nJ}, |c_{ij}| > h\tau_{nJ}), \quad \mathrm{II} = \max_i \sum_{j=1}^{n} |\hat{c}_{ij}| I(|\hat{c}_{ij}| > h\tau_{nJ}, |c_{ij}| \le h\tau_{nJ}),$$

$$\mathrm{III} = \max_i \sum_{j=1}^{n} |c_{ij}| I(|\hat{c}_{ij}| \le h\tau_{nJ}, |c_{ij}| > h\tau_{nJ}).$$

It follows from Lemma 4 that for $0 \le \delta, \epsilon < 1$,

$$E[\mathrm{I}] \le E \max_{1\le i,j\le n} |\hat{c}_{ij} - c_{ij}| \max_i \sum_{j=1}^{n} I(|c_{ij}| > 0) = O(\tau_{nJ})m_n.$$

$$E[\mathrm{II}] \le E \max_{1\le i,j\le n} |\hat{c}_{ij} - c_{ij}| \left(\max_i \sum_{j=1}^{n} I(|\hat{c}_{ij} - c_{ij}| \ge (1-\delta)h\tau_{nJ}) + m_n\right) + h\tau_{nJ}m_n = O(m_n\tau_{nJ}).$$

$$E[\mathrm{III}] \le E \max_i \sum_{j=1}^{n} (|\hat{c}_{ij} - c_{ij}| + |\hat{c}_{ij}|) I(|c_{ij}| > h\tau_{nJ}, |\hat{c}_{ij}| \le h\tau_{nJ}) \le O(\tau_{nJ})m_n + h\tau_{nJ}m_n = O(m_n\tau_{nJ}).$$

Invoking the inequality (15), we have

$$E\|\hat{C}_h - C\| = O(m_n\tau_{nJ}), \quad \|\hat{C}_h - C\| \le O_p(m_n\tau_{nJ}).$$

The remaining part of Lemma 5 can be completed by a similar argument. □

Lemma 6. *Under Conditions C1–C7, if $m_n\tau_{nJ} = o(1)$ as $n \to \infty$ and $J \to \infty$, then for $h > 0$,*

$$\frac{1}{J^2}\sum_{k=1}^{J} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} ((y_{ik} - \bar{y}_i)(y_{jk} - \bar{y}_j) - \hat{c}_{ij})^2 I(|\hat{c}_{ij}| > h\tau_{nJ}) = O_p(m_n/J)$$

For $h = 0$, the equality holds if m_n is replaced by n.

Proof. Without loss of generality, we assume that $\pi(j) = j, 1 \le j \le J$ and that $\bar{y}_i = \bar{y}_j = 0$. By use of Chebyshev's inequality, it suffices to show

$$\frac{1}{Jn}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left[(y_{i1}y_{j1} - \hat{c}_{ij})^2 I(|\hat{c}_{ij}| > h\tau_{nJ})\right] = O(m_n/J).$$

Note that for $0 < \delta < 1$, the above equation follows from the following claims which can be derived from Lemma 4:

$$\begin{aligned}
&\frac{1}{Jn}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left[(y_{i1}y_{j1} - \hat{c}_{ij})^2 |I(|\hat{c}_{ij}| > h\tau_{nJ}) - I(|\hat{c}_{ij}| > h\tau_{nJ})|\right] \\
&\le \frac{1}{Jn}\sum_{i=1}^{n}\sum_{j=1}^{n} E(y_{i1}y_{j1} - \hat{c}_{ij})^2 I(|\hat{c}_{ij}| > h\tau_{nJ}, |c_{ij}| \le \delta h\tau_{nJ}) + E(y_{i1}y_{j1} - \hat{c}_{ij})^2 I(|\hat{c}_{ij}| > h\tau_{nJ}, \delta h\tau_{nJ} < |c_{ij}| \le h\tau_{nJ}) \\
&\quad + \frac{1}{Jn}\sum_{i=1}^{n}\sum_{j=1}^{n} E(y_{i1}y_{j1} - \hat{c}_{ij})^2 I(|c_{ij}| > h\tau_{nJ}, |\hat{c}_{ij}| \le h\tau_{nJ}) \\
&\le \frac{n}{J}\max_{1\le i,j\le n}\left(E(y_{i1}y_{j1} - \hat{c}_{ij})^{2\eta_0}\right)^{1/\eta_0} \max_{1\le i,j\le n} P\left(|\hat{c}_{ij} - c_{ij}| \ge (1-\delta)h\tau_{nJ}\right)^{1-1/\eta_0} + \frac{2m_n}{J}\max_{1\le i,j\le n}(y_{i1}y_{j1} - \hat{c}_{ij})^2 \\
&= \frac{1}{J}\max_{1\le i,j\le n}\left(E(y_{i1}y_{j1} - \hat{c}_{ij})^{2\eta_0}\right)^{1/\eta_0} o(1) + \frac{2m_n}{J}\max_{1\le i,j\le n}(y_{i1}y_{j1} - \hat{c}_{ij})^2 = O(m_n/J). \\
&\frac{1}{J}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left[(y_{i1}y_{j1} - \hat{c}_{ij})^2 I(c_{ij}| > h\tau_{nJ})\right] \le \frac{1}{J}\max_{1\le i,j\le n} E(y_{i1}y_{j1} - \hat{c}_{ij})^2 m_n = O(m_n/J). \quad \square
\end{aligned}$$

