Multivariate Regression Shrinkage and Selection by Canonical Correlation Analysis

Baiguo An[a,b], Jianhua Guo[a,b,*], Hansheng Wang[c]

[a] *Key Laboratory for Applied Statistics of Ministry of Education*
[b] *School of Mathematics and Statistics, Northeast Normal University,Changchun, 130024, Jilin Province, P. R. China*
[c] *Guanghua School of Management, Peking University, Beijing, 100871, P. R. China*

Abstract

The problem of regression shrinkage and selection for multivariate regression is considered. The goal is to consistently identify those variables relevant for regression. This is done not only for predictors but also for responses. To this end, a novel relationship between multivariate regression and canonical correlation is discovered. Subsequently, its equivalent least squares type formulation is constructed, and then the well developed adaptive LASSO type penalty and also a novel BIC-type selection criterion can be directly applied. Theoretical results show that the resulting estimator is selection consistent for not only predictors but also responses. Numerical studies are presented to corroborate our theoretical findings.

Keywords: Adaptive Lasso; Canonical Correlation Analysis; Multivariate Regression; Selection Consistency; Tuning Parameter Selection

1. INTRODUCTION

Due to the fast advance of information technology, a lot of high dimensional datasets have been collected across many different scientific disciplines, such as biology, computer science, engineering, social science, and many others. For those datasets, high dimensionality is a common feature. Then,

*Corresponding author. Tel.: (86) 431 85098576; Fax: (86) 431 85098237.
Email addresses: anbg200@gmail.com (Baiguo An), jhguo@nenu.edu.cn (Jianhua Guo), hansheng@gsm.pku.edu.cn (Hansheng Wang)

statistically how to analyze those high dimensional data becomes an important problem, for which the idea of variable selection has been found very useful.

Under a linear regression setup and also the assumption of sparsity, it has been well understood that correctly identifying sparse solutions can considerably improve the model interpretability and also estimation accuracy. To this end, various shrinkage methods have been developed. Specifically, [18] developed the method of least absolute shrinkage and selection operator (LASSO), which was subsequently further improved by [22], so that its improved version (i.e., the Adaptive LASSO) enjoys the oracle property, in the sense of [7]. To further extend the applicability of the LASSO method to the situation with extra high dimensional predictor, the method of elastic net was developed by [23] and [24]. Despite those pioneer methods' usefulness, none of them is capable of handling high dimensional multivariate responses. This motivates us to develop a novel method to solve this problem.

Specifically, we find that the classical method of canonical correlation analysis (CCA) has been popularly used as a regression method for multivariate responses; see for example [8, 9] and [1]. Recently, its successful applications have been found in computer visual processing [12], gene expression data analysis [15], supervised learning [17], and many others. To further extend CCA's applicability to multivariate regression, we propose here a method of adaptive sparse canonical correlation analysis (ASCCA). The new method is obtained by re-casting the multivariate regression problem as a classical CCA problem, for which a novel least squares type formulation can be constructed. Subsequently, the well developed adaptive LASSO type penalty together with a novel BIC-type selection criterion can be applied directly. We show theoretically that the new method is selection consistent [16] for not only predictors but also responses. Numerical studies are presented to corroborate our theoretical findings.

It is worthwhile to mention that our method is different from the methods of sparse principal component analysis (SPCA) of [25] and [13], where no multivariate responses are involved. Our method is also different from the sparse canonical correlation analysis (SCCA) of [15], where rather restrictive covariance assumptions have been made on both the predictors and responses. Our proposal is also different from the penalized canonical correlation analysis (PCCA) of [20], where no asymptotic theory has been developed.

The article is organized as follows. Next section introduces our ASCCA methodology together with asymptotic theories. Section 3 presents numerical studies. Lastly, the article is concluded with a short conclusion in Section 4.

2. THE ASCCA METHODOLOGY

2.1. A Multivariate Regression Framework

Let (X_i, Y_i) be the observation collected from the ith subject $(1 \le i \le n)$. $X_i = (X_{i1}, \cdots, X_{ip})^\top \in \mathbb{R}^p$ is the predictor and $Y_i = (Y_{i1}, \cdots, Y_{iq})^\top \in \mathbb{R}^q$ is the multivariate response. Assume that $E(X_i) = 0$ and $E(Y_i) = 0$. Next, define $\mathrm{cov}(X_i) = \Sigma_{xx} \in \mathbb{R}^{p\times p}$, $\mathrm{cov}(Y_i) = \Sigma_{yy} \in \mathbb{R}^{q\times q}$, and $\mathrm{cov}(X_i, Y_i) = \Sigma_{xy} = \Sigma_{yx}^\top \in \mathbb{R}^{p\times q}$. Further assume that Σ_{xx}, Σ_{yy} are positive definite. To model the regression relationship between X_i and Y_i, the following model is assumed

$$Y_i = B^\top X_i + \mathcal{E}_i, \tag{1}$$

where $\mathcal{E}_i = (\varepsilon_{i1}, \cdots, \varepsilon_{iq})^\top \in \mathbb{R}^q$ is the random noise and $B = (b_{jk}) \in \mathbb{R}^{p\times q}$ is the coefficient matrix. Write $\alpha_j = (b_{j1}, \cdots, b_{jq})^\top \in \mathbb{R}^q$ as the jth row and $\beta_k = (b_{1k}, \cdots, b_{pk})^\top \in \mathbb{R}^p$ as the kth column. We know immediately $B = (\beta_1, \cdots, \beta_q) = (\alpha_1, \cdots, \alpha_p)^\top \in \mathbb{R}^{p\times q}$. We next use B_0, α_{0j}, and β_{0k} to represent the true values of B, α_j, and β_k, respectively. Obviously, only the predictors with non-zero $\|\alpha_{0j}\|$ are relevant to Y_i, where $\|\cdot\|$ denotes L_2 norm. We thus collect all those indices by a set $\mathcal{M}_T = \{1 \le j \le p : \|\alpha_{0j}\| > 0\}$. We call this as the predictor true model (PTM). Similarly, we can define a predictor full model (PFM) as $\mathcal{M}_F = \{1, 2, \cdots, p\}$.

We next consider what kind of responses can be viewed as "redundant". For convenience, we use $\mathcal{N}_F = \{1, 2, \cdots, q\}$ to denote a response full model (RFM). For an arbitrary subset $\mathcal{N} \subset \mathcal{N}_F$, we use notation $Y_{i(\mathcal{N})}$ to denote the subvector of Y_i corresponding to $\mathcal{N}$. Define $\mathcal{N}^c = \mathcal{N}_F \backslash \mathcal{N}$. We then call $\mathcal{N}$ a "sufficient" response model (SRM) if the conditional distribution of $Y_{i(\mathcal{N}^c)} | (Y_{i(\mathcal{N})}, X_i)$ is the same as that of $Y_{i(\mathcal{N}^c)} | Y_{i(\mathcal{N})}$. Obviously there exists at least one sufficient response model, because $\mathcal{N}_F$ is a SRM. We then define $\mathcal{N}_T$ as the intersection of all SRMs. Under certain regularity conditions, one can easily show that $\mathcal{N}_T$ is also a SRM and thus is the smallest SRM. We then call it the response true model (RTM). Then, our objective is identifying not only the PTM $\mathcal{M}_T$ but also the RTM $\mathcal{N}_T$ consistently.

2.2. Canonical Correlation Analysis

As one can see, the definition of the RTM is very intuitive but less useful practically. This is because a response's relevance and/or irrelevance are not defined according to certain parameter's (e.g., regression coefficient B_0) sparseness. As a consequence, identifying RTM should be much more difficult than identifying PTM. As an interesting solution, we find that whether a response or predictor is relevant is closely related to its loadings on the canonical correlation between X_i and Y_i. Specifically, let $K = \min\{p, q\}$ and then $(\mu_k^\top X_i, \nu_k^\top Y_i)$ with $\mu_k = (\mu_{k1}, \cdots, \mu_{kp})^\top \in \mathbb{R}^p$ and $\nu_k = (\nu_{k1}, \cdots, \nu_{kq})^\top \in \mathbb{R}^q$ be the kth $(1 \leq k \leq K)$ pairs of canonical variables. Then, by [11], μ_k and ν_k are defined as the vectors, which maximize $\lambda_k = \mu_k^\top \Sigma_{xy} \nu_k$ but under the constraint $\mu_k^\top \Sigma_{xx} \mu_{k'} = \nu_k^\top \Sigma_{yy} \nu_{k'} = \delta_{kk'}$ for every $1 \leq k' \leq k$. Here $\delta_{kk'} = 1$ if $k = k'$ and 0 otherwise. Obviously, we are only interested in those positive canonical correlation coefficients, i.e., $\lambda_k > 0$. We thus write $K_0 = \max\{1 \leq k \leq K : \lambda_k > 0\}$, which is referred to a structure dimension. Then, the next theorem gives a useful upper bound for K_0.

Theorem 1. *Under model (1), assume $X_i, \mathcal{E}_i$ are independent normal, we should have $K_0 = rank(\Sigma_{xy}) \leq \min\{|\mathcal{M}_T|, |\mathcal{N}_T|\}$, where $rank(\Sigma_{xy})$ stands for the rank of Σ_{xy}.*

Its proof is given in Appendix A. By Theorem 1 we know that the number of relevant canonical pairs is bounded by the sizes of $\mathcal{M}_T$ and $\mathcal{N}_T$. For real practice, very often we find both p and q are large. As a consequence, the total number of canonical pairs delivered by a standard statistical software (e.g., SAS) should be large, i.e., $K = \min\{p, q\}$. Nevertheless, by Theorem 1, we know that only the first few leading pairs are truly relevant.

We next consider how the model structure, as specified in (1), will affect the sparseness patterns on the canonical loadings, i.e., μ_k and ν_k for every $1 \leq k \leq K_0$.

Theorem 2. *Define $\widetilde{\mu}_j = (\mu_{1j}, \cdots, \mu_{K_0 j})^\top \in \mathbb{R}^{K_0}$ and $\widetilde{\nu}_j = (\nu_{1j}, \cdots, \nu_{K_0 j})^\top \in \mathbb{R}^{K_0}$. Under model (1), assume $X_i, \mathcal{E}_i$ are independent normal. We then should have: (1) $\|\widetilde{\mu}_j\| = 0$ for every $j \notin \mathcal{M}_T$ and $\|\widetilde{\nu}_j\| = 0$ for every $j \notin \mathcal{N}_T$; (2) $\|\widetilde{\mu}_j\| > 0$ for every $j \in \mathcal{M}_T$ and $\|\widetilde{\nu}_j\| > 0$ for every $j \in \mathcal{N}_T$.*

The proof of Theorem 2 can be found in Appendix B. By Theorem 2, we find that whether a variable is relevant is fully determined by its loadings on the canonical coefficients. Because canonical loadings are estimable, thus identifying relevant variables (particularly responses) becomes practically feasible. Subsequently, we are going to develop an estimate method for (μ_k, ν_k) $(1 \leq k \leq K_0)$ to identify the sparse structure of those canonical loadings.

2.3. A Least Squares Formulation

By Theorem 2, both the PTM and RTM can be correctly identified by inferring sparseness about the canonical loadings. With a finite dataset $(X_i, Y_i), 1 \leq i \leq n$, which has been centered, define $\hat{\Sigma}_{xx} = n^{-1} \sum X_i X_i^\top$, $\hat{\Sigma}_{xy} = n^{-1} \sum X_i Y_i^\top$, and $\hat{\Sigma}_{yy} = n^{-1} \sum Y_i Y_i^\top$. Let $\hat{T} = \hat{\Sigma}_{xx}^{-1/2} \hat{\Sigma}_{xy} \hat{\Sigma}_{yy}^{-1/2}$, and its singular value decomposition (SVD) is $\hat{T} = \hat{U}\hat{D}\hat{V}^\top$, where $\hat{U} = (\hat{u}_1, \cdots, \hat{u}_K) \in \mathbb{R}^{p \times K}$ and $\hat{V} = (\hat{v}_1, \cdots, \hat{v}_K) \in \mathbb{R}^{q \times K}$ are matrices with orthogonal columns, and $\hat{D} = \text{diag}\{\hat{\lambda}_1, \cdots, \hat{\lambda}_K\}$ is a diagonal matrix, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_K$ are singular values of $\hat{T}$. Then $\hat{\mu}_k = \hat{\Sigma}_{xx}^{-1/2} \hat{u}_k$ and $\hat{\nu}_k = \hat{\Sigma}_{yy}^{-1/2} \hat{v}_k$ are corresponding sample canonical loadings, which are estimators for μ_k and ν_k respectively. One can also see [1, 11] for more details about sample canonical loadings. However, sample canonical loadings are not sparse in general. Thus, it is of interest to develop a shrinkage method, so that the sparse structure of the canonical loadings can be consistently identified with finite data.

