Sparse CCA via Precision Adjusted Iterative Thresholding

Mengjie Chen  
*University of North Carolina, Chapel Hill, USA*  
Chao Gao  
*Yale University, USA*  
Zhao Ren  
*University of Pittsburgh, USA*  
Harrison H. Zhou  
*Yale University, USA*

**Summary.** Sparse Canonical Correlation Analysis (CCA), as an approach to study the maximal correlation between two sets of random variables, has received considerable attention in high-dimensional data analysis. Despite of its popularity in application, there has been remarkably few theoretical studies on sparse CCA. In this paper, we introduce an elementary sufficient and necessary characterization for sparse CCA. Under this probabilistic model, we propose a computationally efficient procedure, called CAPIT, to estimate the canonical directions. We show that the proposed procedure is rate-optimal under various assumptions on nuisance parameters. We further apply the proposed method to study the association of DNA methylation and gene expression in a breast cancer dataset.

**Keywords:** Canonical Correlation Analysis, Minimax Lower Bound, Optimal Convergence Rate, Single Canonical Pair Model, Sparsity

1. Introduction

Canonical correlation analysis (CCA) is a celebrated technique proposed by Hotelling (1936) to find the linear combinations of two sets of random variables with maximal correlation. Taking an example from integrative cancer genomics studies, CCA can be used to study the relationship between DNA methylation and gene expression (VanderKraats et al., 2013). Let $X \in \mathbb{R}^{p_1}$ be a centered random vector representing $p_1$ methylation probes and $Y \in \mathbb{R}^{p_2}$ be a centered random vector representing $p_2$ genes. The correlation between
DNA methylation and gene expression is defined as
\[
\max_{(a,b)} \{ \text{Cov}(a^T X, b^T Y) : \text{Var}(a^T X) = \text{Var}(b^T Y) = 1 \}.
\]  
(1)
The maximizer \((\theta, \eta)\) of (1) is the linear combinations of \(X\) and \(Y\) that are maximally correlated. They are termed as the canonical directions. Under high dimensional settings where \(p_1\) and \(p_2\) are large, the canonical directions \((\theta, \eta)\) can be sparse and only small subsets of methylation probes and genes are significantly correlated. This gives rise to the sparse CCA problem to be studied in this paper, the aim of which is to find sparse canonical directions between two large sets of random variables.

To formally state the problem, we write the covariance matrix of \((X^T, Y^T)^T\) as follows,
\[
\text{Cov} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{pmatrix},
\]  
(2)
where \(\Sigma_1 = \mathbb{E}XX^T\) is the covariance matrix of \(X\), \(\Sigma_2 = \mathbb{E}YY^T\) is the covariance matrix of \(Y\), and \(\Sigma_{21} = \Sigma_{12}^T = \mathbb{E}XY^T\) is the cross covariance between \(X\) and \(Y\). Then, (1) can be recast as
\[
(\theta, \eta) = \arg \max_{(a,b)} \{ a^T \Sigma_{12} b : a^T \Sigma_1 a = 1, b^T \Sigma_2 b = 1 \}.
\]  
(3)
The optimization problem (3) can be solved by singular value decomposition (SVD) on the matrix \(\Sigma_1^{-1/2} \Sigma_{12} \Sigma_2^{-1/2}\). In practice, Hotelling (1936) proposed to replace \(\Sigma_1^{-1/2} \Sigma_{12} \Sigma_2^{-1/2}\) by the sample version \(\hat{\Sigma}_1^{-1/2} \hat{\Sigma}_{12} \hat{\Sigma}_2^{-1/2}\). This leads to consistent estimation of the canonical directions \((\theta, \eta)\) when the dimensions \(p_1\) and \(p_2\) are fixed and the sample size \(n\) increases.

However, in the high-dimensional setting, when the dimensions \(p_1\) and \(p_2\) are larger than the sample size \(n\), this SVD approach may not work due to the fact that the inverse of the sample covariance does not exist.

This difficulty motivates researchers to impose sparsity constraints on the canonical directions (Wiesel et al., 2008; Witten et al., 2009; Parkhomenko et al., 2009; Hardoon and Shawe-Taylor, 2011; Lê Cao et al., 2009; Waaijenborg and Zwinderman, 2009; Avants et al., 2010). The sparsity assumptions can effectively reduce the dimensionality and at the same time improve the interpretability in real application. However, to our best knowledge, there is no full characterization of a probabilistic CCA model under which the canonical directions are indeed sparse. Thus theoretical studies on sparse CCA are scarcely populated in the literature.

In this paper, we present a sufficient and necessary condition on the covariance structure (2) with which the solution of CCA is sparse. We show in Section 2, when the rank of \(\Sigma_{12}\) equals one, \((\theta, \eta)\) is the solution of (3) if and only if (2) satisfies \(\Sigma_{12} = \lambda \Sigma_1 \theta \eta^T \Sigma_2\)
with $\lambda \in (0,1]$, which we term as the single canonical pair (SCP) model. Under the SCP model, the canonical directions are sparse if and only if $\theta$ and $\eta$ are sparse, and the marginal covariance matrices $(\Sigma_1, \Sigma_2)$ are nuisance parameters.

Motivated by the SCP model, we propose a method to estimate the sparse canonical directions, called CAPIT, which stands for Canonical correlation Analysis via Precision adjusted Iterative Thresholding. Our basic idea is simple. First, we obtain a good estimator of the precision matrices $(\Omega_1, \Omega_2) = (\Sigma_1^{-1}, \Sigma_2^{-1})$. Then, we transform the data $(X, Y)$ by the estimated precision matrices as $(\hat{\Omega}_1X, \hat{\Omega}_2Y)$ to adjust the influence of the nuisance parameters $(\Sigma_1, \Sigma_2)$. Finally, we apply iterative thresholding on the transformed data to obtain canonical directions. This method achieves the optimal statistical accuracy in only finite steps of iterations.

Rates of convergence for the proposed estimating procedure are obtained under various sparsity assumptions on canonical directions and covariance structure assumptions on $(\Sigma_1, \Sigma_2)$. In Section 4.2, we establish the minimax lower bound for the sparse CCA problem. The rates of convergence match the minimax lower bound as long as the convergence rate of estimating the nuisance parameters $(\Sigma_1, \Sigma_2)$ is not larger than that of estimating the canonical directions. To the best of our knowledge, this is the first sparse CCA method with theoretical guarantee.