Lemma 7. *Under Conditions C1–C7, if $m_n\tau_{nJ} = o(1)$ as $n \to \infty$ and $J \to \infty$, then for $h > 0$,*

$$\hat{\mu}_n = \mu_n + O_p(\tau_{nJ}), \quad \delta_n^2 = \|C - \mu_n I_n\|_F^2 + O(m_n\tau_{nJ}), \quad d_n^2 = \delta_n^2 + O_p(m_n\tau_{nJ}), \quad b_n^2 = O_p(m_n/J).$$

For $h > 0$, the above equalities continue to hold if m_n is replaced by n.

Proof. Without loss of generality, we assume $\pi(j) = j$, $1 \le j \le J$. It follows from Lemmas 4 and 5 that

$$\hat{\mu}_n = \frac{1}{n}\sum_{i}^{n}\hat{c}_{ii} = \frac{1}{n}\sum_{i}^{n} c_{ii} + O_p(\tau_{nJ}) = \mu_n + O_p(\tau_{nJ}). \quad \bar{b}_n^2 = O(m_n/J)$$

It follows from Ledoit and Wolf (2004) that $\|\mathrm{C}\|_F^2 = O(1)$, since $\max_i E[y_{i1}^4] < \infty$. We have

$$|E\left(\|\hat{\mathrm{C}}_h - \mu_n I_n\|_F^2 - \|\mathrm{C} - \mu_n I_n\|_F^2\right)| \le E\|\hat{\mathrm{C}}_h - \mathrm{C}\|_F^2 + 2\|\mathrm{C} - \mu_n I_n\|_F E\|\hat{\mathrm{C}}_h - \mathrm{C}\|_F = O(m_n\tau_{nJ}).$$

Note that

$$|d_n - \|\hat{\mathrm{C}}_h - \mu_n I_n\|_F| \le \|(\hat{\mu}_n - \mu_n)I_n\|_F = |\hat{\mu}_n - \mu_n| = O_p(\tau_{nJ}),$$

which implies that

$$d_n^2 = \left(\|\hat{\mathrm{C}}_h - \mu_n I_n\|_F + O_p(\tau_{nJ})\right)^2 = (\|\mathrm{C} - \mu_n I_n\|_F + O_p(m_n\tau_{nJ}) + O_p(\tau_{nJ}))^2 = \|\mathrm{C} - \mu_n I_n\|_F^2 + O_p(m_n\tau_{nJ}).$$

Therefore, $d_n^2 = \delta_n^2 + O_p(m_n\tau_{nJ})$. The proof is completed. $\square$

Lemma 8. *Under Conditions C1–C7, if $m_n\tau_{nJ} = o(1)$ and $\|C - \mu_n I_n\|_F$ is bounded below from zero as $n \to \infty$ and $J \to \infty$, then*

$$\|\hat{C}_{hs} - C\| = O_p(m_n\tau_{jn}), \quad \|\hat{C}_{hs}^{-1} - C^{-1}\| = O_p(m_n\tau_{nJ}), \quad \|\hat{C}_{hs}^{-2} - C^{-2}\| = O_p(m_n\tau_{nJ}).$$