To this end, a least squares type objective function is needed. In particular, one can verify that

$$\mu_k \propto \mu_k^* = \arg\min_{\mu} E(\mu^\top X - \nu_k^\top Y)^2, \tag{2}$$

$$\nu_k \propto \nu_k^* = \arg\min_{\nu} E(\mu_k^\top X - \nu^\top Y)^2. \tag{3}$$

By (2) and (3), we know that the sparse structure of (μ_k, ν_k) is the same as that of (μ_k^*, ν_k^*). With a finite dataset, one can estimate μ_k^* and ν_k^* by $\hat{\mu}_k^* = \arg\min_\mu \sum_{i=1}^n (X_i^\top \mu - Y_i^\top \hat{\nu}_k)^2$ and $\hat{\nu}_k^* = \arg\min_\nu \sum_{i=1}^n (X_i^\top \hat{\mu}_k - Y_i^\top \nu)^2$, here $\hat{\mu}_k$ and $\hat{\nu}_k$ are the corresponding sample canonical loadings. To identify the sparse solutions [22, 21] in (μ_k^*, ν_k^*), we propose the following two penalized least squares functions

$$Q_\lambda^a(\mu) = \sum_{i=1}^n \left(X_i^\top \mu - Y_i^\top \hat{\nu}_k \right)^2 + \lambda \sum_{j=1}^p |\mu_j| / |\hat{\mu}_{kj}|, \tag{4}$$

$$Q_\tau^b(\nu) = \sum_{i=1}^n \left(X_i^\top \hat{\mu}_k - Y_i^\top \nu \right)^2 + \tau \sum_{j=1}^q |\nu_j| / |\hat{\nu}_{kj}|. \tag{5}$$

Here, $\mu_j (1 \leq j \leq p)$ and $\nu_j (1 \leq j \leq q)$ stand for the jth components of μ and ν, respectively. Subsequently, shrinkage estimators for μ_k^* and ν_k^* are given by $\hat{\mu}_{\lambda,k}^* = \text{argmin}_\mu Q_\lambda^a(\mu)$ and $\hat{\nu}_{\tau,k}^* = \text{argmin}_\nu Q_\tau^b(\nu)$, respectively. Furthermore, the shrinkage estimators for canonical loadings μ_k and ν_k can be achieved by $\hat{\mu}_{\lambda,k} = \hat{\mu}^*_{\lambda,k}/\sqrt{\hat{\mu}^{*\top}_{\lambda,k}\hat{\Sigma}_{xx}\hat{\mu}^*_{\lambda,k}}$ and $\hat{\nu}_{\tau,k} = \hat{\nu}^*_{\tau,k}/\sqrt{\hat{\nu}^{*\top}_{\tau,k}\hat{\Sigma}_{yy}\hat{\nu}^*_{\tau,k}}$, respectively. As one can see, both (4) and (5) are very standard Adaptive LASSO-type objective functions. Thus, the well developed LARS algorithm [6] can be directly used to obtain their piece-wise linear solution paths. Subsequently, the best model can be selected out from the transitional points by minimizing the following BIC criteria

$$\mathrm{BIC}^a_{\lambda,k} = \log\left\{n^{-1}\sum_{i=1}^n\left(X_i^\top\hat{\mu}^*_{\lambda,k} - Y_i^\top\hat{\nu}_k\right)^2\right\} + \hat{df}^{\,a}_{\lambda,k}\times\frac{\log n}{n} \quad \text{for} \quad \hat{\mu}^*_{\lambda,k}, \tag{6}$$

$$\mathrm{BIC}^b_{\tau,k} = \log\left\{n^{-1}\sum_{i=1}^n\left(X_i^\top\hat{\mu}_k - Y_i^\top\hat{\nu}^*_{\tau,k}\right)^2\right\} + \hat{df}^{\,b}_{\tau,k}\times\frac{\log n}{n} \quad \text{for} \quad \hat{\nu}^*_{\tau,k}. \tag{7}$$

Here $\hat{df}^{\,a}_{\lambda,k}$ and $\hat{df}^{\,b}_{\tau,k}$ stand for the numbers of nonzero loadings in $\hat{\mu}^*_{\lambda,k}$ and $\hat{\nu}^*_{\tau,k}$ respectively, which are simple estimators for the degrees of freedom. [26] showed that the number of nonzero parameter estimation is an unbiased estimator for the degrees of freedom of the lasso. The resulting optimal tuning parameters are $\hat{\lambda}^a_k = \arg\min_\lambda \mathrm{BIC}^a_{\lambda,k}$, and $\hat{\tau}^b_k = \arg\min_\tau \mathrm{BIC}^b_{\tau,k}$. Further, the final shrinkage estimators are $\hat{\mu}^*_{\hat{\lambda}^a_k,k}$ and $\hat{\nu}^*_{\hat{\tau}^b_k,k}$. With a slight abuse of the notation, we still use $\hat{\mu}^*_k$ and $\hat{\nu}^*_k$ to represent them for convenience. Then, the models identified by those shrinkage estimators are given by $\widehat{\mathcal{M}} = \{1 \le j \le p : \hat{\mu}^*_{kj} \ne 0 \text{ for some } 1 \le k \le K_0\}$ and $\widehat{\mathcal{N}} = \{1 \le j \le q : \hat{\nu}^*_{kj} \ne 0 \text{ for some } 1 \le k \le K_0\}$. In the next subsection, one can see that $P(\widehat{\mathcal{M}} = \mathcal{M}_T) \to 1$ and $P(\widehat{\mathcal{N}} = \mathcal{N}_T) \to 1$. This means the proposed estimators are selection consistent for both PTM and RTM.

2.4. Consistency Property

Some asymptotic properties are established in this subsection. Selection consistency of our method is also included. We here assume variable dimensions p and q are both fixed, and the situation that p, q grow together with sample size will be considered in next subsection. we make the following two technical assumptions in this subsection.

(a1) $\sqrt{n}(\hat{\Sigma}_{xx(k,l)} - \Sigma_{xx(k,l)})$, $\sqrt{n}(\hat{\Sigma}_{xy(k,l)} - \Sigma_{xy(k,l)})$, and $\sqrt{n}(\hat{\Sigma}_{yy(k,l)} - \Sigma_{yy(k,l)})$ are all asymptotically normal distributed for every k, l. Here for an arbitrary matrix Ω, $\Omega_{(k,l)}$ denotes the element corresponding to the kth row and the lth column of Ω.

(a2) For every $k \leq K_0$, $\sqrt{n}(\hat{\mu}_k - \mu_k)$ and $\sqrt{n}(\hat{\nu}_k - \nu_k)$ are both asymptotically normal distributed.

When X_i and Y_i are normal distributed, the moments $E(X_{ij}^2 X_{ij'}^2)$, $E(X_{ij}^2 Y_{ij'}^2)$ and $E(Y_{ij}^2 Y_{ij'}^2)$ for every j and j' are all bounded, hence by the central limit theorem(CLT) one can see that the claim of the assumption (a1) is true. Further, if the nonzero population canonical correlations are all distinct (i.e. $\lambda_1 > \cdots > \lambda_{K_0}$), the canonical loadings $\mu_1, \cdots, \mu_{K_0}, \nu_1, \cdots, \nu_{K_0}$ are uniquely determined except for multiplication by -1. To eliminate this indeterminacy we can require that for every $k \leq K_0$, the first nonzero element of μ_k is positive and denote it by μ_{kj_k}. And for sample canonical loadings $\hat{\mu}_k$, we also require that $\hat{\mu}_{kj_k}$ is positive. One can refer to [2, 5, 10] for the asymptotic properties of sample canonical loadings. [2] showed that for every $k \leq K_0$, $\sqrt{n}(\hat{\mu}_k - \mu_k)$ and $\sqrt{n}(\hat{\nu}_k - \nu_k)$ are asymptotically normal distributed and the asymptotic covariances are also obtained in [2]. Hence the assumption (a2) is reasonable. The two assumptions (a1) and (a2) will be used in proof details. For every $1 \leq k \leq K_0$, let $\mathcal{A}_k = \{j : \mu_{kj}^* \neq 0\}$ and $\widehat{\mathcal{A}}_{\lambda,k} = \{j : \hat{\mu}_{\lambda,kj}^* \neq 0\}$. Then we have the following theorem.

Theorem 3. *If $\lambda \to \infty$, and $n^{-1/2}\lambda \to 0$, then for every $k \leq k_0$, we have (1) $\hat{\mu}_{\lambda,k}^* - \mu_k^* = O_p(n^{-1/2})$, and (2) $P(\widehat{\mathcal{A}}_{\lambda,k} = \mathcal{A}_k) \to 1$.*

The proof of Theorem 3 is given in Appendix C. The similar technical proof has appeared in much literature, one can also refer to [7, 22, 21], and so on. By Theorem 3, we know that as long as λ satisfies the conditions in Theorem 3, the resulting estimator $\hat{\mu}_{\lambda,k}^*$ is $\sqrt{n}-$consistent, and all nonzero loadings can be identified consistently. Next we will show the selection consistency of BIC criteria. Let $\widehat{\mathcal{A}}_k = \{j : \hat{\mu}_{kj}^* \neq 0\}$, then we have the following theorem.

Theorem 4. *For every $k \leq K_0$, we have $P(\widehat{\mathcal{A}}_k = \mathcal{A}_k) \to 1$.*

We put the proof of Theorem 4 in Appendix D. One can also refer to [13] for similar proof idea. This theorem implies that all nonzero loadings can be identified consistently. Obviously, it is a direct corollary of Theorem 4 that $P(\widehat{\mathcal{M}} = \mathcal{M}_T) \to 1$. This achieves the selection consistency for predictor model. For response model, the selection consistency (i.e., $P(\widehat{\mathcal{N}} = \mathcal{N}_T) \to 1$) is also true. The proof details are very similar to those for predictor model, hence we omit it here.

2.5. Asymptotic properties with growing p and q

In this subsection, we study the asymptotic properties of our method under the situation that p and q, the dimensions of X_i and Y_i respectively, grow with the sample size n, but the structure dimension K_0 is fixed. Define $\widehat{\mathcal{M}}_\lambda = \{1 \le j \le p : \hat{\mu}^*_{\lambda,kj} \ne 0 \text{ for some } 1 \le k \le K_0\}$ and $\widehat{\mathcal{N}}_\tau = \{1 \le j \le q : \hat{\nu}^*_{\tau,kj} \ne 0 \text{ for some } 1 \le k \le K_0\}$. In our theoretical analysis, we will show that for every $k \le K_0$, $P(\widehat{\mathcal{A}}_{\lambda,k} = \mathcal{A}_k) \to 1$, which implies that $P(\widehat{\mathcal{M}}_\lambda = \mathcal{M}_T) \to 1$. The following regularity conditions are assumed for theoretical analysis throughout this subsection.

(A1) $p \to \infty, q \to \infty, n^{-1}(p+q)^{2(1+\kappa)} \to 0$, where κ is a positive constant;

(A2) Assume that X_i and Y_i are normal distributed, and $\Sigma_{xx} = I_p$, $\Sigma_{yy} = I_q$;

(A3) Denote the minimum and maximum eigenvalues of a positive definite matrix M by $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$, respectively. Then assume that there exist positive constants b and B, satisfying that $b \le \lambda_{\min}(\hat{\Sigma}_{xx}) \le \lambda_{\max}(\hat{\Sigma}_{xx}) \le B$, and $b \le \lambda_{\min}(\hat{\Sigma}_{yy}) \le \lambda_{\max}(\hat{\Sigma}_{yy}) \le B$;

(A4) For population canonical correlation coefficients $\lambda_1, \cdots, \lambda_{K_0}$, assume that for every $1 \le k < K_0$, $\lambda_k - \lambda_{k+1} \ge l$, where l is a positive constant;

(A5)

$$\lim_{n\to\infty} \frac{\lambda\sqrt{p+q}}{n} = 0, \qquad \lim_{n\to\infty} \frac{\lambda}{n}\sqrt{\frac{\sqrt{n}}{(p+q)^{1+\kappa}}} = \infty;$$

(A6) For every $k \le K_0$,

$$\lim_{n\to\infty} \min\left\{\sqrt{\frac{\sqrt{n}}{(p+q)^{1+\kappa}}}, \frac{n}{\lambda\sqrt{(p+q)}}\right\}\left(\min_{j\in\mathcal{A}_k} |\mu_{kj}|\right) \to \infty.$$

Condition (A1) restricts the growing rates of p and q. Condition (A3) assumes that the sample covariance matrix has a reasonably good behavior, and similar condition is considered in [24]. In fact, under conditions (A1) and (A2), it is true with probability 1 that $\lambda_{\min}(\hat{\Sigma}_{xx}) \to 1$,$\lambda_{\max}(\hat{\Sigma}_{xx}) \to 1$, $\lambda_{\min}(\hat{\Sigma}_{yy}) \to 1$,$\lambda_{\max}(\hat{\Sigma}_{yy}) \to 1$[3]. Hence, with probability 1 it is true that $b \le \lambda_{\min}(\hat{\Sigma}_{xx}) \le \lambda_{\max}(\hat{\Sigma}_{xx}) \le B$, and $b \le \lambda_{\min}(\hat{\Sigma}_{yy}) \le \lambda_{\max}(\hat{\Sigma}_{yy}) \le B$ for some positive constants b, B. But we still make assumption (A3) here for the purpose of simplified proof. Condition (A4) guarantees that the canonical loadings $\mu_1, \cdots, \mu_{K_0}$, $\nu_1, \cdots, \nu_{K_0}$ are uniquely determined except for multiplication by -1. Condition (A6) is similar to condition (A6) in [24], which allows the nonzero elements in μ_k to vanish but the rate is restricted.

Theorem 5. *Under conditions (A1)–(A6),we have that for every* $1 \leq k \leq K_0$, $P(\widehat{\mathcal{A}}_{\lambda,k} = \mathcal{A}_k) \to 1$.

