Electronic Journal of Statistics
Vol. 15 (2021) 3428–3477
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1859

Multivariate variable selection by means of null-beamforming*

Jian Zhang[1] **and Elaheh Oftadeh**[1]

[1]*School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7FS, U.K.*
e-mail: jz79@kent.ac.uk; ela.oftadeh@gmail.com

Abstract: This article aims to use beamforming, a covariate-assisted data projection method to solve the problem of variable selection for multivariate random-effects regression models. The new approach attempts to explore the covariance structure in the data with a small number of random-effects covariates. The basic premise behind the proposal is to scan through a covariate space with a series of forward filters named null-beamformers; each is tailored to a particular covariate in the space and resistant to interference effects originating from other covariates. Applying the proposed method to simulated and real multivariate regression data, we show that it can substantially outperform the existing methods of multivariate variable selection in terms of sensitivity and specificity. A theory on selection consistency is established under certain regularity conditions.

MSC2020 subject classifications: Primary 62G08, 62M10; secondary 62P05.
Keywords and phrases: Multivariate random-effects regression models, principal variable analysis, multivariate variable selection, null-beamforming.

Received July 2020.

Contents

*The research of the second author was supported by a Graduate Teaching Assistant (GTA) scholarship at the University of Kent.

1. Introduction

The advance of high-throughput technology in science has generated various types of correlated data. Integrative analysis holds great promises for uncovering hidden links between these data. For this purpose, a class of multivariate random-effects regression models are investigated in this paper, where subject-specific random effects are introduced to account for the variations among subjects [11]. Suppose that there are J subjects (or responses in the terminology of multivariate regression) under study, each depending on the same set of random-effects covariates indexed by $\{1, ..., p\}$. Let $\mathbf{y}_j$ and $\mathbf{x}_k$ are column vectors of n measurements on subject j and on covariate k respectively. Then, a multivariate random-effects regression model can be written as

$$\mathbf{y}_j = \mathbf{x}_1\mu_1 + \cdots + \mathbf{x}_p\mu_p + \mathbf{x}_1(\beta_{1j} - \mu_1) + \cdots + \mathbf{x}_p(\beta_{pj} - \mu_p) + \boldsymbol{\varepsilon}_j, 1 \le j \le J,$$

with fixed-effects $\mu_k \in \mathrm{R}$, random-effects coefficients $\beta_{kj} \in \mathrm{R}$ and error vectors $\boldsymbol{\varepsilon}_j \in \mathrm{R}^n$. We assume that coefficient vectors $\boldsymbol{\beta}_j = (\beta_{1j}, ..., \beta_{pj})^T$ have mean $\mu = (\mu_1, ..., \mu_p)^T$ and covariance matrix Σ and that error vectors $\boldsymbol{\varepsilon}_j$ have mean zero and unknown covariance matrix Λ. We further assume that the random-effects coefficients and the error vectors are independent of each other. As estimation of the fixed-effects has already been studied in [17], we focus on inference of the random-effects coefficients in the above model.

Denote by $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ the observations on all covariates. Then the conditional covariance matrix $\mathrm{cov}(\mathbf{y}_j|\mathbf{X})$ admits the following decomposition

$$\mathrm{cov}(\mathbf{y}_j|\mathbf{X}) = \sum_{k=1}^{p} \sigma_{kk}\mathbf{x}_k\mathbf{x}_k^T + 2\sum_{1\le k<l\le p} \sigma_{kl}\mathbf{x}_k\mathbf{x}_l^T + \Lambda,$$

where σ_{kk} is called variance component associated with covariate k. This shows that the above random-effects model provides a way to describe the covariance structure of the data. However, when the number of covariates, p, is larger than both n and J, the problem of estimating either these variance components or random-effects coefficients is ill-posed, where the commonly used multivariate least squares criterion does not provide a unique solution. To tackle this issue, we impose a sparsity assumption on the model that only a small number of variance components in the above covariance matrix decomposition have positive values. We are interested in the problem of identifying these non-zero variance components. We refer to covariate k as an non-active covariate when $\beta_{k1} = \cdots = \beta_{kJ} =$ constant. By definition, $\sigma_{kk} = 0$ holds if and only if

$P(\text{covariate } k \text{ is non-active}) = 1$. So the above problem is equivalent to a multivariate variable selection problem where we want to infer active covariates for a multivariate regression model [17]. A conventional remedy for variable selection is to penalize the magnitudes of regression coefficients in the least squares. When the penalty is increasing, estimates are zeroed out, and a subset model is then identified and estimated. Such a remedy is particularly of interest when the dimension p is large and candidate covariates contain many redundant or irrelevant variables. The variable selection procedure LASSO [19] followed this remedy. Over the past two decades, much progress has been made along this direction [6, 26], among others. As the recent research on variable selection mainly focuses on a univariate response setting, limited research has been done on multiple responses settings, e.g., [2, 15, 16, 4, 17, 24, 13].

Despite of the above progress, a few issues remain to be addressed. First, most of these methods have been developed for independent measurements. There are various applications in which measurements on each subject are dependent. For instance, sensitivity measurements of a drug can be dependent as cell lines used in these measurements exhibit genetic relatedness when they are associated with the same types of cancers [10]. In multiple genome-wide association studies, individual genotypes in a subject group are correlated [25]. In neuroimaging, measurements from different sensors outside a brain are dependent as they are generated from the same neuronal sources inside the brain [20]. In finance, returns of different stocks are correlated due to the so-called cross-sectional dependence [9]. Secondly, the existing methods mentioned above are mainly for multivariate fixed-effects regression models, where given the values of covariates $\mathbf{X}$, the response covariance structure is determined only by error terms. In contrast, multivariate random-effects regression models are hierarchical, where conditional on the values of covariates, the responses depend not only on error terms but also on random-effects coefficients [11]. Although, in principle, multivariate regression data can be fitted by either a fixed-effects model or a random-effects model, a comparison between these two approaches has not been made in literature [2]. Finally, most of the existing inference procedures are not computationally scalable to large-scale data with many subjects or many responses. This prohibits their applications to big data.

Here, we address these issues by generalizing the idea of beamforming, a covariate-assisted data projection method [23] to multivariate regression settings. Our contributions are three-fold. First, we develop a novel algorithm called principal variable analysis (PVA) to identify important covariates by covariate-interference-adjusted data projections (called forward beamforming or null-beamforming) that account for the maximum amount of variation in the data. Such a procedure provides a principled way to extract information about covariates from the multivariate regression data. In the PVA, unlike the existing methods, we gauge the importance of each covariate with respect to the multivariate response by its information index (called power), which is defined by its variance component. The higher the power of a covariate, the more amount of variations in the response data it can account for. We estimate the power of each covariate by performing null-beamforming on the data. To adjust for varying

background noises, we replace the power by signal-to-noise ratio (SNR), a relative information index for each covariate. In each forward step, after nulling the previously selected covariates, we are able to adjust the SNR values of the remaining covariates and to conduct iteratively screening for these covariates. The iteration will be terminated once no covariate significantly stands out in the current step. The above procedure produces a list of highly ranked covariates called principal variables along with their estimated regression coefficients. Based on these selected covariates, the response covariance matrix can be decomposed into two parts: one for the selected variance components and the other for noises. In this sense, the PVA is viewed as a covariate-assisted principal component analysis. As the beamforming can be implemented through parallel computing, the PVA is scalable to large-scale data. Secondly, we establish a global property of selection consistency for the PVA under some regularity conditions. In particular, a sufficient condition for a consistent selection was imposed on the number of subjects, J, the number of covariates, p, the number of measurements per subject, n, and the number of non-zero entries in $\mathrm{cov}(\mathbf{y}_j|X)$, m_n, that is, $n^{-\alpha_0}\log(p)$ is bounded for some positive constant α_0 and $m_n\sqrt{\log(n)/J} \to 0$, as n, J and p tend to infinity. This implies that the proposed procedure can handle the variable selection problem in an ultra-high dimensional covariate space. Finally, we conduct a set of simulation studies to evaluate the performance of the PVA compared to the existing variable selection methods. The numerical results demonstrate that in terms of sensitivity and specificity the PVA can substantially outperform the existing methods such as the multivariate group LASSO, the multivariate elastic-net, the multivariate LASSO, the multivariate sparse group LASSO, among others. We also apply our method to some anti-cancer drug data, identifying a novel set of genes for predicting drug sensitivity in cancer cell lines. Using the information extracted from the Human Protein Atlas Portal at `http://www.proteinatlas.org/cancer`, we show that most of the identified genes have significantly high protein staining levels at least in one or more than one of common cancers.

The remaining of the paper is organized as follows. The details of the proposed methodology and algorithm are provided in Section 2. An asymptotic theory on the proposed procedure is developed in Section 3. The simulation studies and a real data application are presented in Section 4. The discussion and conclusion are made in Section 5. The technical details, proofs, and extra theorems can be found in the Appendices A and B. 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 F_n, we define the spectral norm $||\mathrm{F}_n||$ as $\lambda^{1/2}_{\max}(\mathrm{F}_n^T\mathrm{F}_n)$. For a sequence of real numbers $\{u_n\}$, we say $\mathrm{F}_n = O(u_n)$ if $||\mathrm{F}_n||/|u_n|$ is bounded from above and $\mathrm{F}_n = o(u_n)$ if $||\mathrm{F}_n||/|u_n|$ tends to zero as n tends to infinity.

2. Methodology

Let $\mathbf{Y} = \mathbf{Y}_{n\times J} = (y_{ij})_{n\times J} = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_J)$ and $\mathbf{X} = \mathbf{X}_{n\times p} = (x_{ik})_{n\times p} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$. We reformulate the multivariate random-effects regression model

in the previous section in the following matrix form:

$$\mathbf{Y} = \mathbf{X}\mathrm{B} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where unknown random regression coefficient matrix $\mathrm{B} = \mathrm{B}_{p\times J} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_J)$ and $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_{n\times J} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \cdots, \boldsymbol{\varepsilon}_J)$ with $\boldsymbol{\beta}_j$ and $\boldsymbol{\varepsilon}_j$, respectively, containing the regression coefficients and the error terms related to the jth subject. As usual, we start with a least squares-based regression analysis. It can be shown that when $p < n$, the least square solution, $\hat{\mathrm{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ gives the same coefficients as fitting univariate multiple regression models to $(\mathbf{y}_j, \mathbf{X}), 1 \le j \le J$ separately. Note that treating $\beta_{kj}, 1 \le j \le J$ as correlated random coefficients allows us to explore the dependence between $\mathbf{y}_j, 1 \le j \le J$. As pointed out before, when $p > n$, $\mathbf{X}^T\mathbf{X}$ is not invertible and thus the above least squares solution is not unique. To tackle the problem, we make the assumption that variance components in a decomposition of $\mathrm{cov}(\mathbf{y}_j|\mathbf{X})$ are sparse, that is, $\sigma_{kk} = 0$ for the majority of covariates $k \in \{1, 2, ..., p\}$. The goal of this paper is to identify these covariates of non-zero variance component and to estimate their regression coefficients given observations $(\mathbf{Y}, \mathbf{X})$.

2.1. Power and signal-to-noise ratio

To rank covariates, we define an information index called power for each covariate by projecting the response data to the covariate space. The concept of power, defined as the variance of a signal, is borrowed from the research field of signal processing, where sensor observations $\mathbf{y}_j, 1 \le j \le J$ are often assumed weakly stationary [20, 23]. In genetics, the above concept describes the so-called pleiotropic genetic effect of a single gene on multiple phenotypic traits, where multivariate linear models have been developed to connect genetic variant data to multiple quantitative traits [5]. In the multivariate random-effects regression setting, the power is the variance component of a covariate in conditional covariance matrix $\mathrm{C} = \mathrm{cov}(\mathbf{y}_j|\mathbf{X})$ given $\mathbf{X}$, where we model regression coefficients of the multiple responses to each covariate as realizations from a random variable with a finite second moment. Then, the amount of information on each covariate in these regression coefficients can be measured by variability in these coefficients. The larger the variability, the higher degree of variation in the response data is accounted for by this covariate. In practice, the regression coefficients $(\beta_{kj})_{1\le j\le J}$ at covariate k (therefore its approximate power $\sum_{j=1}^{J}(\beta_{kj} - \bar{\beta}_k)^2/J$ with $\bar{\beta}_k = \sum_{j=1}^{J}\beta_{kj}/J$ as J tends to infinity) are unknown. We estimate $(\beta_{kj})_{1\le j\le J}$ by projecting response data into the coefficient space of the kth covariate along the direction $\mathbf{w}$ that can minimize interferences with the other covariates and with the background noise. That is, for the kth covariate, we estimate its regression coefficients by the projected data $\mathbf{w}^T\mathbf{Y}$ in which $\mathrm{var}(\mathbf{w}^T\mathbf{y}_j|\mathbf{X}) = \mathbf{w}^T\mathrm{C}\mathbf{w}$ attains the minimum, subject to $\mathbf{w}^T\mathbf{x}_k = 1$. To this end, we consider the Lagrange multiplier $L(\mathbf{w}, \lambda) = \mathbf{w}^T\mathrm{C}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{x}_k - 1)$

and solve partial derivative equations

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = 2\mathrm{C}\mathbf{w} - \lambda \mathbf{x}_k = 0, \quad \frac{\partial L(\mathbf{w})}{\partial \lambda} = \mathbf{w}^T \mathbf{x}_k - 1 = 0.$$

We have the solution $\mathbf{w}_k = \mathrm{C}^{-1}\mathbf{x}_k / \mathbf{x}_k^T \mathrm{C}^{-1}\mathbf{x}_k$. We project $\mathbf{Y}$ along the direction $\mathbf{w}_k$ to give an estimator $\mathbf{w}_k^T \mathbf{Y} = \mathbf{x}_k^T \mathrm{C}^{-1}\mathbf{Y} / \mathbf{x}_k^T \mathrm{C}^{-1}\mathbf{x}_k$ for $(\beta_{kj})_{1 \le j \le J}$ [23]. If let $\mathrm{C} = \text{constant} \times I_n$, where I_n is an $n \times n$ identity matrix (i.e., ignoring correlations in the measurements on each subject), then the above estimator reduces to a marginal multivariate least squares estimator. To explain why the above approach can provide an interference-minimized estimator of the underlying power, we assume that error term $\boldsymbol{\varepsilon}_j$ is independent of the p-dimensional regression coefficient $\boldsymbol{\beta}_j$ and with $\mathrm{cov}(\boldsymbol{\varepsilon}_j) = \Lambda$ and $\mathrm{cov}(\boldsymbol{\beta}_j) = \Sigma = (\sigma_{i_1 j_1})_{p \times p}$. Then, we have $\mathrm{C} = \mathbf{X}\Sigma\mathbf{X}^T + \Lambda$. Note that under the constraint $\mathbf{w}^T \mathbf{x}_k = 1$, we have

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

which yields

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

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

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

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

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

$\mathbf{w}^T\mathbf{x}_\nu = I_m\}$, where $\mathrm{tr}(\cdot)$ is the trace operator and I_m is an $m \times m$ identity matrix. Note that $\mathbf{w}^T\mathbf{x}_\nu = I_m$ define m linear filters which null each other. The projection of $\mathbf{Y}$ along interference-minimized directions gives an estimator, $(\mathbf{x}_\nu^T \mathrm{C}^{-1}\mathbf{x}_\nu)^{-1}\mathbf{x}_\nu^T\mathrm{C}^{-1}\mathbf{Y}$, for random coefficient matrix B. The above estimator will reduce to a marginal least squares estimator if let $\mathrm{C} = \text{constant} \times I_n$. However, in practice, C is unknown and often not diagonal. We need estimate C by the data.

Covariates can be correlated. For example, in the cancer genomic data, genes as covariates can be highly correlated if they are located in the same pathway. Consequently, the finite sample power estimator of a covariate may have a bias due to interferences with other covariates. To address this problem, we further null the previously identified covariates by adding more constraints on the linear filter in each step as follows. Let ω and ν be two disjoint subsets of the covariates with sizes m_1 and m respectively. To define a ω-nulled power matrix of ν, adding null constraints $\mathbf{w}^T\mathbf{x}_\omega = 0_{m\times m_1}$ into the linear filters $\mathbf{w}^T\mathbf{x}_\nu = I_m$, we consider the following optimization problem:

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

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

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

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

2.2. Estimation of response covariance matrix

Note that the power estimation needs an estimator of the response covariance matrix, for example, the sample covariance matrix $\hat{\mathrm{C}} = \sum_{j=1}^J \mathbf{y}_j\mathbf{y}_j^T/J - \bar{\mathbf{y}}\bar{\mathbf{y}}^T = (\hat{c}_{ij})$, where $\bar{\mathbf{y}} = \sum_{j=1}^J \mathbf{y}_j/J = (\bar{y}_1, ..., \bar{y}_n)^T$ and $\hat{c}_{ij} = \sum_{t=1}^J (y_{it} - \bar{y}_i)(y_{jt} - \bar{y}_j)/J$. As the sample covariance matrix can be inconsistent with the true one when the dimension n is larger than J, [1] and [3] amended it by thresholding its entries: $\hat{\mathrm{C}}_h = \hat{\mathrm{C}}(\tau_{nJ}) = (\hat{c}_{ij}I(|\hat{c}_{ij}| > h\tau_{nJ}))$, where $I(\cdot)$ is the indicator and $\tau_{nJ} = \sqrt{\log(n)/J}$ with the tuning constant $h \geq 0$. Under certain mixing conditions, [23] showed that the thresholded sample covariance matrix was consistent with the true one with dependent sample. For a finite sample, the thresholded covariance matrix may still be degenerate when the number of subjects is close to or smaller than the number of measurements per subject. So, following [12], we further shrink the thresholded covariance estimator to a diagonal matrix as follows:

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

where

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

$$\hat{\mu}_n = <\hat{\mathrm{C}}_h, I_n>, \quad d_n^2 = <\hat{\mathrm{C}}_h-\hat{\mu}_n I_n, \hat{\mathrm{C}}_h-\hat{\mu}_n I_n>, \quad b_n^2 = \min\{\bar{b}_n^2, d_n^2\},$$

and $<\mathrm{D}_1, \mathrm{D}_2> = \mathrm{tr}(\mathrm{D}_1\mathrm{D}_2^T)/n$ for any $n\times n$ matrices D_1 and D_2. Having defined $\hat{\mathrm{C}}_{hs}$, we estimate the power matrices γ_ν and $\gamma_{\nu|\omega}$ by $\hat{\gamma}_\nu = (\mathbf{x}_\nu\hat{\mathrm{C}}_{hs}\mathbf{x}_\nu)^{-1}$ and $\hat{\gamma}_{\nu|\omega} = \phi_{\nu|\omega}^T(\mathbf{x}_{\nu\cup\omega}^T\hat{\mathrm{C}}_{hs}\mathbf{x}_{\nu\cup\omega})^{-1}\phi_{\nu|\omega}$ respectively. Similarly, the ω-nulled SNR can be estimated by

$$\hat{\mathrm{SNR}}_{\nu|\omega} \propto \mathrm{tr}\left(\hat{\gamma}_{\nu|\omega}(\hat{\mathbf{w}}_{\nu|\omega}^T\hat{\mathbf{w}}_{\nu|\omega})^{-1}\right), \tag{2.2}$$

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

2.3. Principal variable analysis

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

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

Nulling: In the iteration m, $m \ge 2$, let ω_{m-1} denote the set of covariates selected in the first $m-1$ iterations. For any covariate k not in ω_{m-1}, using the formula (2.2), we calculate the nulled SNR, $\hat{\mathrm{SNR}}_{\{k\}|\omega_{m-1}}$, as well as an estimated optimal projection direction $\hat{\mathbf{w}}$. We then find $k_m \notin \omega_{m-1}$ in which $\hat{\mathrm{SNR}}_{\{k\}|\omega_{m-1}}$ attains the maximum.