We point out that existing sparse CCA methods in the literature may have both computational and statistical drawbacks. Computationally, regularized versions of (3) such as methods in Waaijenborg and Zwinderman (2009) and Wiesel et al. (2008) are based on heuristics to avoid the non-convex nature of (3), but there is no guarantee whether these algorithms would lead to consistent estimators. Statistically, existing methods do not explicitly take into account the influence of the nuisance parameters. For example, Witten et al. (2009) and Parkhomenko et al. (2009) use diagonal matrix or identity matrix to approximate the unknown covariance matrices $(\Sigma_1, \Sigma_2)$. Such approximation is valid only when the covariance matrices $(\Sigma_1, \Sigma_2)$ are nearly diagonal. Otherwise, it does not lead to a consistent estimator. We illustrate this by a numerical example with non-diagonal covariance structure. In this simulation, we consider two multivariate Gaussian distributions with sparse precision matrices and sparse canonical directions. The sparsity assumption on precision matrices corresponds to a sparse graphical interpretation of $X$ and $Y$. We apply the proposed method and the Penalized Matrix Decomposition (PMD) method by Witten et al. (2009). The results are shown in Figure 1. By taking into account the structure of the nuisance parameters, CAPIT accurately recovers the canonical
Fig. 1. Visualization of the simulation results of estimating $(\theta, \eta)$ for a replicate from Scenario II in Section 5.2 when $p_1 = p_2 = 200$ and $n = 500$. The $\{1, 2, ..., 200\}$-th coordinates represent $\theta$ and the $\{201, 202, ..., 400\}$-th coordinates represent $\eta$.

directions, while PMD is not consistent. See Section 5.2 for more details.

In addition, we apply CAPIT to a genomic data example mentioned at the beginning, to study the correlation between methylation and gene expression. CAPIT explicitly takes into account the sparse graphical structure among genes by estimating the sparse precision matrix. Interestingly, we identify methylation probes that are associated with genes that are previously characterized as prognosis signatures of the metastasis of breast cancer. This example suggests CAPIT provides a new approach for exploratory and interpretive analysis of multiple datasets under high-dimensional settings.

To summarize, the contribution of this paper is two-fold. First, we characterize the sparse CCA problem by proposing a probabilistic model and establish the minimax lower bound under certain sparsity class. Second, we propose the CAPIT method that adapts to both sparsity of the canonical directions and the nuisance structure. The CAPIT procedure is computationally efficient and attains optimal rate of convergence under various conditions. The paper is organized as follows. We first provide a full characterization of the sparse CCA model in Section 2. The CAPIT method and algorithm are presented in Section 3. Section 4 is devoted to a theoretical analysis of our method. This section also presents the minimax lower bound. Section 5 and Section 6 investigate the numerical performance of our procedure by simulation studies and a real data example. The proofs of all technical results are given in the supplementary material.
1.1. Notations

For a matrix $A = (a_{ij})$, we use $||A||$ to denote its largest singular value and call it the spectral norm of $A$. The Frobenius norm is defined as $||A||_F = \sqrt{\sum_{ij} a_{ij}^2}$. The matrix $l_1$ norm is defined as $||A||_{l_1} = \max_j \sum_i |a_{ij}|$. The norm $|| \cdot ||$, when applied to a vector, is understood as the usual Euclidean $l_2$ norm. For any two real numbers $a$ and $b$, we use notations $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For any set $S$, $1_S$ is the corresponding indicator function. Other notations will be introduced along with the text.

2. The Sparse CCA Model

To study CCA from a theoretical point of view, we need a probabilistic model that naturally characterizes the optimization procedure (3). Given the covariance structure (2), we present the following proposition.

**Proposition 2.1.** When $\Sigma_{12}$ is of rank 1, the solution (up to sign jointly) of (3) is $(\theta, \eta)$ if and only if the cross covariance between $X$ and $Y$ can be written as $\Sigma_{12} = \lambda \Sigma_1^1/2 \theta \eta^T \Sigma_2$, where $0 < \lambda \leq 1$, $\theta^T \Sigma_1 \theta = 1$ and $\eta^T \Sigma_2 \eta = 1$. As a consequence, the correlation between $a^T X$ and $b^T Y$ are maximized by $\text{corr}(\theta^T X, \eta^T Y)$, and $\lambda$ is the canonical correlation between $X$ and $Y$.

The proposition above is just an elementary consequence of SVD after transforming the parameters $\theta$ and $\eta$ into $\Sigma_{12}^{1/2} \theta$ and $\Sigma_{12}^{1/2} \eta$ respectively. Due to the page limits, the proof is given in the supplement. For general $\Sigma_{12}$ with rank $r \geq 1$, it is a routine extension that the unique (up to sign jointly) solution of Equation (3) is $(\theta_1, \eta_1)$ if and only if the covariance structure between $X$ and $Y$ can be written as $\Sigma_{12} = \Sigma_1 \left( \sum_{i=1}^r \lambda_i \theta_i \eta_i^T \right) \Sigma_2$, where $\lambda_i$ decreases $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_r > 0$, $r = \text{rank}(\Sigma_{12})$, and $\{(\theta_i, \eta_i)\}$ are orthonormal w.r.t. metric $\Sigma_1$ and $\Sigma_2$ respectively, i.e., $\theta_i^T \Sigma_1 \theta_j = 1_{\{i = j\}}$ and $\eta_i^T \Sigma_2 \eta_j = 1_{\{i = j\}}$.

In this paper, we focus on the case where $r = 1$. Inspired by Proposition 2.1, we propose a probabilistic model of $(X, Y)$, where the canonical directions $(\theta, \eta)$ are explicitly modeled in the joint distribution of $(X, Y)$. 
The Single Canonical Pair Model

\[
\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \lambda \Sigma_1 \theta \eta^T \Sigma_2 \\ \lambda \Sigma_2 \eta \theta^T \Sigma_1 & \Sigma_2 \end{pmatrix} \right),
\]

(4)

with \( \Sigma_1 > 0, \Sigma_2 > 0, \theta^T \Sigma_1 \theta = 1, \eta^T \Sigma_2 \eta = 1 \) and \( 0 < \lambda \leq 1 \).

Just as the single-spike model in PCA (Tipping and Bishop, 1999; Johnstone and Lu, 2009), the model (4) explicitly models \((\lambda, \theta, \eta)\) in the form of the joint distribution of \((X, Y)\). Besides, it can be generalized to multiple canonical pairs structure as in the multi-spike model (Birnbaum et al., 2013). However, the PCA and CCA problems are fundamentally different. This can be seen from the covariance structure in the single canonical pair (SCP) model. The covariance structure of SCP is summarized as

\[ \Sigma_{12} = \lambda \Sigma_1 \theta \eta^T \Sigma_2. \]

(5)

Correspondingly, the rank-one model in PCA is called the single-spike model, whose covariance structure can be written as

\[ \Sigma = \lambda \theta \theta^T + I, \]

(6)

where \( \theta \) is the principal component of the random variable. A comparison of (5) and (6) reveals that estimation in the CCA problem is more involved than that in the PCA problem because of the presence of the nuisance parameters \((\Sigma_1, \Sigma_2)\). The difficulty of estimating covariance matrices and precision matrices is well known in high-dimensional statistics (Cai and Zhou, 2012; Ren et al., 2015). In the sparse PCA setting, the absence of nuisance parameter in (6) leads to algorithms directly applied on the sample covariance matrix \( \hat{\Sigma} \), and the corresponding theoretical analysis is more tractable. In contrast, in the sparse CCA setting, not only do we need to adapt to the underlying sparsity of \((\theta, \eta)\), but we also need to adapt to the unknown covariance structure \((\Sigma_1, \Sigma_2)\). We are going to show in Section 4 how various structures of \((\Sigma_1, \Sigma_2)\) influence the convergence rate of the proposed method.