The proof of Theorem 5 can be found in Appendix E. By Theorem 5, one can see that as long as the assumed conditions (A1)–(A6) are true, the nonzero loadings of μ_k^* can be identified consistently (i.e. $P(\widehat{\mathcal{A}}_{\lambda,k} = \mathcal{A}_k) \to 1$). Because the structure dimension K_0 is fixed, one can easily obtain that it is a direct corollary of Theorem 5 that $P(\widehat{\mathcal{M}}_\lambda = \mathcal{M}_T) \to 1$. For the response model, if similar conditions are also assumed, it is also true that $P(\widehat{\mathcal{N}}_\tau = \mathcal{N}_T) \to 1$, and we omit its proof details here because of similar proof process with predictor model. This demonstrates that the selection consistency of our ASCCA method is still valid when variable dimensions p and q grow with the sample size n.

2.6. Structure Dimension Determination

For a practical implementation, it is important to get an accurate estimator for the structure dimension K_0. Otherwise, computing shrinkage estimators for every (μ_k, ν_k) with $1 \leq k \leq K$ is computationally expensive and also statistically inefficient. In fact, as we mentioned earlier, usually only a very few number of the canonical pairs are truly relevant, which is determined by the structure dimension K_0. To practically decide its value, we follow the idea of [14] and propose the following maximum eigenvalue ratio criterion (MERC). More specifically, let $\hat{U}\hat{D}\hat{V}^\top = \hat{\Sigma}_{xy}$ be the singular value decomposition of the sample covariance matrix $\hat{\Sigma}_{xy} = n^{-1}\sum_{i=1}^n X_i Y_i^\top$. Here $\hat{D} = \mathrm{diag}(\hat{\lambda}_1, \cdots, \hat{\lambda}_K)$ is a diagonal matrix with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_K \geq 0$. Intuitively, for every $j \leq K_0$, the value of $\hat{\lambda}_j$ is expected to converge to its population value $\lambda_j > 0$. As a consequence, their ratio $\hat{r}_j = \hat{\lambda}_j/\hat{\lambda}_{j+1} = O_p(1)$ for every $j < K_0$. On the other hand, for every $j > K_0$, we should have $\hat{\lambda}_j \to_p 0$ but under a comparable speed [5]. Consequently, we should have $\hat{r}_j = O_p(1)$ also for every $j > K_0$. Nevertheless, if $j = K_0$, we should have $\hat{\lambda}_j \to_p \lambda_{K_0} > 0$ but $\hat{\lambda}_{j+1} \to_p 0$. Consequently, we should have $\hat{r}_j \to_p \infty$ for $j = K_0$. Such an interesting observation suggests that we can estimate K_0 by $\widehat{K}_0 = \mathrm{argmax}_j \hat{r}_j$. See also [14] for a more detailed discussion.

3. NUMERICAL STUDIES

3.1. Simulation Study

Example 1. This is an example revised from [18]. In particular, the predictor X_i is generated from a multivariate normal distribution with mean 0 and covariance $\mathrm{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1-j_2|}$ for any $1 \le j_1, j_2 \le p$. We next generate the first response as $Y_{i1} = X_i^\top \beta_{01} + \varepsilon_{i1}$, where β_{01} is a p-dimensional vector with its 1st, 4th, and 7th components given by 3, 1.5, and 2, respectively. Other components of β_{01} are all fixed to be 0. Given X_i and Y_{i1}, other responses are generated as $Y_{ij} = Y_{i1} + \epsilon_{ij}$ for every $1 < j \le q$, where ϵ_{ij} $(1 \le j \le q)$ are independent noises from the standard normal distribution. As one can see, for this case there is a total of 3 relevant predictors as $\mathcal{M}_T = \{1, 4, 7\}$ and only one relevant response as $\mathcal{N}_T = \{1\}$. Furthermore, because we have only one relevant response here, we know that the structure dimension $K_0 = 1$. Hence only the first pair of canonical variables is considered in this example.

Example 2. For this example, X_i is generated in the same way as the previous example. The first two responses are generated as $Y_{ij} = X_i^\top \beta_{0j} + \varepsilon_{ij}, j = 1, 2$, where β_{01} takes the same value with β_{01} in previous example. β_{02} is another p-dimensional vector with its first three positions fixed to be 5 and others to be 0. Given X_i, Y_{i1}, and Y_{i2}, other responses are generated according to $Y_{ij} = Y_{i1} + Y_{i2} + \varepsilon_{ij}$ for every $2 < j \le q$. Once again, ϵ_{ij} $(1 \le j \le q)$ are independent noises from the standard normal distribution. For this case there is a total of 5 relevant predictors as $\mathcal{M}_T = \{1, 2, 3, 4, 7\}$, and 2 relevant responses as $\mathcal{N}_T = \{1, 2\}$. Moreover, from the detailed proof of Theorem 1, we know that the structure dimension $K_0 = \mathrm{rank}\big((\beta_{01}, \beta_{02})\big) = 2$. Hence we only consider the first two pairs of canonical variables in this example.

Example 3. This example is designed to see the performance of ASCCA method when it works on nonnormal distributed data. In particular, we firstly draw $\pi_j (1 \le j \le p)$ independently from uniform distribution $U(0, 1)$, then predictors $X_{ij} (1 \le j \le p)$ are drawn independently from bernoulli distributions $B(1, \pi_j)$. The first two responses are generated from the following distribution $Y_{ij} \sim B\big(1, e^{X_i^\top \beta_{0j}}/(1+e^{X_i^\top \beta_{0j}})\big), j = 1, 2$, where β_{01} and β_{02} are fixed the same as example 2. Given X_i, Y_{i1}, and Y_{i2}, the rest responses are generated independently from bernoulli distribution $B\big(1, e^{(Y_{i1}+Y_{i2})}/(1+e^{(Y_{i1}+Y_{i2})})\big)$. As one can see, there are 5 relevant predictors as $\mathcal{M} = \{1, 2, 3, 4, 7\}$, and 2 relevant responses as $\mathcal{N} = \{1, 2\}$. In this example, we do not know the exact value of the structure dimension K_0, but we think it is enough to only consider the first two pairs of canonical variables, because there are only two relevant responses.

Every example is randomly replicated for a total of 500 times with various parameter specifications for n, p and q. For each simulated dataset, the method of ASCCA is used to identify a set of relevant predictors and responses, which are denoted as $\widehat{\mathcal{M}}$ and $\widehat{\mathcal{N}}$, respectively. Then the percentages of the experiments with correctly fitted PTM (i.e., $\widehat{\mathcal{M}} = \mathcal{M}_T$), and RTM (i.e., $\widehat{\mathcal{N}} = \mathcal{N}_T$) are computed and summarized in Table 1. As one can see, for an arbitrary p and q specification, large sample size n always leads to better performances. In fact, as long as the sample size is large enough, we should find the correct fit percentages would be very close to 100%. This numerically confirms that our ASCCA methods is selection consistent. For comparison, the lasso type sparse method which is called SCCA here is also considered. From the simulation results, one can see that SCCA method performs worse than ASCCA method, and does not have selection consistency. One can also see that it seems more difficulty to identify the true response model $\mathcal{N}_T$. From the results of example 3, we find that the ASCCA method can also be used in some nonnormal data analysis, and performs better than SCCA method. Essentially, the response true model $\mathcal{N}_T$ is defined through conditional independence, and independence is equivalent with irrelevance in the normal case. Consequently, the proposed ASCCA method can correctly identify the true model through studying the relevance between variables. In the nonnormal case, if independence and irrelevance are also equivalent with each other, then the ASCCA method can also perform well. This is corroborated by example 3.

Example 4. We use this example to test the performance of our method when variable dimensions p and q grow with the sample size n. In particular, X_i and Y_i are generated in the same way as example 2, but $p = [10n^{1/3}]$ and $q = [p/2]$, which both increase with sample size. Hence we know here there are 5 relevant predict variables as $\mathcal{M} = \{1, 2, 3, 4, 7\}$, and 2 relevant response variables as $\mathcal{N} = \{1, 2\}$. We only need to consider the first two pairs of canonical variables. In this example, we still use BIC criterion to select the tuning parameters. The example is also randomly replicated for a total of 500 times with various sample size $n = 100, 200, 400, 800$. We use $(I_{\mathcal{M}_T}, II_{\mathcal{N}_T}, I_{\mathcal{N}_T}, II_{\mathcal{N}_T})$ to measure the model selection performance, where

Table 1: Detailed simulation results of Examples 1-3 based on 500 simulation iterations

			Percentage of the Correct Fit			
			ASCCA		SCCA	
	$(p,\ q)$	n	$\mathcal{M}_T$	$\mathcal{N}_T$	$\mathcal{M}_T$	$\mathcal{N}_T$
Ex.1	(8,3)	100	0.866	0.922	0.254	0.262
		200	0.916	0.960	0.262	0.286
		500	0.952	0.984	0.368	0.266
	(20,10)	100	0.658	0.568	0.108	0.002
		200	0.836	0.816	0.238	0.006
		500	0.924	0.942	0.318	0.008
	(50,20)	100	0.310	0.014	0.070	0.000
		200	0.692	0.450	0.208	0.000
		500	0.858	0.838	0.284	0.000
Ex.2	(8,3)	100	0.698	0.732	0.084	0.456
		200	0.812	0.844	0.102	0.480
		500	0.920	0.962	0.128	0.508
	(20,10)	100	0.312	0.142	0.010	0.016
		200	0.548	0.324	0.018	0.000
		500	0.764	0.646	0.050	0.000
	(50,20)	100	0.052	0.000	0.000	0.000
		200	0.342	0.082	0.006	0.000
		500	0.658	0.362	0.030	0.000
Ex.3	(8,3)	400	0.346	0.866	0.306	0.606
		600	0.560	0.914	0.396	0.654
		800	0.680	0.928	0.410	0.670
		1000	0.746	0.952	0.372	0.704
	(20,10)	400	0.408	0.554	0.160	0.296
		600	0.454	0.724	0.182	0.350
		800	0.590	0.768	0.196	0.434
		1000	0.632	0.852	0.212	0.482
	(50,20)	400	0.030	0.146	0.028	0.084
		600	0.132	0.352	0.082	0.212
		800	0.280	0.530	0.196	0.324
		1000	0.404	0.622	0.252	0.374

Table 2: Detailed simulation results of Example 4 based on 500 simulation iterations

			ASCCA				SCCA			
n	p	q	$I_{\mathcal{M}_T}$	$II_{\mathcal{M}_T}$	$I_{\mathcal{N}_T}$	$II_{\mathcal{N}_T}$	$I_{\mathcal{M}_T}$	$II_{\mathcal{M}_T}$	$I_{\mathcal{N}_T}$	$II_{\mathcal{N}_T}$
100	46	23	5	15.64	1.89	15.22	5	19.94	1.93	17.03
200	58	29	5	2.90	1.97	5.53	5	8.37	1.99	7.53
400	73	36	5	1.12	1.99	2.87	5	5.66	2.00	8.76
800	92	46	5	0.45	2.00	1.87	5	4.19	2.00	11.68

$I_{\mathcal{M}_T}$ denotes the average number of relevant predictors correctly identified, and $II_{\mathcal{M}_T}$ is the average number of irrelevant predictors but incorrectly identified as relevant variables. $I_{\mathcal{N}_T}$ and $II_{\mathcal{N}_T}$ are also similarly defined for response variables. The simulation results are summarized in Table 2. From Table 2, one can see that the ASCCA method performs significantly better than SCCA method. Both ASCCA and SCCA can identify nearly all relevant variables, but SCCA method incorrectly identifies more irrelevant variables as relevant variables than ASCCA method.

3.2. A Real Example

To further demonstrate its practical usefulness, we apply our method to a teaching evaluation dataset. This dataset contains a total of 341 records. Each of them corresponds to one particular course instructed at Peking University. The responses of interest are various evaluation scores given by students. Specifically, it measures the students average agreements towards the following nine statements: (Q1) I think this is a good course, (Q2) the course improves my knowledge, (Q3) the schedule is reasonable, (Q4) the course is difficult, (Q5) the course pace is too fast, (Q6) the course load is very heavy, (Q7) the text book is good, (Q8) the reference book is helpful, and (Q9) open this course is necessary.

To explain those responses, the following explanatory variables are also collected. They include (P1) whether the students are undergraduate (1=yes, 0=no) and (P2) whether the students are graduate (1=yes, 0=no). If the students are neither undergraduate nor graduate, then they must be MBA students. We also considered (P3) whether the instructor is an associate professor (1=yes, 0=no) and (P4) whether the instructor is a full professor (1=yes, 0=no). If an instructor is neither an associate nor a full professor, he/she must be an assistant professor. Lastly, the instructor's gender is also recorded (1=male and 0=female).

We firstly apply the MERC method to estimate the structure dimension, which gives $\widehat{K}_0 = 1$. Subsequently, the method of ASCCA is used. We find that P2 (i.e., whether the student is graduate) is the only relevant predictor with positive coefficient estimates. Such a result implies that teaching a graduate course is more likely to generate higher teaching evaluation scores, as compared with undergraduate and MBA courses. In contrast, no significant difference is observed between undergraduate and MBA courses. Such a result matches our teaching experiences very well. In addition to P2, ASCCA also identifies 4 relevant responses, which are given by Q3, Q4, Q5, and Q8. Such a result implies that a student's satisfaction towards on a course is mainly determined by this course's schedule (Q3), difficulty (Q4), pace (Q5), and also reference book (Q8). Once those 4 responses are determined, all other responses, particularly a student's overall satisfaction (i.e., Q1), is also determined on average. Such an insightful finding suggests that various evaluation scores of a course can be simultaneously improved if the students' satisfaction towards Q3, Q4, Q5, and Q8 can be enhanced.