Proof. Note that

$$\begin{aligned}
\|\hat{\mathrm{C}}_{hs} - \mathrm{C}\| &= \|\frac{b_n^2}{d_n^2}(I_n - \mathrm{C}) + \frac{d_n^2 - b_n^2}{d_n^2}(\hat{\mathrm{C}}_h - \mathrm{C})\| \le \frac{b_n^2}{d_n^2}\|I_n - \mathrm{C}\| + \frac{d_n^2 - b_n^2}{d_n^2}\|\hat{\mathrm{C}}_h - \mathrm{C}\| \\
&= O(m_n/J) + (1 - O(m_n/J))O_p(m_n\tau_{nJ}) = O_p(m_n\tau_{nJ}).
\end{aligned}$$

the remaining proofs are similar to the proof of Lemma 3 in Zhang and Liu (2015). The details are omitted. $\square$

In the remaining of the Appendix, we describe the MIS and the linear PVA as follows. In the MIS, we treat the multivariate additive random-effects model as multiple univariate additive models. We conducts the iterative univariate

nonparametric independence screening of Fan et al. [3] for active covariates in each of these univariate models. This gives rise to J potentially overlapped estimated subsets of active covariates. We then takes the union of these subsets as an estimated set of active covariates. The MIS can be implemented in the following steps.

- Step 1: for each $k \in [1:p]$, compute $\hat{f}_{kj} = \arg\min\{\|\mathbf{y}_j - \mathbf{f}_{kj}(\mathbf{x}_k)\|_2^2/n : \mathbf{f}_{kj} \in \mathcal{S}\}$, where $\mathcal{S}$ is the space of polynomial splines of degree $s \geq 1$ and s is pre-specified. Randomly permute the rows of $\mathbf{X}$, yielding $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_p)$. Let $\hat{f}_{kj}^* = \arg\min\{\|\mathbf{y}_j - \mathbf{f}_{kj}(\tilde{\mathbf{x}}_k)\|_2^2/n : \mathbf{f}_{kj} \in \mathcal{S}\}$. Then, we select the active set $\mathcal{A}_{j1} = \{k : \|\hat{f}_{kj}\|^2 \geq \min_k \|\hat{f}_{kj}^*\|^2\}$.
- Step 2: Apply the penalized procedure, penGAM of Meier, Geer, and Bühlmann [12] to the jth univariate additive model with the covariate set, say $\mathcal{A}_{j1}$, resulting in an estimated set of active covariates, say $\mathcal{M}_{j1}$.
- Step 3: For each $i \in \mathcal{M}_{j1}^c = \{1, \ldots, p\} \backslash \mathcal{M}_{j1}$, minimize $\|\mathbf{y}_j - \sum_{k \in \mathcal{M}_{j1}} f_{kj}(\mathbf{x}_k) - f_{ij}(\mathbf{x}_i)\|_2^2/n$ with respect to $f_{kj} \in \mathcal{S}$ for all $i \in \mathcal{M}_{j1}$ and $f_{ij} \in \mathcal{S}$. Following the same marginally screening procedure in Steps 1 and 2, except that only the rows of $(\mathbf{x}_k : k \in \mathcal{M}_{j1}^c)$ are randomly permuted, to select an additional subset of active covariates, $\mathcal{A}_{j2}$. Then, the penGAM is applied to the jth univariate additive model with the covariates restricted to $\mathcal{M}_{j1} \cup \mathcal{A}_{j2}$, giving rise to an updated estimate of active covariates, $\mathcal{M}_{j2}$.
- Step 4: Repeat Step 3 until $\mathcal{M}_{jl} = \mathcal{M}_{j,l-1}$ for some l. Let $\mathcal{M}_j = \mathcal{M}_{jl}$.
- Step 5: Take the union of $\mathcal{M}_j, j = 1, \ldots, J$ as the final estimate of active covariates.

In our simulation studies, we adjust the screening level in the MIS so that its specificity is close to a required level. For example, to make the specificity of MIS no less than α, we consider the following. We will take the union of $\mathcal{M}_j, j = 1, \ldots, J$ as the final estimate of active covariates when the cardinality of union of $\mathcal{M}_j, j = 1, \ldots, J \leq \lfloor(1-\alpha)p\rfloor$. Otherwise, we need to replace Step 5 by Step 5' below.