Forward selection and stopping criteria: After a number of iterations, the nulled SNR values will start leveling off, which indicates that the remaining covariates have no significant contributions to the covariance structure of the response. This motivates us to set the following stopping criteria in each iteration: For $m \ge 2$, at the end of the mth iteration, we make a scree plot of the nulled SNR values and identify an elbow point. To find the elbow point, we consider the vector which links the highest and the lowest points on the scree plot. Then we find the orthogonal distance from each point on the plot to this vector. The point on the plot with the largest distance is selected as the elbow point. The elbow point partitions the remaining covariates into two subsets, namely upper set and lower set. The lower set, containing those covariates with SNR values lower than the elbow point, is uninformative about the responses. To test the hypothesis that the upper set is uninformative, we calculate the mean μ_l and standard deviation θ_l for the lower subset. The hypothesis is accepted if the maximum nulled SNR value, $\hat{\mathrm{SNR}}_{\max} = \max\{\hat{\mathrm{SNR}}_{k|\omega_{m-1}} : k \notin \omega_{m-1}\}$, of the upper set falls into the following confidence interval, $|\hat{\mathrm{SNR}}_{\max} - \mu_l| \le c_0\theta_l$, where c_0 is a tuning constant. We set the default value $c_0 = 5$. Applying the central limit theorem to the SNR values in the lower set, the above interval can be shown to

have the asymptotic confidence level of $1 - 5.73 \times 10^{-7}$ after multiple testing adjustment. The iteration will be terminated when the upper subset is uninformative. Otherwise, we update ω_{m-1} and $\mathbf{x}_{\omega_{m-1}}$ by letting $\omega_m = \{k_m\} \cup \omega_{m-1}$ and $\mathbf{x}_{\omega_m} = (\mathbf{x}_{k_m}, \mathbf{x}_{\omega_{m-1}})$, and the iteration will continue. Note that our simulations (not shown here) did indicate that the performance of PVA was not very sensitive to the choice of c_0 when it took values between 3 and 5.

2.4. Covariate network

Statistical connectivity patterns in the selected covariates are a hallmark feature for connecting pleiotropic traits such as drug inhibitory concentrations to genetic variants in genetics and for studying functional networks in neuroscience [5, 14]. Here, to quantify such patterns, we compute the regression coefficient-based Pearson correlation coefficient for each pair of the selected covariates. The details are as follows. Suppose that q covariates are selected by the PVA. Based on the multivariate least squares, we obtain $\hat{\mathrm{B}}_0$, an estimator of the $q \times J$ regression coefficient matrix for these covariates. For any pair of rows (i, j) in $\hat{\mathrm{B}}_0$, we calculate Fisher's z-transformation of their correlation coefficient r_{ij}, $z_{ij} = 0.5 \ln((1 - r_{ij})/(1 + r_{ij}))$. For rows $i < j$, we want to test whether z_{ij} (i.e., r_{ij}) is significantly away from 0. There are $q(q-1)/2$ such tests in total. Note that if the underlying correlation coefficient is zero, then $z_{ij} \approx N(0, 1/(J-3))$ in distribution. After Bonferroni correction to multiple testing, we can claim that z_{ij} is significantly away from zero if $\sqrt{J-3}|z_{ij}| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the critical value of $N(0,1)$ at the level $\alpha/2 = 0.01/q(q-1)$. For example, for our cancer data in Section 4 where $q = 37, J = 131$, we obtained $z_{\alpha/2} = 4.33$. We are now ready to construct a network with q nodes, each stands for a selected covariate (a row in $\hat{\mathrm{B}}_0$). We assign an edge to link nodes i and j if z_{ij} is significantly away from zero.

3. Theory

In literature, no general asymptotic theory was provided on variable selection for high dimensional multivariate regression models with the exception of [17]. In [17], Sofer et al. developed a selection consistency theory for a special class of multivariate fixed-effects regression models, where regression coefficients did not change across responses (i.e., $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_J$). In this section, we develop a general theory on selection consistency of the proposed procedure PVA for multivariate random-effects regression models. We divide the theory into two parts according to whether C is known or not. Here, we present only the case where C is estimated. The remaining is deferred into the Appendix B.

As before, assume that regression coefficient matrix B and error terms $\boldsymbol{\varepsilon}$ in the model (2.1) are independent and that given $\mathbf{X}$, the covariance matrices of $\mathbf{y}_j$, $\boldsymbol{\beta}_j$ and $\boldsymbol{\varepsilon}_j$, denoted by $\mathrm{C} = (c_{ik})_{n\times n}$, $\Sigma = (\sigma_{ik})_{p\times p}$ and Λ respectively, are independent of index j. Then, we have $\mathrm{C} = \mathbf{X}\Sigma\mathbf{X}^T + \Lambda$. Assume that Λ is positively definite. For ease of presentation, we consider the special case,

where $\Lambda = \sigma^2 I_n$ and $\mathbf{x}_k^T \mathbf{x}_k = n, 1 \le k \le n$. If $\Lambda \ne \sigma^2 I_n$, we can change $\mathbf{Y}$ and $\mathbf{X}$ by the transformations $\Lambda^{-1/2}\mathbf{Y}$ and $\Lambda^{-1/2}\mathbf{X}$ (under which the power is invariant), followed by rescaling $\Lambda^{-1/2}\mathbf{X}$ and B (see [23]). Then a general theory can be derived from the special case. We denote the full set of covariates by $[1:p] = \{1, 2, \cdots, p\}$ corresponding to $\mathbf{x}_1, \cdots, \mathbf{x}_p$, and the true covariate set by ν_0. Let $\nu = \{k_1, \cdots, k_{p_1}\}$ denote any subset of $[1:p]$ with size $|\nu|$. The $(k_1, \cdots, k_{p_1})$th columns of $\mathbf{X}$ forms a data matrix $\mathbf{x}_\nu$ for the covariate set ν. If let $\mathbf{e}_\nu$ be a $p \times p_1$ selection matrix in which for $1 \le j \le p_1$, its (k_j, j)th entry takes value of 1 and the other entries take values of 0, then we can write $\mathbf{x}_\nu = \mathbf{X}\mathbf{e}_\nu$. Let σ_k^2 denote σ_{kk} in Σ, which shows the underlying power at the kth covariate. Let ν_0 be the underlying set of covariates. Let $A_{\nu_0} = \mathrm{C} - \mathbf{x}_{\nu_0}\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\mathbf{x}_{\nu_0}^T$, the underlying noise covariance matrix. For any subset ν, if A_{ν_0} is invertible, then we define the coherence (i.e., collinearity) matrices within and between $\mathbf{x}_\nu$ and $\mathbf{x}_{\nu_0}$: $\mathrm{R}_{\nu\nu} = \mathbf{x}_\nu^T A_{\nu_0}^{-1}\mathbf{x}_\nu/n$, $\mathrm{R}_{\nu\nu_0} = \mathbf{x}_\nu^T A_{\nu_0}^{-1}\mathbf{x}_{\nu_0}/n$, $\mathrm{R}_{\nu_0\nu_0} = \mathbf{x}_{\nu_0}^T A_{\nu_0}^{-1}\mathbf{x}_{\nu_0}/n$. Suppose that for $\nu_0 = \{k_1, ..., k_{p_0}\}$ and for any $\nu \subseteq \nu_0$, we can find $j_{[1:m]} = \{j_1, ..., j_m\} \subseteq \{1, ..., p_0\}$ such that $\nu = \{k_j : j \in j_{[1:m]}\}$. Let $\mathbf{e}_{\nu \lhd \nu_0}$ be a $|\nu_0| \times |\nu|$ indicator matrix with the (j_l, l)th entry equal to 1, $1 \le l \le |\nu|$ and with other entries equal to zeros. Using $\mathbf{e}_{\nu \lhd \nu_0}$, we select sub-columns from $\mathbf{x}_{\nu_0}$ to form $\mathbf{x}_\nu$, namely $\mathbf{x}_\nu = \mathbf{x}_{\nu_0}\mathbf{e}_{\nu \lhd \nu_0}$.

To identify active covariates, we impose the following regularity conditions on the covariance structures of the multivariate response variable and covariates, where $\mathbf{X}$ is treated as deterministic. If we treat $\mathbf{X}$ as a random design matrix, some parallel conditions can be assumed through replacing $O(\cdot)$ by $O_p(\cdot)$ in the following conditions.

(C0). There exists a permutation on $\mathbf{y}_j, 1 \le j \le J$ so that the resulted sequence is strictly stationary with covariance matrix C and that $(\mathbf{y}_j, \mathbf{X}), 1 \le j \le J$ follow the model (2.1). The error term $\boldsymbol{\varepsilon}_j$ and the p-dimensional regression coefficient $\boldsymbol{\beta}_j$ are independent of each other.

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

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

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

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

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

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

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

In the last condition, we assume that there exists a permutation π on $\{1, ..., J\}$ so that $\mathbf{y}_{\pi(j)}, 1 \le j \le J$ are strong mixing. Let $\mathcal{F}_0^{k_0}$ and $\mathcal{F}_k^\infty$ denote the σ-

algebras generated by $\{\mathbf{y}_{\pi(j)} : 0 \le j \le k_0\}$ and $\{\mathbf{y}_{\pi(j)} : j \ge k\}$ respectively. Define the mixing coefficient

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

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

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

Remark 3.1. *Note that under Condition* $(C0)$, $\boldsymbol{y}_j$*'s (therefore* $\boldsymbol{\varepsilon}_j$*'s) can be mutually dependent on each other. Condition (C1) says that there are no redundant covariates in* ν_0*. Condition (C2) implies that* $||R_{\nu\nu_0}|| \le ||R_{\nu\nu}||^{1/2}||R_{\nu_0\nu_0}||^{1/2} = O(n^{\alpha_0})$*. Note that for* $\nu \subseteq [1:p] \setminus \nu_0$*,* $R_{\nu\nu}^{-1}/n = (\boldsymbol{x}_\nu^T A_{\nu_0}^{-1} \boldsymbol{x}_\nu)^{-1}$ *is the power of* ν *after adjusting the influence of* ν_0*. So, Condition (C2) says that the adjusted power of* ν *is of order* $n^{-1+\alpha_0} = o(1)$*, which is negligible. This is natural as* ν *may contain noisy covariates. Similarly, Condition (C3) says that* ν_0*-adjusted power of* ν *is also negligible. Conditions (C2) to (C3) are the assumptions commonly used in the large sample theory for linear regression models (e.g., [21]). To verify Conditions (C1)∼(C3), we refer readers to [8, 21] under the assumptions that* $\sigma_k^2 = 0$*,* $k \notin \nu_0$ *and that* $\boldsymbol{X}$ *is assumed to be a random matrix satisfying some moment conditions and that the growth of the dimension* p *is not too fast compared to the number of measurements per response,* n*. For example, following [8], we assume that* $\boldsymbol{X}$ *has a concentration property, i.e., for some constant* c_1*, any* $u > 0$ *and* $\nu \subseteq [1:p]$*,* $|\nu| \le rn$*,*

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

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

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

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

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

This implies that Condition (C2) holds with an overwhelming probability. Analogously, Condition (C3) holds if $\boldsymbol{x}_\nu$ *and* $\boldsymbol{x}_{\nu_0}$ *are asymptotically, uniformly non coherent with respect to* $\nu \subseteq [1:p] \setminus \nu_0$*, in the sense that* $R_{\nu\nu_0} = o(1)$*.*

Condition (C4) is a technical condition which holds when $\sigma_k^2 = 0$ (or sufficiently close to zero in a sense), $k \notin \nu_0$. Note that (C5) holds if y_{ij}'s are Gaussian. And (C6) holds if there exist $1 = j_0 < j_1 < \cdots < j_m = J$ such that $\{\boldsymbol{y}_j\}_{1\le j\le J}$ can be divided into mutually independent segments $\{\boldsymbol{y}_j\}_{j_{k-1}\le j<j_k}, 1 \le k \le m$.

Letting $\nu_1 \subseteq \nu_0$ and $\nu_2 \subseteq [1:p] \setminus \nu_0$, in the next theorem, we show that the sparsistency property holds for the estimated nulled powers. Recall that $\tau_{nJ} = \sqrt{\log(n)/J}$.

Theorem 3.1. *Suppose that Conditions (C0)∼(C6) hold and that $\tau_{nJ} n^2 = o(1)$ as both n and J tend to infinity. Then, we have:*

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

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

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

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

where

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

The above theorem implies that uniformly for $a \in \nu_0\setminus\nu_1$ and $|\nu_1\cup\nu_2| < rn$, the $(\nu_1 \cup \nu_2)$-nulled power of a admits the form $\hat{\gamma}_{a|\nu_1\cup\nu_2} = (\mathbf{e}_{a\lhd\nu_0}^T\Sigma_{\nu_0\setminus\nu_1}^{-1}\mathbf{e}_{a\lhd\nu_0})^{-1} + O_p(n^{-1+3\alpha_0+2\alpha_1} + n^2\tau_{nJ})$. For $a \notin \nu_0$, $\hat{\gamma}_{a|\nu_1\cup\nu_2} = o_p(1)$. Note that it can be seen from the proofs in the Appendix B that $n^2\tau_{nJ}$=o(1) in Theorem 3.1 can be replaced by $m_n\tau_{nJ} = o(1)$ which depends on m_n, the number of non-zero entries

in C. We therefore show that the so-called sparsistency property holds for the nulled-power-based PVA. We further show that the sparsistency property also holds for the SNR-based PVA as follows.

Theorem 3.2. *Suppose that Conditions (C0)∼(C6) hold and that* $\tau_{nJ}n^2 = o(1)$ *as both* n *and* J *tend to infinity. Then, we have:*

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

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

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

where $\Sigma_{\nu_0\setminus\nu_1}^{-1}$, $\boldsymbol{\Psi}$ *and* $\boldsymbol{\Phi}$ *are defined in Theorem 3.1.*

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

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

which is bounded below from zero as $\lambda_{\max}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right) = O(1)$. It can also be shown that $\sigma^2\zeta_0\mathbf{e}_{a\lhd\nu_0}^T\Sigma_{\nu_0\setminus\nu_1}^{-1}\boldsymbol{\Phi}\Sigma_{\nu_0\setminus\nu_1}^{-1}\mathbf{e}_{a\lhd\nu_0} = O(n^{3\alpha_0+2\alpha_1})$. Consequently, the leading term in Theorem 3.2 (ii) tends to infinity as $n^{1-3\alpha_0-2\alpha_1}$ tends to infinity. In contrast, for $a \notin \nu_0$, $\hat{\text{SNR}}_{a|\nu_1\cup\nu_2}$ converges to a constant as stated in Theorem 3.2 (i). Compared to Theorem B.4 in the Appendix B, we can see that Theorem 3.2 provides a sharper contrast between active and non-active covariates.

Let $\omega_{\hat{m}}$ denote the set of covariates derived from the (SNR-based) PVA. We have the following selection consistency for $\omega_{\hat{m}}$.

Corollary 3.1. *Under the conditions in Theorem 3.2, as both* n *and* J *tend to infinity, we have the selection consistency in the sense that* $P(\omega_{\hat{m}} = \nu_0) \to 1$.

The above corollary, together with Remark 3.1, implies that along with other regularity conditions, if the condition on J, p and n that $n^{-\alpha_0}\log(p) = O(1)$ for some positive constant α_0 and that $n^2\tau_{nJ} = o(1)$ is satisfied, then the PVA-based variable selection is consistent. As pointed out before, the condition $n^2\tau_{nJ} = o(1)$ can be replaced by a sparsity condition where $m_n\tau_{nJ} = o(1)$.

4. Numerical results

In this section, we assess the performance of the PVA in identifying active covariates using synthetic and real data. As our simulations suggest that the SNR-based PVA performs better than the power-based PVA, we consider four versions of the SNR-based PVA with the four different estimators of C, namely, Ledoit-Wolf's shrinkage estimator and the optimal shrinkage of thresholded estimator $\hat{\mathrm{C}}_{hs}$ with $h = 0.01, 0.005, 0.001$. They are denoted by PVA(sh0), PVA(hs1), PVA(hs2) and PVA(hs3) respectively.

4.1. Synthetic data

We compare the performance of the PVA to those implemented in the R-packages 'glmnet' (Friedman, Hastie, Simon, Tibshirani, version 2.1), 'lsgl' (Vincent, version 1.3.5) and 'mrce' (Rothman, version 2.1): the multivariate group LASSO (MGL), the multivariate elastic-net (MENET), the multivariate LASSO (ML), the multivariate group sparse LASSO (MGSL) and multivariate regression with covariance estimation (MRCE) when all these procedures fix their specificity values approximately at the same level as the PVA. A brief introduction to these methods can be found in the Appendix A. The Bayesian method of [2] is excluded from our comparison as it is computationally infeasible for the large scale data considered here.

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

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

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

Setting 4.1 (B was uncorrelated both within rows and between rows): Modifying a simulation setting in [16], we simulated 50 data sets of $(\mathbf{Y}, \mathbf{X})$ from the model (2.1). Each dataset was generated in the following steps. First, we drew an i.i.d. sample of size np from the standard normal $N(0, 1)$ to form an $n \times p$ matrix $\mathbf{X}$. Secondly, we drew n independent auto-regressive row-vectors from the J-dimensional multivariate normal $N_J(0, E_0)$, where $E_0 = (0.7^{|i-j|})_{J \times J}$. We stacked these row vectors to generate an $n \times J$ error term matrix $\boldsymbol{\varepsilon}$. Thirdly, we generated $\mathrm{B} = (\beta_{kj})_{p \times J} = s_0 \mathrm{B}_0$, where s_0 was a scale factor, $\mathrm{B}_0 = (b_{kj})_{p \times J}$, $b_{kj} = \eta_{kj} u_{kj}$, with η_{kj} and u_{kj} independently sampled from the Bernoulli distribution Bin(0.1) (0.1 is the success probability) and the uniform distribution $U(s_1, s_2)$ respectively. We considered combinations of $(n, p, J, p_0, \alpha, s_0, s_1, s_2)$ with $n = 50$, $p = 100, 1000$, $J = 20$, $p_0 = 5$, $\alpha = 0, 1$, $s_0 = 0.45, 0.6,$

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

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

Setting 4.3 (B was weakly correlated within rows): We generated 50 data sets of $(\mathbf{Y}, \mathbf{X})$ from the model (2.1) for each combination of (n, p, J, p_0), where $n = 42, 88, 150$, $p = 2000$, $J = 20, 34, 131$, and $p_0 = 37, 50, 70$ is the number of true active covariates underpinning the model. Each dataset was generated in the following steps. We began with calculating a $J \times J$ sample covariance matrix Ω by using the $n\times J$ weakly correlated sub-data matrix of the imputed IC50 data. Given Ω, we randomly generated p row-vectors from a J-dimensional normal $N_J(\mathbf{0}, \Omega)$, stacking them together to form a matrix B. We then modified entries of B so that the resulting matrix contained exactly p_0 non-zero rows which would be taken as p_0 active covariates later. The details were omitted. To obtain matrix $\mathbf{X}$, we let F_0 be the $p\times p$ sample covariance matrix of the gene expressions in our cancer drug data which were obtained in the next section. Given F_0, we then generated n iid row vectors from a multivariate normal $N_p(\mathbf{0}, \mathrm{F}_0)$, stacking them together to form matrix $\mathbf{X}$. We generated the error term matrix, $\boldsymbol{\varepsilon}$, by sampling from $N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ J times as its column vectors, where $\sigma^2 = 0.1$. Finally, we obtained $\mathbf{Y}$ by setting $\mathbf{Y} = \mathbf{X}\mathrm{B} + \boldsymbol{\varepsilon}$.

Setting 4.4 (B was strongly correlated within rows): Similar to Setting 4.3, we generated 50 data sets of $(\mathbf{Y}, \mathbf{X})$ from the model (2.1) for each combination of (n, p, J, p_0), where $n = 42, 88, 150$, $p = 2000$, $J = 20, 34, 131$, and $p_0 = 37, 50, 70$ is the number of true active covariates underpinning the model. Each dataset was generated in the same steps as Setting 4.3, except that matrix Ω was replaced by one with high correlation coefficients. The further details were omitted.

Setting 4.5 (B was moderately correlated within rows): Similar to Setting 4.3, we generated 50 data sets of $(\mathbf{Y}, \mathbf{X})$ from the model (2.1) for each combination of (n, p, J, p_0), where $n = 20, 42$, $p = 2000$, $J = 131$, and $p_0 = 20, 37$. Here, Ω was generated from the n non-missing rows of the IC50 data while $\mathbf{X}$ was produced by use of the gene expression data corresponding to the above n non-missing rows. The error term matrix was generated by sampling from $N_n(0, \sigma^2 I_n)$ J times as before but with $\sigma^2 = 0.0645$. The further details were omitted.