4. CONCLUSIONS

In this paper, we introduce a variable selection method for multivariate regression. This method can identify relevant variables consistently not only for predictors but also for multivariate responses. Both theory results and numerical studies show that our method behaves well. However, much future work is still need to be done. Especially when the variable dimension is much higher than sample size, how to identify important variables consistently is a more interesting issue. To the best of our knowledge, there are less studies about the theoretical property of SCCA when variable dimension is much higher than sample size, which can be our future work to study.

ACKNOWLEDGEMENTS

We are very grateful to the editor, an associate editor, and two anonymous referees for their helpful and constructive comments, which led to a much improved manuscript. Research by Baiguo An and Jianhua Guo was supported in part by the National Natural Science Foundation of China (No. 11025102, No. 11226216), the Program Innovative Research Team (PCSIRT) in University (# IRT0519) and a Jilin Project (No. 20100401). Research by Hansheng Wang was supported in part by the Fox Ying Tong Education Foundation, the National Natural Science Foundation of China (No. 11131002), Fundamental Research Funds for the Central Universities, Research Funds of Renmin University of China, and the Center for Statistical Science at Peking University.

APPENDIX

Before presenting the detailed proof, we first introduce some notations. Let $A = (a_{ij}) \in \mathbb{R}^{d_1 \times d_2}$ be an arbitrary matrix, then $I_1 \subset \{1, \cdots, d_1\}$ and $I_2 \subset \{1, \cdots, d_2\}$ are two arbitrary index sets. We then define $A_{(I_1,I_2)} = (a_{ij} : i \in I_1, j \in I_2) \in \mathbb{R}^{|I_1| \times |I_2|}$ to be the submatrix according to I_1 and I_2. Next define $A_{[I_1]} = (a_{ij} : i \in I_1, 1 \leq j \leq d_2)$ to be a matrix collecting A's row vectors according to I_1, and $A_{<I_2>} = (a_{ij} : 1 \leq i \leq d_1, j \in I_2)$ to be a matrix collecting A's column vectors according to I_2.

Appendix A. Proof of Theorem 1

By the definition of K_0, one can see that $K_0 = \mathrm{rank}(\Sigma_{xy})$. By model (1), we have $\Sigma_{xy} = \Sigma_{xx} B_0$, thus $K_0 = \mathrm{rank}(B_0)$. Because $B_{0[\mathcal{M}_T^c]} = 0$, we must have $K_0 = \mathrm{rank}(B_{0[\mathcal{M}_T]}) \leq |\mathcal{M}_T|$. We next prove $K_0 \leq |\mathcal{N}_T|$. By normality assumption and the definition of RTM $\mathcal{N}_T$, we know that $0 = \mathrm{cov}(X_i, Y_{i(\mathcal{N}_T^c)} \mid Y_{i(\mathcal{N}_T)}) = \Sigma_{xx}(B_{0<\mathcal{N}_T^c>} - B_{0<\mathcal{N}_T>} \Sigma^{-1}_{yy(\mathcal{N}_T,\mathcal{N}_T)} \Sigma_{yy(\mathcal{N}_T,\mathcal{N}_T^c)})$. This implies that $B_{0<\mathcal{N}_T^c>} = B_{0<\mathcal{N}_T>} \Sigma^{-1}_{yy(\mathcal{N}_T,\mathcal{N}_T)} \Sigma_{yy(\mathcal{N}_T,\mathcal{N}_T^c)}$. Consequently, every column vector of $B_{0<\mathcal{N}_T^c>}$ is a linear combination of column vectors of $B_{0<\mathcal{N}_T>}$. As a result, we should also have $K_0 = \mathrm{rank}(B_{0<\mathcal{N}_T>}) \leq |\mathcal{N}_T|$. This completes the proof.

Appendix B. Proof of Theorem 2

Let $T = \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} = \Sigma_{xx}^{1/2} B_0 \Sigma_{yy}^{-1/2}$, and its singular value decomposition is denoted by $T = UDV^\top$. Here $U \in \mathbb{R}^{p \times K_0}$ and $V \in \mathbb{R}^{q \times K_0}$ are two matrices with orthogonal column vectors. $D = \mathrm{diag}\{\lambda_1, \cdots, \lambda_{K_0}\} \in \mathbb{R}^{K_0 \times K_0}$ is a diagonal matrix, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{K_0} > 0$ are the non-zero singular values of T. By [11], we know that the solutions of canonical loadings are given by $\mathbb{U} = (\mu_1, \cdots, \mu_{K_0}) = \Sigma_{xx}^{-1/2} U$ and $\mathbb{V} = (\nu_1, \cdots, \nu_{K_0}) = \Sigma_{yy}^{-1/2} V$. Moreover, $\lambda_1, \cdots, \lambda_{K_0}$ are the corresponding non-zero canonical correlations. Subsequently, the claim of Theorem 2(1) can be proved in two steps.

`The 1st step:` In this step, we will prove that for every $j \notin \mathcal{M}_T$, $\|\tilde{\mu}_j\| = 0$. This is equivalent to $\mathbb{U}_{[\mathcal{M}_T^c]} = 0$. Because $\Sigma_{xx}^{-1/2} T = B_0 \Sigma_{yy}^{-1/2}$, we have $\Sigma_{xx[\mathcal{M}_T^c]}^{-1/2} T = B_{0[\mathcal{M}_T^c]} \Sigma_{yy}^{-1/2}$. Furthermore, note that $B_{0[\mathcal{M}_T^c]} = 0$, which implies $\Sigma_{xx[\mathcal{M}_T^c]}^{-1/2} T = 0$. Consequently, the row vectors of $\Sigma_{xx[\mathcal{M}_T^c]}^{-1/2}$ are orthogonal to the column vectors of T. As a result, they should also be orthogonal to $\mathrm{span}(T)$, which is the linear subspace spanned by the column vectors of T. Furthermore, note that $\mathrm{span}(T) = \mathrm{span}(U)$. As a result, the row vectors of $\Sigma_{xx[\mathcal{M}_T^c]}^{-1/2}$ are orthogonal to the column vectors of U, and thus $\mathbb{U}_{[\mathcal{M}_T^c]} = \Sigma_{xx[\mathcal{M}_T^c]}^{-1/2} U = 0$.

`The 2nd step:` In this step, we will prove that for every $j \notin \mathcal{N}_T$, $\|\tilde{\nu}_j\| = 0$. To this end, it suffices to show that $\mathbb{V}_{[\mathcal{N}_T^c]} = 0$. By normality assumption and the definition of RTM $\mathcal{N}_T$, we know that $0 = \mathrm{cov}(X_i, Y_{i(\mathcal{N}_T^c)} \mid Y_{i(\mathcal{N}_T)}) = \Sigma_{xx}(B_{0<\mathcal{N}_T^c>} - B_{0<\mathcal{N}_T>} \Sigma_{yy(\mathcal{N}_T,\mathcal{N}_T)}^{-1} \Sigma_{yy(\mathcal{N}_T,\mathcal{N}_T^c)})$. Denote Σ_{yy}^{-1} by Ω. One can verify that $\Omega_{(\mathcal{N}_T^c,\mathcal{N}_T)} = -\Omega_{(\mathcal{N}_T^c,\mathcal{N}_T^c)} \Sigma_{yy(\mathcal{N}_T^c,\mathcal{N}_T)} \Sigma_{yy(\mathcal{N}_T,\mathcal{N}_T)}^{-1}$. In the meanwhile, $\Sigma_{yy}^{-1/2} T^\top = \Sigma_{yy}^{-1} B_0^\top \Sigma_{xx}^{1/2}$. Consequently, we have $\Sigma_{yy[\mathcal{N}_T^c]}^{-1/2} T^\top = \Omega_{[\mathcal{N}_T^c]} B_0^\top \Sigma_{xx}^{1/2}$. Because

$$
\begin{aligned}
\Omega_{[\mathcal{N}_T^c]} B_0^\top \Sigma_{xx}^{1/2} &= \Omega_{(\mathcal{N}_T^c,\mathcal{N}_T^c)} (I, -\Sigma_{yy(\mathcal{N}_T^c,\mathcal{N}_T)} \Sigma_{yy(\mathcal{N}_T,\mathcal{N}_T)}^{-1}) (B_{0<\mathcal{N}_T^c>}, B_{0<\mathcal{N}_T>})^\top \Sigma_{xx}^{1/2} \\
&= \Omega_{(\mathcal{N}_T^c,\mathcal{N}_T^c)} \mathrm{cov}(X_i, Y_{i(\mathcal{N}_T^c)} \mid Y_{i(\mathcal{N}_T)})^\top \Sigma_{xx}^{-1/2} = 0,
\end{aligned}
$$

we have $\Sigma_{yy[\mathcal{N}_T^c]}^{-1/2} T^\top = 0$. As a consequence, the row vectors of $\Sigma_{yy[\mathcal{N}_T^c]}^{-1/2}$ are orthogonal to $\mathrm{span}(T^\top)$. Furthermore, we know that $\mathrm{span}(T^\top) = \mathrm{span}(V)$. This implies that the row vectors of $\Sigma_{yy[\mathcal{N}_T^c]}^{-1/2}$ are orthogonal to the column vectors of V. This proves that $\mathbb{V}_{[\mathcal{N}_T^c]} = \Sigma_{yy[\mathcal{N}_T^c]}^{-1/2} V = 0$.

To establish the claim of Theorem 2(2), we consider what would happen if it is not correct. Specifically, we consider the following two steps.

`The 3rd step:` In this step, we will prove that for every $j \in \mathcal{M}_T$, $\|\widetilde{\mu}_j\| > 0$. If this is not true, then there must exist a $j_1 \in \mathcal{M}_T$, such that $\mu_{kj_1} = 0$ for every $1 \le k \le K_0$. This means that $\mathbb{U}_{[\{j_1\}]} = \Sigma_{xx[\{j_1\}]}^{-1/2} U = 0$. As a consequence, we have $\Sigma_{xx[\{j_1\}]}^{-1/2} T = 0$. Because $\Sigma_{xx}^{-1/2} T = B_0 \Sigma_{yy}^{-1/2}$, we have $\Sigma_{xx[\{j_1\}]}^{-1/2} T = \alpha_{0j_1}^\top \Sigma_{yy}^{-1/2} = 0$. Furthermore, $\Sigma_{yy}^{-1/2}$ is of full rank, α_{0j_1} must be equal to 0. This is contradictory to the assumption $j_1 \in \mathcal{M}_T$. Hence for every $j \in \mathcal{M}_T$, there must exist at least one $1 \le k_1 \le K$, such that $\mu_{k_1 j} \ne 0$, that is $\|\widetilde{\mu}_j\| > 0$.

`The 4th step:` In this step, we will prove that for every $j \in \mathcal{N}_T$, $\|\widetilde{\nu}_j\| > 0$. If this is not true, then there must exist a $j_2 \in \mathcal{N}_T$, such that $\nu_{kj_2} = 0$ for every $1 \le k \le K_0$. This means that $\mathbb{V}_{[\{j_2\}]} = \Sigma_{yy[\{j_2\}]}^{-1/2} V = 0$. As a consequence, we have that $\Sigma_{yy[\{j_2\}]}^{-1/2} T^\top = 0$. Denote $\mathcal{N}_T \setminus \{j_2\}$ by $\mathcal{N}_{T\setminus j_2}$. We then have $\Sigma_{yy[\mathcal{N}_{T\setminus j_2}^c]}^{-1/2} T^\top = 0$. Recall that $\Omega = \Sigma_{yy}^{-1}$. Similar to the proof of Theorem 2(1), one can show that $0 = \Sigma_{yy[\mathcal{N}_{T\setminus j_2}^c]}^{-1/2} T^\top = \Omega_{(\mathcal{N}_{T\setminus j_2}^c, \mathcal{N}_{T\setminus j_2}^c)} \mathrm{cov}(X_i, Y_{i(\mathcal{N}_{T\setminus j_2}^c)} | Y_{i(\mathcal{N}_{T\setminus j_2})})^\top \Sigma_{xx}^{-1/2}$. Combining the fact that both $\Omega_{(\mathcal{N}_{T\setminus j_2}^c, \mathcal{N}_{T\setminus j_2}^c)}$ and $\Sigma_{xx}^{-1/2}$ are of full rank, we have $\mathrm{cov}(X_i, Y_{i(\mathcal{N}_{T\setminus j_2}^c)} | Y_{i(\mathcal{N}_{T\setminus j_2})}) = 0$. Under the normality assumption, this implies that given $Y_{i(\mathcal{N}_{T\setminus j_2})}$, X_i and $Y_{i(\mathcal{N}_{T\setminus j_2}^c)}$ are conditional independent. This is contradictory to the definition of RTM $\mathcal{N}_T$. Hence for every $j \in \mathcal{N}_T$, there must exist at least one $1 \le k_2 \le K_0$, such that $\nu_{k_2 j} \neq 0$, that is $\|\widetilde{\nu}_j\| > 0$.

This completes the proof of the whole theorem.

Appendix C. Proof of Theorem 3

`The 1st step:` In this step, we will prove the claim of Theorem 3(1).