3. Methodology

In this section, we introduce the CAPIT algorithm to estimate the sparse canonical direction pair \((\theta, \eta)\) in the single canonical pair model in details. The procedure is motivated by the power method, a standard technique to compute the leading eigenvector of a given
symmetric matrix $S$ (Golub and Van Loan, 1996). Let $S$ be a $p \times p$ symmetric matrix. We compute its leading eigenvector. Starting with a vector $v^{(0)}$ non-orthogonal to the leading eigenvector, the power method generates a sequence of vectors $v^{(i)}$, $i = 1, 2, \ldots$, by alternating the multiplication step $w^{(i)} = S v^{(i-1)}$ and the normalization step $v^{(i)} = w^{(i)}/\|w^{(i)}\|$ until convergence. The limit of the sequence, denoted by $v^{(\infty)}$, is the leading eigenvector. The power method can be generalized to compute the leading singular vectors of any $p_1 \times p_2$ dimensional rectangular matrix $M$. Suppose the SVD of a rank $d$ matrix $M$ is $M = U D V^T$, where $D$ is the $d$ dimensional diagonal matrix with singular values on the diagonal. Given an initial pair $(u^{(0)}, v^{(0)})$ non-orthogonal to the leading singular vectors, to compute the leading singular vectors, power method alternates the following steps until $(u^{(0)}, v^{(0)})$ converges to $(u^{(\infty)}, v^{(\infty)})$, which are the left and right leading singular vectors.

(a) Right Multiplication: $w^{(i)}_l = M v^{(i-1)}$,  
(b) Left Normalization: $u^{(i)} = w^{(i)}_l / \|w^{(i)}_l\|$,  
(c) Right Multiplication: $w^{(i)}_r = M^T u^{(i)}$,  
(d) Right Normalization: $v^{(i)} = w^{(i)}_r / \|w^{(i)}_r\|$.

Our goal is to estimate the canonical direction pair $(\theta, \eta)$. The power method above motivates us to find a matrix $\hat{A}$ close to $\lambda \theta \eta^T$ of which $(\theta/\|\theta\|, \eta/\|\eta\|)$ is the leading pair of singular vectors. Note that the covariance structure is $\Sigma_{12} = \lambda \Sigma_1 \theta \eta^T \Sigma_2$. Suppose we know the marginal covariance structures of $X$ and $Y$, i.e. $\Omega_1 = \Sigma_1^{-1}$ and $\Omega_2 = \Sigma_2^{-1}$ are given, it is very natural to consider $\Omega_1 \hat{\Sigma}_{12} \Omega_2$ as the target matrix, where $\hat{\Sigma}_{12}$ is the sample cross-covariance between $X$ and $Y$. Unfortunately, the covariance structures $\Omega_1$ and $\Omega_2$ are unknown as nuisance parameters, but a rate-optimal estimator of $\Omega_j$ ($j = 1, 2$) usually can be obtained under various assumptions on the covariance or precision structures of $X$ and $Y$ in many high-dimensional settings. In the literature, some commonly used structures are sparse precision matrix, sparse covariance matrix, bandable covariance matrix and Toeplitz covariance matrix structures. Later we will discuss the estimators of the precision matrices and their influences to the final estimation error of canonical direction pair $(\theta, \eta)$.

We consider the idea of data splitting. Suppose we have $2n$ i.i.d. copies $(X_i, Y_i)_{1 \leq i \leq 2n}$. We use the first half to compute the sample covariance $\hat{\Sigma}_{12} = \frac{1}{n} \sum_{i=1}^{n} X_i Y_i^T$, and use the second half to estimate the precision matrices by $\hat{\Omega}_1$ and $\hat{\Omega}_2$. Hence the matrix $\hat{A} = \hat{\Omega}_1 \hat{\Sigma}_{12} \hat{\Omega}_2$ is available to us. The reason for data splitting is that we can write the matrix $\hat{A}$ in an alternative form. That is,

$$\hat{A} = \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i \tilde{Y}_i^T,$$
where \( \hat{X}_i = \hat{\Omega}_1 X_i \) and \( \hat{Y}_i = \hat{\Omega}_2 Y_i \) for all \( i = 1, ..., n \). Conditioning on \((X_{n+1}, Y_{n+1}), ..., (X_{2n}, Y_{2n})\), the transformed data \((\hat{X}_i, \hat{Y}_i)_{1 \leq i \leq n}\) are still independently identically distributed. This feature allows us to explore some useful concentration results in the matrix \( \hat{A} \) to prove theoretical results. Conditioning on the second half of data, the expectation of \( \hat{A} \) is 
\[
\beta = \hat{\Omega}_2 \Sigma_2 \eta.
\]
Therefore, as long as the estimators \((\hat{\Omega}_1, \hat{\Omega}_2)\) are accurate in the sense that
\[
||\alpha - \theta|| \vee ||\beta - \eta|| = ||(\hat{\Omega}_1 \Sigma_1 - I) \theta|| \vee ||(\hat{\Omega}_2 \Sigma_2 - I) \eta||
\]
is small, the final rate of convergence is also small.

If we naively apply the power method above to \( \hat{A} = \hat{\Omega}_1 \hat{\Sigma}_1 \hat{\Omega}_2 \) in high-dimensional setting, the variance accumulated across all \( p_1 \) and \( p_2 \) coordinates of left and right singular vectors goes very large and it is possible that we can never obtain a consistent estimator of the space spanned by the singular vectors. Johnstone and Lu (2009) proved that when \( p/n \to 0 \), the leading eigenspace estimated directly from the sample covariance matrix can be nearly orthogonal to the truth under the PCA setting in which \( \hat{A} \) is the sample covariance matrix with dimension \( p_1 = p_2 = p \). Under the sparsity assumption of \((\theta, \eta)\), a natural way of reducing the variance is to only estimate those coordinates with large values in \( \theta \) and \( \eta \) respectively and simply estimate the rest coordinates by zero. Although bias is caused by this thresholding idea, the overall estimation error is reduced due to the bias-variance tradeoff. The idea of combining the power method and the iterative thresholding procedure leads to the algorithm in the next section, which was also proposed by Yang et al. (2013) for a general data matrix \( \hat{A} \) without a theoretical analysis.

### 3.1. Iterative Thresholding

Define a thresholding function \( T(a, t) \) in which \( a \) is a vector and \( t \) is the thresholding level. This paper assumes that \( T(a, t) = \left( a_k 1_\{|a_k| \geq t\} \right) \) is the hard-thresholding function, but any function serves the same purpose in theory as long as it satisfies (i) \(|T(a, t)_k - a_k| \leq t\) and (ii) \( T(a, t)_k = 0 \) whenever \(|a_k| < t\). Therefore the thresholding function can be hard-thresholding, soft-thresholding or SCAD (Fan and Li, 2001). The algorithm is summarized below.