For each combination of $(n, p, J, p_0, s_0, s_1, s_2)$ in Settings 4.1 and 4.2, we applied the PVA, MGL, MENET, ML, MSGL and MRCE to each of 50 data sets respectively and calculated their sensitivity values when the specificity value was fixed approximately at the same level. Note that in Settings 4.3 to 4.5, it

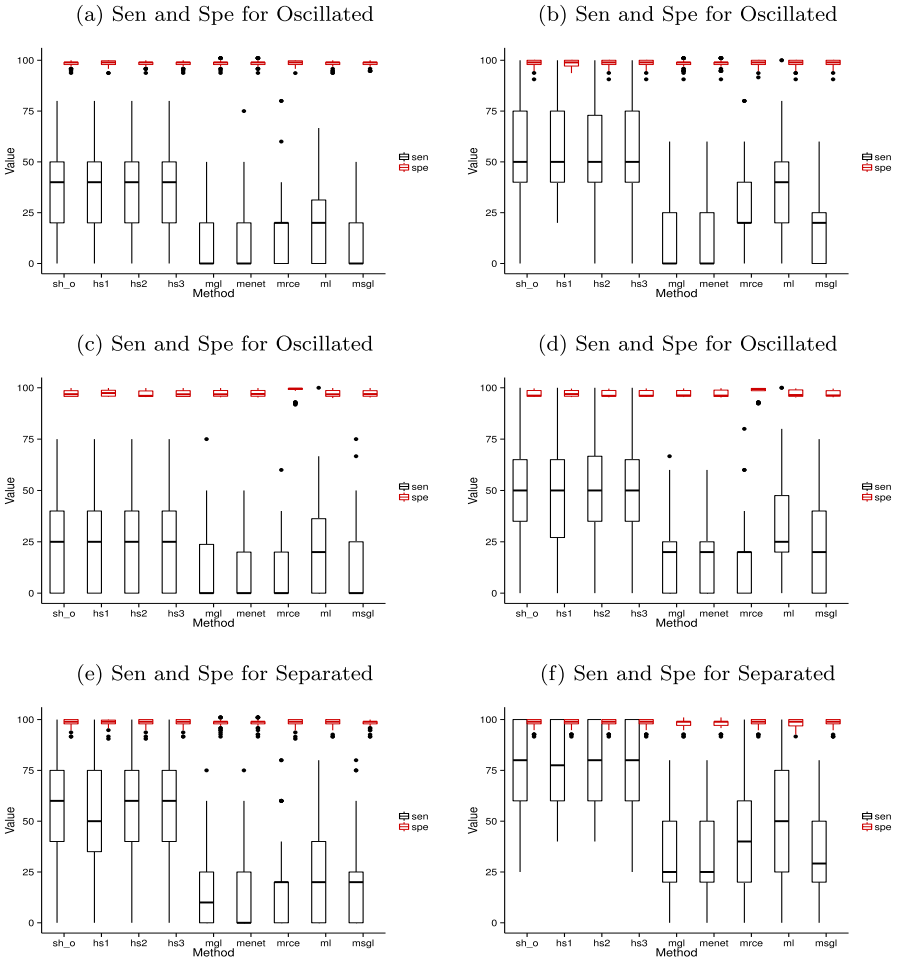

FIG 1. *Box plots of sensitivity (short for Sen) and specificity (short for Spe) values are for Settings 4.1 with* $(n, p, J, p_0, s_0, s_1, s_2)$ *taking the following values. (a):* $(50, 100, 20, 5, 0.45, -1, 1)$. *(b):* $(50, 100, 20, 5, 0.6, -1, 1)$. *(c):* $(50, 1000, 20, 5, 0.45, -1, 1)$. *(d):* $(50, 1000, 20, 5, 0.6, -1, 1)$. *(e):* $(50, 100, 20, 5, 0.45, 0.5, 1)$. *(f):* $(50, 100, 20, 5, 0.6, 0.5, 1)$. *In each panel, from the left to the right, the odd columns are for sensitivity while the even columns are for specificity. In each panel, box-plots from the left to the right are for* sh_0*, hs1, hs2, hs3, mgl, menet, mrce, ml and msgl respectively. B was called "Oscillated" if its non-zero entries were oscillates around* 0*; "Separated" if non-zero entries were uniformly bigger than* 0.

was too time-consuming to run MRCE on a PC. In light of this, we skipped MRCE in our comparison in these settings. For the MGL, MENET, ML, MSGL and MRCE, we adjusted their penalty coefficients to achieve approximately the same specificity as that of the PVA. These sensitivity and specificity values were summarized using box-plots as shown in Figures 1∼7. In these figures sh_0, hs1, hs2 and hs3 correspond to PVA based on the shrunk and thresholded covari-

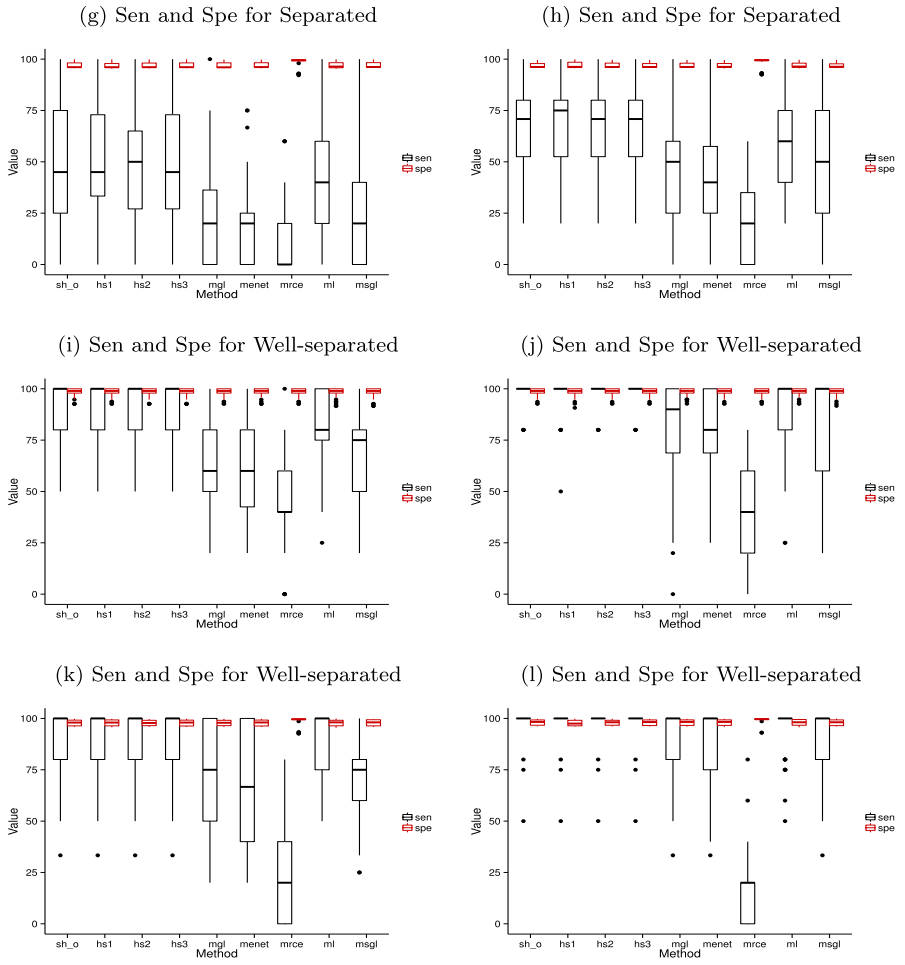

FIG 2. *(Continuation of Fig. 1) box plots of sensitivity (short for Sen) and specificity (short for Spe) values are for Settings 4.1 with* $(n, p, J, p_0, s_0, s_1, s_2)$ *taking the following values. (g):* $(50, 1000, 20, 5, 0.45, 0.5, 1)$. *(h):* $(50, 1000, 20, 5, 0.6, 0.5, 1)$. *(i):* $(50, 100, 20, 5, 0.45, 1, 2)$. *(j):* $(50, 100, 20, 5, 0.6, 1, 2)$. *(k):* $(50, 1000, 20, 5, 0.45, 1, 2)$. *(l):* $(50, 1000, 20, 5, 0.6, 1, 2)$. *B was called "Well-separated" if non-zero entries in B were uniformly bigger than* 0 *by amounts not less than a constant.*

ance estimators with tuning constants $h = 0, 0.01, 0.005, 0.001$ respectively. And mgl, menet, mrce, ml, msgl stand for the multivariate group LASSO, the multivariate elastic-net, the multivariate regression with covariance estimation, the multivariate LASSO and the multivariate sparse group LASSO respectively.

The results indicated that the PVA substantially outperformed the MGL, MENET, ML, MSGL and MRCE in terms of sensitivity and specificity in all the scenarios under consideration. In Settings 4.1 and 4.2, the results suggested that the performances of the MGL, MENET, ML, MGSL and MRCE

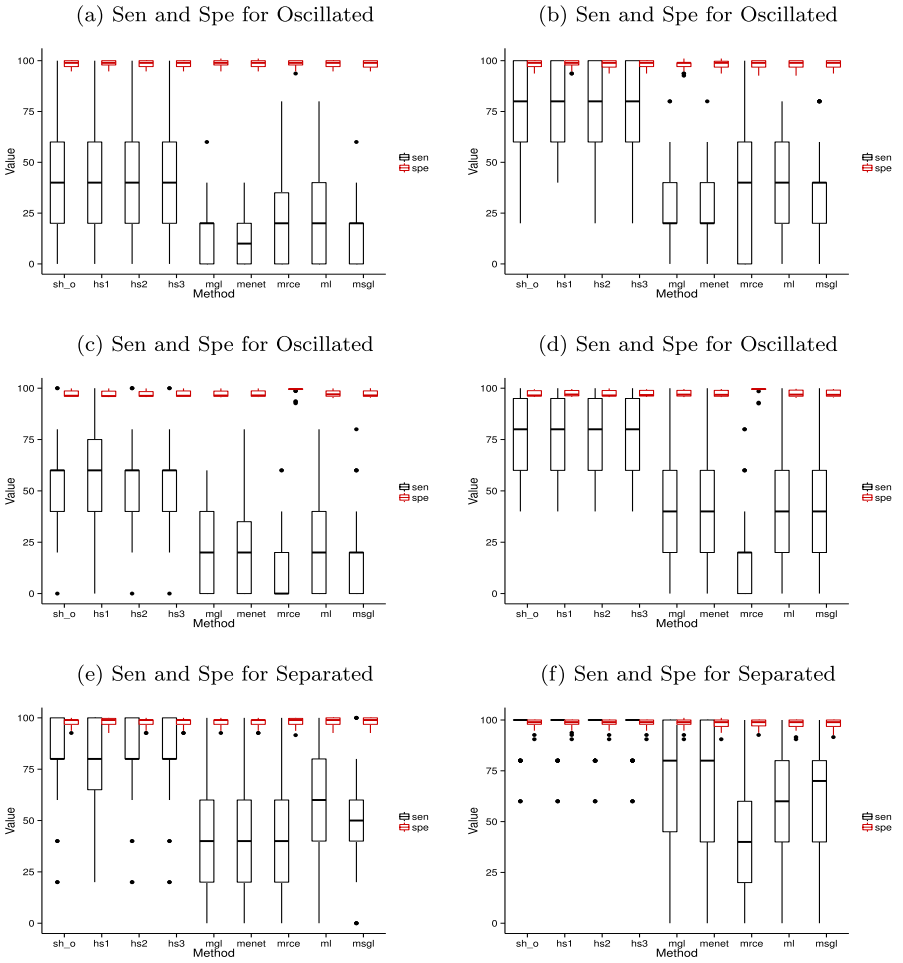

FIG 3. *Box plots of sensitivity (short for Sen and specificity (short for Spe) values are for Setting 4.2 with* $(n, p, J, p_0, s_0, s_1, s_2)$ *taking the following values. (a):* $(50, 100, 20, 5, 0.45, -1, 1)$. *(b):* $(50, 100, 20, 5, 0.6, -1, 1)$. *(c):* $(50, 1000, 20, 5, 0.45, -1, 1)$. *(d):* $(50, 1000, 20, 5, 0.6, -1, 1)$. *(e):* $(50, 100, 20, 5, 0.45, 0.5, 1)$. *(f):* $(50, 100, 20, 5, 0.6, 0.5, 1)$. *In each panel, from the left to the right, the odd columns are for sensitivity while the even columns are for specificity. In each panel, box-plots from the left to the right are for* sh_0*, hs1, hs2, hs3, mgl, menet, mrce, ml and msgl respectively. Adopt the same notations in Figures 1 and 2.*

had deteriorated sharply when the separation between active and non-active covariates, in terms of regression coefficients, was decreasing. In contrast, the performance of the PVA was much more robust than the other procedures to interferences between active and non-active covariates. This was due to interferences being minimized through the optimization in the null-beamforming. This explained why the PVA substantially outperformed the other procedures as the separation between active and non-active covariates was decreasing.

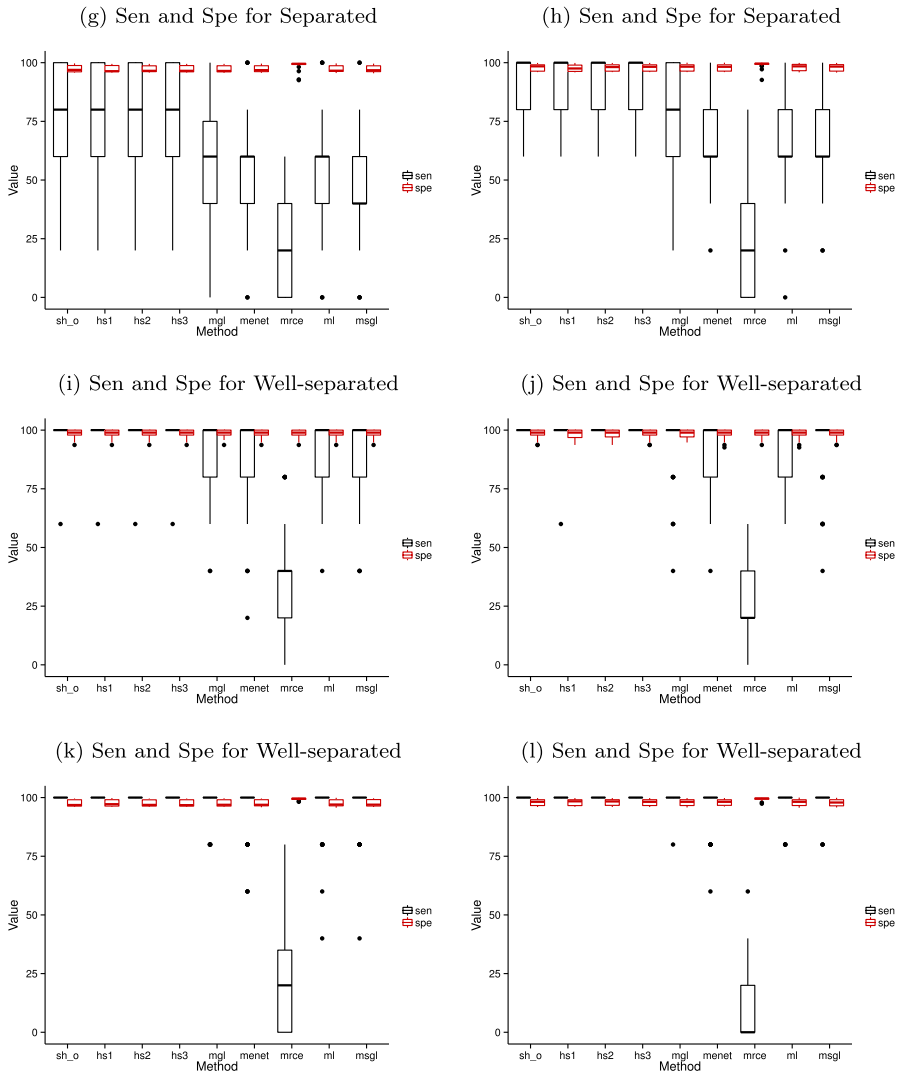

FIG 4. *(Continuation of Fig. 3) box plots of sensitivity and specificity values are for Setting 4.2 with* $(n, p, J, p_0, s_0, s_1, s_2)$ *taking the following values. (g):* $(50, 1000, 20,\ 5, 0.45, 0.5, 1)$. *(h):* $(50, 1000, 20, 5, 0.6, 0.5, 1)$. *(i):* $(50, 100, 20, 5, 0.45, 1, 2)$. *(j):* $(50, 100, 20, 5, 0.6, 1, 2)$. *(k):* $(50, 1000, 20, 5, 0.45, 1, 2)$. *(l):* $(50, 1000, 20, 5,\ 0.6, 1, 2)$. *Here, we adopt the same notations in Figures 1 and 2.*

For example, for the oscillated case where $p = 1000$, $(n, J, p_0, s_0, s_1, s_2) = (50, 20, 5, 0.45, -1, 1)$, the average percentage improvements of PVA(hs3) in sensitivity over the MGL, MENET, MRCE, ML and MSGL were respectively 130%, 190%, 202%, 343% and 853% when the specificity values were fixed roughly at the same level. In contrast, for the well-separated case where $p = 1000$,

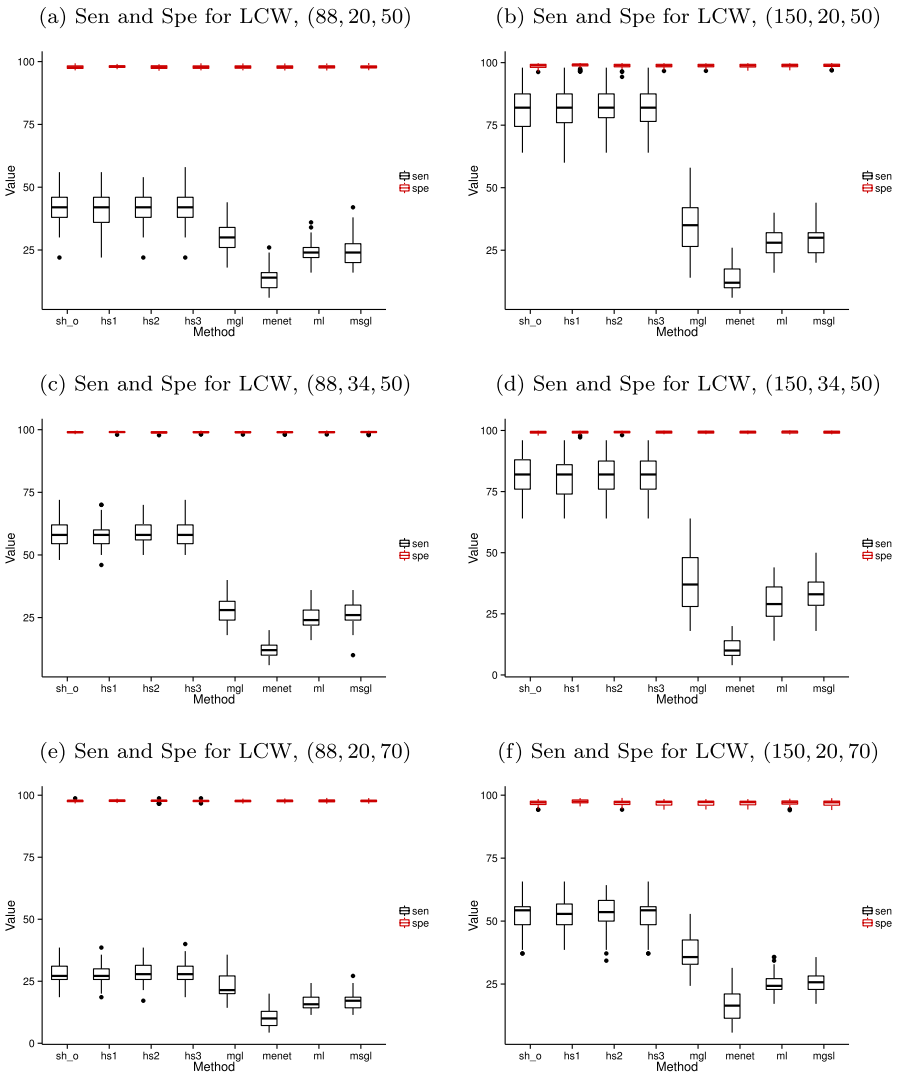

FIG 5. *Box plots of sensitivity (short for Sen) and specificity (short for Spe) values for Setting 4.3 (Low correlations within rows, short for LCW) with* (n, J, p_0) *indicated in the title of each plot,* $p = 2000$ *and* $c_0 = 5$. *Here, we adopt the same notations as in Figure 1.*

$(n, J, p_0, s_0, s_1, s_2) = (50, 20, 5, 0.45, 1, 2)$, the average percentage sensitivity improvements of the PVA(hs3) over the MGL, MENET, MRCE, ML and MSGL were respectively 42%, 44%, 98%, 12% and 35% when the specificity values were also fixed roughly at the same level. Only in the well-separated case, the other five procedures had competitive performances with the PVA.