Let $u = n^{1/2}(\mu - \mu_k^*)$, and $V_n(u) = Q_\lambda^a(\mu_k^* + n^{-1/2}u) - Q_\lambda^a(\mu_k^*) = u^\top \hat{\Sigma}_{xx} u + 2u^\top \big(n^{1/2}\hat{\Sigma}_{xx}\mu_k^* - n^{1/2}\hat{\Sigma}_{xy}\hat{\nu}_k\big) + \lambda \sum_{j=1}^p |\hat{\mu}_{kj}|^{-1}(|\mu_{kj}^* + n^{-1/2}u_j| - |\mu_{kj}^*|)$. Denote $\hat{u} = \arg\min_u V_n(u)$, then $\hat{u} = n^{1/2}(\hat{\mu}_{\lambda,k}^* - \mu_k^*)$. By laws of large numbers, we have that $\hat{\Sigma}_{xx} \to_P \Sigma_{xx}$. Let $W_n = n^{1/2}\hat{\Sigma}_{xx}\mu_k^* - n^{1/2}\hat{\Sigma}_{xy}\hat{\nu}_k = n^{1/2}(\hat{\Sigma}_{xx} - \Sigma_{xx})\mu_k^* - n^{1/2}(\hat{\Sigma}_{xy} - \Sigma_{xy})\nu_k - \hat{\Sigma}_{xy} n^{1/2}(\hat{\nu}_k - \nu_k)$. By assumptions in subsection 2.4, we can have that $W_n \to_d W$ for some random variable W. If $j \in \mathcal{A}_k$, it is easy to show that $\lambda |\hat{\mu}_{kj}|^{-1}(|\mu_{kj}^* + n^{-1/2}u_j| - |\mu_{kj}^*|) \to 0$, if $j \in \mathcal{A}_k^C$, $\lambda |\hat{\mu}_{kj}|^{-1}(|\mu_{kj}^* + n^{-1/2}u_j| - |\mu_{kj}^*|) = \lambda n^{-1/2}|\hat{\mu}_{kj}|^{-1}|u_j| \to \infty$, when $u_j \neq 0$. Thus we can gain that $V_n(u) \to_d V(u)$, where $V(u) = u^\top \Sigma_{xx} u + 2u^\top W$, if $u_{(\mathcal{A}_k^C)} = 0$, and $V(u) = \infty$, otherwise. Because $V_n(u)$ is convex, and $V(u)$ has the unique minimizer $(-W_{(\mathcal{A}_k)}^\top \Sigma_{xx(\mathcal{A}_k, \mathcal{A}_k)}, 0)^\top$, following [22], we have $\hat{u} \to_p \arg\min V(u)$. Consequently, $\hat{u} = O_p(1)$, which implies that $\hat{\mu}_{\lambda,k}^* - \mu_k^* = O_p(n^{-1/2})$.

`The 2nd step:` In this step, we will prove the claim of Theorem 3(2).

Due to that $\hat{\mu}_{\lambda,k}^* \to_P \mu_k^*$, for every $j \in \mathcal{A}_k$, $P(j \in \widehat{\mathcal{A}}_{\lambda,k}) \ge P(|\hat{\mu}_{\lambda,kj}^* - \mu_{kj}^*| < |\mu_{kj}^*|/2) \to 1$. Consequently, it suffices to show that for every $j \notin \mathcal{A}_k$, $P(j \in \widehat{\mathcal{A}}_{\lambda,k}) \to 0$. If $j \in \widehat{\mathcal{A}}_{\lambda,k}$, by the Karush-Kuhn-Tucker (KKT) conditions[4], we have that

$$n^{-1/2}\frac{\partial Q_\lambda^a(\mu)}{\partial \mu_j}|_{\hat{\mu}_{\lambda,k}^*} = 2n^{1/2}\big(\hat{\Sigma}_{xx[j]}\hat{\mu}_{\lambda,k}^* - \hat{\Sigma}_{xy[j]}\hat{\nu}\big) + \lambda n^{-1/2}|\hat{\mu}_{kj}|^{-1} sgn(\hat{\mu}_{\lambda,kj}^*) = 0,$$

which implies that $2n^{1/2}|\big(\hat{\Sigma}_{xx[j]}\hat{\mu}^*_{\lambda,k} - \hat{\Sigma}_{xy[j]}\hat{\nu}\big)| = \lambda n^{-1/2}|\hat{\mu}_{kj}|^{-1}$. However, it is easy to show that $2n^{1/2}|\big(\hat{\Sigma}_{xx[j]}\hat{\mu}^*_{\lambda,k} - \hat{\Sigma}_{xy[j]}\hat{\nu}\big)| = O_p(1)$, and $\lambda n^{-1/2}|\hat{\mu}_{kj}|^{-1} \to_P \infty$, hence $P(j \in \widehat{\mathcal{A}}_{\lambda,k}) \leq P(2n^{1/2}|\big(\hat{\Sigma}_{xx[j]}\hat{\mu}^*_{\lambda,k} - \hat{\Sigma}_{xy[j]}\hat{\nu}\big)| = \lambda n^{-1/2}|\hat{\mu}_{kj}|^{-1}) \to 0$. This completes the proof.

Appendix D. Proof of Theorem 4

Let $\lambda_n = \log n$, which satisfies the conditions in Theorem 3, hence it follows that $\widehat{\mathcal{A}}_{\lambda_n,k} = \mathcal{A}_k$ with probability tending to one. we next partition $\mathbb{R}^+$ into the following three mutually exclusive regions:$\mathbb{R}_- = \{\lambda \in \mathbb{R}^+, \widehat{\mathcal{A}}_{\lambda,k} \not\supset \mathcal{A}_k\}, \mathbb{R}_0 = \{\lambda \in \mathbb{R}^+, \widehat{\mathcal{A}}_{\lambda,k} = \mathcal{A}_k\}$, and $\mathbb{R}_+ = \{\lambda \in \mathbb{R}^+, \widehat{\mathcal{A}}_{\lambda,k} \supset \mathcal{A}_k, \widehat{\mathcal{A}}_{\lambda,k} \neq \mathcal{A}_k\}$. To prove theorem 4, it suffices to show that $P(\inf_{\lambda \in \mathbb{R}_- \bigcup \mathbb{R}_+} \mathrm{BIC}^a_{\lambda,k} > \mathrm{BIC}^a_{\lambda_n,k}) \to 1$.

For every $\lambda \in \mathbb{R}_-$,

$$\begin{aligned}
& e^{\mathrm{BIC}^a_{\lambda,k}} - e^{\mathrm{BIC}^a_{\lambda_n,k}} \\
= \quad & e^{\hat{df}^a_{\lambda,k}\log n/n}\left\{n^{-1}\sum_{i=1}^n\Big(X_i^\top\hat{\mu}^*_{\lambda,k} - Y_i^\top\hat{\nu}_k\Big)^2\right\} \\
& -e^{\hat{df}^a_{\lambda_n,k}\log n/n}\left\{n^{-1}\sum_{i=1}^n\Big(X_i^\top\hat{\mu}^*_{\lambda_n,k} - Y_i^\top\hat{\nu}_k\Big)^2\right\} \\
\geq \quad & e^{\hat{df}^a_{\lambda,k}\log n/n}\min_{\mathcal{A}\not\supset\mathcal{A}_k}\min_{\{\mu:\mu_j=0,\forall j\notin\mathcal{A}\}}\left\{n^{-1}\sum_{i=1}^n\Big(X_i^\top\mu - Y_i^\top\hat{\nu}_k\Big)^2\right\} \\
& -e^{\hat{df}^a_{\lambda_n,k}\log n/n}\left\{n^{-1}\sum_{i=1}^n\Big(X_i^\top\hat{\mu}^*_{\lambda_n,k} - Y_i^\top\hat{\nu}_k\Big)^2\right\} \\
\to_P \quad & \min_{\mathcal{A}\not\supset\mathcal{A}_k}\min_{\{\mu:\mu_j=0,\forall j\notin\mathcal{A}\}} E\big(X\mu - Y\nu_k\big)^2 - \min E\big(X\mu - Y\nu_k\big)^2 > 0,
\end{aligned}$$

where the last inequality is due to that $\mathcal{A} \not\supset \mathcal{A}_k$. This implies that

$$P(\inf_{\lambda\in\mathbb{R}^-}\mathrm{BIC}^a_{\lambda,k} > \mathrm{BIC}^a_{\lambda_n,k}) \to 1.$$

Next, we will consider the case that $\lambda \in \mathbb{R}_+$. For arbitrary $\mathcal{A} \supset \mathcal{A}_k$, we define $\hat{\mu}^*_{\mathcal{A},k} = \mathrm{argmin}_{\{\mu\in\mathbb{R}^p:\mu_j=0,\forall j\notin\mathcal{A}\}}\sum_{i=1}^n\Big(X_i^\top\mu - Y_i^\top\hat{\nu}_k\Big)^2$, and $\mu^*_{\mathcal{A},k} = \mathrm{argmin}_{\{\mu\in\mathbb{R}^p:\mu_j=0,\forall j\notin\mathcal{A}\}}E\big(\mu^\top X - \nu_k^\top Y\big)^2$. It is easy to show that $\hat{\mu}^*_{\mathcal{A},k} - \mu^*_{\mathcal{A},k} = O_p(n^{-1/2})$. And because that $\mathcal{A} \supset \mathcal{A}_k$, it is obvious that $\mu^*_{\mathcal{A},k} = \mu^*_k$. Hence

$\hat{\mu}^*_{\mathcal{A},k} - \mu^*_k = O_p(n^{-1/2})$. As a result,

$$\sum_{i=1}^n \big(X_i^\top \hat{\mu}^*_{\mathcal{A},k} - Y_i^\top \hat{\nu}_k\big)^2 - \sum_{i=1}^n \big(X_i^\top \hat{\mu}^*_{\lambda_n,k} - Y_i^\top \hat{\nu}_k\big)^2$$

$$= n\Big\{ \hat{\mu}^{*\top}_{\mathcal{A},k}\hat{\Sigma}_{xx}\hat{\mu}^*_{\mathcal{A},k} - \hat{\mu}^{*\top}_{\lambda_n,k}\hat{\Sigma}_{xx}\hat{\mu}^*_{\lambda_n,k} - 2\big(\hat{\mu}^*_{\mathcal{A},k} - \hat{\mu}^*_{\lambda_n,k}\big)^\top \hat{\Sigma}_{xy}\hat{\nu}_k \Big\}$$

$$= n\big(\hat{\mu}^*_{\mathcal{A},k} - \mu^*_k\big)^\top \hat{\Sigma}_{xx}\big(\hat{\mu}^*_{\mathcal{A},k} - \mu^*_k\big) - n\big(\hat{\mu}^*_{\lambda_n,k} - \mu^*_k\big)^\top \hat{\Sigma}_{xx}\big(\hat{\mu}^*_{\lambda_n,k} - \mu^*_k\big)$$

$$+2n\big(\hat{\mu}^*_{\mathcal{A},k} - \hat{\mu}^*_{\lambda_n,k}\big)^\top \big\{\hat{\Sigma}_{xx}\mu^*_k - \hat{\Sigma}_{xy}\hat{\nu}_k\big\}.$$

We know that $\hat{\mu}^*_{\mathcal{A},k}, \hat{\mu}^*_{\lambda_n,k}$ are both $\sqrt{n}-$consistent estimators for μ^*_k, hence we have that $n\big(\hat{\mu}^*_{\mathcal{A},k} - \mu^*_k\big)^\top \hat{\Sigma}_{xx}\big(\hat{\mu}^*_{\mathcal{A},k} - \mu^*_k\big)$, and $n\big(\hat{\mu}^*_{\lambda_n,k} - \mu^*_k\big)^\top \hat{\Sigma}_{xx}\big(\hat{\mu}^*_{\lambda_n,k} - \mu^*_k\big)$ are both $O_p(1)$. And also that $\sqrt{n}\big(\hat{\mu}^*_{\mathcal{A},k} - \hat{\mu}^*_{\lambda_n,k}\big) = O_p(1)$. Because $\sqrt{n}\big(\hat{\Sigma}_{xx}\mu^*_k - \hat{\Sigma}_{xy}\hat{\nu}_k\big) = \sqrt{n}(\hat{\Sigma}_{xx} - \Sigma_{xx})\mu^*_k - \sqrt{n}(\hat{\Sigma}_{xy} - \Sigma_{xy})\nu_k - \hat{\Sigma}_{xy}\sqrt{n}(\hat{\nu}_k - \nu_k)$, and $\hat{\Sigma}_{xx}, \hat{\Sigma}_{xy}, \hat{\nu}_k$ are $\sqrt{n}-$consistent estimators for $\Sigma_{xx}, \Sigma_{xy}, \nu_k$, respectively, hence we have $\sqrt{n}\big(\hat{\Sigma}_{xx}\mu^*_k - \hat{\Sigma}_{xy}\hat{\nu}_k\big) = O_p(1)$. This implies that $n\big(\hat{\mu}^*_{\mathcal{A},k} - \hat{\mu}^*_{\lambda_n,k}\big)^\top \big\{\hat{\Sigma}_{xx}\mu^*_k - \hat{\Sigma}_{xy}\hat{\nu}_k\big\} = O_p(1)$. Consequently, we gain that $\sum_{i=1}^n \Big(X_i^\top \hat{\mu}^*_{\mathcal{A},k} - Y_i^\top \hat{\nu}_k\Big)^2 - \sum_{i=1}^n \Big(X_i^\top \hat{\mu}^*_{\lambda_n,k} - Y_i^\top \hat{\nu}_k\Big)^2 = O_p(1)$.