**Remark 3.1.** In Algorithm 1, we do not provide specific stopping rule such as that the difference between successive iterations is small enough. For the single canonical pair model, we show in Section 4 that the optimal rate is achieved in one step. The intuition is simple: when \( A \) is of rank one, we can simply obtain the left singular vector via right multiplying \( A \) by any vector non-orthogonal to the right singular vector. Although in the
current setting \( \hat{A} \) is not a rank one matrix, the effect caused by the second singular value does not change the statistical performance of our final estimator.

Algorithm 1: CAPIT: Iterative Thresholding

**Input:** Sample cross covariance matrix \( \hat{\Sigma}_{12} \);
Estimators of precision matrices \( \hat{\Omega}_1, \hat{\Omega}_2 \);
Initialization pair \( \alpha^{(0)}, \beta^{(0)} \);
Thresholding level \( \gamma_1, \gamma_2 \).

**Output:** Canonical direction estimator \( \alpha^{(\infty)}, \beta^{(\infty)} \).

Set \( \hat{A} = \hat{\Omega}_1 \hat{\Sigma}_{12} \hat{\Omega}_2 \);

repeat

Right Multiplication: \( w^{r,(i)}_l = \hat{A} \beta^{(i-1)} \);
Left Thresholding: \( w^{l,(i)}_{th} = T \left( w^{l,(i)}, \gamma_1 \right) \);
Left Normalization: \( \alpha^{(i)} = w^{l,(i)}_{th} / \left\| w^{l,(i)}_{th} \right\| \);
Left Multiplication: \( w^{r,(i)} = \hat{A}^T \alpha^{(i)} \);
Right Thresholding: \( w^{r,(i)}_{th} = T \left( w^{r,(i)}, \gamma_2 \right) \);
Right Normalization: \( \beta^{(i)} = w^{r,(i)}_{th} / \left\| w^{r,(i)}_{th} \right\| \);

until Convergence of \( \alpha^{(i)} \) and \( \beta^{(i)} \);

Remark 3.2. The thresholding level \( (\gamma_1, \gamma_2) \) are user-specified. Theoretically, they should be set at the level \( O \left( \sqrt{\log(p_1 p_2) / n} \right) \). In Section A.2 in the supplementary material, we present a fully data-driven \( (\gamma_1, \gamma_2) \) along with the proof.

Remark 3.3. The estimator \( (\alpha^{(\infty)}, \beta^{(\infty)}) \) does not directly estimate \((\theta, \eta)\) because the former are unit vectors while the latter satisfies \( \theta^T \Sigma_1 \theta = \eta^T \Sigma_2 \eta = 1 \). We will prove they are almost in the same direction by considering the loss function \( |\sin \angle(a, b)|^2 \) in Johnstone and Lu (2009). Details are presented in Section 4.

Remark 3.4. The estimators of precision matrices \( \hat{\Omega}_1 \) and \( \hat{\Omega}_2 \) depend on the second half of the data and the estimator \( \hat{\Sigma}_{12} \) depends on the first half of the data. In practice, after we apply Algorithm 1, we will swap two parts of the data and use the first half to get \( \hat{\Sigma}_{12} \) and the second half to obtain \( \hat{\Omega}_1, \hat{\Omega}_2 \). Then, Algorithm 1 is run again on the new estimators. The final estimator can be calculated by averaging the two.

3.2. Initialization by Coordinate Thresholding

In Algorithm 1, we need to provide an initializer \( (\alpha^{(0)}, \beta^{(0)}) \) as input. We generate a sensible initialization in this section which is similar to the “diagonal thresholding” sparse
PCA method proposed by Johnstone and Lu (2009). Specifically, we apply a thresholding step to pick index sets $B_1$ and $B_2$ of the coordinates of $\theta$ and $\eta$ respectively. Those index sets can be thought as strong signals. Then a standard SVD is applied on the submatrix of $\hat{A}$ with rows and columns indexed by $B_1$ and $B_2$. The dimension of this submatrix is relatively low such that the SVD on it is fairly accurate. The leading pair of singular vectors is of dimensions $|B_1|$ and $|B_2|$, where $|\cdot|$ denotes the cardinality. In the end, we zero-pad the leading pair of singular vectors into dimensions $p_1$ and $p_2$ respectively to provide our initializer $(\alpha(0), \beta(0))$. This method is summarized in Algorithm 2.

**Algorithm 2: CAPIT: Initialization by Coordinate Thresholding**

**Input:** Sample covariance matrix $\hat{\Sigma}_{12}$; Estimators of precision matrices $\hat{\Omega}_1, \hat{\Omega}_2$; Thresholding level $t_{ij}$.

**Output:** Initializer $\alpha(0)$ and $\beta(0)$.

1. Coordinate selection: pick the index sets $B_1$ and $B_2$ of the coordinates of $\theta$ and $\eta$ respectively as follows,
   $$B_1 = \left\{ i, \max_j |\hat{a}_{ij}| / t_{ij} \geq \sqrt{\frac{\log p_1}{n}} \right\}, B_2 = \left\{ j, \max_i |\hat{a}_{ij}| / t_{ij} \geq \sqrt{\frac{\log p_2}{n}} \right\};$$

2. Reduced SVD: compute the leading pair of singular vectors $(\alpha(0), \beta(0), B_1, B_2)$ on the submatrix $\hat{A}_{B_1, B_2}$;

3. Zero-Padding procedure: construct the initializer $(\alpha(0), \beta(0))$ by zero-padding $(\alpha(0), \beta(0), B_1, B_2)$ on index sets $B_1^c$ and $B_2^c$ respectively,
   $$\alpha_{B_1}^{(0)} = \alpha(0), \beta_{B_1}^{(0)} = 0, \alpha_{B_2}^{(0)} = \beta(0), \beta_{B_2}^{(0)} = 0.$$

The thresholding level $t_{ij}$ in Algorithm 2 is a user specified constant and allowed to be adaptive to each location $(i, j)$. The theoretical data-driven constant for each $t_{ij}$ is provided in Section A.2 in the supplementary material. It is clear the initializer is not unique since if $(\alpha(0), \beta(0))$ serves as the output, $(-\alpha(0), -\beta(0))$ is also a solution of Algorithm 2. However, either pair works as an initializer and provides the same result because we estimate the sub-spaces spanned by the leading pair of singular vectors.

### 3.3. Precision Estimation

Algorithms 1 and 2 start with precision estimators $\hat{\Omega}_1$ and $\hat{\Omega}_2$. As we mentioned, we apply the second half of the data to estimate the precision matrix $\hat{\Omega}_1$ and $\hat{\Omega}_2$. In this section, we discuss four commonly assumed covariance structures of $X$ itself and provide corresponding estimators. We apply the same procedure to $Y$. 
3.3.1. Sparse Precision Matrices

Precision matrix is closely connected to the undirected graphical model, which is a powerful tool to model the relationships among a large number of random variables in a complex system. It is known that recovering the structure of an undirected Gaussian graph is equivalent to recovering the support of the precision matrix. Thus sparse graphical structure among variables in $X$ can be imposed by assuming sparse precision matrices $\Omega_1$. Many algorithms to estimate sparse precision matrix were proposed in the literature. See, e.g., Meinshausen and Bühlmann (2006), Friedman et al. (2008), Cai et al. (2011) and Ren et al. (2015). In this paper, we apply the CLIME method to estimate $\Omega_1$. For details of the algorithm, we refer to Cai et al. (2011).