A similar conclusion can be made for the other settings. For example, for $p = 2000$, $(n, J, p_0) = (88, 20, 50)$, $(150, 20, 50)$, $(88, 34, 50)$, $(150, 34, 50)$ in Setting

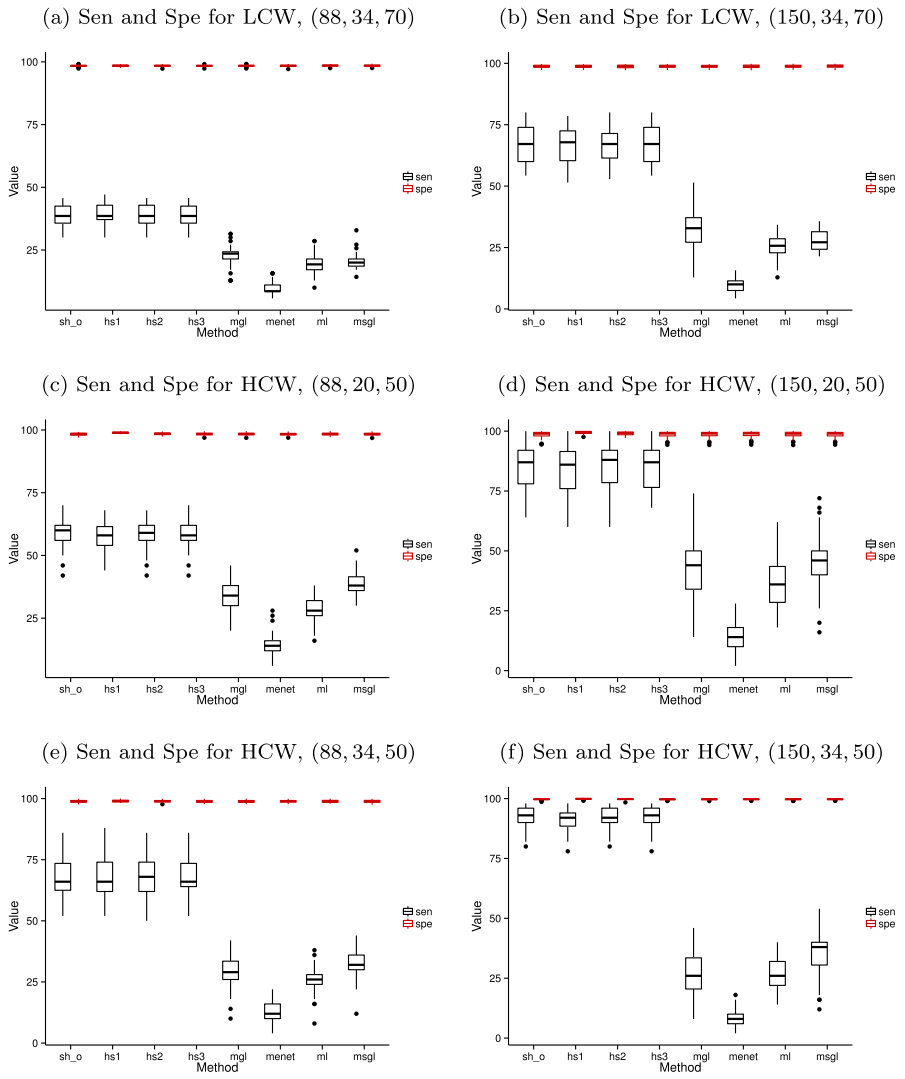

FIG 6. *(Continuation of Fig. 5) box plots of sensitivity (short for Sen) and specificity (short for Spe) values for Setting 4.3 (Low correlations within rows, short for LCW) and Setting 4.4 (High correlations within rows, short for HCW) with* (n, J, p_0) *indicated in the title of each plot,* $p = 2000$ *and* $c_0 = 5$*. Here, we adopt the same notations as in Figure 1.*

4.4, when the specificity values were fixed roughly at the same level, compared to the MGL, on average the sensitivity values of the PVA(hs3) were increased by 74%, 97%, 136%, and 237% respectively. Compared to the MENET, on average the sensitivity values were increased by 312%, 478%, 443% and 968% respectively. In comparison to the ML, on average the sensitivity values were increased by 103%, 133%, 163% and 250% respectively. In comparison to the

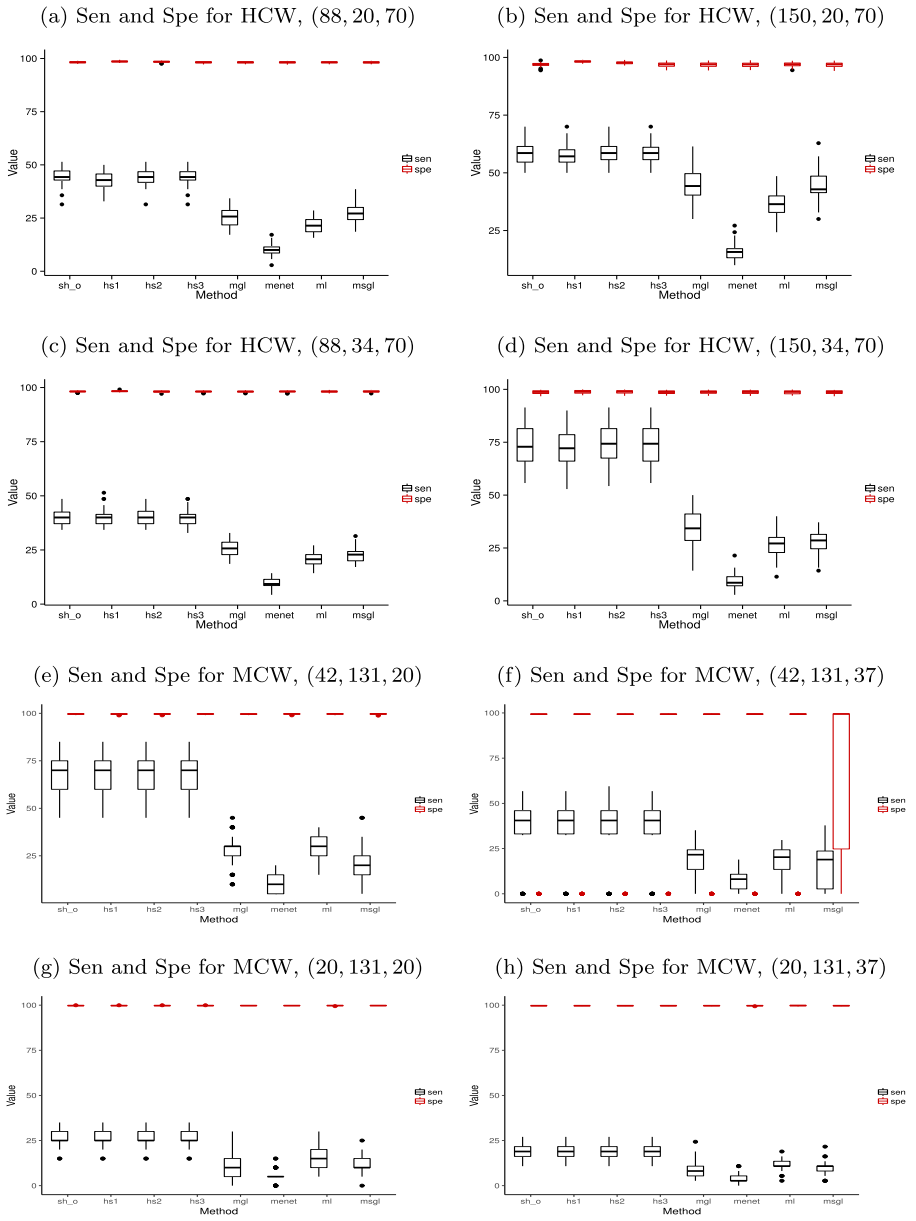

FIG 7. *Box plots of sensitivity and specificity values for Setting 4.4 (HCW) and Setting 4.5 (Moderated correlations within rows, short for MCW) with* (n, J, p_0) *indicated in the title of each plot,* $p = 2000$ *and* $c_0 = 5$. *Here, we adopt the same notations as in Figure 1.*

MSGL, on average the sensitivity values were increased by 53%, 85%, 110% and 169%. The results also suggested that the improvements of the PVA(hs3) over the other procedures in sensitivity were decreasing when p_0 changed from 50 to 70, although they were still large. This was expected as the model complexity increased but the number of measurements per response did not increase.

In Setting 4.3, we considered a weakly correlated regression coefficient matrix B. With the same combinations of (n, p, J, p_0) as before, compared to highly correlated B setting, the improvements over the other procedures reduced but they were still substantial. This reflected a fact that the higher the correlations in columns or rows of B, the stronger intra-correlations the response variable would receive. Therefore, more accurate variable selection would be derived from the PVA as it could explore correlation structures in the data better than the other methods. The results also indicated that the sensitivity improvements of the PVA over the other procedures were increasing in J and n. The similar result was also obtained in Setting 4.5.

We recorded the running times of performing the above procedures on each of the 50 data sets in each setting. The results showed that on average the PVA was run much faster than the ML and MSGL and was also very competitive with the MGL and MENET when we applied them to these data sets in terms of log-CPU-times in seconds. The details were omitted.

4.2. Anti-cancer drug data

Cancer drugs exert their function through binding to one or more protein targets [22]. Early "one gene, one drug, one cancer" paradigm considers the role of individual genes and their changes in drug-perturbed states, which largely ignore a target's cellular and physiological context. Meanwhile, cancer gene-centric methods largely ignore the multi-factor-driven attribute of cancer diseases at the cellular level. With the generation of rich data resources for genome-wide gene expressions and drug- and cancer-induced perturbations, data integrative approaches such as PVA try to provide systematic insights into mechanisms of drugs and cancers in a "multiple genes, multiple drugs, multiple types of cancers" paradigm.

In this section, we focus on the following two data sets: IC50 values of drug sensitivity in cancer cell lines and the corresponding gene expression DNA microarrays [10]. According to cancer encyclopedia, IC50 is a concentration of drug that reduces a biochemical activity such as cell multiplication to 50 percent of its normal value in the absence of the inhibitor. The data sets contain gene expression levels of 13321 genes and median inhibitory concentrations (IC50s) of 131 drugs across 586 cell lines. Among these cell lines, only 42 had complete records of their response to 131 drugs. Here, we considered only the 42 completed cell lines. The challenging problem of imputing remaining cell lines will be addressed in a separate work. We aim to identify biomarkers (a set of genes) that underpin the drug sensitivity in cancer cell lines. Multivariate random-effects regression models can be used to recover these biomarkers, where we treat drugs as subjects (or responses), IC50 values of each drug on cell lines as measurements and

genes as random-effects covariates. Note that, in the above regression, multiple drugs are simultaneously linked to the same set of covariates. So, the higher the number of drugs, the more information about these covariates can be extracted from the drug sensitivity data.

Letting $\mathbf{X}$ be log-gene-expression levels and $\mathbf{Y}$ be IC50 values of 42 completely observed cell lines, we considered the model (2.1) for $(\mathbf{Y}, \mathbf{X})$ with the number of measurements per response $n = 42$, the number of covariates $p = 13321$ and the number of the responses $J = 131$. As $p \gg n$ and $p \gg J$, the model estimation was ill-posed. To reduce the number of covariates, we performed PVA(hs3)-based variable selection on the dataset, identifying 37 active covariates (i.e., genes) for the response variable (i.e., IC50s) as follows: C18ORF24 (SKA1), IARS, CLASP1, STAMBPL1, GSTM3, EML1, TRIM6-TRIM34, DECR1, EP400, RPL39L, FAIM3, CD1A, CIDEB, TP53, QKI, SNTB1, SEMA4C, NUDT2, RFX2, GPSN2 (TECR), C21ORF45 (MIS18A), COL5A1, RP1.153G14.3 (ZNF391), MKL1, FKSG44, KIAA1856, HDGF2, CROCC, WDR76, RPS14, MAP3K6, MAP3K6, LY6E, SLCO2B1, NR1D2, RHBDD3 and STX7. We then fitted a reduced multivariate regression model to the dataset by restricting the covariates to the selected, obtaining an estimated vector of the 131-dimensional regression coefficients for each selected gene.

We constructed a network, displayed in Figure 8, for the selected genes based on their regression coefficients across 131 drugs. The network was strongly connected as there always existed a path from any node to any other node. Surprisingly, although by the iterative nulling, the selected genes were uncorrelated in their expression levels, they were strongly correlated when they reacted to cancer drugs as shown in Figure 8. This suggests that these genes are potentially correlated in a high function level (e.g., protein level).

To reveal the potential roles of these selected genes played in cancer drug sensitivity, as their protein products would dictate their functions, we investigated their protein stainings in the following 20 common cancers [18]: Breast, Carcinoid, Cervical, Colorectal, Endometrial, Glioma, Hand and neck, Liver, Lung, Lymphoma, Melanoma, Ovarian, Pancreatic, Prostate, Renal, Skin, Stomach, Testis, Thyroid and Urothelial. We extracted such information from the Human Protein Atlas Portal at `http://www.proteinatlas.org/cancer`. As in the Portal, we classified the protein expression/staining levels into 4 categories: high, medium, low and not detected. We assigned the scores of $3, 2, 1$ and 0 to the above categories respectively. If a gene did not play a role in a cancer, it would receive a score of zero as its protein staining at that cancer would be hardly detectable. We found 34 of the selected genes, which had positive staining levels on at least one of these cancers. This implied that these genes might play certain functional roles in growths of some of these cancers. In the Portal, there was no information available on the remaining 3 of the selected genes.

5. Conclusion

High dimensional multivariate regression data, where columns stand for measurements on responses (or subjects) can be fitted by both fixed-effects models

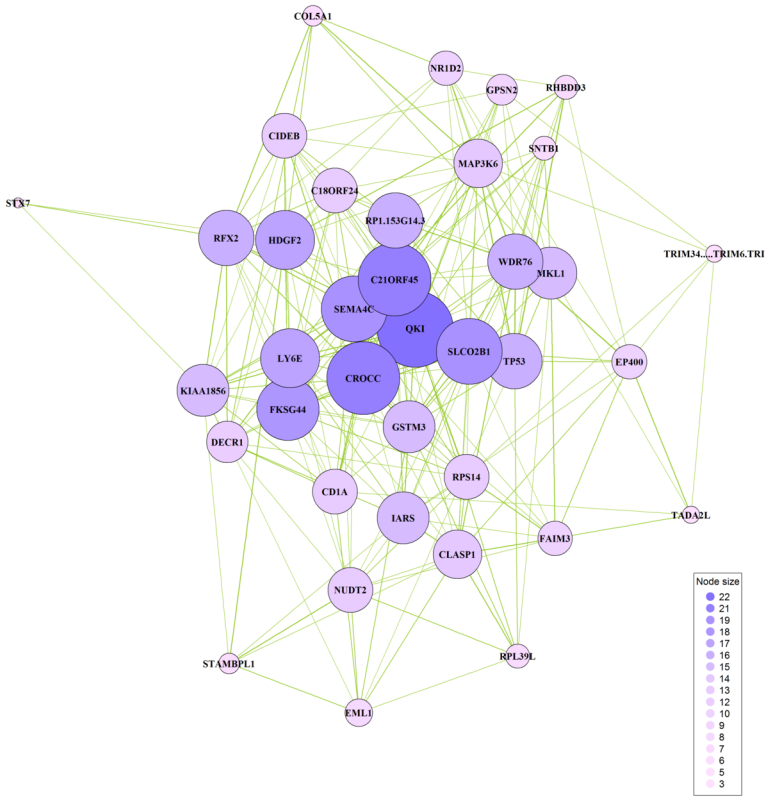

FIG 8. *A network of the* 37 *selected genes based on their regression coefficients across* 131 *drugs. The size (called degree) of each node is proportional to the number of connections of that node with other nodes. The thickness of each edge represents the magnitude of the correlation coefficient between the nodes linked by this edge. The higher the correlation coefficient, the thicker the edge is. The largest degree of* 22 *and the smallest degree of* 3 *were attained by gene QKI and gene STX7 respectively.*

and random-effects models with helps of variable selection. The existing multivariate variable selection methods have been put forward mainly for fixed-effects models. In this paper, we have developed a novel approach called PVA for selecting random-effects covariates in multivariate random-effects regression models. PVA is covariate-assisted, in which we project the response data matrix into the space spanned by each covariate and define a relative information index SNR by the variance component ratio between this covariate and the background noise. The resulting SNR values are then used to rank covariates. The highly ranked covariates are called principal variables. By the PVA, we try to find a small number of principal variables to explain the maximum amount of variation in the data. Our approach allows us to consider correlations between measurements and between responses (between rows and columns in the response data matrix) while the existing methods are only able to deal with correlation structures be-

tween responses. In a multivariate fixed-effects model with many responses, for each covariate, we need to estimate many regression coefficients, which is a high-dimensional problem when the number of responses (or subjects) is very large. In contrast, in a multivariate random-effects model, for each covariate, we only need to estimate its variance component, which is a low-dimensional problem. This difference provides a foundation for the PVA approach to multivariate variable selection. In multivariate regression models, all responses are related to the same set of covariates, which implies that the larger the number of responses, the more information on covariates can be extracted from the response data. Therefore, the accuracy of random-effects covariate selection is expected to increase as the number of responses is increasing. However, when all covariate variance components are zeros, the models reduce to a class of fixed-effects models, where the methods in [17] can be employed while the PVA approach is not applicable.

We have established a novel theory on selection consistency for the proposed method when along with other regularity conditions, the number of covariates p, the number of non-zero entries m_n in covariance matrix. the number of measurements per response n and the number of responses J satisfy $m_n\sqrt{\log(n)/J} = o(1)$ and $\log(p) = O(n^{\alpha_0})$. In particular, we have shown that under these regularity conditions, true active covariates are asymptotically separable from non-active covariates in terms of their power or SNR values as n and J tend to infinite. We have also shown that the nulled power has a higher value than a non-nulled power and is adaptive to response covariance structures. We have conducted a wide range of simulation studies to compare the PVA with the multivariate group LASSO, the multivariate elastic-net, the multivariate LASSO, the multivariate sparse group LASSO and the MRCE. This has explained why the PVA can outperform the existing multivariate variable selection procedures in the literature as these methods are not adaptive to response covariance structures. The simulation results have shown that the PVA can substantially perform better than its competitors in all the scenarios under considerations while the PVA is scalable to the data size by iteratively calculating the power or SNR values. The simulation studies in Settings 4.1∼4.5 have shown that even when the response covariance matrix is not sparse or when J is much smaller than n, PVA can still have a superior performance than the existing methods.

To demonstrate the usage of the PVA in practice, we have conducted PVA on a cancer drug dataset and identified a list of principal genes and the related network to predict the drug's sensitivity to cancers in a "multiple genes, multiple drugs, multiple types of cancers" paradigm. The correlations of the selected genes in the RNA expression levels are largely different from those in their functional levels (their contributions to the IC50 values). The results have been further validated by the protein expression levels of these genes in 20 common cancers. We should mention that we have applied the cross-validation-based multivariate group LASSO and the multivariate elastic-net to the same dataset. Unfortunately, we have ended up with a few thousand genes being selected, which were very difficult to interpret in practice.