For every $\lambda \in \mathbb{R}_+$, we have

$$n\Big(\mathrm{BIC}^a_{\lambda,k} - \mathrm{BIC}^a_{\lambda_n,k}\Big)$$

$$= n\Big\{ \log\{n^{-1}\sum_{i=1}^n \big(X_i^\top \hat{\mu}^*_{\lambda,k} - Y_i^\top \hat{\nu}_k\big)^2\} - \log\{n^{-1}\sum_{i=1}^n \big(X_i^\top \hat{\mu}^*_{\lambda_n,k} - Y_i^\top \hat{\nu}_k\big)^2\}\Big\}$$

$$+(\hat{df}^a_{\lambda,k} - \hat{df}^a_{\lambda_n,k})\log n$$

$$\geq \Big(n^{-1}\sum_{i=1}^n \big(X_i^\top \hat{\mu}^*_{\lambda_n,k} - Y_i^\top \hat{\nu}_k\big)^2\Big)^{-1}\Big\{\sum_{i=1}^n \big(X_i^\top \hat{\mu}^*_{\lambda,k} - Y_i^\top \hat{\nu}_k\big)^2$$

$$-\sum_{i=1}^n \big(X_i^\top \hat{\mu}^*_{\lambda_n,k} - Y_i^\top \hat{\nu}_k\big)^2\Big\} + o_p(1) + \log n$$

$$\geq \Big(n^{-1}\sum_{i=1}^n \big(X_i^\top \hat{\mu}^*_{\lambda_n,k} - Y_i^\top \hat{\nu}_k\big)^2\Big)^{-1}\Big\{\sum_{i=1}^n \big(X_i^\top \hat{\mu}^*_{\widehat{\mathcal{A}}_{\lambda,k},k} - Y_i^\top \hat{\nu}_k\big)^2$$

$$-\sum_{i=1}^{n}\left(X_i^\top\hat{\mu}^*_{\lambda_n,k}-Y_i^\top\hat{\nu}_k\right)^2\Big\}+o_p(1)+\log n$$

$$\geq\Big(n^{-1}\sum_{i=1}^{n}\left(X_i^\top\hat{\mu}^*_{\lambda_n,k}-Y_i^\top\hat{\nu}_k\right)^2\Big)^{-1}\min_{\mathcal{A}\supset\mathcal{A}_k}\Big\{\sum_{i=1}^{n}\left(X_i^\top\hat{\mu}^*_{\mathcal{A},k}-Y_i^\top\hat{\nu}_k\right)^2$$

$$-\sum_{i=1}^{n}\left(X_i^\top\hat{\mu}^*_{\lambda_n,k}-Y_i^\top\hat{\nu}_k\right)^2\Big\}+o_p(1)+\log n$$

$$=O_p(1)+\log n.$$

Hence, $\inf_{\lambda\in\mathbb{R}_+}n\Big(\mathrm{BIC}^a_{\lambda,k}-\mathrm{BIC}^a_{\lambda_n,k}\Big)\geq O_p(1)+\log n$. The right-hand side of above formula diverges to $+\infty$ as $n\to\infty$, which implies $P(\inf_{\lambda\in\mathbb{R}_+}\mathrm{BIC}^a_{\lambda,k}>\mathrm{BIC}^a_{\lambda_n,k})\to 1$. Consequently, $P(\inf_{\lambda\in\mathbb{R}_-\bigcup\mathbb{R}_+}\mathrm{BIC}^a_{\lambda,k}>\mathrm{BIC}^a_{\lambda_n,k})\to 1$. this completes the proof.

Appendix E. Proof of Theorem 5

Denote $(X_i^\top,Y_i^\top)^\top$ by $Z_i\in\mathbb{R}^{(p+q)}$, define $\mathrm{cov}(Z_i)=\Sigma$, and the corresponding sample covariance is $\hat{\Sigma}=n^{-1}\sum_{i=1}^n Z_iZ_i^\top$. For an arbitrary matrix $S=(s_{ij})$, $\|S\|_F=\sqrt{\Sigma_{i,j}s_{ij}^2}$ denotes its Frobenius norm. The following 6 lemmas compose the whole proof detail.

Lemma 1. *Under conditions (A1) and (A2), we have that*

$$n/(p+q)^{2+2\kappa}E\|\hat{\Sigma}-\Sigma\|_F^2\to 0.$$

Proof.

$$\frac{n}{(p+q)^{2+2\kappa}}E\|\hat{\Sigma}-\Sigma\|_F^2=\frac{n}{(p+q)^{2+2\kappa}}\sum_{1\leq j,k\leq(p+q)}E(\hat{\Sigma}_{(j,k)}-\Sigma_{(j,k)})^2$$

$$=\frac{n}{(p+q)^{2+2\kappa}}\sum_{1\leq j,k\leq(p+q)}E\Big(\frac{1}{n}\sum_{i=1}^{n}Z_{ij}Z_{ik}-\Sigma_{(j,k)}\Big)^2$$

$$=\frac{n}{(p+q)^{2+2\kappa}}\sum_{1\leq j,k\leq(p+q)}E\Big(\frac{1}{n^2}\big(\sum_{i=1}^{n}Z_{ij}Z_{ik}\big)^2-\Sigma^2_{(j,k)}\Big)$$

$$=\frac{n}{(p+q)^{2+2\kappa}}\sum_{1\leq j,k\leq(p+q)}E\Big(\frac{1}{n^2}\sum_{i=1}^{n}Z_{ij}^2Z_{ik}^2+\frac{1}{n^2}\sum_{i\neq i'}Z_{ij}Z_{ik}Z_{i'j}Z_{i'k}-\Sigma^2_{(j,k)}\Big)$$

$$= \frac{n}{(p+q)^{2+2\kappa}} \sum_{1\le j,k\le (p+q)} \Big(\frac{1}{n}E\big(Z_{ij}^2 Z_{ik}^2\big) + \frac{n-1}{n}\Sigma_{(j,k)}^2 - \Sigma_{(j,k)}^2\Big)$$

$$= \frac{1}{(p+q)^{2+2\kappa}} \sum_{1\le j,k\le (p+q)} \Big(E\big(Z_{ij}^2 Z_{ik}^2\big) - \Sigma_{(j,k)}^2\Big)$$

$$= \frac{1}{(p+q)^{2+2\kappa}} \sum_{1\le j,k\le (p+q)} \Big(\Sigma_{(k,k)}\Sigma_{(j,j)} + \Sigma_{(j,k)}^2\Big)$$

$$\le \frac{1}{(p+q)^{2+2\kappa}} \sum_{1\le j,k\le (p+q)} 2 = \frac{2}{(p+q)^{2\kappa}} \to 0.$$

The proof of Lemma 1 is completed.

Recall that $\hat{T} = \hat{\Sigma}_{xx}^{-1/2}\hat{\Sigma}_{xy}\hat{\Sigma}_{yy}^{-1/2}$, and $T = \Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2} = \Sigma_{xy}$.

Lemma 2. *Under conditions (A1), (A2) and (A3), we have that*

$$n/(p+q)^{2+2\kappa} E\|\hat{T} - T\|_F^2 \to 0.$$

Proof.

$$\|\hat{T}-T\|_F^2 \le 2\|\hat{\Sigma}_{xx}^{-1/2}(\hat{\Sigma}_{xy}\hat{\Sigma}_{yy}^{-1/2} - \Sigma_{xy}\Sigma_{yy}^{-1/2})\|_F^2 + 2\|(\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2})\Sigma_{xy}\Sigma_{yy}^{-1/2}\|_F^2$$

$$\le 2\lambda_{\max}(\hat{\Sigma}_{xx}^{-1})\|\hat{\Sigma}_{xy}\hat{\Sigma}_{yy}^{-1/2} - \Sigma_{xy}\Sigma_{yy}^{-1/2}\|_F^2 + 2\|(\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2})\Sigma_{xy}\|_F^2$$

$$\le 4\lambda_{\max}(\hat{\Sigma}_{xx}^{-1})\big\{\|(\hat{\Sigma}_{xy} - \Sigma_{xy})\hat{\Sigma}_{yy}^{-1/2}\|_F^2 + \|\Sigma_{xy}(\hat{\Sigma}_{yy}^{-1/2} - \Sigma_{yy}^{-1/2})\|_F^2\big\}$$

$$+2\lambda_{\max}(\Sigma_{xy}\Sigma_{yx})\|\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2}\|_F^2$$

$$\le 4\lambda_{\max}(\hat{\Sigma}_{xx}^{-1})\big\{\lambda_{\max}(\hat{\Sigma}_{yy}^{-1})\|\hat{\Sigma}_{xy} - \Sigma_{xy}\|_F^2 + \lambda_{\max}(\Sigma_{xy}\Sigma_{yx})\|\hat{\Sigma}_{yy}^{-1/2} - \Sigma_{yy}^{-1/2}\|_F^2\big\}$$

$$+2\lambda_{\max}(\Sigma_{xy}\Sigma_{yx})\|\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2}\|_F^2.$$

It is easy to show that $\lambda_{\max}(\Sigma_{xy}\Sigma_{yx}) \le 1$, hence we have

$$\|\hat{T} - T\|_F^2 \le 4b^{-2}\|\hat{\Sigma}_{xy} - \Sigma_{xy}\|_F^2 + 4b^{-1}\|\hat{\Sigma}_{yy}^{-1/2} - \Sigma_{yy}^{-1/2}\|_F^2$$

$$+2\|\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2}\|_F^2. \tag{8}$$

From Lemma 1, we know that $E\|\hat{\Sigma}_{xy} - \Sigma_{xy}\|_F^2 = (p+q)^{2+2\kappa}n^{-1}o(1)$. Next we consider the third term in the right part of (8), $\|\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2}\|_F^2$. By

Condition (A2), we know $\Sigma_{xx} = I_p$, hence it is easy to show that

$$
\begin{aligned}
(\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2})(\hat{\Sigma}_{xx}^{-1/2} + \Sigma_{xx}^{-1/2}) &= (\hat{\Sigma}_{xx}^{-1/2} - I_p)(\hat{\Sigma}_{xx}^{-1/2} + I_p) \\
&= \hat{\Sigma}_{xx}^{-1} - I_p = -\hat{\Sigma}_{xx}^{-1}(\hat{\Sigma}_{xx} - I_p).
\end{aligned}
$$

Consequently, $\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2} = -\hat{\Sigma}_{xx}^{-1}(\hat{\Sigma}_{xx} - I_p)(\hat{\Sigma}_{xx}^{-1/2} + I_p)^{-1}$. Further, we can have that

$$
\begin{aligned}
\|\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2}\|_F^2 &= \|\hat{\Sigma}_{xx}^{-1}(\hat{\Sigma}_{xx} - I_p)(\hat{\Sigma}_{xx}^{-1/2} + I_p)^{-1}\|_F^2 \\
&\le \lambda_{\max}^2(\hat{\Sigma}_{xx}^{-1})\lambda_{\max}^2\big((\hat{\Sigma}_{xx}^{-1/2} + I_p)^{-1}\big)\|\hat{\Sigma}_{xx} - I_p\|_F^2 \\
&\le b^{-2}\frac{B}{(1+\sqrt{B})^2}\|\hat{\Sigma}_{xx} - I_p\|_F^2.
\end{aligned}
$$

Hence, we obtain $E\|\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2}\|_F^2 \le b^{-2}B(1+\sqrt{B})^{-2}E\|\hat{\Sigma}_{xx} - I_p\|_F^2 = (p+q)^{2+2\kappa}n^{-1}o(1)$. Similarly, we can also have that $E\|\hat{\Sigma}_{yy}^{-1/2} - \Sigma_{yy}^{-1/2}\|_F^2 = (p+q)^{2+2\kappa}n^{-1}o(1)$. Consequently, we obtain that $E\|\hat{T} - T\|_F^2 \le 4b^{-2}E\|\hat{\Sigma}_{xy} - \Sigma_{xy}\|_F^2 + 4b^{-1}E\|\hat{\Sigma}_{yy}^{-1/2} - \Sigma_{yy}^{-1/2}\|_F^2 + 2E\|\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2}\|_F^2 = (p+q)^{2+2\kappa}n^{-1}o(1)$, which implies that $n(p+q)^{-(2+2\kappa)}E\|\hat{T} - T\|_F^2 \to 0$. This completes the proof of Lemma 2.

The singular value decomposition of T is denoted by $T = UDV^\top$, where $U = (u_1, \cdots, u_p) \in \mathbb{R}^{p\times p}$ and $V = (v_1, \cdots, v_q) \in \mathbb{R}^{q\times q}$ are orthogonal matrices, and $D \in \mathbb{R}^{p\times q}$ is a matrix with $D_{(i,i)} = \lambda_i$, and $D_{(i,j)} = 0$ for $i \ne j$. $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_K$ are not only singular values, but also the population canonical correlation coefficients. By Condition (A4), we know that $\lambda_1 > \lambda_2 > \cdots > \lambda_{K_0} > 0$, and $\lambda_k = 0$ for $k > K_0$. For $\hat{T}$, we also make the singular value decomposition $\hat{T} = \hat{U}\hat{D}\hat{V}^\top$, where $\hat{U} = (\hat{u}_1, \cdots, \hat{u}_p) \in \mathbb{R}^{p\times p}$ and $\hat{V} = (\hat{v}_1, \cdots, \hat{v}_q) \in \mathbb{R}^{q\times q}$ are orthogonal matrices, and $\hat{D} \in \mathbb{R}^{p\times q}$ is a matrix with $\hat{D}_{(i,i)} = \hat{\lambda}_i$, and $\hat{D}_{(i,j)} = 0$ for $i \ne j$. $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_K \ge 0$ are not only singular values of $\hat{T}$, but also the sample canonical correlation coefficients. Here we require that for every $k \le K_0$ $\hat{u}_k^\top u_k \ge 0$. If $\hat{u}_k$ does not satisfy this requirement, we can replace $\hat{u}_k, \hat{v}_k$ by $-\hat{u}_k$ and $-\hat{v}_k$, respectively. By von Neumann's trace inequality [19], one can easily obtain that for every

$1 \leq k \leq K = \min(p, q)$,

$$\frac{n}{(p+q)^{2+2\kappa}} E(\hat{\lambda}_k - \lambda_k)^2 \leq \frac{n}{(p+q)^{2+2\kappa}} \sum_{k=1}^{K} E(\hat{\lambda}_k - \lambda_k)^2$$

$$\leq \frac{n}{(p+q)^{2+2\kappa}} E\|\hat{T} - T\|_F^2 \to 0.$$

Recall that $\mu_k = \Sigma_{xx}^{-1/2} u_k$, $\nu_k = \Sigma_{yy}^{-1/2} v_k$ are population canonical loadings of the k-th pair canonical variables, and sample canonical loadings $\hat{\mu}_k = \hat{\Sigma}_{xx}^{-1/2} \hat{u}_k$, $\hat{\nu}_k = \hat{\Sigma}_{yy}^{-1/2} \hat{v}_k$ are the corresponding estimators.