3.3.2. Bandable Covariance Matrices

The bandable class of covariance matrices was proposed by Bickel and Levina (2008a) to model variables that are collected in a natural order, such as time series data. In this setting, $\sigma_{ij}$ decays to zero at certain rate as $|i - j|$ goes away from the diagonal. Bandable matrix can be estimated by banding or tapering the sample covariance matrix. We apply the tapering method proposed in Cai et al. (2010). Let $\omega = (\omega_m)_{0 \leq m \leq p-1}$ be a weight sequence with $\omega_m$ given by

$$\omega_m = \begin{cases} 
1, & \text{when } m \leq k/2 \\
2 - \frac{2m}{k}, & \text{when } k/2 < m \leq k \\
0, & \text{Otherwise}
\end{cases}$$

where $k$ is the bandwidth. The tapering estimator $\hat{\Sigma}_1$ of the covariance matrix of $X$ is given by $\hat{\Sigma}_1 = (\hat{\omega}_{ij}^{\text{sam}} \omega_{|i-j|})$, where $\hat{\omega}_{ij}^{\text{sam}}$ is the $(i, j)$-th entry of the sample covariance matrix. The bandwidth $k$ is chosen through cross-validation in practice. An alternative adaptive method was proposed by Cai and Yuan (2012). The final estimator of $\Omega_1$ is then defined as $\hat{\Omega}_1 = \hat{\Sigma}_1^{-1}$.

3.3.3. Toeplitz Covariance Matrices

Toeplitz matrix is a covariance matrix in which each descending diagonal from left to right is constant. Class of Toeplitz covariance matrices arises naturally in the analysis of stationary stochastic processes. If $X$ is a stationary process with autocovariance sequence $(\alpha_m) \equiv (\alpha_0, \alpha_1, \cdots, \alpha_{p-1}, \cdots)$, then the covariance matrix $\Sigma_1 = (\sigma_{ij})_{p_1 \times p_1}$ has a Toeplitz structure $\sigma_{ij} = \alpha_{|i-j|}$. We apply the following tapering method proposed in Cai et al. (2013) to estimate Toeplitz Covariance Matrices. Define $\bar{\sigma}_m = \frac{1}{p-m} \sum_{s-t=m} \hat{\omega}_{st}^{\text{sam}}$,
the average of sample covariance along each off-diagonal. Then the tapering estimator 
\( \hat{\Sigma}_1 = (\hat{\sigma}_{st}) \) with bandwidth \( k \) is defined as 
\( \hat{\sigma}_{st} = \omega_{|s-t|} |\hat{\sigma}_{s-t}| \), where \( \omega = (\omega_m)_{0 \leq m \leq p-1} \) is defined in Equation (7). In practice, we pick bandwidth \( k \) using cross-validation. The final estimator of \( \Omega_1 \) is then defined as \( \hat{\Omega}_1 = \hat{\Sigma}_1^{-1} \).

3.3.4. Sparse Covariance Matrices

When no information such as bandable or Toeplitz structure is available in real application, we can impose general sparsity assumption on the covariance matrix, i.e., most of entries in each row/column of the covariance matrix are zero or negligible. We apply a hard thresholding procedure proposed in Bickel and Levina (2008b) to estimate sparse covariance matrices. Let \( \hat{\sigma}_{ij}^{\text{sam}} \) be the \((i, j)\)-th entry of the sample covariance matrix of \( X \). The thresholding estimator \( \hat{\Sigma}_1 = (\hat{\sigma}_{st}) \) is given by 
\[ \hat{\sigma}_{ij} = \hat{\sigma}_{ij}^{\text{sam}} I (|\hat{\sigma}_{ij}^{\text{sam}}| \geq \gamma \sqrt{\log p / n}) \]
for some constant \( \gamma \) which is chosen through cross-validation. The final estimator of \( \Omega_1 \) is then defined as \( \hat{\Omega}_1 = \hat{\Sigma}_1^{-1} \).

4. Statistical Properties and Optimality

In this section, we present the statistical properties and optimality of the proposed estimator. We first present the convergence rates of our procedure, and then a minimax lower bound for a wide range of parameter spaces.

4.1. Convergence Rates

Note that the SCP model is fully determined by the parameter \((\Sigma_1, \Sigma_2, \lambda, \theta, \eta)\), among which we are interested in estimating \((\theta, \eta)\). To achieve statistical consistency, we need some assumptions on the interesting part \((\theta, \eta)\) and the nuisance part \((\Sigma_1, \Sigma_2, \lambda)\).

Assumption A - Sparsity Condition on \((\theta, \eta)\):

We assume \( \theta \) and \( \eta \) are in the weak \( l_q \) ball, with \( 0 \leq q < 2 \), i.e.
\[ |\theta_{(k)}|^q \leq s_1 k^{-1}, \quad |\eta_{(k)}|^q \leq s_2 k^{-1}, \]
where \( \theta_{(k)} \) is the \( k \)-th largest coordinate by magnitude. Let \( p = p_1 \lor p_2 \) and \( s = s_1 \lor s_2 \). The sparsity levels \( s_1 \) and \( s_2 \) satisfy the following condition,
\[ s = o \left( \left( \frac{n}{\log p} \right)^{\frac{1}{2} - \frac{q}{4}} \right). \tag{8} \]

Remark 4.1. In general, we can allow \( \theta \) to be in the weak \( l_{q_1} \) ball and \( \eta \) to be in the weak \( l_{q_2} \) ball with \( q_1 \neq q_2 \). In that case, we require \( s_i = o \left( (n/\log p)^{\frac{1}{2} - \frac{q_i}{4}} \right) \) for \( i = 1, 2 \).
There is no fundamental difference in the analysis and procedures. For simplicity, in the paper we only consider \( q_1 = q_2 \).

**Assumption B - General Conditions on \((\Sigma_1, \Sigma_2, \lambda)\):**

(a) We assume there exists constants \( w \) and \( W \), such that

\[
0 < w \leq \lambda_{\min}(\Sigma_i) \leq \lambda_{\max}(\Sigma_i) \leq W < \infty,
\]

for \( i = 1, 2 \).

(b) To make sure the signals do not vanish, we assume the canonical correlation is bounded below by a positive constant \( C_i \), i.e., \( 0 \leq C_i \leq \lambda \leq 1 \).

(c) We require the existence of consistent estimators \((\hat{\Omega}_1, \hat{\Omega}_2)\) in the sense that

\[
x_{\Omega} = \|\hat{\Omega}_1 \Sigma_1 - I\| \vee \|\hat{\Omega}_2 \Sigma_2 - I\| = o(1),
\]

with probability at least \( 1 - O(p^{-2}) \).