Appendix A: The existing approaches to multivariate variable selection

To introduce the multivariate group LASSO (MGL) and the multivariate elastic-net (MENET), we consider the following penalization problem

$$\min_{\mathrm{B}} \left\{ \frac{1}{2n} ||\mathbf{Y} - \mathbf{X}\mathrm{B}||_F^2 + \lambda \left[(1-\alpha) \parallel \mathbf{B} \parallel_F^2 + \alpha \sum_{k=1}^{p} \sqrt{\sum_{j=1}^{J} \beta_{kj}^2} \right] \right\},$$

where $||\mathbf{B}||_F = \sqrt{\sum_{kj} \beta_{kj}^2}$ denotes the standard Forbinus norm of $\mathbf{B} = (\beta_{kj})$, $||\mathbf{Y} - \mathbf{X}\mathrm{B}||_F$ is the Forbinus norm of $\mathbf{Y} - \mathbf{X}\mathrm{B}$, n is the number of observations, $0 \leq \alpha \leq 1$ is a pre-determined constant and $\lambda \geq 0$ is the penalty coefficient. The solution defines the multivariate group LASSO when $\alpha = 1$, ridge estimator when $\alpha = 0$ corresponds to the ridge estimation, and the multivariate elastic net when $\alpha = 0.5$.

The multivariate LASSO (ML) and the multivariate sparse LASSO (MSGL) can be derived by the following penalization problem with

$$\min_{\mathrm{B}} \left\{ \frac{1}{2n} ||\mathbf{Y} - \mathbf{X}\mathrm{B}||_F^2 + \lambda \left[(1-\alpha)|\mathbf{B}| + \alpha \sum_{k=1}^{p} \sqrt{\sum_{j=1}^{J} \beta_{kj}^2} \right] \right\},$$

by setting $\alpha = 0$ and $\alpha = 0.5$ respectively, where $|\mathrm{B}| = \sum_{kj} |\beta_{kj}|$ is the L_1 norm of B. The sparse multivariate regression with covariance estimation (MRCE) is referred to Rothman et al. (2010).

Appendix B: Extra theorems, technical details and proofs

In this appendix, for a $n \times n$ square matrix D, let $||D||$ be the operator norm, the square root of the largest eigenvalue of DD^T. Slightly abusing the notation, now let $||D||_F$ denote the size-normalized Forbinus norm, $\sqrt{\mathrm{tr}(DD^T)/n}$, where $\mathrm{tr}(\cdot)$ is the trace.

B.1. Theory on principal variable analysis with known covariance

We begin with an ideal setting where C is known. This includes the case of $J = \infty$ in which we can estimate C exactly. We establish lower bounds for the SNRs below.

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

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

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

Proposition B.2. *If both $\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}$ and A_{ν_0} are invertible and $\boldsymbol{x}_{\nu_0}$ has a full column rank, then the predictive power matrix*

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

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

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

Theorem B.1. *Suppose that there exist constants $0 \leq \alpha_1 \leq (1 - 3\alpha_0)/2$ and $c_2 > 0$, $c_2 n^{-\alpha_1} \leq \lambda_{min}(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}) \leq \lambda_{max}(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}) = O(1)$. Let $\Sigma_{\nu\triangleleft\nu_0} = (\boldsymbol{e}_{\nu\triangleleft\nu_0}^T(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0})^{-1}\boldsymbol{e}_{\nu\triangleleft\nu_0})^{-1}$, a partial covariance matrix of ν with respect to ν_0. Then, under Conditions (C0)$\sim$(C3), as n tends to infinity, we have:*

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

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

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

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

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

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

with

$$\gamma_\nu^{11} = \Sigma_{\nu_1\triangleleft\nu_0} + O(n^{-1+3\alpha_0+2\alpha_1}),$$

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

The above theorem also indicates that compared to the underlying power matrix, $\mathbf{e}_\nu^T\Sigma\mathbf{e}_\nu$, the predictive power matrix γ_ν may be not consistent if the collinearity between a pair of the predictors does not converge to zero as n tends infinity. This can be seen from the derivation of the predictive power at the predictor $k_j \in \nu_0$ below. Let $\sigma_{k_j[-k_j]}$ denote $(\sigma_{k_jk_i} : i \neq j)$, the jth row in $\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}$ excluding the jth coordinate. Let $\sigma_{[-k_j]k_j}$ denote $(\sigma_{k_ik_j} : i \neq j)$, the jth column in Σ excluding the jth coordinate. Let $\sigma_{[-k_j][-k_j]}$ denote the remaining matrix after removing the jth row and the jth column from $\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}$. Then, the (j,j)th entry in $(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0})^{-1}$ is equal to $(\sigma_{k_j}^2-\sigma_{k_j[-k_j]}\sigma_{[-k_j][-k_j]}^{-1}\sigma_{[-k_j]k_j})^{-1}$. The following corollary says that under Condition (C1), the predictor k_j does have positive predictive power although the power has deteriorated due to the interferences with other predictors. Therefore, if we employ the estimated predictive power to screen predictors and if $\hat{\mathrm{C}}$ is consistent with C, then under Conditions (C0)$\sim$(C3), the screening procedure can have a sure screening property that for an appropriately chosen threshold, all predictors in ν_0 can be detected with a probability approaching to one.

Corollary B.1. *Under the conditions in Theorem B.1, as n tends infinity, we have:*

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

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

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

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

Theorem B.2. *Suppose that there exist constants* $0 \le \alpha_1 \le (1-6\alpha_0)/5$ *and* $c_2 > 0$, $c_2 n^{-\alpha_1} \le \lambda_{min}\left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right) \le \lambda_{max}\left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right) = O(1)$. *Then, under Conditions (C0)∼(C3), as* n *tends to infinity, we have:*

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

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

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

where

$$\begin{aligned}\Sigma_{\nu_1\lhd\nu_0} &= \left(\boldsymbol{e}_{\nu_1\lhd\nu_0}^T \left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right)^{-1} \boldsymbol{e}_{\nu_1\lhd\nu_0}\right)^{-1}, \\ \Sigma_{\nu_0\backslash\nu_1}^{-1} &= \left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right)^{-1/2} P_{\nu_0\backslash\nu_1} \left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right)^{-1/2}, \\ P_{\nu_0\backslash\nu_1} &= I_{|\nu_0|} - \left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right)^{-1/2} \boldsymbol{e}_{\nu_1\lhd\nu_0} \Sigma_{\nu_1\lhd\nu_0} \boldsymbol{e}_{\nu_1\lhd\nu_0}^T \left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right)^{-1/2}, \\ F_{\nu_2} &= R_{\nu_2\nu_2} - R_{\nu_2\nu_0} R_{\nu_0\nu_0}^{-1} R_{\nu_0\nu_2}, \\ \boldsymbol{\Phi} &= R_{\nu_0\nu_0}^{-1} + R_{\nu_0\nu_0}^{-1} R_{\nu_0\nu_2} F_{\nu_2}^{-1} R_{\nu_2\nu_0} R_{\nu_0\nu_0}^{-1}, \\ \boldsymbol{\Psi} &= \left(I_{|\nu_0|} - \boldsymbol{e}_{\nu_1\lhd\nu_0} \Sigma_{\nu_1\lhd\nu_0} \boldsymbol{e}_{\nu_1\lhd\nu_0}^T \left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right)^{-1}\right) \boldsymbol{\Phi} \\ &\quad \left(I_{|\nu_0|} - \boldsymbol{e}_{\nu_1\lhd\nu_0} \Sigma_{\nu_1\lhd\nu_0} \boldsymbol{e}_{\nu_1\lhd\nu_0}^T \left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right)^{-1}\right)^T.\end{aligned}$$

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

$$\begin{aligned}\left(\mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\backslash\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0}\right)^{-1} &= \lambda_{\min}\left(\left(\mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\backslash\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0}\right)^{-1}\right) \\ &\ge \lambda_{\min}\left(\left(\mathbf{e}_{\{a\}\cup\nu_1}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\{a\}\cup\nu_1}\right)^{-1}\right) \\ &= \left(\lambda_{\max}\left(\mathbf{e}_{\{a\}\cup\nu_1}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\{a\}\cup\nu_1}\right)\right)^{-1} \\ &\ge \left(\lambda_{\max}\left(\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1}\right)\right)^{-1} \\ &= \lambda_{\min}\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right) \ge c_2 n^{-\alpha_1}.\end{aligned}$$

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

$\sigma_{k_j[-k_j]}\sigma^{-1}_{[-k_j][-k_j]}\sigma_{[-k_j]k_j]}$. Therefore, it follows from the definition of $\gamma_{a|\nu_1\cup\nu_2}$, Corollary B.1 and Theorem B.1 that $\gamma_a \leq \gamma_{a|\nu_1\cup\nu_2}$ and that both γ_a and $\gamma_{a|\nu_1\cup\nu_2}$ can be asymptotically less than or equal to $\sigma^2_{k_j}$ due to interferences with other predictors. Furthermore, we have a sharp result as follows.

Corollary B.2. *Under conditions in Theorem B.2, as n tends to infinity, we have:*

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

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

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

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

Theorem B.3. *Under the conditions in Theorem B.2 and Condition (C4), as n tends to infinity, we have:*

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

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

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

where $\Sigma^{-1}_{\nu_0\setminus\nu_1}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are defined in Theorem B.2.

B.2. Theory on principal variable analysis with estimated covariance

We will show later in Lemma B.4 that under Conditions (C1)∼(C6), the optimal shrinkage covariance estimator $\hat{\mathrm{C}}_{hs}$ is consistent with the true covariance C. This allows us to state the following theorems for the case where unknown C is estimated by $\hat{\mathrm{C}}_{hs}$.

Theorem B.4. *Suppose that Conditions (C0)∼(C6) hold and that $\tau_{nJ}n^2 = o(1)$ as both n and J tend to infinity. Then, we have:*

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

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

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

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

where

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

where $\Sigma_{\nu_1 \lhd \nu_0} = (\boldsymbol{e}_{\nu_1|\nu_0}^T (\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0})^{-1} \boldsymbol{e}_{\nu_1|\nu_0})^{-1}$.

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

Corollary B.3. *Under the conditions in Theorem B.4, if* $\alpha_1 < \min\{(1-\alpha_0)/3, (1-3\alpha_0)/2\}$, $n^{2+\alpha_0}\tau_{nJ} = o(1)$ *and* $n^{2+\alpha_1}\tau_{nJ} = o(1)$, *then as both* n *and* J *tend to infinity,* $P(\nu_0 = \hat{\nu}_d) \to 1$.

B.3. Proofs

Proof of Proposition B.1. Note that $\mathrm{C} = \mathbf{X}\Sigma\mathbf{X}^T + \sigma^2 I_n \geq \sigma^2 I_n$. Therefore,

$$\begin{aligned} \sigma^2 \mathbf{w}_{\nu|\omega}^T \mathbf{w}_{\nu|\omega} &= \sigma^2 e_{\nu|\omega}^T \left(\mathbf{x}_{\nu\cup\omega}^T \mathrm{C}^{-1} \mathbf{x}_{\nu\cup\omega}\right)^{-1} \mathrm{C}^{-2} \mathbf{x}_{\nu\cup\omega} \left(\mathbf{x}_{\nu\cup\omega}^T \mathrm{C}^{-1} \mathbf{x}_{\nu\cup\omega}\right)^{-1} e_{\nu|\omega} \\ &\leq e_{\nu|\omega}^T \left(\mathbf{x}_{\nu\cup\omega}^T \mathrm{C}^{-1} \mathbf{x}_{\nu\cup\omega}\right)^{-1} \mathbf{x}_{\nu\cup\omega}^T \mathrm{C}^{-1} I_n \mathbf{x}_{\nu\cup\omega} \left(\mathbf{x}_{\nu\cup\omega}^T \mathrm{C}^{-1} \mathbf{x}_{\nu\cup\omega}\right)^{-1} e_{\nu|\omega} \\ &= \gamma_{\nu|\omega}, \end{aligned}$$

which implies $\mathrm{SNR}_{\nu|\omega} \geq 1$. When all predictors in $1:p = \{1, 2, ..., p\}$ are not active and $\mathrm{C} = \sigma^2 I_n$, $\gamma_{\nu|\omega} = \sigma^2 \mathbf{w}_{\nu|\omega}^T \mathbf{w}_{\nu|\omega}$. Therefore, $\mathrm{SNR}_{\nu|\omega} = 1$, which attains its lower bound. This completes the proof. □

Proof of Proposition B.2. It follows from the Woodbury matrix identity. □

Proof of Corollary B.1. It follows from Theorem B.1. □

Proof of Corollary B.2. It follows from Theorem B.2. □

To prove Theorems B.1∼B.3, we needs the following lemma which gives a $\mathbf{x}_{\nu_0}$-projection based decomposition of the quadratic form $\mathbf{x}_\nu^T \mathrm{C}^{-1} \mathbf{x}_\nu$.

Lemma B.1. *Under Conditions (C0)∼(C3), for any $\nu \subseteq 1:p$, if $\alpha_1 < (1-3\alpha_0)/2$, then*

$$\begin{aligned}\boldsymbol{x}_\nu^T C^{-1}\boldsymbol{x}_\nu &= n\left(R_{\nu\nu} - R_{\nu\nu_0}R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu}\right) + R_{\nu\nu_0}R_{\nu_0\nu_0}^{-1}\left(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu}\\ &\quad +O(n^{-1+3\alpha_0+2\alpha_1}).\end{aligned}$$

In particular, for $\nu \subseteq \nu_0$, if $\alpha_1 < (1-\alpha_0)/2$, then

$$\boldsymbol{x}_\nu^T C^{-1}\boldsymbol{x}_\nu = \boldsymbol{e}_{\nu\lhd\nu_0}^T\left(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}\right)^{-1}\boldsymbol{e}_{\nu\lhd\nu_0} + O(n^{-1+\alpha_0+2\alpha_1}).$$

For $a \in \nu_0$ and $\nu_1 \subseteq \nu_0$, if $\alpha_1 < (1-\alpha_0)/2$, then

$$\begin{aligned}\boldsymbol{x}_a^T C^{-1}\boldsymbol{x}_{\nu_1} &= \boldsymbol{e}_{a\lhd\nu_0}^T\left(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}\right)^{-1}\boldsymbol{e}_{\nu_1\lhd\nu_0}\\ &\quad -n^{-1}\boldsymbol{e}_{a\lhd\nu_0}^T\left(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1}\left(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}\right)^{-1}\boldsymbol{e}_{\nu_1\lhd\nu_0}\\ &\quad +O(n^{-2+5\alpha_0+3\alpha_1}).\end{aligned}$$

For $\nu_1 \subseteq \nu_0$ and $\nu_2 \subseteq 1:p\setminus\nu_0$, if $\alpha_1 < (1-2\alpha_0)/2$, then

$$\begin{aligned}\boldsymbol{x}_{\nu_1}^T C^{-1}\boldsymbol{x}_{\nu_2} &= \boldsymbol{e}_{\nu_1\lhd\nu_0}^T\left(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu_2} + O(n^{-1+2\alpha_0+2\alpha_1}),\\ \boldsymbol{x}_{\nu_2}^T C^{-1}\boldsymbol{x}_{\nu_1} &= R_{\nu_2\nu_0}R_{\nu_0\nu_0}^{-1}(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0})^{-1}\boldsymbol{e}_{\nu_1\lhd\nu_0} + O(n^{-1+2\alpha_0+2\alpha_1}).\end{aligned}$$

Proof of Lemma B.1. Note that

$$||\mathrm{R}_{\nu_0\nu_0}^{-1/2}\mathrm{R}_{\nu_0\nu}|| \le ||\mathrm{R}_{\nu_0\nu_0}^{-1/2}\mathrm{R}_{\nu_0\nu}\mathrm{R}_{\nu\nu}^{-1/2}||||\mathrm{R}_{\nu\nu}^{1/2}|| \le ||\mathrm{R}_{\nu\nu}^{1/2}|| = O(n^{\alpha_0/2}).$$

When $\nu \subseteq \nu_0$, we have $\mathbf{x}_\nu = \mathbf{x}_{\nu_0}\mathbf{e}_{\nu\lhd\nu_0}$, $\mathrm{R}_{\nu\nu_0} = \mathbf{e}_{\nu\lhd\nu_0}^T\mathrm{R}_{\nu_0\nu_0}$, $\mathrm{R}_{\nu_0\nu} = \mathrm{R}_{\nu_0\nu_0}\mathbf{e}_{\nu\lhd\nu_0}$ and $\mathrm{R}_{\nu\nu} = \mathbf{e}_{\nu\lhd\nu_0}^T\mathrm{R}_{\nu_0\nu_0}\mathbf{e}_{\nu\lhd\nu_0}$. These together with the Woodbury matrix identity and a Taylor expansion complete the proof. □

Proof of Theorem B.1. To prove (i), let $\Sigma_{\nu\lhd\nu_0} = (\mathbf{e}_{\nu\lhd\nu_0}^T(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0})^{-1}\mathbf{e}_{\nu\lhd\nu_0})^{-1}$. Then, by using the assumption,

$$\lambda_{\max}\left(\Sigma_{\nu\lhd\nu_0}^{1/2}\right) \le \left(\lambda_{\min}\left(\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\right)\right)^{-1/2} = \lambda_{\max}^{1/2}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right) = O(1).$$

This together with Lemma B.1 yields (i).

To prove (ii), we apply Lemma B.1 to calculate $\mathbf{x}_\nu^T\mathrm{C}^{-1}\mathbf{x}_\nu$. We have

$$\begin{aligned}\mathbf{x}_\nu^T\mathrm{C}^{-1}\mathbf{x}_\nu &= n\mathrm{F}_\nu^{1/2}\Big(I_{|\nu|} + n^{-1}\mathrm{F}_\nu^{-1/2}\mathrm{R}_{\nu\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0})^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu}\mathrm{F}_\nu^{-1/2}\\ &\quad +\mathrm{F}_\nu^{-1/2}O(n^{-2+3\alpha_0+2\alpha_1})\mathrm{F}_\nu^{-1/2}\Big)\mathrm{F}_\nu^{1/2},\end{aligned}$$

where $\mathrm{F}_\nu = (\mathrm{R}_{\nu\nu} - \mathrm{R}_{\nu\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu})$. Therefore, when $\alpha_1 \le 1-3\alpha_0$,

$$\gamma_\nu = n^{-1}\mathrm{F}_\nu^{-1/2}\left(I_{|\nu|} - O(n^{-1+3\alpha_0+\alpha_1}) - O(n^{-2+4\alpha_0+2\alpha_1})\right)\mathrm{F}_\nu^{-1/2} = O(n^{-1+\alpha_0}),$$

which completes the proof of (ii).