Lemma 3. *Under Conditions (A1)–(A4), we have that for every $k \leq K_0$,*

$$\frac{\sqrt{n}}{(p+q)^{1+\kappa}} E\|\hat{\mu}_k - \mu_k\|^2 \to 0, \qquad \frac{\sqrt{n}}{(p+q)^{1+\kappa}} E\|\hat{\nu}_k - \nu_k\|^2 \to 0.$$

Proof. Let $\tilde{T} = \hat{U} D \hat{V}^\top$, one can easily obtain that $n(p+q)^{-(2+2\kappa)} E\|\tilde{T} - T\|_F^2 \to 0$, from which we have that $n(p+q)^{-(2+2\kappa)} E\|\tilde{T}\tilde{T}^\top - TT^\top\|_F^2 \to 0$, which is equivalent to $n(p+q)^{-(2+2\kappa)} E\|U^\top \hat{U} D^2 \hat{U}^\top U - D^2\|_F^2 \to 0$. Denote $\hat{U}^\top U$ by $\Gamma = (\gamma_1, \cdots, \gamma_p)$, which is an orthogonal matrix, then we have that $n(p+q)^{-(2+2\kappa)} E(\gamma_1^\top D^2 \gamma_1 - \lambda_1^2)^2 \to 0$. we expand the left part of above formula and obtain that

$$\frac{n}{(p+q)^{2+2\kappa}} E(\gamma_1^\top D^2 \gamma_1 - \lambda_1^2)^2 = \frac{n}{(p+q)^{2+2\kappa}} E\Big((1 - \gamma_{11}^2)\lambda_1^2 - \sum_{i \neq 1} \lambda_i^2 \gamma_{1i}^2\Big)^2$$

$$\geq \frac{n}{(p+q)^{2+2\kappa}} E\Big((1 - \gamma_{11}^2)\lambda_1^2 - \lambda_2^2 \sum_{i \neq 1} \gamma_{1i}^2\Big)^2 = \frac{n}{(p+q)^{2+2\kappa}} E\Big((1 - \gamma_{11}^2)^2 (\lambda_1^2 - \lambda_2^2)^2\Big).$$

In Condition (A4), we assume that $\lambda_1 - \lambda_2 \geq l$, hence we can have $(\lambda_1^2 - \lambda_2^2)^2 \geq l^4$, Consequently, we can obtain that

$$\frac{n}{(p+q)^{2+2\kappa}} E\Big((1 - \gamma_{11}^2)^2\Big) \leq \frac{n}{l^4 (p+q)^{2+2\kappa}} E\Big((1 - \gamma_{11}^2)^2 (\lambda_1^2 - \lambda_2^2)^2\Big)$$

$$\leq \frac{n}{l^4 (p+q)^{2+2\kappa}} E(\gamma_1^\top D^2 \gamma_1 - \lambda_1^2)^2 \to 0.$$

This implies that $\sqrt{n}(p+q)^{-(1+\kappa)}E\big(1-\gamma_{11}^2\big) \to 0$, and further recall that we require that $\gamma_{11} = \hat{u}_1^\top u_1 \geq 0$, hence $\sqrt{n}(p+q)^{-(1+\kappa)}E\Big(1-\gamma_{11}\Big) \leq \sqrt{n}(p+q)^{-(1+\kappa)}E\Big(1-\gamma_{11}^2\Big) \to 0$. Consequently, $\sqrt{n}(p+q)^{-(1+\kappa)}E\|\hat{u}_1-u_1\|^2 = 2\sqrt{n}(p+q)^{-(1+\kappa)}E(1-\gamma_{11}) \to 0$. Sequentially, one can use similar skill easily to obtain that for every $1 \leq k \leq K_0$, $\sqrt{n}(p+q)^{-(1+\kappa)}E\|\hat{u}_k - u_k\|^2 \to 0$. Moreover, for every $1 \leq k \leq K_0$ we have

$$
\begin{aligned}
E\|\hat{\mu}_k - \mu_k\|^2 &= E\|\hat{\Sigma}_{xx}^{-1/2}\hat{u}_k - \Sigma_{xx}^{-1/2}u_k\|^2 \\
&\leq 2E\|(\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2})\hat{u}_k\|^2 + 2E\|\Sigma_{xx}^{-1/2}(\hat{u}_k - u_k)\|^2 \\
&\leq 2E\|(\hat{\Sigma}_{xx}^{-1/2} - \Sigma_{xx}^{-1/2})\|_F^2 + 2E\|(\hat{u}_k - u_k)\|^2 \\
&= \frac{(p+q)^{2+2\kappa}}{n}o(1) + \frac{(p+q)^{1+\kappa}}{\sqrt{n}}o(1) \\
&= \frac{(p+q)^{1+\kappa}}{\sqrt{n}}o(1).
\end{aligned}
$$

This implies that $\sqrt{n}(p+q)^{-(1+\kappa)}E\|\hat{\mu}_k-\mu_k\|^2 \to 0$. By similar procedure, one can also obtain that for every $1 \leq k \leq K_0$, $\sqrt{n}(p+q)^{-(1+\kappa)}E\|\hat{\nu}_k - \nu_k\|^2 \to 0$. This completes the proof of Lemma 3.

Denote $\mathbb{X} = (X_1, \cdots, X_n)^\top$, and $\mathbb{Y} = (Y_1, \cdots, Y_n)^\top$, then the function $Q_\lambda^a(\mu)$ can be rewritten as $Q_\lambda^a(\mu) = \Big\{\|\mathbb{Y}\hat{\nu}_k - \mathbb{X}\mu\|^2 + \lambda\sum_{j=1}^p |\mu_j|/|\hat{\mu}_{kj}|\Big\}$, where μ_j is the jth component of μ.

Lemma 4. *Under Conditions (A2) and (A3), we have*

$$
\|\hat{\mu}_{\lambda,k}^* - \mu_k^*\|^2 \leq 2\lambda^2 n^{-2} b^{-2}\sum_{j=1}^p \hat{\mu}_{kj}^{-2} + 4\|\hat{\mu}_k - \mu_k\|^2 + 4(\hat{\lambda}_k - \lambda_k)^2.
$$

Proof. It is easy to see that

$$
\|\mathbb{Y}\hat{\nu}_k - \mathbb{X}\hat{\mu}_{\lambda,k}^*\|^2 + \lambda\sum_{j=1}^p \frac{|\hat{\mu}_{\lambda,kj}^*|}{|\hat{\mu}_{kj}|} \leq \|\mathbb{Y}\hat{\nu}_k - \mathbb{X}\hat{\mu}_{0,k}^*\|^2 + \lambda\sum_{j=1}^p \frac{|\hat{\mu}_{0,kj}^*|}{|\hat{\mu}_{kj}|}.
$$

Hence, we have $\lambda\sum_{j=1}^p |\hat{\mu}_{kj}|^{-1}(|\hat{\mu}_{0,kj}^*| - |\hat{\mu}_{\lambda,kj}^*|) \geq \|\mathbb{Y}\hat{\nu}_k - \mathbb{X}\hat{\mu}_{\lambda,k}^*\|^2 - \|\mathbb{Y}\hat{\nu}_k - \mathbb{X}\hat{\mu}_{0,k}^*\|^2 = \|\mathbb{X}(\hat{\mu}_{\lambda,k}^* - \hat{\mu}_{0,k}^*)\|^2 \geq n\lambda_{\min}(\hat{\Sigma}_{xx})\|\hat{\mu}_{\lambda,k}^* - \hat{\mu}_{0,k}^*\|^2 \geq nb\|\hat{\mu}_{\lambda,k}^* - \hat{\mu}_{0,k}^*\|^2$. Moreover, we have that $\lambda\sum_{j=1}^p |\hat{\mu}_{kj}|^{-1}(|\hat{\mu}_{0,kj}^*| - |\hat{\mu}_{\lambda,kj}^*|) \leq \lambda\sum_{j=1}^p |\hat{\mu}_{kj}|^{-1}|\hat{\mu}_{0,kj}^* -$

$\hat{\mu}^*_{\lambda,kj}| \le \lambda\sqrt{\sum_{j=1}^p \hat{\mu}_{kj}^{-2}}\|\hat{\mu}^*_{\lambda,k} - \hat{\mu}^*_{0,k}\|$. Hence, we can obtain that $nb\|\hat{\mu}^*_{\lambda,k} - \hat{\mu}^*_{0,k}\|^2 \le \lambda\sqrt{\sum_{j=1}^p \hat{\mu}_{kj}^{-2}}\|\hat{\mu}^*_{\lambda,k} - \hat{\mu}^*_{0,k}\|$, which implies that $\|\hat{\mu}^*_{\lambda,k} - \hat{\mu}^*_{0,k}\| \le (nb)^{-1}\lambda\sqrt{\sum_{j=1}^p \hat{\mu}_{kj}^{-2}}$. For $\hat{\mu}^*_{0,k}$, it is easy to gain that $\hat{\mu}^*_{0,k} = \hat{\lambda}_k\hat{\mu}_k$. Hence

$$\|\hat{\mu}^*_{0,k} - \mu^*_k\|^2 = \|\hat{\lambda}_k\hat{\mu}_k - \lambda_k\mu_k\|^2$$

$$\begin{aligned}
&\le 2\hat{\lambda}_k^2\|\hat{\mu}_k - \mu_k\|^2 + 2\|\mu_k\|^2(\hat{\lambda}_k - \lambda_k)^2 \\
&\le 2\|\hat{\mu}_k - \mu_k\|^2 + 2\lambda_{\min}^{-1}(\Sigma_{xx})(\hat{\lambda}_k - \lambda_k)^2 \\
&\le 2\|\hat{\mu}_k - \mu_k\|^2 + 2(\hat{\lambda}_k - \lambda_k)^2.
\end{aligned}$$

Consequently, we have that $\|\hat{\mu}^*_{\lambda,k} - \mu^*_k\|^2 \le 2\lambda^2 n^{-2}b^{-2}\sum_{j=1}^p \hat{\mu}_{kj}^{-2} + 4\|\hat{\mu}_k - \mu_k\|^2 + 4(\hat{\lambda}_k - \lambda_k)^2$. This completes the proof of Lemma 4.

Lemma 5. *Under Conditions (A1)–(A6), we have $P(\hat{\mathcal{A}}_{\lambda,k} \subset \mathcal{A}_k) \to 1$.*

Proof. For an arbitrary index set $I \subseteq \{1, \cdots, p\}$, define $\mathbb{X}_I$ to be the matrix collecting $\mathbb{X}$'s column vectors according to I. Let $\tilde{\mu}^*_{\lambda,k} = (\tilde{\mu}^{*\top}_{\lambda,\mathcal{A}_k}, 0^\top)^\top$, where

$$\tilde{\mu}^*_{\lambda,\mathcal{A}_k} = \arg\min_{\mu}\Big\{\|\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\mu\|^2 + \lambda\sum_{j\in\mathcal{A}_k}|\mu_j|/|\hat{\mu}_{kj}|\Big\}. \tag{9}$$

To prove the claim of the Lemma, it only need to prove that with probability tending to 1, $\tilde{\mu}^*_{\lambda,k}$ is also the minimal value point of $Q^a_\lambda(\mu)$, which is equivalent to that $\tilde{\mu}^*_{\lambda,k}$ satisfies the Karush-Kuhn-Tucker (KKT) conditions [4] of the optimization problem $\min_\mu Q^a_\lambda(\mu)$. By the definition of $\tilde{\mu}^*_{\lambda,k}$, It suffices to show that $P(\forall j \in \mathcal{A}^c_k, |-2\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})| \le \lambda|\hat{\mu}_{kj}|^{-1}) \to 1$, or equivalently, $P(\exists j \in \mathcal{A}^c_k, |-2\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})| > \lambda|\hat{\mu}_{kj}|^{-1}) \to 0$. Let $M = \sqrt{n^{-1/2}(p+q)^{1+\kappa}}$, we have

$$P(\exists j \in \mathcal{A}^c_k, |-2\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})| > \lambda|\hat{\mu}_{kj}|^{-1})$$

$$\le \sum_{j\in\mathcal{A}^c_k} P(|-2\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})| > \lambda|\hat{\mu}_{kj}|^{-1})$$

$$\le \sum_{j\in\mathcal{A}^c_k} P(|-2\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})| > \lambda|\hat{\mu}_{kj}|^{-1}, |\hat{\mu}_{kj}| \le M) + \sum_{j\in\mathcal{A}^c_k} P(|\hat{\mu}_{kj}| > M).$$