**Loss Function**

For two vectors \( a, b \), a natural way to measure the discrepancy of their directions is the sin of the angle \( |\sin \angle(a, b)| \), see Johnstone and Lu (2009). We consider the loss function

\[
L(a, b)^2 = 2|\sin \angle(a, b)|^2.
\]

It is easy to calculate that

\[
L(a, b) = \frac{a a^T}{||a||^2} - \frac{b b^T}{||b||^2}.
\]

The convergence rate of the CAPIT procedure is presented in the following theorem.

**Theorem 4.1.** Assume the Assumptions A and B above hold. Let \((\alpha^{(k)}, \beta^{(k)})\) be the sequence from Algorithm 1, with the initializer \((\alpha^{(0)}, \beta^{(0)})\) calculated by Algorithm 2. The thresholding levels are

\[
t_{ij}, \quad \gamma_1 = c_1 \sqrt{\frac{\log p}{n}}, \quad \gamma_2 = c_2 \sqrt{\frac{\log p}{n}},
\]

for sufficiently large constants \((t_{ij}, c_1, c_2)\). Then with probability at least \( 1 - O(p^{-2}) \), we have

\[
L(\alpha^{(k)}, \theta)^2 \vee L(\beta^{(k)}, \eta)^2 \leq C \left( s \left( \frac{\log p}{n} \right)^{1-\eta/2} + \|\hat{\Omega}_1 \Sigma_1 - I\| \theta^2 \vee \|\hat{\Omega}_2 \Sigma_2 - I\| \eta^2 \right),
\]

for all \( k = 1, 2, ..., K \) with \( K = O(1) \) and some constant \( C > 0 \).

**Remark 4.2.** The thresholding levels depend on some unknown constants \((t_{ij}, c_1, c_2)\) for the simplicity of presentation. A more involved fully data-driven choice of thresholding levels are presented in Section A.2 in the supplementary material along with the proof.
The upper bound in Theorem 4.1 implies that the error of estimating the nuisance parameters $\hat{\Omega}_i$ contribute to the overall error by $||((\hat{\Omega}_1\Sigma_1 - I)\theta||^2 \lor ||((\hat{\Omega}_2\Sigma_2 - I)\eta||^2$. In the literature, optimal rates of convergence for estimating $\Omega_i$ with various structural assumptions under spectral norm have been established and can be applied here through the inequality $||((\hat{\Omega}_1\Sigma_1 - I)\theta||^2 \leq ||((\hat{\Omega}_1 - \Omega_1)||^2 ||\Sigma_1\theta||^2 \leq W||((\hat{\Omega}_1 - \Omega_1)||^2$. Due to the limited space, we only discuss one setting in which we assume sparse precision matrix structure on $\Omega_i$. Besides the first general condition in Assumption B, we assume each row/column of $\Omega_i$ is in a weak $l_{q_0}$ ball with $0 \leq q_0 \leq 1$. i.e. $\Omega_i \in \mathcal{G}_{q_0}(s_0, p_i)$ for $i = 1, 2$, where

$$\mathcal{G}_{q_0}(s_0, p) = \left\{ \Omega = (\omega_{ij})_{p \times p} : \max_j |\omega_{j(k)}|^{q_0} \leq s_0 k^{-1} \text{ for all } k \right\},$$

and the matrix $l_1$ norm of $\Omega_i$ is bounded by some constant $\|\Omega_i\|_{l_1} \leq w^{-1}$. The notation $\omega_{j(k)}$ means the $k$-th largest coordinate of $j$-th row of $\Omega$ in magnitude. Recall that $p = p_1 \lor p_2$. Under the assumptions that $s_0^2 = O \left( (n/\log p)^{1-q_0} \right)$, Theorem 2 in Cai et al. (2011) implies that CLIME estimator with an appropriate tuning parameter attaining the following rate of convergence $||((\hat{\Omega}_1\Sigma_1 - I)\theta||^2$ with probability at least $1 - O(\frac{p^{-2}}{})$,

$$||((\hat{\Omega}_1\Sigma_1 - I)\theta||^2 \leq W||((\hat{\Omega}_1 - \Omega_1)||^2 \leq C s_0^2 \left( \frac{\log p}{n} \right)^{1-q_0}.$$ 

Therefore we obtain the following corollary.

**Corollary 4.1.** Assume the Assumptions A and B holds, $\Omega_i \in \mathcal{G}_{q_0}(s_0, p_i)$ $i = 1, 2$, $\|\Omega_i\|_{l_1} \leq w^{-1}$ and $s_0^2 = O \left( (n/\log p)^{1-q_0} \right)$. Let $(\alpha^{(k)}, \beta^{(k)})$ be the sequence from Algorithm 1, with the initializer $(\alpha^{(0)}, \beta^{(0)})$ calculated by Algorithm 2 and $\hat{\Omega}_i$ obtained by applying CLIME procedure in Cai et al. (2011). The thresholding levels are the same as those in Theorem 4.1. Then with probability at least $1 - O(p^{-2})$, we have

$$L(\alpha^{(k)}, \theta)^2 \lor L(\beta^{(k)}, \eta)^2 \leq C \left( s \left( \frac{\log p}{n} \right)^{1-q/2} + s_0^2 \left( \frac{\log p}{n} \right)^{1-q_0} \right),$$

for all $k = 1, 2, ..., K$ with $K = O(1)$ and some constant $C > 0$.

**Remark 4.3.** It can be seen from the analysis that similar upper bounds hold in Corollary 4.1 with probability $1 - O(p^{-h})$ by picking different thresholding constants in Algorithms 1, 2 and CLIME procedure for any $h > 0$. Assuming that $n = o(p^h)$, the boundedness of loss function implies that the result Corollary 4.1 also holds in expectation.

### 4.2. Minimax Lower Bound

In this section, we establish a minimax lower bound in a simpler setting with known covariance matrices $\Sigma_1$ and $\Sigma_2$. We further assume $\Sigma_i = I_{p_i \times p_i}$ for $i = 1, 2$ for simplicity.
Otherwise, we can transfer the data accordingly to make $\Sigma_i = I_{p_i \times p_i}$. Before the precise statements, we introduce the parameter space of $(\theta, \eta, \lambda)$. Define

$$F_{q_1, q_2}^{p_1, p_2} (s_1, s_2, C_\lambda) = \left\{ N \left( 0, \Sigma \right) : \Sigma \text{ is specified in (4) }, \lambda \in (C_\lambda, 1), \begin{array}{l} \Sigma_i = I_{p_i \times p_i}, i = 1, 2, \\ |\theta|_q^{(k)} \leq s_1 k^{-1}, |\eta|_q^{(k)} \leq s_2 k^{-1}, \text{ for all } k. \end{array} \right\}. \quad (10)$$

In the sparsity class (10), the covariance matrices $\Sigma_i = I_{p_i \times p_i}$ for $i = 1, 2$ are known and unit vectors $\theta, \eta$ are in the weak $l_q$ ball, with $0 \leq q < 2$. We allow the dimensions of two random vectors $p_1$ and $p_2$ to be very different and only require that $\log p_1$ and $\log p_2$ are comparable with each other,

$$\log p_1 \asymp \log p_2. \quad (11)$$

Remember $s = s_1 \lor s_2$ and $p = p_1 \lor p_2$.