To prove (iii), we note that

$$\mathbf{x}_\nu^T \mathrm{C}^{-1} \mathbf{x}_\nu = \begin{pmatrix} \mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_1} & \mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2} \\ \mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_1} & \mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2} \end{pmatrix}.$$

Invoking Lemma B.1, we have

$$\begin{aligned} \mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_1} &= \mathbf{e}_{\nu_1 \lhd \nu_0}^T (\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0})^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0} + O(n^{-1+\alpha_0+2\alpha_1}), \\ \mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2} &= n \left(\mathrm{R}_{\nu_2\nu_2} - \mathrm{R}_{\nu_2\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1} \mathrm{R}_{\nu_0\nu_2} \right) + O(n^{2\alpha_0+\alpha_1}). \end{aligned}$$

Similarly, we can show that

$$\begin{aligned} \mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2} &= \mathbf{e}_{\nu_1 \lhd \nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0} \right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1} \mathrm{R}_{\nu_0\nu_2} + O(n^{-1+2\alpha_0+2\alpha_1}), \\ \mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_1} &= \mathrm{R}_{\nu_2\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1} (\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0})^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0} + O(n^{-1+2\alpha_0+2\alpha_1}). \end{aligned}$$

Let $\Sigma_{\nu_1 \lhd \nu_0} = (\mathbf{e}_{\nu_1 \lhd \nu_0}^T (\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0})^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0})^{-1}$. Then,

$$\lambda_{\max}(\Sigma_{\nu_1 \lhd \nu_0}) \leq \lambda_{\max}(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}) = O(1).$$

The proof is completed by applying Lemma B.1 to each block. □

Proof of Theorem B.2. To prove (i), let ν denote $\nu_1 \cup \nu_2 \cup \{a\}$. We partition $\mathbf{x}_\nu^T \mathrm{C}^{-1} \mathbf{x}_\nu$ into the block matrix below:

$$\mathbf{x}_\nu^T \mathrm{C}^{-1} \mathbf{x}_\nu = \begin{pmatrix} \mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_1} & \mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2 \cup \{a\}} \\ \mathbf{x}_{\nu_2 \cup \{a\}}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_1} & \mathbf{x}_{\nu_2 \cup \{a\}}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2 \cup \{a\}} \end{pmatrix}$$

and that $\mathbf{x}_{\nu_1} = \mathbf{x}_{\nu_0} \mathbf{e}_{\nu_1 \lhd \nu_0}$. Invoking Lemma B.1, we have

$$\begin{aligned} \mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_1} &= \mathbf{e}_{\nu_1 \lhd \nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0} \right)^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0} + O(n^{-1+\alpha_0+2\alpha_1}), \\ \mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2 \cup \{a\}} &= \mathbf{e}_{\nu_1 \lhd \nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0} \right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1} \mathrm{R}_{\nu_0(\nu_2 \cup \{a\})} + O(n^{-1+2\alpha_0+2\alpha_1}), \\ \mathbf{x}_{\nu_2 \cup \{a\}}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_1} &= \mathrm{R}_{(\nu_2 \cup \{a\})\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0} \right)^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0} + O(n^{-1+2\alpha_0+2\alpha_1}), \\ \mathbf{x}_{\nu_2 \cup \{a\}}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2 \cup \{a\}} &= n \left(\mathrm{R}_{(\nu_2 \cup \{a\})(\nu_2 \cup \{a\})} - \mathrm{R}_{(\nu_2 \cup \{a\})\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1} \mathrm{R}_{\nu_0(\nu_2 \cup \{a\})} \right) \\ &\quad + O(n^{2\alpha_0+\alpha_1}). \end{aligned}$$

We partition $(\mathbf{x}_\nu^T \mathrm{C}^{-1} \mathbf{x}_\nu)^{-1}$ in the same as we did above for $\mathbf{x}_\nu^T \mathrm{C}^{-1} \mathbf{x}_\nu$:

$$(\mathbf{x}_\nu^T \mathrm{C}^{-1} \mathbf{x}_\nu)^{-1} = \begin{pmatrix} \mathrm{A}^{11} & \mathrm{A}^{12} \\ \mathrm{A}^{21} & \mathrm{A}^{22} \end{pmatrix}.$$

Then, by definition we have

$$\gamma_{a|\nu_1 \cup \nu_2} = (\mathbf{0}_{|\nu_2|}^T, 1) \mathrm{A}^{22} (\mathbf{0}_{|\nu_2|}^T, 1)^T. \tag{B.1}$$

And let $\mathrm{F}_{\nu_2\cup\{a\}} = \mathrm{R}_{(\nu_2\cup a)(\nu_2\cup a)} - \mathrm{R}_{(\nu_2\cup a)\nu_0}\mathrm{R}^{-1}_{\nu_0\nu_0}\mathrm{R}_{\nu_0(\nu_2\cup a)}$ and define Σ_{ν_1} as in the proof of Theorem B.1. We have

$$\begin{aligned}\mathrm{A}^{22} &= \frac{1}{n}\begin{pmatrix} \mathrm{R}_{\nu_2\nu_2} - \mathrm{R}_{\nu_2\nu_0}\mathrm{R}^{-1}_{\nu_0\nu_0}\mathrm{R}_{\nu_0\nu_2} & \mathrm{R}_{\nu_2 a} - \mathrm{R}_{\nu_2\nu_0}\mathrm{R}^{-1}_{\nu_0\nu_0}\mathrm{R}_{\nu_0 a} \\ \mathrm{R}_{a\nu_2} - \mathrm{R}_{a\nu_0}\mathrm{R}^{-1}_{\nu_0\nu_0}\mathrm{R}_{\nu_0\nu_2} & \mathrm{R}_{aa} - \mathrm{R}_{a\nu_0}\mathrm{R}^{-1}_{\nu_0\nu_0}\mathrm{R}_{\nu_0 a} \end{pmatrix}^{-1} \\ &\quad + O(n^{-2+4\alpha_0+2\alpha_1}). \end{aligned} \tag{B.2}$$

Note that any main block matrix has a larger smallest eigenvalue than that of the whole matrix. Therefore, by Condition (C3) we have

$$\begin{aligned}&\lambda_{\max}\left(\left(\mathrm{R}_{aa} - \mathrm{R}_{a\nu_0}\mathrm{R}^{-1}_{\nu_0\nu_0}\mathrm{R}_{\nu_0 a} - (\mathrm{R}_{a\nu_2} - \mathrm{R}_{a\nu_0}\mathrm{R}^{-1}_{\nu_0\nu_0}\mathrm{R}_{\nu_0\nu_2})\right.\right. \\ &\quad \left.\left.\times(\mathrm{R}_{\nu_2\nu_2} - \mathrm{R}_{\nu_2\nu_0}\mathrm{R}^{-1}_{\nu_0\nu_0}\mathrm{R}_{\nu_0\nu_2})^{-1}(\mathrm{R}_{\nu_2 a} - \mathrm{R}_{\nu_2\nu_0}\mathrm{R}^{-1}_{\nu_0\nu_0}\mathrm{R}_{\nu_0 a})\right)^{-1}\right) \\ &\le \lambda_{\max}\left(\mathrm{F}^{-1}_{\nu_2\cup\{a\}}\right) = O(n^{\alpha_0}).\end{aligned}$$

This together with equations (B.1) and (B.2) shows that $\gamma_{a|\nu_1\cup\nu_2} = O(n^{-1+\alpha_0})$. The proof of (i) is completed.

To prove (ii), we let $\nu = \{a\}\cup\nu_1\cup\nu_2$ and write $\mathbf{x}_\nu^T\mathrm{C}^{-1}\mathbf{x}_\nu$ as the block matrix below:

$$\mathbf{x}_\nu^T\mathrm{C}^{-1}\mathbf{x}_\nu = \begin{pmatrix} \mathbf{x}_a^T\mathrm{C}^{-1}\mathbf{x}_a & \mathbf{x}_a^T\mathrm{C}^{-1}\mathbf{x}_{\nu_1} & \mathbf{x}_a^T\mathrm{C}^{-1}\mathbf{x}_{\nu_2} \\ \mathbf{x}_{\nu_1}^T\mathrm{C}^{-1}\mathbf{x}_a & \mathbf{x}_{\nu_1}^T\mathrm{C}^{-1}\mathbf{x}_{\nu_1} & \mathbf{x}_{\nu_1}^T\mathrm{C}^{-1}\mathbf{x}_{\nu_2} \\ \mathbf{x}_{\nu_2}^T\mathrm{C}^{-1}\mathbf{x}_a & \mathbf{x}_{\nu_2}^T\mathrm{C}^{-1}\mathbf{x}_{\nu_1} & \mathbf{x}_{\nu_2}^T\mathrm{C}^{-1}\mathbf{x}_{\nu_2} \end{pmatrix}.$$

Applying Lemma B.1 to each block, we have

$$\begin{aligned}\mathbf{x}_a^T\mathrm{C}^{-1}\mathbf{x}_a &= \mathbf{e}_{a\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{a\triangleleft\nu_0} \\ &\quad - n^{-1}\mathbf{e}_{a\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}^{-1}_{\nu_0\nu_0}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{a\triangleleft\nu_0} \\ &\quad + O(n^{-2+2\alpha_0+3\alpha_1}), \\ \mathbf{x}_a^T\mathrm{C}^{-1}\mathbf{x}_{\nu_1} &= \mathrm{R}_{a\nu_0}\mathrm{R}^{-1}_{\nu_0\nu_0}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0} \\ &\quad - n^{-1}\mathbf{e}_{a\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}^{-1}_{\nu_0\nu_0}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0} \\ &\quad + O(n^{-2+2\alpha_0+3\alpha_1}), \\ \mathbf{x}_a^T\mathrm{C}^{-1}\mathbf{x}_{\nu_2} &= \mathbf{e}_{a\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}^{-1}_{\nu_0\nu_0}\mathrm{R}_{\nu_0\nu_2} \\ &\quad - n^{-1}\mathbf{e}_{a\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}^{-1}_{\nu_0\nu_0}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}^{-1}_{\nu_0\nu_0}\mathrm{R}_{\nu_0\nu_2} \\ &\quad + O(n^{-2+5\alpha_0+3\alpha_1}), \\ \mathbf{x}_{\nu_1}^T\mathrm{C}^{-1}\mathbf{x}_a &= \mathbf{e}_{\nu_1\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}^{-1}_{\nu_0\nu_0}\mathrm{R}_{\nu_0 a} \\ &\quad - n^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}^{-1}_{\nu_0\nu_0}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{a\triangleleft\nu_0} \\ &\quad + O(n^{-2+2\alpha_0+3\alpha_1}), \\ \mathbf{x}_{\nu_1}^T\mathrm{C}^{-1}\mathbf{x}_{\nu_1} &= \mathbf{e}_{\nu_1\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0} \\ &\quad - n^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}^{-1}_{\nu_0\nu_0}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0} \\ &\quad + O(n^{-2+2\alpha_0+3\alpha_1}),\end{aligned}$$

$$
\begin{aligned}
\mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2} &= \mathbf{e}_{\nu_1 \lhd \nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1} \mathrm{R}_{\nu_0\nu_2} \\
&\quad - n^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1} \mathrm{R}_{\nu_0\nu_2} \\
&\quad + O(n^{-2+5\alpha_0+3\alpha_1}). \\
\mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_a &= \mathrm{R}_{\nu_2\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{a \lhd \nu_0} \\
&\quad - n^{-1} \mathrm{R}_{\nu_2\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{a \lhd \nu_0} \\
&\quad + O(n^{-2+5\alpha_0+3\alpha_1}), \\
\mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_1} &= \mathrm{R}_{\nu_2\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0} \\
&\quad - n^{-1} \mathrm{R}_{\nu_2\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0} \\
&\quad + O(n^{-2+5\alpha_0+3\alpha_1}), \\
\mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2} &= n \left(\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}\right) \\
&\quad + \mathrm{R}_{\nu_2\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1} \mathrm{R}_{\nu_0\nu_2} \\
&\quad + O(n^{-1+3\alpha_0+2\alpha_1}), \\
\left(\mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2}\right)^{-1} &= n^{-1} \Big(\mathrm{F}_{\nu_2}^{-1} - n^{-1} \mathrm{F}_{\nu_2}^{-1} \mathrm{R}_{\nu_2\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathrm{R}_{\nu_0\nu_0}^{-1} \mathrm{R}_{\nu_0\nu_2} \mathrm{F}_{\nu_2}^{-1} \\
&\quad + O(n^{-2+5\alpha_0+2\alpha_1}) \Big)
\end{aligned}
$$

with $\mathrm{F}_{\nu_2} = \mathrm{R}_{\nu_2\nu_2} - \mathrm{R}_{\nu_2\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1} \mathrm{R}_{\nu_0\nu_2}$. Therefore, we have

$$
\begin{aligned}
&\mathbf{x}_a^T \mathrm{C}^{-1} \mathbf{x}_a - \mathbf{x}_a^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2} (\mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2})^{-1} \mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_a \\
&\quad = \mathbf{e}_{a \lhd \nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0} \\
&\qquad - n^{-1} \mathbf{e}_{a \lhd \nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \left(\mathrm{R}_{\nu_0\nu_0}^{-1} + \mathrm{R}_{\nu_0\nu_0}^{-1} \mathrm{R}_{\nu_0\nu_2} \mathrm{F}_{\nu_2}^{-1} \mathrm{R}_{\nu_2\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1}\right) \\
&\qquad \times \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0} + O(n^{-2+6\alpha_0+3\alpha_1}). \\
&\mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_a - \mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2} (\mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2})^{-1} \mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_a \\
&\quad = \mathbf{e}_{\nu_1 \lhd \nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{a \lhd \nu_0} - n^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \\
&\qquad \times \left(\mathrm{R}_{\nu_0\nu_0}^{-1} + \mathrm{R}_{\nu_0\nu_0}^{-1} \mathrm{R}_{\nu_0\nu_2} \mathrm{F}_{\nu_2}^{-1} \mathrm{R}_{\nu_2\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1}\right) \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{a \lhd \nu_0} \\
&\qquad + O(n^{-2+6\alpha_0+3\alpha_1}).
\end{aligned}
$$

Letting $\Sigma_{\nu_1 \lhd \nu_0} = (\mathbf{e}_{\nu_1 \lhd \nu_0}^T (\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0})^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0})^{-1}$, by the conditions in Theorem B.2, we have $\Sigma_{\nu_1 \lhd \nu_0} = O(1)$ and

$$
\begin{aligned}
&\left\{\mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_1} - \mathbf{x}_{\nu_1}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2} (\mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_2})^{-1} \mathbf{x}_{\nu_2}^T \mathrm{C}^{-1} \mathbf{x}_{\nu_1}\right\}^{-1} \\
&\quad = \Big\{ \Sigma_{\nu_1 \lhd \nu_0}^{-1} - n^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \left(\mathrm{R}_{\nu_0\nu_0}^{-1} + \mathrm{R}_{\nu_0\nu_0}^{-1} \mathrm{R}_{\nu_0\nu_2} \mathrm{F}_{\nu_2}^{-1} \mathrm{R}_{\nu_2\nu_0} \mathrm{R}_{\nu_0\nu_0}^{-1}\right) \\
&\qquad \times \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\nu_1 \lhd \nu_0} + O(n^{-2+6\alpha_0+3\alpha_1}) \Big\}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&= \Sigma_{\nu_1\lhd\nu_0} \\
&\quad + n^{-1}\Sigma_{\nu_1\lhd\nu_0}\mathbf{e}_{\nu_1\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\left(\mathrm{R}_{\nu_0\nu_0}^{-1}+\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu_2}\mathrm{F}_{\nu_2}^{-1}\mathrm{R}_{\nu_2\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\right) \\
&\quad \times\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\lhd\nu_0}\Sigma_{\nu_1\lhd\nu_0}+O(n^{-2+6\alpha_0+3\alpha_1}).
\end{aligned}
$$

The above asymptotic expressions together with the definition of $\gamma_{a|\nu_1\cup\nu_2}$ entail

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

where $\Sigma_{\nu_0\setminus\nu_1}^{-1}$ is defined in Theorem B.2 and $\boldsymbol{\Psi}$ is equal to

$$
\begin{aligned}
&\left(I_{|\nu_0|}-\mathbf{e}_{\nu_1\lhd\nu_0}\Sigma_{\nu_1\lhd\nu_0}\mathbf{e}_{\nu_1\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\right)\left(\mathrm{R}_{\nu_0\nu_0}^{-1}+\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu_2}\mathrm{F}_{\nu_2}^{-1}\mathrm{R}_{\nu_2\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\right) \\
&\quad\times\left(I_{|\nu_0|}-\mathbf{e}_{\nu_1\lhd\nu_0}\Sigma_{\nu_1\lhd\nu_0}\mathbf{e}_{\nu_1\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\right)^T.
\end{aligned}
$$

The proof of (ii) is completed. □

To prove Theorem B.3, we need a more lemma as follows.

Lemma B.2. *Suppose that there exist constants* $0\le\alpha_1\le(1-3\alpha_0)/2$ *and* $c_2>0$, $c_2n^{-\alpha_1}\le\lambda_{min}\left(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}\right)\le\lambda_{max}\left(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}\right)=O(1)$. *Under Conditions* ***(C1)*** *∼* ***(C4)***, *for any set* $\nu\subseteq 1:p$, *we have:*

(i) If $\nu\subseteq\nu_0$, *then*

$$
\begin{aligned}
\boldsymbol{x}_\nu^T C^{-2}\boldsymbol{x}_\nu &= \frac{1}{n}\boldsymbol{e}_{\nu\lhd\nu_0}^T\left(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}\right)^{-1}R_{\nu_0\nu_0}^{-1} \\
&\quad\times(\boldsymbol{x}_{\nu_0}^T A_{\nu_0}^{-2}\boldsymbol{x}_{\nu_0}/n)R_{\nu_0\nu_0}^{-1}\left(\boldsymbol{e}_{\nu_0}^T\Sigma\boldsymbol{e}_{\nu_0}\right)^{-1}\boldsymbol{e}_{\nu\lhd\nu_0} \\
&\quad+O(n^{-2+3\alpha_0+3\alpha_1}).
\end{aligned}
$$

(ii) If $\nu\subseteq 1:p\setminus\nu_0$, *then*

$$
\begin{aligned}
\boldsymbol{x}_\nu^T C^{-2}\boldsymbol{x}_\nu &= n\left\{\boldsymbol{x}_\nu^T A_{\nu_0}^{-2}\boldsymbol{x}_\nu/n-R_{\nu\nu_0}R_{\nu_0\nu_0}^{-1}\boldsymbol{x}_{\nu_0}^T A_{\nu_0}^{-2}\boldsymbol{x}_\nu/n\right. \\
&\quad-\left(\boldsymbol{x}_\nu^T A_{\nu_0}^{-2}\boldsymbol{x}_{\nu_0}/n\right)R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu}+R_{\nu\nu_0}R_{\nu_0\nu_0}^{-1}\left(\boldsymbol{x}_{\nu_0}^T A_{\nu_0}^{-2}\boldsymbol{x}_{\nu_0}/n\right) \\
&\quad\left.\times R_{\nu_0\nu_0}^{-1}R_{\nu_0\nu}\right\}+O(n^{-1+4\alpha_0+\alpha_1}).
\end{aligned}
$$

(iii) If $\nu=\nu_1\cup\nu_2$ *with* $\nu_1\subseteq\nu_0$ *and* $\nu_2\subseteq 1:p\setminus\nu_0$, *then*

$$
\boldsymbol{x}_\nu^T C^{-2}\boldsymbol{x}_\nu=\begin{pmatrix}\boldsymbol{x}_{\nu_1}^T C^{-2}\boldsymbol{x}_{\nu_1} & \boldsymbol{x}_{\nu_1}^T C^{-2}\boldsymbol{x}_{\nu_2}\\ \boldsymbol{x}_{\nu_2}^T C^{-2}\boldsymbol{x}_{\nu_1} & \boldsymbol{x}_{\nu_2}^T C^{-2}\boldsymbol{x}_{\nu_2}\end{pmatrix},
$$

where

$$\begin{aligned}
\boldsymbol{x}_{\nu_1}^T C^{-2} \boldsymbol{x}_{\nu_1} &= \frac{1}{n} \boldsymbol{e}_{\nu_1 \lhd \nu_0}^T \left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right)^{-1} R_{\nu_0\nu_0}^{-1} (\boldsymbol{x}_{\nu_0}^T A_{\nu_0}^{-2} \boldsymbol{x}_{\nu_0}/n) R_{\nu_0\nu_0}^{-1} \\
&\quad \times \left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right)^{-1} \boldsymbol{e}_{\nu_1 \lhd \nu_0} + O(n^{-2+3\alpha_0+3\alpha_1}). \\
\boldsymbol{x}_{\nu_1}^T C^{-2} \boldsymbol{x}_{\nu_2} &= \frac{1}{n} \boldsymbol{e}_{\nu_1 \lhd \nu_0}^T \left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right)^{-1} R_{\nu_0\nu_0}^{-1} \left(\left(\boldsymbol{e}_{\nu_0}^T \Sigma \boldsymbol{e}_{\nu_0}\right)^{-1} \zeta_0 + O(1)\right) \\
&\quad \times R_{\nu_0\nu_0}^{-1} R_{\nu_0\nu_2} + O(n^{-2+4\alpha_0+3\alpha_1}). \\
\boldsymbol{x}_{\nu_2}^T C^{-2} \boldsymbol{x}_{\nu_2} &= n \left(\boldsymbol{x}_{\nu_2}^T A_{\nu_0}^{-2} \boldsymbol{x}_{\nu_2}/n - R_{\nu_2\nu_0} R_{\nu_0\nu_0}^{-1} \boldsymbol{x}_{\nu_0}^T A_{\nu_0}^{-2} \boldsymbol{x}_{\nu_2}/n \right. \\
&\quad -(\boldsymbol{x}_{\nu_2}^T A_{\nu_0}^{-2} \boldsymbol{x}_{\nu_0}/n) R_{\nu_0\nu_0}^{-1} R_{\nu_0\nu_2} + R_{\nu_2\nu_0} R_{\nu_0\nu_0}^{-1} (\boldsymbol{x}_{\nu_0}^T A_{\nu_0}^{-2} \boldsymbol{x}_{\nu_0}/n) \\
&\quad \left. \times R_{\nu_0\nu_0}^{-1} R_{\nu_0\nu_2}\right) + O(n^{-1+4\alpha_0+\alpha_1}).
\end{aligned}$$