Step 5': Sort all the indices in the $\mathcal{M}_j, j = 1, \ldots, J$ in descending order by number of occurrences and take the $\lfloor(1-\alpha)p\rfloor$th index as the final estimate of active covariates.

In the linear PVA, consider the following multivariate regression model

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

where $\mathbf{Y}$, $\mathbf{X}$, B and E are the $n \times J$ response data matrix, the $n \times p$ design matrix, the $p \times J$ random-effects matrix and the $n \times J$ error matrix. The linear PVA is a SNR-based screening procedure for the above model. See Zhang and Oftadeh [21]. The details are omitted.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2022.105069. More details on numerical results can be found in the Online Supplementary Material. Part A: The boxplots of the simulated sensitivities and specificities show robustness of choosing tuning parameters in the fPVA. The boxplots of the CPU times show the advantage of the fPVA over the MIS. Part B: Tables 1 and 2 present the p-values of the Wilcoxon signed test for the significant cancer-gene pairs.

References

[1] P. Bickel, E. Levina, Covariance regularization by thresholding, Ann. Statist. 36 (2008) 2577–2604.
[2] C. de Boor, A Practical Guide To Splines, Springer, New York, 1978.
[3] J. Fan, Y. Feng, R. Song, Nonparametric independence screening in sparse ultra-high-dimensional additive models, J. Amer. Statist. Assoc. 106 (2011) 544–557.
[4] M.J. Garnett, et al., Systematic identification of genomic markers of drug sensitivity in cancer cells, Nature 483 (2012) 570–575.
[5] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. 6 (1979) 65–70.
[6] J. Huang, J.L. Horowitz, F. Wei, Variable selection in nonparametric additive models, Ann. Statist. 38 (2010) 2282–2323.
[7] V. Koltchinskii, M. Yuan, Sparsity in multiple kernel learning, Ann. Statist. 38 (2010) 3660–3695.
[8] N.M. Laird, J.H. Ware, Random-effects models for longitudinal data, Biometrics 38 (1982) 963–974.
[9] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multi. Analy. 88 (2004) 365–411.
[10] X. Lin, D. Zhang, Inference in generalized additive mixed models by using smoothing splines, J. Roy. Stat. Soc. B 61 (1999) 381–400.
[11] Y. Lin, H. Zhang, Component selection and smoothing in multivariate nonparametric regression, Ann. Statist. 34 (2006) 2272–2297.
[12] L. Meier, S. van de Geer, P. Bühlmann, High-dimensional additive modeling, Ann. Statist. 37 (2009) 3779–3821.
[13] P. Ravikumar, H. Liu, J. Lafferty, J. Wasserman, Sparse additive models, J. Roy. Stat. Soc. B 71 (2009) 1009–1030.
[14] R.A. Rigby, D.M. Stasinopoulos, Generalized additive models for location, scale and shape, J. Roy. Stat. Soc. C 54 (2005) 507–554.
[15] B.W. Stewart, C.P. Wild, International Agency for Research on Cancer. World Health Organization, WHO Press, New York, 2014.
[16] C.J. Stone, Additive regression and other nonparametric models, Ann. Statist. 13 (1985) 689–705.
[17] T.W. Yee, Vector Generalized Linear and Additive Models: With an Implementation in R, Springer, New York, 2015.
[18] T.W. Yee, C.J. Wild, Vector generalized additive models, J. Roy. Stat. Soc. B 58 (1996) 481–493.
[19] J. Zhang, C. Liu, On linearly constrained minimum variance beamforming, J. Mach. Learn. Res. 16 (2015) 2099–2145.
[20] J. Zhang, C. Liu, G. Green, Source localization with MEG data: A beamforming approach based on covariance thresholding, Biometrics 70 (2014) 121–131.
[21] J. Zhang, E. Oftadeh, Multivariate variable selection by means of null-beamforming, Electron. J. Stat. 15 (2021) 3428–3477.
[22] H. Zhang, G. Wahba, Y. Lin, M. Voelker, M. Ferris, M. Klein, B. Klein, Variable selection and model building via likelihood basis pursuit, J. Amer. Statist. Assoc. 99 (2004) 659–672.