**Theorem 4.2.** For any $q \in [0, 2)$, we assume that $s_1 \left( \frac{n}{\log p} \right)^{q/2} = o(p_i)$ for $i = 1, 2$ and (11) holds. Moreover, we assume $s \left( \frac{\log p}{n} \right)^{1 - \frac{q}{2}} \leq c_0$, for some constant $c_0 > 0$. Then we have

$$\inf_{(\hat{\theta}, \hat{\eta})} \sup_{P \in F} \mathbb{E}_P \left( L^2(\hat{\theta}, \theta) \vee L^2(\hat{\eta}, \eta) \right) \geq C_s \left( \frac{\log p}{n} \right)^{1 - q/2},$$

where $F = F_{q_1, q_2}^{p_1, p_2} (s_1, s_2, C_\lambda)$ and $C$ is a constant only depending on $q$ and $C_\lambda$.

Theorem 4.2 implies the minimaxity for the sparse CCA problem when the covariance matrices $\Sigma_1$ and $\Sigma_2$ are unknown. The lower bound directly follows from Theorem 4.2 and the upper bound follows from Theorem 4.1. Define the parameter space

$$F_{q_1, q_0}^{p_1, p_2} (s_0, s_1, s_2, C_\lambda, w, W) = \left\{ N \left( 0, \Sigma \right) : \Sigma \text{ is specified in (4) }, \lambda \in (C_\lambda, 1), \begin{array}{l} \Sigma_i^{-1} \in G_{q_0} (s_0, p_i), W^{-1} \leq \lambda_{\min}(\Sigma_i^{-1}), ||\Sigma_i^{-1}||_i \leq w^{-1}, \\ |\theta|_q^{(k)} \leq s_1 k^{-1}, |\eta|_q^{(k)} \leq s_2 k^{-1} \text{ for all } k. \end{array} \right\}.$$

Since $F_{q_1, q_2}^{p_1, p_2} (s_1, s_2, C_\lambda) \subset F_{q_1, q_0}^{p_1, p_2} (s_0, s_1, s_2, C_\lambda, w, W)$, the lower bound for the smaller space holds for the larger one. Combining the Corollary 4.1 and the minimax lower bound in Theorem 4.2, we obtain that the minimax rate of convergence of estimating canonical directions over parameter spaces $F_{q_1, q_0}^{p_1, p_2} (s_0, s_1, s_2, C_\lambda, w, W)$.

**Corollary 4.2.** Under the assumptions in Corollary 4.1 and Theorem 4.2 and assume $n = o(p^h)$ for some $h > 0$, we have

$$\inf_{(\hat{\theta}, \hat{\eta})} \sup_{P \in F} \mathbb{E}_P \left( L(\hat{\theta}, \theta)^2 \vee L(\hat{\eta}, \eta)^2 \right) \asymp \frac{\log p}{n}^{1 - q/2},$$

for $F = F_{q_1, q_0}^{p_1, p_2} (s_0, s_1, s_2, C_\lambda, w, W)$, provided that $s_0^2 \left( \frac{\log p}{n} \right)^{1 - q_0} \leq C s \left( \frac{\log p}{n} \right)^{1 - q/2}$ for some constant $C > 0$. 

\textbf{Sparse CCA} 15
5. Simulation Studies

In this section, we present the simulation studies for two scenarios, sparse covariance structure and sparse precision structure, respectively. A short discussion about the results is presented at the end.

5.1. Scenario I: Sparse Covariance Matrix

In the first scenario, the covariance matrices $\Sigma_1$ and $\Sigma_2$ are sparse. More specifically, the covariance matrix $\Sigma_1 = \Sigma_2 = (\sigma_{ij})_{1 \leq i, j \leq p}$ takes the form

$$\sigma_{ij} = \rho^{|i-j|} \quad \text{with} \quad \rho = 0.3.$$  

The canonical pair $(\theta, \eta)$ is generated by normalizing a vector taking the same value at the coordinates $(1, 6, 11, 16, 21)$ and zero elsewhere such that $\theta^T \Sigma_1 \theta = 1$ and $\eta^T \Sigma_2 \eta = 1$. The canonical correlation $\lambda$ is set as $0.9$. We generate the $2n \times p$ data matrices $X$ and $Y$ jointly from (4). As described in the methodology section, we split the data into two halves. In the first step, we estimate the precision matrices $\Omega_1$ and $\Omega_2$ using the first half of the data.

To select the tuning parameters for different procedures, we further split the first part of the data into a $2 : 1$ training set and tuning set. We select the tuning parameters by minimizing the distance of the estimated covariance from the training set and sample covariance matrix of the tuning set in term of the Frobenius norm. More specifically, the bandwidths in the Toeplitz method and in the tapering method are selected through a screening on numbers in the interval of $(1, p)$. The tuning parameter in the thresholding method is selected through a screening on 50 numbers in the interval of $[0.01, 0.5]$.

After obtaining the estimators $\hat{\Omega}_1$ and $\hat{\Omega}_2$, we perform Algorithms 1 and 2 by using $\hat{\Sigma}_{12}$ estimated from the second half of the data. The thresholding parameters $\gamma_1$ and $\gamma_2$ are set to be $2.5 \sqrt{\frac{\log p}{n}}$ for the tapering and thresholding methods, while the thresholding parameter $t_{ij}$ is set to be $2.5$ for all $(i, j)$. For the Toeplitz method, the thresholding
parameters $\gamma_1 = \gamma_2 = 2\sqrt{\frac{\log p}{n}}$ while parameter $t_{ij} = 2$ for all $(i,j)$. The resulted estimator is denoted as $(\hat{\theta}_1, \hat{\eta}_1)$. Then we swap the data, repeat the above procedures and obtain $(\hat{\theta}_2, \hat{\eta}_2)$. The final CAPIT estimator $(\hat{\theta}, \hat{\eta})$ is the average of $(\hat{\theta}_1, \hat{\eta}_1)$ and $(\hat{\theta}_2, \hat{\eta}_2)$.

**Remark 5.1.** Since we use hard thresholding for the CAPIT estimator, $(\hat{\theta}, \hat{\eta})$ is sparse. Denote the supports by $\hat{S}$ and $\hat{T}$, respectively. We propose the following bias correction step by ordinary least squares. That is,

$$\hat{\eta}_{OLS}^T = \arg \min_b \frac{1}{n} \sum_{i=1}^{n} \left( X_{i,S}^T \hat{\theta}_S - Y_{i,T}^T b \right)^2,$$

$$\hat{\theta}_{OLS}^S = \arg \min_a \frac{1}{n} \sum_{i=1}^{n} \left( Y_{i,T}^T \hat{\eta}_{OLS}^T - X_{i,S}^T a \right)^2.$$  

The CAPIT estimators after bias correction are $\hat{\theta}_{OLS} = (\hat{\theta}_{OLS}^S, 0_{c_S})$ and $\hat{\eta}_{OLS} = (\hat{\eta}_{OLS}^T, 0_{c_T})$. The least squares refinement step (12) and (13) is motivated by the equivalent regression formulation of the CCA problem. Namely, under the SCP model, we have

$$\lambda_\eta = \arg \min_b \mathbb{E}(X^T \theta - Y^T b)^2, \quad \lambda_\theta = \arg \min_a \mathbb{E}(Y^T \eta - X^T a)^2.$$  

The equations (12) and (13) can be viewed as the empirical versions of (14) on the estimated supports. Since the CAPIT estimator is already rate-optimal, the refined estimator $(\hat{\theta}_{OLS}, \hat{\eta}_{OLS})$ does not improve the convergence rate, but the constant before the rate, thus leading to better finite sample performance.