Proof of Lemma B.2. Note that

$$\begin{aligned}
\mathrm{C}^{-2} &= \left(\mathrm{A}_{\nu_0}^{-1} - \mathrm{A}_{\nu_0}^{-1} \mathbf{x}_{\nu_0} \left(\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} + \mathbf{x}_{\nu_0}^T \mathrm{A}_{\nu_0}^{-1} \mathbf{x}_{\nu_0}\right)^{-1} \mathbf{x}_{\nu_0}^T \mathrm{A}_{\nu_0}^{-1}\right)^2 \\
&= \mathrm{A}_{\nu_0}^{-2} - \mathrm{A}_{\nu_0}^{-1} \mathbf{x}_{\nu_0} \left(\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} + \mathbf{x}_{\nu_0}^T \mathrm{A}_{\nu_0}^{-1} \mathbf{x}_{\nu_0}\right)^{-1} \mathbf{x}_{\nu_0}^T \mathrm{A}_{\nu_0}^{-2} \\
&\quad - \mathrm{A}_{\nu_0}^{-2} \mathbf{x}_{\nu_0} \left(\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} + \mathbf{x}_{\nu_0}^T \mathrm{A}_{\nu_0}^{-1} \mathbf{x}_{\nu_0}\right)^{-1} \mathbf{x}_{\nu_0}^T \mathrm{A}_{\nu_0}^{-1} \\
&\quad + \mathrm{A}_{\nu_0}^{-1} \mathbf{x}_{\nu_0} \left(\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} + \mathbf{x}_{\nu_0}^T \mathrm{A}_{\nu_0}^{-1} \mathbf{x}_{\nu_0}\right)^{-1} \mathbf{x}_{\nu_0}^T \mathrm{A}_{\nu_0}^{-2} \mathbf{x}_{\nu_0} \\
&\quad \times \left(\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} + \mathbf{x}_{\nu_0}^T \mathrm{A}_{\nu_0}^{-1} \mathbf{x}_{\nu_0}\right)^{-1} \mathbf{x}_{\nu_0}^T \mathrm{A}_{\nu_0}^{-1}.
\end{aligned}$$

The proof is completed by some algebraic manipulation. □

Proof of Theorem B.3. To prove the part (i), let $\nu = \nu_1 \cup \{a\} \cup \nu_2$ and abusing the notation, let $\mathbf{e}_{a|\nu_1\cup\nu_2}^T = (0_{|\nu_1|}^T, 0_{|\nu_2|}^T, 1)$ and $\mathbf{e}_{a|\nu_2}^T = (0_{|\nu_2|}^T, 1)$, where $0_{|\nu_1|}$ and $0_{|\nu_2|}$ are $|\nu_1|$-dimensional and $|\nu_2|$-dimensional vectors of zeros. Then, we have

$$\mathbf{w}_{a|\nu_1\cup\nu_2}^T \mathbf{w}_{a|\nu_1\cup\nu_2} = \mathbf{e}_{a|\nu_1\cup\nu_2}^T \gamma_\nu \mathbf{x}_\nu^T \mathrm{C}^{-2} \mathbf{x}_\nu \gamma_\nu \mathbf{e}_{a|\nu_1\cup\nu_2}. \tag{B.3}$$

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

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

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

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

Substituting them into the equation (B.3), we have

$$\begin{aligned}
\mathbf{w}_{a|\nu_1\cup\nu_2}^T \mathbf{w}_{a|\nu_1\cup\nu_2} &= \mathbf{e}_{a|\nu_2}^T \left(\mathrm{A}^{21} \mathrm{B}_{11} \mathrm{A}^{12} + \mathrm{A}^{22} \mathrm{B}_{21} \mathrm{A}^{12} + \mathrm{A}^{21} \mathrm{B}_{12} \mathrm{A}^{22} \right. \\
&\quad \left. + \mathrm{A}^{22} \mathrm{B}_{22} \mathrm{A}^{22}\right) \mathbf{e}_{a|\nu_2},
\end{aligned} \tag{B.4}$$

where it follows from Lemmas B.1 and B.2 that

$$\begin{aligned}
\mathrm{A}^{11} &= \left(\mathbf{e}_{\nu_1\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0}\right)^{-1}+O(n^{-1+3\alpha_0+2\alpha_1}).\\
\mathrm{A}^{12} &= -n^{-1}\left(\mathbf{e}_{\nu_1\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\\
&\quad\times\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0(\nu_2\cup\{a\})}\mathrm{F}_{\nu_2\cup\{a\}}^{-1}+O(n^{-2+5\alpha_0+3\alpha_1}),\\
\mathrm{A}^{21} &= -n^{-1}\mathrm{F}_{\nu_2\cup\{a\}}^{-1}\mathrm{R}_{(\nu_2\cup\{a\})\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0}\left(\mathbf{e}_{\nu_1\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\right.\\
&\quad\left.\times\mathbf{e}_{\nu_1\triangleleft\nu_0}\right)^{-1}+O(n^{-2+5\alpha_0+3\alpha_1}).\\
\mathrm{A}^{22} &= n^{-1}\mathrm{F}_{\nu_2\cup\{a\}}^{-1}+O(n^{-2+4\alpha_0+2\alpha_1}).
\end{aligned}$$

with $\mathrm{F}_{\nu_2\cup\{a\}}=\mathrm{R}_{(\nu_2\cup\{a\})(\nu_2\cup\{a\})}-\mathrm{R}_{(\nu_2\cup\{a\})\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0(\nu_2\cup\{a\})}$. And

$$\begin{aligned}
\mathrm{B}_{11} &= n^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}(\mathbf{x}_{\nu_0}^T\mathrm{A}_{\nu_0}^{-2}\mathbf{x}_{\nu_0}/n)\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0}\\
&\quad+O(n^{-2+3\alpha_0+3\alpha_1}),\\
\mathrm{B}_{12} &= n^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\zeta_0\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}+O(1)\right)\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0(\nu_2\cup\{a\})}\\
&\quad+O(n^{-2+4\alpha_0+3\alpha_1}),\\
\mathrm{B}_{21} &= n^{-1}\mathrm{R}_{(\nu_2\cup\{a\})\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\zeta_0\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}+O(1)\right)\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\triangleleft\nu_0}\\
&\quad+O(n^{-2+4\alpha_0+3\alpha_1}),\\
\mathrm{B}_{22} &= n\left\{\mathbf{x}_{\nu_2\cup\{a\}}^T\mathrm{A}_{\nu_0}^{-2}\mathbf{x}_{\nu_2\cup\{a\}}/n-\mathrm{R}_{(\nu_2\cup\{a\})\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathbf{x}_{\nu_0}^T\mathrm{A}_{\nu_0}^{-2}\mathbf{x}_{\nu_2\cup\{a\}}/n\right.\\
&\quad-(\mathbf{x}_{\nu_2\cup\{a\}}^T\mathrm{A}_{\nu_0}^{-2}\mathbf{x}_{\nu_0}/n)\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0(\nu_2\cup\{a\})}\\
&\quad\left.+\mathrm{R}_{(\nu_2\cup\{a\})\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}(\mathbf{x}_{\nu_0}^T\mathrm{A}_{\nu_0}^{-2}\mathbf{x}_{\nu_0}/n)\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0(\nu_2\cup\{a\})}\right\}+O(n^{-1+4\alpha_0+\alpha_1}).
\end{aligned}$$

Using the above asymptotic expressions, we obtain $\mathrm{A}^{21}\mathrm{B}_{11}\mathrm{A}^{12}=O(n^{-3+7\alpha_0+4\alpha_1})$, $\mathrm{A}^{22}\mathrm{B}_{21}\mathrm{A}^{12}=O(n^{-3+5\alpha_0+3\alpha_1})$, $\mathrm{A}^{21}\mathrm{B}_{12}\mathrm{A}^{22}=O(n^{-3+5\alpha_0+3\alpha_1})$, and $\mathrm{A}^{22}\mathrm{B}_{22}\mathrm{A}^{22}$ is equal to

$$\begin{aligned}
&n^{-1}\mathrm{F}_{\nu_2\cup\{a\}}^{-1}\left(\mathbf{x}_{\nu_2\cup\{a\}}^T\mathrm{A}_{\nu_0}^{-2}\mathbf{x}_{\nu_2\cup\{a\}}/n-\mathrm{R}_{(\nu_2\cup\{a\})\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathbf{x}_{\nu_0}^T\mathrm{A}_{\nu_0}^{-2}\mathbf{x}_{\nu_2\cup\{a\}}/n\right.\\
&-(\mathbf{x}_{\nu_2\cup\{a\}}^T\mathrm{A}_{\nu_0}^{-2}\mathbf{x}_{\nu_0}/n)\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0(\nu_2\cup\{a\})}\\
&\left.+\mathrm{R}_{(\nu_2\cup\{a\})\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}(\mathbf{x}_{\nu_0}^T\mathrm{A}_{\nu_0}^{-2}\mathbf{x}_{\nu_0}/n)\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0(\nu_2\cup\{a\})}+O(n^{-2+4\alpha_0+\alpha_1})\right)\mathrm{F}_{\nu_2\cup\{a\}}^{-1}\\
&+O(n^{-2+6\alpha_0+2\alpha_1}).
\end{aligned}$$

Combining these equations with the equation (B.4) and Condition (C4), we show that $\mathbf{w}_{a|\nu_1\cup\nu_2}^T\mathbf{w}_{a|\nu_1\cup\nu_2}$ is equal to

$$\begin{aligned}
&\zeta_0 n^{-1}\mathbf{e}_{a|\nu_2}^T\mathrm{F}_{\nu_2\cup\{a\}}^{-1}\mathbf{e}_{a|\nu_2}+O(n^{-2+4\alpha_0+\alpha_1})\\
&\quad=\zeta_0 n^{-1}\left(\mathrm{R}_{aa}-\mathrm{R}_{a\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0 a}-(\mathrm{R}_{a\nu_2}-\mathrm{R}_{a\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu_2})\right.
\end{aligned}$$

$$\times(\mathrm{R}_{\nu_2\nu_2}-\mathrm{R}_{\nu_2\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu_2})^{-1}(\mathrm{R}_{\nu_2 a}-\mathrm{R}_{\nu_2\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0 a})\big)^{-1}$$
$$+O(n^{-2+4\alpha_0+\alpha_1}).$$

This together with Theorem B.2 yields

$$\mathrm{SNR}_{a|\nu_1\cup\nu_2}=\frac{\gamma_{a|\nu_1\cup\nu_2}}{\sigma^2\mathbf{w}_{a|\nu_1\cup\nu_2}^T\mathbf{w}_{a|\nu_1\cup\nu_2}}=\frac{1}{\zeta_0\sigma^2}+O(n^{-2+5\alpha_0+2\alpha_1}).$$

This completes the proof of the part (i).

To prove the part (ii), let $\nu=\{a\}\cup\nu_1\cup\nu_2$. Let $\mathbf{e}_{a|\nu_1\cup\nu_2}^T=(1,0_{|\nu_1|}^T,0_{|\nu_2|}^T)$ and $\mathbf{e}_{a|\nu_1}^T=(1,0_{|\nu_1|}^T)$, where $0_{|\nu_1|}$ and $0_{|\nu_2|}$ are $|\nu_1|$-dimensional and $|\nu_2|$-dimensional vectors of zeros. Then, we have

$$\mathbf{w}_{a|\nu_1\cup\nu_2}^T\mathbf{w}_{a|\nu_1\cup\nu_2}=\mathbf{e}_{a|\nu_1\cup\nu_2}^T\gamma_\nu\mathbf{x}_\nu^T\mathrm{C}^{-2}\mathbf{x}_\nu\gamma_\nu\mathbf{e}_{a|\nu_1\cup\nu_2}. \tag{B.5}$$

We partition $\mathbf{x}_\nu^T\mathrm{C}^{-1}\mathbf{x}_\nu$ into the following block matrix:

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

Following the same block dimensions as above, we partition $\gamma_\nu=\left(\mathbf{x}_\nu^T\mathrm{C}^{-1}\mathbf{x}_\nu\right)^{-1}$ and $\mathbf{x}_\nu^T\mathrm{C}^{-2}\mathbf{x}_\nu$ into the block matrices as follows:

$$\gamma_\nu=\begin{pmatrix}\mathrm{F}^{11} & \mathrm{F}^{12}\\ \mathrm{F}^{21} & \mathrm{F}^{22}\end{pmatrix},\quad \mathbf{x}_\nu^T\mathrm{C}^{-2}\mathbf{x}_\nu=\begin{pmatrix}\mathrm{G}_{11} & \mathrm{G}_{12}\\ \mathrm{G}_{21} & \mathrm{G}_{22}\end{pmatrix}.$$

To derive the asymptotic expressions for these block matrices, we applied the matrix inverse formula for each block matrix. We have

$$\mathrm{F}^{11}=\Sigma_{\{a\}\cup\nu_1}-O(n^{-1+3\alpha_0+2\alpha_1}),$$

where

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

A similar derivation gives

$$\mathrm{F}^{22}=\left(n\mathrm{F}_{\nu_2}+O(n^{2\alpha_0+2\alpha_1})\right)^{-1}=n^{-1}\mathrm{F}_{\nu_2}^{-1}+O(n^{-2+4\alpha_0+2\alpha_1})=O(n^{-1+\alpha_0}),$$

where $\mathrm{F}_{\nu_2}=\mathrm{R}_{\nu_2\nu_2}-\mathrm{R}_{\nu_2\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu_2}$ and

$$\begin{aligned}\mathrm{F}^{12} &= -n^{-1}\Sigma_{(\{a\}\cup\nu_1)}\mathbf{e}_{(\{a\}\cup\nu_1)\lhd\nu_0}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu_2}\mathrm{F}_{\nu_2}^{-1}\\ &\quad +O(n^{-2+5\alpha_0+3\alpha_1}),\end{aligned}$$

$$\begin{aligned}\mathrm{F}^{21} &= -n^{-1}\mathrm{F}_{\nu_2}^{-1}\mathrm{R}_{\nu_2\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{(\{a\}\cup\nu_1)\lhd\nu_0}^T\Sigma_{\{a\}\cup\nu_1}\\ &\quad +O(n^{-2+5\alpha_0+3\alpha_1}).\end{aligned}$$

Using Condition (C4), we can show that

$$\begin{aligned}\mathrm{G}_{11} &= n^{-1}\mathbf{e}_{(\{a\}\cup\nu_1)\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}(\mathbf{x}_{\nu_0}^T\mathrm{A}_{\nu_0}^{-2}\mathbf{x}_{\nu_0}/n)\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\\ &\quad\times\mathbf{e}_{(\{a\}\cup\nu_1)\lhd\nu_0}+O(n^{-2+3\alpha_0+3\alpha_1})\\ &= \zeta_0 n^{-1}\mathbf{e}_{(\{a\}\cup\nu_1)\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{(\{a\}\cup\nu_1)\lhd\nu_0}\\ &\quad +O(n^{-2+3\alpha_0+3\alpha_1}),\\ \mathrm{G}_{12} &= \zeta_0 n^{-1}\mathbf{e}_{(\{a\}\cup\nu_1)\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu_2}\\ &\quad +O(n^{-2+4\alpha_0+3\alpha_1}),\\ \mathrm{G}_{21} &= \zeta_0 n^{-1}\mathrm{R}_{\nu_2\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathrm{R}_{\nu_0\nu_0}^{-1}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{(\{a\}\cup\nu_1)\lhd\nu_0}\\ &\quad +O(n^{-2+4\alpha_0+3\alpha_1}),\\ \mathrm{G}_{22} &= \zeta_0 n\left(\mathrm{F}_{\nu_2}+O(n^{-1+2\alpha_0})\right).\end{aligned}$$

Consequently, substituting the above expressions into the equation (B.5), we have

$$\begin{aligned}\mathbf{w}_{a|\nu_1\cup\nu_2}^T\mathbf{w}_{a|\nu_1\cup\nu_2} &= \mathbf{e}_{a|\nu_1}^T\left(\mathrm{F}^{11}\mathrm{G}_{11}\mathrm{F}^{11}+\mathrm{F}^{12}\mathrm{G}_{12}\mathrm{F}^{11}+\mathrm{F}^{11}\mathrm{G}_{12}\mathrm{F}^{21}\right.\\ &\quad \left.+\mathrm{F}^{12}\mathrm{G}_{22}\mathrm{F}^{21}\right)\mathbf{e}_{a|\nu_1}\end{aligned}\tag{B.6}$$

with

$$\begin{aligned}&\mathrm{F}^{11}\mathrm{G}_{11}\mathrm{F}^{11}+\mathrm{F}^{12}\mathrm{G}_{12}\mathrm{F}^{11}+\mathrm{F}^{11}\mathrm{G}_{12}\mathrm{F}^{21}+\mathrm{F}^{12}\mathrm{G}_{22}\mathrm{F}^{21}\\ &\quad=\zeta_0 n^{-1}\Sigma_{\{a\}\cup\nu_1}\mathbf{e}_{(\{a\}\cup\nu_1)\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)\mathbf{\Phi}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{(\{a\}\cup\nu_1)\lhd\nu_0}\Sigma_{\{a\}\cup\nu_1}\\ &\times(1+o(1)),\end{aligned}$$

where

$$\mathbf{\Phi}=\mathrm{R}_{\nu_0\nu_0}^{-1}+\mathrm{R}_{\nu_0\nu_0}^{-1}\mathrm{R}_{\nu_0\nu_2}\mathrm{F}_{\nu_2}^{-1}\mathrm{R}_{\nu_2\nu_0}\mathrm{R}_{\nu_0\nu_0}^{-1}.$$

We partition $\Sigma_{\{a\}\cup\nu_1}^{-1}$ into a block matrix, namely

$$\Sigma_{\{a\}\cup\nu_1}^{-1}=\begin{pmatrix}\mathbf{e}_{a\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{a\lhd\nu_0} & \mathbf{e}_{a\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\lhd\nu_0}\\ \mathbf{e}_{\nu_1\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{a\lhd\nu_0} & \mathbf{e}_{\nu_1\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{\nu_1\lhd\nu_0}\end{pmatrix},$$

where its inverse can be parallelly written as

$$\Sigma_{\{a\}\cup\nu_1}=\begin{pmatrix}D^{11} & D^{12}\\ D^{21} & D^{22}\end{pmatrix}.$$

Similarly, we write

$$\mathbf{e}_{(\{a\}\cup\nu_1)\lhd\nu_0}^T\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{\Phi}\left(\mathbf{e}_{\nu_0}^T\Sigma\mathbf{e}_{\nu_0}\right)^{-1}\mathbf{e}_{(\{a\}\cup\nu_1)\lhd\nu_0}=\begin{pmatrix}\mathrm{H}_{11} & \mathrm{H}_{12}\\ \mathrm{H}_{21} & \mathrm{H}_{22}\end{pmatrix},$$

where

$$\begin{aligned}
\mathrm{H}_{11} &= \mathbf{e}_{a\lhd\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{\Phi} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{a\lhd\nu_0},\\
\mathrm{H}_{12} &= \mathbf{e}_{a\lhd\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{\Phi} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\nu_1\lhd\nu_0},\\
\mathrm{H}_{21} &= \mathbf{e}_{\nu_1\lhd\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{\Phi} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{a\lhd\nu_0},\\
\mathrm{H}_{22} &= \mathbf{e}_{\nu_1\lhd\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{\Phi} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\nu_1\lhd\nu_0}.
\end{aligned}$$

Combining these partitions with the equation (B.6), we have

$$\begin{aligned}
\mathbf{w}_{a|\nu_1\cup\nu_2}^T \mathbf{w}_{a|\nu_1\cup\nu_2} &= \zeta_0 n^{-1}(\mathrm{D}^{11}\mathrm{H}_{11}\mathrm{D}^{11} + \mathrm{D}^{12}\mathrm{H}_{21}\mathrm{D}^{11} + \mathrm{D}^{11}\mathrm{H}_{12}\mathrm{D}^{21}\\
&\quad + \mathrm{D}^{12}\mathrm{H}_{22}\mathrm{D}^{21})(1+o(1)). \qquad \text{(B.7)}
\end{aligned}$$

Note that

$$\begin{aligned}
\mathrm{D}^{11} &= \left(\mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0}\right)^{-1},\\
\mathrm{D}^{12} &= -\mathrm{D}^{11}\mathbf{e}_{a\lhd\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{\nu_1\lhd\nu_0} \Sigma_{\nu_1\lhd\nu_0},\\
\mathrm{D}^{21} &= -\Sigma_{\nu_1\lhd\nu_0} \mathbf{e}_{\nu_1\lhd\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{a\lhd\nu_0} \mathrm{D}^{11}.
\end{aligned}$$

Therefore, the equation (B.7) implies that

$$\mathbf{w}_{a|\nu_1\cup\nu_2}^T \mathbf{w}_{a|\nu_1\cup\nu_2} = \zeta_0 n^{-1} (\mathrm{D}^{11})^2 \mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{\Phi} \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0} (1+o(1)).$$

Finally, by definition, we have

$$\begin{aligned}
\mathrm{SNR}_{a|\nu_1\cup\nu_2} &= \frac{\mathrm{D}^{11} + n^{-1}\mathbf{e}_{a\lhd\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{\Psi} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{a\lhd\nu_0}}{\sigma^2\zeta_0 n^{-1}(\mathrm{D}^{11})^2 \mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{\Phi} \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0} (1+o(1))}\\
&\quad + \frac{O(n^{-2+6\alpha_0+5\alpha_1})}{\sigma^2\zeta_0 n^{-1}(\mathrm{D}^{11})^2 \mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{\Phi} \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0} (1+o(1))}\\
&= \frac{n\mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0} + O(n^{-2+6\alpha_0+5\alpha_1})}{\sigma^2\zeta_0 \mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{\Phi} \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0} (1+o(1))}\\
&\quad + \frac{\left(\mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0}\right)^2 \mathbf{e}_{a\lhd\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{\Psi} \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{a\lhd\nu_0}}{\sigma^2\zeta_0 \mathbf{e}_{a\lhd\nu_0}^T \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{\Phi} \Sigma_{\nu_0\setminus\nu_1}^{-1} \mathbf{e}_{a\lhd\nu_0} (1+o(1))}.
\end{aligned}$$

The proof is completed. □

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

Let $\kappa = \max\{2(2/\kappa_1 + 1/\kappa_2) - 1, (4/3)(1/\kappa_1 + 1/\kappa_2) - 1/3, 1\}$. 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 $||D||_F \leq ||D||$.