By the claim of Lemma 3, we have

$$\sum_{j\in\mathcal{A}_k^c} P(|\hat{\mu}_{kj}|>M) \le \frac{1}{M^2}E\sum_{j\in\mathcal{A}_k^c}|\hat{\mu}_{kj}|^2 \le \frac{1}{M^2}E\|\hat{\mu}_k-\mu_k\|^2 = o(1).$$

Moreover, let $\eta_k = \min_{j\in\mathcal{A}_k}(\lambda_k^{-1}|\mu_{kj}^*|)$ and $\hat{\eta}_k = \min_{j\in\mathcal{A}_k}(|\hat{\mu}_{kj}|)$, then we have

$$\sum_{j\in\mathcal{A}_k^c} P(|-2\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})| > \lambda|\hat{\mu}_{kj}|^{-1}, |\hat{\mu}_{kj}|\le M)$$

$$\le \sum_{j\in\mathcal{A}_k^c} P(|-2\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})| > \lambda|\hat{\mu}_{kj}|^{-1}, |\hat{\mu}_{kj}|\le M, \hat{\eta}_k \ge \eta/2)$$
$$+P(\hat{\eta}_k\le \eta_k/2),$$

and

$$P(\hat{\eta}_k\le\eta_k/2) \le P(\|\hat{\mu}_k-\mu_k\|\ge\eta_k/2) \le \frac{4E\|\hat{\mu}_k-\mu_k\|^2}{\eta_k^2} = \frac{(p+q)^{1+\kappa}}{\sqrt{n}\eta_k^2}o(1).$$

By condition (A6), we know that $(p+q)^{1+\kappa}n^{-1/2}\eta_k^{-2} = o(1)$, hence we have that $P(\hat{\eta}_k\le\eta_k/2) = o(1)$.

$$\sum_{j\in\mathcal{A}_k^c} P(|-2\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})| > \lambda|\hat{\mu}_{kj}|^{-1}, |\hat{\mu}_{kj}|\le M, \hat{\eta}_k\ge\eta_k/2)$$

$$\le \sum_{j\in\mathcal{A}_k^c} P(|\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})| > \frac{\lambda}{2M}, \hat{\eta}_k\ge\eta_k/2)$$
$$\le \frac{4M^2}{\lambda^2}E\Big(\sum_{j\in\mathcal{A}_k^c}|\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})|^2 I(\hat{\eta}_k\ge\eta_k/2)\Big).$$

Moreover,

$$\sum_{j\in\mathcal{A}_k^c}|\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})|^2 = \|\mathbb{X}_{\mathcal{A}_k^c}^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})\|_F^2$$

$$\le \|\mathbb{X}^\top(\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})\|_F^2 \le n\lambda_{\max}(\hat{\Sigma}_{xx})\|\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k}\|^2$$
$$\le nB\|\mathbb{Y}\hat{\nu}_k - \mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k}\|^2$$

$$\leq 2nB\Big\{\|\mathbb{Y}(\hat{\nu}_k-\nu_k)\|^2+\|\mathbb{Y}\nu_k-\mathbb{X}\tilde{\mu}^*_{\lambda,k}\|^2\Big\}$$

$$\leq 2nB\Big\{n\lambda_{\max}(\hat{\Sigma}_{yy})\|\hat{\nu}_k-\nu_k\|^2+2\|\mathbb{Y}\nu_k-\mathbb{X}\mu^*_k\|^2+2\|\mathbb{X}(\tilde{\mu}^*_{\lambda,k}-\mu^*_k)\|^2\Big\}$$

$$\leq 2nB\Big\{nB\|\hat{\nu}_k-\nu_k\|^2+2nB\|\tilde{\mu}^*_{\lambda,k}-\mu^*_k\|^2+2\|\mathbb{Y}\nu_k-\mathbb{X}\mu^*_k\|^2\Big\}.$$

Hence,

$$\frac{4M^2}{\lambda^2}E\Big(\sum_{j\in\mathcal{A}_k^c}|\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k-\mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})|^2I(\hat{\eta}_k\geq\eta_k/2)\Big)$$

$$\leq\frac{8BnM^2}{\lambda^2}\Big(nBE(\|\hat{\nu}_k-\nu_k\|^2)+2nBE\big(\|\tilde{\mu}^*_{\lambda,k}-\mu^*_k\|^2I(\hat{\eta}_k\geq\eta_k/2)\big)$$

$$+2E(\|\mathbb{Y}\nu_k-\mathbb{X}\mu^*_k\|^2)\Big)$$

$$\leq\frac{16Bn^2M^2}{\lambda^2}\Big(\frac{(p+q)^{1+\kappa}}{\sqrt{n}}o(1)+BE\big(\|\tilde{\mu}^*_{\lambda,k}-\mu^*_k\|^2I(\hat{\eta}_k\geq\eta_k/2)\big)+1\Big).$$

From the claims of Lemma 3 and Lemma 4, we can obtain that

$$E\big(\|\tilde{\mu}^*_{\lambda,k}-\mu^*_k\|^2I(\hat{\eta}_k\geq\eta_k/2)\big)$$

$$\leq\frac{2\lambda^2E\Big(\sum_{j\in\mathcal{A}_k}\hat{\mu}_{kj}^{-2}\Big)}{n^2b^2}+4E\|\hat{\mu}_k-\mu_k\|^2+4E(\hat{\lambda}_k-\lambda_k)^2$$

$$\leq\frac{2\lambda^2p}{b^2n^2\eta_k^2}+\frac{(p+q)^{1+\kappa}}{\sqrt{n}}o(1)=o(1).$$

Consequently, we have

$$\frac{4M^2}{\lambda^2}E\Big(\sum_{j\in\mathcal{A}_k^c}|\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k-\mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})|^2I(\hat{\eta}_k\geq\eta_k/2)\Big)$$

$$\leq\frac{16Bn^2M^2}{\lambda^2}\big(1+o(1)\big).$$

Combining above results and Condition (A6), one can obtain that

$$P(\exists j\in\mathcal{A}_k^c,|-2\mathbb{X}_j^\top(\mathbb{Y}\hat{\nu}_k-\mathbb{X}_{\mathcal{A}_k}\tilde{\mu}^*_{\lambda,\mathcal{A}_k})|>\lambda|\hat{\mu}_{kj}|^{-1})$$

$$\leq \frac{16Bn^2M^2}{\lambda^2}\big(1+o(1)\big) + \frac{(p+q)^{1+\kappa}}{\sqrt{n}\eta_k^2}o(1) + o(1) \to 0.$$

This completes the proof of this lemma.

Lemma 6. *Under Conditions (A1)–(A6), we have $P(\mathcal{A}_k \subset \hat{\mathcal{A}}_{\lambda,k}) \to 1$.*

Proof. It suffices to show that $P(\min_{j\in\mathcal{A}_k} |\tilde{\mu}^*_{\lambda,kj}| > 0) \to 1$. By similar prove process of Lemma 4, we can have

$$\Big|\min_{j\in\mathcal{A}_k} |\tilde{\mu}^*_{\lambda,kj}| - \min_{j\in\mathcal{A}_k} |\tilde{\mu}^*_{0,kj}|\Big| \leq \|\tilde{\mu}^*_{\lambda,k} - \tilde{\mu}^*_{0,k}\| \leq \frac{\lambda\sqrt{\sum_{j\in\mathcal{A}_k} \hat{\mu}_{kj}^{-2}}}{nb} < \frac{\lambda\sqrt{p}}{bn\hat{\eta}_k},$$

hence,

$$\min_{j\in\mathcal{A}_k} |\tilde{\mu}^*_{\lambda,kj}| > \min_{j\in\mathcal{A}_k} |\tilde{\mu}^*_{0,kj}| - \frac{\lambda\sqrt{p}}{bn\hat{\eta}_k}.$$

One can also see that $\min_{j\in\mathcal{A}_k} |\tilde{\mu}^*_{0,kj}| > \min_{j\in\mathcal{A}_k} |\mu^*_{kj}| - \|\tilde{\mu}^*_{0,k} - \mu^*_k\|$. Consequently, we have that

$$\min_{j\in\mathcal{A}_k} |\tilde{\mu}^*_{\lambda,kj}| > \min_{j\in\mathcal{A}_k} |\mu^*_{kj}| - \|\tilde{\mu}^*_{0,k} - \mu^*_k\| - \frac{\lambda\sqrt{p}}{bn\hat{\eta}_k}.$$

$$= \lambda_k\eta_k - \frac{\lambda\sqrt{p}}{bn\hat{\eta}_k} + o_p(1) = \lambda_k\eta_k - \frac{\lambda\sqrt{p}}{bn\eta_k}\frac{\eta_k}{\hat{\eta}_k} + o_p(1)$$

$$= \lambda_k\eta_k - \frac{\eta_k}{\hat{\eta}_k}o(1) + o_p(1).$$

Moreover,$E\Big((\hat{\eta}_k-\eta_k)^2\Big) \leq E\|\hat{\mu}_k-\mu_k\|^2 = (p+q)^{1+\kappa}n^{-1/2}o(1)$. Hence, we have that $\sqrt{n}\eta_k^2(p+q)^{-(1+\kappa)}E\Big((\hat{\eta}_k/\eta_k-1)^2\Big) = o(1)$. This implies that $E\Big((\hat{\eta}_k/\eta_k-1)^2\Big) = o(1)$. Hence, $\eta_k/\hat{\eta}_k = O_p(1)$, thus we have $\min_{j\in\mathcal{A}_k} |\tilde{\mu}^*_{\lambda,kj}| > \lambda_k\eta_k - O_p(1)o(1) + o_p(1) = \lambda_k\eta_k + o_p(1)$. Consequently,

$$P(\min_{j\in\mathcal{A}_k} |\tilde{\mu}^*_{\lambda,kj}| > 0) \geq P(\min_{j\in\mathcal{A}_k} |\tilde{\mu}^*_{\lambda,kj}| > \lambda_k\eta_k) \to 1.$$

This completes the proof of Lemma 6.

Combining the results of Lemma 5 and Lemma 6, we obtain that $P(\hat{\mathcal{A}}_{\lambda,k} = \mathcal{A}_k) \to 1$. This completes the whole proof of Theorem 5.

REFERENCES

[**1**] Anderson, T. W., An Introduction to Multivariate Statistical Analysis, John Wiley: New York. 1984.

[**2**] Anderson, T. W., Asymptotic theory for canonical correlation analysis, Journal of Multivariate Analysis, 70 (1999) 1–29.

[**3**] Bai, Z.D., Yin, Y.Q., Limit of the smallest eigenvalue of a large dimensional sample covariance matrix, The Annals of Probability, 21 (1993) 1275–1294.

[**4**] Boyd, S., Vandenberghe, L., Convex Optimization, Cambridge University Press: Cambridge. 2004.

[**5**] Eaton, M. L., Tyler, D. E., The asymptotic distribution of singular values with applications to canonical correlations and correspondence analysis, Journal of Multivariate Analysis, 34 (1994) 439–446.

[**6**] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., Least angle regression (with discussion), The Annals of Statistics, 32 (2004) 407–489.

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

[**8**] Hotelling, H., The most predictable criterion, Educational Psychology, 26 (1935) 139–142.

[**9**] Hotelling, H., Relations between two sets of variates, Biometrika, 28 (1936) 321–377.

[**10**] Izenman, A. J., Reduced-rank regression for the multivariate linear model, Journal of Multivariate Analysis , 5 (1975) 248–264.

[**11**] Johnson, R. A., Wichern, D. W., Applied Multivariate Statistical Analysis (5th Ed.), Pearson Education, (2003).

[**12**] Kidron, E., Schechner, Y. Y., Elad, M., Pixels that sound, Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1 (2005) 88–95.

[**13**] Leng, C.,Wang, H., On general adaptive sparse principal component analysis, Journal of Computational and Graphical Statistics, 18(1) (2009) 201-215.

[**14**] Luo, R., Wang, H., Tsai, C. L., Contour projected dimension reduction, The Annals of Statistics, 37 (2009) 3743–3778.

[**15**] Parkhomenko, E., Tritchler, D., Beyene, J., Genome-wide sparse canonical correlation of gene expression with genotypes, BMC Proceedings, 1(Suppl 1): S119, (2007).

[**16**] Shao, J., An asymptotic theory for linear model selection, Statistica Sinica, 7 (1997) 221–264.

[**17**] Sun, J., Ji, S., Ye, J., A least squares formulation for canonical correlation analysis, Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, (2008).

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

[**19**] Von Neumann, J., Some matrix inequalities and metrization of matric space, Tomsk Univ. Rev, 1 (1937) 286–300.

[**20**] Waaijenborg, S., Zwinderman, A. H., Penalized canonical correlation analysis to quantify the association between gene expression and DNA markers, BMC Proceedings, 1(Suppl 1) S122, (2007).

[**21**] Zhang, H. H., Lu, W., Adaptive lasso for Cox's proportional hazard model, Biometrika, 94 (2007) 691–703.

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

[**23**] Zou, H., Hastie, T., Regression shrinkage and selection via the elastic net, Journal of the Royal Statistical Society, Series B, 67 (2005) 301–320.

[**24**] Zou, H., Zhang, H., On the adaptive elastic-net with a diverging number of parameters, The Annals of Statistics, 37 (2009) 1733–1751.

[**25**] Zou, H., Hastie, T., Tibshirani, R., Sparse principal component analysis, Journal of Computational and Graphical Statistics, 15 (2006) 265–286.

[**26**] Zou, H., Hastie, T. and Tibshirani, R., On the degrees of freedom of the lasso, The Annals of Statistics, 35 (2007) 2173–2192.