Lemma B.3. *Under Conditions (C0)∼(C3), if* $\tau_{nJ} = o(1)$ *as* $n \to \infty$ *and* $J \to \infty$,

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

Proof of Lemma B.3. Without loss of generality, we assume $\pi(j) = j, 1 \le j \le J$. We use the methods developed in [7, 23] to prove the lemma. Following the proof in [23], we can find constants $d_t, t = 1, ..., 5$, such that for any $u > 0$,

$$\begin{aligned} &P\left(\max_{1\le i,j\le n} \left|\frac{1}{J}\sum_{k=1}^{J} y_{ik}y_{kj} - c_{ij}\right| > u\right) \\ &\le n^2 J \exp\left(-\frac{(Ju)^\kappa}{d_1}\right) + n^2 \exp\left(-\frac{(Ju)^2}{d_2(1+Jd_3)}\right) \\ &+n^2 \exp\left(-\frac{(Ju)^2}{d_4 J} \exp\left(\frac{(Ju)^{\kappa(1-\kappa)}}{d_5(\log(Ju))^\kappa}\right)\right). \end{aligned} \tag{B.8}$$

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 (B.8), we have

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

We also have

$$\begin{aligned} E[\max_{1\le i,j\le n} Q_{ij}^2] &\le 2(h_n\tau_{nJ})^2 + \int_{(h_n\tau_{nJ})^2}^{\infty} P\left(\max_{1\le i,j\le n} Q_{ij} > \sqrt{u}\right) du \\ &= 2(h_n\tau_{nJ})^2 + 2\int_{h_n\tau_{nJ}}^{\infty} vP\left(\max_{1\le i,j\le n} Q_{ij} > v\right) dv \\ &\le (\tau_{nJ})^2(2h_n^2 + o(1)) = O(\tau_{nJ}^2). \end{aligned}$$

Finally, for as ϵ, n and J tend to infinity,

$$P\left(\max_{1\le i,j\le n} Q_{ij} > \epsilon\tau_{nJ}\right) \le E[\max_{1\le i,j\le n} Q_{ij}]/(\epsilon\tau_{nJ}) \to 0,$$

which implies $\max_{1\le i,j\le n} Q_{ij} = O_p(\tau_{nJ})$.

Similarly, we can show

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

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

Let m_n be the number of non-zero entries in covariance matrix C. In the next lemma, we show the convergence rates of the threshold estimator.

Lemma B.4. *Under Conditions (C0)∼(C6), if $m_n\tau_{nJ} = o(1)$ as $n \to \infty$ and $J \to \infty$, then for $h > 0$,*

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

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

Proof of Lemma B.4. Without loss of generality, we assume $\pi(j) = j, 1 \le j \le J$. We consider $h > 0$. Denote

$$\begin{aligned} T_1 &= ||(\hat{c}_{ij}I(|\hat{c}_{ij}| > h\tau_{nJ})) - (c_{ij}I(|c_{ij}| > h\tau_{nJ}))||, \\ \mathrm{I} &= \max_i \sum_{j=1}^n |\hat{c}_{ij} - c_{ij}|I(|\hat{c}_{ij}| > h\tau_{nJ}, |c_{ij}| > h\tau_{nJ}), \\ \mathrm{II} &= \max_i \sum_{j=1}^n |\hat{c}_{ij}|I(|\hat{c}_{ij}| > h\tau_{nJ}, |c_{ij}| \le h\tau_{nJ}), \\ \mathrm{III} &= \max_i \sum_{j=1}^n |c_{ij}|I(|\hat{c}_{ij}| \le h\tau_{nJ}, |c_{ij}| > h\tau_{nJ}). \end{aligned}$$

Then,

$$\begin{aligned} ||\hat{\mathrm{C}}_h - \mathrm{C}|| &\le T_1 + \max_i \sum_{j=1}^n |c_{ij}I(|c_{ij}| \le h\tau_{nJ}) \le T_1 + h\tau_{nJ}m_n \\ &\le \mathrm{I} + \mathrm{II} + \mathrm{III} + h\tau_{nJ}m_n. \end{aligned} \tag{B.9}$$

By Lemma B.3, we have

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

Note that for $0 \leq \delta < 1$,

$$\begin{aligned} \text{II} &\leq \max_{1\leq i,j\leq n} |\hat{c}_{ij} - c_{ij}| \left(\max_i \sum_{j=1}^n I(|\hat{c}_{ij} - c_{ij}| \geq (1-\delta)h\tau_{nJ}) + m_n \right) \\ &\quad + h\tau_{nJ} m_n. \end{aligned} \tag{B.10}$$

And

$$\begin{aligned} &E \max_{1\leq i,j\leq n} |\hat{c}_{ij} - c_{ij}| \max_i \sum_{j=1}^n I(|\hat{c}_{ij} - c_{ij}| \geq (1-\delta)h\tau_{nJ}) \\ &\leq (m_n + \epsilon) E \max_{1\leq i,j\leq n} |\hat{c}_{ij} - c_{ij}| = O(m_n \tau_{nJ}). \end{aligned}$$

This, together with the inequality in (B.10), implies that

$$E[\text{II}] \leq O(m_n\tau_n) + O(\tau_{nJ})m_n + h\tau_n m_n = O(m_n\tau_{nJ}).$$

Similarly,

$$\begin{aligned} E[\text{III}] &\leq E \max_i \sum_{j=1}^n |\hat{c}_{ij} - c_{ij}| \sum_{j=1}^n I(|c_{ij}| > h\tau_{nJ}) + h\tau_{nJ} m_n \\ &\leq O(\tau_{nJ})m_n + h\tau_{nJ} m_n = O(\tau_{nJ} m_n). \end{aligned}$$

Consequently,

$$E[T_1] \leq E[\text{I}] + E[\text{II}] + E[\text{III}] = O(m_n\tau_{nJ}),$$

which, together with the inequality in (B.9), implies that

$$E||\hat{\text{C}}_h - \text{C}|| = O(m_n\tau_{nJ}).$$

Invoking the Chebyshev's inequality, we have

$$||\hat{\text{C}}_h - \text{C}|| \leq O_p(m_n\tau_{nJ}).$$

We now turn to the second inequality. Note that

$$E||\hat{\text{C}}_h - \text{C}||^2 \leq 6(E[\text{I}^2] + E[\text{II}^2] + E[\text{III}^2]) + 2h^2(\tau_{nJ}m_n)^2. \tag{B.11}$$

$$\begin{aligned} E[\text{I}^2] &\leq 2E \max_{1\leq i,j\leq n} |\hat{c}_{ij} - c_{ij}|^2 m_n^2 = O(m_n\tau_{nJ})^2. \\ E[\text{II}^2] &\leq 2E \max_{1\leq i,j\leq n} |\hat{c}_{ij} - c_{ij}|^2 (m_n + \epsilon + m_n)^2 + 2(hm_n\tau_{nJ})^2 \\ &= O(m_n\tau_{nJ})^2. \\ E[\text{III}^2] &\leq 2E \max_{1\leq i,j\leq n} |\hat{c}_{ij} - c_{ij}|^2 m_n^2 + 2(h\tau_{nJ}m_n)^2 = O(m_n\tau_{nJ})^2. \end{aligned}$$

These with the inequality in (B.11) implies that

$$E||\hat{\text{C}}_h - \text{C}||^2 \leq O(m_n\tau_{nJ})^2 + 2(hm_n\tau_{nJ})^2 = O(m_n\tau_{nJ})^2.$$

When $h = 0$, the results can be proved similarly. The proof is completed. □

Lemma B.5. *Under Conditions (C0)∼(C6), 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 of Lemma B.5. 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

$$\frac{1}{Jn}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[(y_{i1}y_{j1}-\hat{c}_{ij})^2 I(|\hat{c}_{ij}|>h\tau_{nJ})\right] = O(m_n/J). \tag{B.12}$$

For this purpose, we first show that

$$\begin{aligned}&\frac{1}{Jn}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[(y_{i1}y_{j1}-\hat{c}_{ij})^2 I(|\hat{c}_{ij}|>h\tau_{nJ})\right|\\ &\quad=\frac{1}{Jn}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[(y_{i1}y_{j1}-\hat{c}_{ij})^2 I(c_{ij}|>h\tau_{nJ})\right].\end{aligned} \tag{B.13}$$

Note that for constant $0 < \delta < 1$,

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

The last equality follows from the following facts

$$\begin{aligned}E(y_{i1}y_{j1}-\hat{c}_{ij})^2 &\le 2\max_i E[y_{i1}^4]+o(1),\\ E(y_{i1}y_{j1}-\hat{c}_{ij})^{2\eta_0} &\le 2^{2\eta_0}\max_i E[y_{i1}^{4\eta_0}]+o(1).\end{aligned}$$

The equation (B.13) is proved. Finally, note that

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

This together with (B.13) shows the equation (B.12). The proof is completed. □

Lemma B.6. *Under Conditions (C0)∼(C6), if $m_n\tau_{nJ}=o(1)$ as $n\to\infty$ and $J\to\infty$, then for $h>0$,*

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

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

Proof of Lemma B.6. Without loss of generality, we assume $\pi(j)=j, 1\leq j\leq J$. We consider $h>0$. The proof is similar when $h=0$. It follows from Lemma B.3 that

$$\begin{aligned}\hat{\mu}_n &= \frac{1}{n}\sum_{i}^{n}\hat{c}_{ii} = \frac{1}{n}\sum_{i}^{n}c_{ii} + O_p(\tau_{nJ})\\ &= \mu_n + O_p(\tau_{nJ}).\end{aligned}$$

It follows from [12] that $||\mathrm{C}||_F^2 = O(1)$, if $\max_i E[y_{i1}^4]<\infty$. We have

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

Note that

$$\begin{aligned}|d_n - ||\hat{\mathrm{C}}_h-\mu_n I_n||_F| &\leq ||(\hat{\mu}_n-\mu_n)I_n||_F\\ &= |\hat{\mu}_n-\mu_n| = O_p(\tau_{nJ}),\end{aligned}$$

which implies that

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

Therefore,

$$d_n^2 = \delta_n^2 + O_p(m_n\tau_{nJ}).$$

It follows from Lemma B.5 that $\bar{b}_n^2 = O(m_n/J)$. The proof is completed. □

Lemma B.7. *Under Conditions (C0)∼(C6), 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* $||\hat{C}_{hs}-C|| = O_p(m_n\tau_{jn})$, $||\hat{C}_{hs}^{-1} - C^{-1}|| = O_p(m_n\tau_{nJ})$, $||\hat{C}_{hs}^{-2} - C^{-2}|| = O_p(m_n\tau_{nJ})$.

Proof of Lemma B.7. Note that

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

The remaining proofs are similar to the proof of Lemma 7.3 in [23]. The details are omitted. □

Proof of Theorems B.4, 3.1 and 3.2: Invoking Theorems B.1∼B.3, the proofs are similar to the proof of Theorem 2 in [23] by using Lemmas B.1∼B.7 and thus omitted. □

Proof of Corollary B.3. For $a \in \nu_0$, we have

$$\begin{aligned} \hat{\gamma}_a &= \left(\mathbf{e}_{a|\nu_0}^T \left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1} \mathbf{e}_{a|\nu_0}\right)^{-1} + O(n^{-1+\alpha_0+2\alpha_1} + n^2\tau_{nJ}) \\ &\geq \left(\lambda_{\max}\left(\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right)^{-1}\right)\right)^{-1} + O(n^{-1+\alpha_0+2\alpha_1} + n^2\tau_{nJ}) \\ &= \lambda_{\min}\left(\mathbf{e}_{\nu_0}^T \Sigma \mathbf{e}_{\nu_0}\right) + O(n^{-1+\alpha_0+2\alpha_1} + n^2\tau_{nJ}) \\ &\geq \zeta_0 n^{-\alpha_1} + O(n^{-1+\alpha_0+2\alpha_1} + n^2\tau_{nJ}) \\ &= \zeta_0(1+o(1))n^{-\alpha_1} > n^{-1+\alpha_0}\log(n). \end{aligned}$$

For $a \notin \nu_0$, we have $\hat{\gamma}_a = O_p(n^{-1+\alpha_0} + n^2\tau_{nJ}) = O_p(n^{-1+\alpha_0})$, which implies that with probability tending to one, $a \notin \hat{\nu}_d$ as n and J tend to infinity. □

Proof of Corollary 3.1. Let $\omega_0 = \emptyset$ and $\hat{\mathrm{SNR}}_{k|\omega_0} = \hat{\mathrm{SNR}}_k$. According to the definition of PVA in Subsection 2.3, $w_m = \{k_m\} \cup \omega_{m-1}$, where $k_m = \mathrm{argmax}\{\hat{\mathrm{SNR}}_{k|\omega_{m-1}} : k \notin \omega_{m-1}\}$, $0 \leq m \leq \hat{m}$, with the value of $\mathrm{SNR}_{\max}$ at $\hat{m}$ falling into the confidence interval defined in the stopping criteria in the first time. Under the conditions in Theorem B.3, we first prove that $P(\omega_{\hat{m}} \subseteq \nu_0) \to 1$ as both n and J both tend to infinity. Note that

$$\begin{aligned} & P(\omega_{\hat{m}} \cap (1:p \setminus \nu_0) \neq \emptyset) \\ \leq\ & P\left(\exists k_m \notin \nu_0 \text{ such that } k_m = \mathrm{argmax}\{\hat{\mathrm{SNR}}_{k|\omega_{m-1}} : k \notin \omega_{m-1}\}\right) \\ \leq\ & P\left(\exists m \max\{\hat{\mathrm{SNR}}_{k|\omega_{m-1}} : k \in G_2\} \geq \min\{\hat{\mathrm{SNR}}_{k|\omega_{m-1}} : k \in G_1\}\right) \end{aligned} \quad \text{(B.14)}$$

where $G_1 = \{k \notin \omega_{m-1}, k \in \nu_0\}$ and $G_2 = \{k \notin \omega_{m-1}, k \notin \nu_0\}$. By Theorem 3.2, uniformly for m, $\max\{\hat{\mathrm{SNR}}_{k|\omega_{m-1}} : k \in G_2\}$ converges to $(\zeta_0\sigma^2)^{-1} \geq 1$ in probability uniformly for m, while $\min\{\hat{\mathrm{SNR}}_{k|\omega_{m-1}} : k \in G_1\}$ converges to infinity in probability. Combining these with the inequalities in (B.14), we show that $P(\omega_{\hat{m}} \cap (1:p \setminus \nu_0) \neq \emptyset)$ tends to zero, which is equivalent to $P(\omega_{\hat{m}} \subseteq \nu_0) \to 1$. To complete the proof, we show $P(\nu_0 \subseteq \omega_{\hat{m}}) \to 1$ by contradiction as follows. If there exists $k \in \nu_0 \setminus \omega_{\hat{m}}$, then we can divide $\{k \notin \omega_{\hat{m}}\}$ into two

non-empty groups $G_1 = \{k \notin \omega_{\hat{m}}, k \in \nu_0\}$ and $G_2 = \{k \notin \omega_{\hat{m}}, k \notin \nu_0\}$. By Theorem 3.2, we have $\min\{\hat{\text{SNR}}_{k|\omega_{\hat{m}}} : k \in G_1\}$ tends to infinity in probability, whereas $\max\{\hat{\text{SNR}}_{k|\omega_{\hat{m}}} : k \in G_2\}$ converges to $(\zeta_0\sigma^2)^{-1}$ in probability. This is in contradiction with the definition of $\hat{m}$. □

Acknowledgments

We are grateful to the Editor, an associate editor and two reviewers for their valuable comments on the manuscript that have helped to improve the paper. We are grateful to Professor Martin Michaelis from School of Bioscience, University of Kent for discussions on cancer drug studies.

References

[1] Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Stat.*, **36**, 2577–2604. MR2485008

[2] Brown, P. J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Jour. Roy. Stat. Soc. Ser. B*, **60**, 627–641. MR1626005

[3] Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Jour. Ameri. Stat. Assoc.*, **106**, 672–684. MR2847949

[4] Chen, L. and Huang, J. (2011). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Jour. Ameri. Stat. Assoc.*, **107**, 1533–1545. MR3036414

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

[6] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Jour. Ameri. Stat. Assoc.*, **96**, 1348–1360. MR1946581

[7] Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Ann. Stat.*, **39**, 3320–3356. MR3012410

[8] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Jour. Roy. Stat. Soc. Ser. B*, **70**, 849–911. MR2530322

[9] Froot, K. A. (1989). Consistent covariance matrix estimation with cross-sectional dependence and heteroskedasticity in cross-sectional financial data. *Jour. Finan. and Quant. Analy.*, **24**, 333–355.

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

[11] Laird, N. M. and Ware, J. H. (2002). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

[12] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Jour. Multi. Analy.*, **88**, 365–411. MR2026339

[13] Li, Y., Nan, B., and Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, **71**, 354–363. MR3366240

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

[15] Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R., and Wang, P. (2010). Regularized multivariate regression for identifying master covariates with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* **4**, 53–77. MR2758084

[16] Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Jour. Comput. Graph. Stat.*, **19**, 947–962. MR2791263

[17] Sofer, T., Dicker, L., and Lin, X. (2014). Variable selection for high dimensional multivariate outcomes. *Stat. Sinica*, **24**, 1633–1654. MR3308655

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

[19] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Jour. Royal. Statist. Soc. Ser. B*, **58**, 267–288. MR1379242

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

[21] Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Jour. Ameri. Stat. Assoc.*, **104**, 1512–1524. MR2750576

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

[23] Zhang, J. and Liu, C. (2015). On linearly constrained minimum variance beamforming. *Jour. Mach. Learn. Res.*, **16**, 2099–2145. MR3450503

[24] Zhang, J., Liu, C., and Green, G. (2014). Source localization with MEG data: A beamforming approach based on covariance thresholding. *Biometrics*, **70**, 121–131. MR3251673

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

[26] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Jour. Roy. Stat. Soc. Ser. B*, **67**, 301–320. MR2137327