We compare our method with the penalized matrix decomposition proposed by Witten et al. (2009) (denoted as PMD), the convex programming method proposed by Gao et al. (2014) (denoted as CoLaR), and the vanilla singular vector decomposition method for CCA (denoted as SVD). For PMD, we use the R function implemented by the authors (Witten et al., 2013), which performs sparse CCA by $l_1$-penalized matrix decomposition and selects the tuning parameters using a permutation scheme. The implementation details and the choice of tuning parameters of CoLaR follow Gao et al. (2014).

We evaluate the performance of different methods by the loss function $L(\hat{\theta}, \theta) \lor L(\hat{\eta}, \eta)$. The results from 100 independent replicates are summarized in Table 1. When CAPIT (Toep), CAPIT (Tap) and CAPIT (Thresh) are combined with the OLS refinement, the errors are the same for the first few digits. Therefore, the results are reported in the same row (CAPIT+OLS) of Table 1.
Table 1. Scenario I: Sparse covariance matrix. Estimation errors for \((\theta, \eta)\) as measured by \(L(\hat{\theta}, \theta) \lor L(\hat{\eta}, \eta)\) based on the median of 100 replicates. Numbers in parentheses are the simulation median absolute deviations.

| \(p_1 = p_2\) | 200  | 300  | 200  | 500  |
|----------------|------|------|------|------|
| \(n\)          | 750  | 750  | 1000 | 1000 |
| CAPIT (Toep)    | 0.11(0.03) | 0.11(0.03) | 0.1(0.02) | 0.09(0.03) |
| CAPIT (Tap)     | 0.12(0.06) | 0.13(0.07) | 0.1(0.05) | 0.09(0.04) |
| CAPIT (Thresh)  | 0.11(0.03) | 0.11(0.03) | 0.09(0.03) | 0.1(0.02) |
| CAPIT+OLS       | 0.04(0.01) | 0.04(0.01) | 0.03(0.01) | 0.04(0.01) |
| CoLaR           | 0.05(0.01) | 0.05(0.02) | 0.04(0.01) | 0.04(0.01) |
| PMD             | 0.16(0.03) | 0.36(0.02) | 0.14(0.02) | 0.11(0.03) |
| SVD             | 0.32(0.01) | 0.44(0.01) | 0.27(0.01) | 0.53(0.02) |

5.2. Scenario II: Sparse Precision Matrix

In the second scenario, we consider that the precision matrices \(\Omega_1\) and \(\Omega_2\) are sparse. In particular, \(\Omega_1 = \Omega_2 = (\omega_{ij})_{1 \leq i, j \leq p}\) take the form:

\[
\omega_{ij} = \begin{cases} 
1 & \text{if } i = j \\
0.5 & \text{if } |i - j| = 1 \\
0.4 & \text{if } |i - j| = 2 \\
0 & \text{otherwise}.
\end{cases}
\]

The canonical pair \((\hat{\theta}, \hat{\eta})\) is the same as described in Scenario I. We generate the \(2n \times p\) data matrices \(X\) and \(Y\) jointly from (4).

As described in the methodology section, we split the data into two halves. In the first step, we estimate the precision matrices by the CLIME proposed in Cai et al. (2011) (denoted as CAPIT (CLIME)). The tuning parameter is selected by maximizing the log-likelihood function. In the second step, we perform Algorithms 1 and 2 with \(\hat{\Sigma}_{12}\) estimated from the second half. The thresholding parameter \(\gamma_1\) and \(\gamma_2\) are set to be \(1.5\sqrt{\frac{\log p}{n}}\) and \(t_{ij}\) is set to be 1.5. The resulted estimator is denoted as \((\hat{\theta}_{[1]}, \hat{\eta}_{[1]})\). Then we swap the data, repeat the above procedures and obtain \((\hat{\theta}_{[2]}, \hat{\eta}_{[2]})\). The final CAPIT estimator \((\hat{\theta}, \hat{\eta})\) is the average of \((\hat{\theta}_{[1]}, \hat{\eta}_{[1]})\) and \((\hat{\theta}_{[2]}, \hat{\eta}_{[2]})\). The refinement via (12) and (13) is denoted as CAPIT+OLS.

For comparison, we also apply PMD, CoLaR and SVD in this case. The results from 100 independent replicates are summarized in Table 2. A visualization of the estimation from a replicate in from the case \(n = 500, p = 200\) under Scenario II is shown in Figure 1.
Table 2. Scenario II: Sparse precision matrix. Estimation errors for $(\theta, \eta)$ as measured by $L(\hat{\theta}, \theta) \lor L(\hat{\eta}, \eta)$ based on the median of 100 replications. Numbers in parentheses are the simulation median absolute deviations.

| $p_1 = p_2$ | 200  | 200  | 500  |
|-------------|------|------|------|
| $n$         | 500  | 750  | 750  |
| CAPIT (CLIME) | 0.41(0.35) | 0.2(0.05) | 0.21(0.12) |
| CAPIT+OLS  | 0.07(0.06) | 0.05(0.03) | 0.03(0.01) |
| CoLaR       | 0.06(0.02) | 0.05(0.01) | 0.05(0.01) |
| PMD         | 1.41(0) | 1.19(0.33) | 1.41(0) |
| SVD         | 0.52(0.03) | 0.39(0.02) | 0.84(0.03) |

5.3. Discussion on the Simulation Results

The above results (Table 1 and Table 2) show that CAPIT+OLS gives the best performance in most cases, and the result of CoLaR is very similar. The PMD method proposed by Witten et al. (2009) and the vanilla SVD method (Hotelling, 1936) are outperformed by the others. It is not surprising that the SVD method does not perform better than our method because of the sparsity assumption in the signals. We shall focus our discussion on the comparison of our method with PMD and CoLaR.

Comparison with PMD. The PMD method is defined by the solution of the following optimization problem

$$(\hat{\theta}_{PMD}, \hat{\eta}_{PMD}) = \arg \max_{(u,v)} \left\{ u^T \Sigma_{12} v : ||u|| \leq 1, ||v|| \leq 1, ||u||_1 \leq c_1, ||v||_1 \leq c_2 \right\}.$$  

As noted by Witten et al. (2009), the PMD method approximates the covariance $\Sigma_1$ and $\Sigma_2$ by the identity matrices $I_{p_1 \times p_1}$ and $I_{p_2 \times p_2}$. If we ignore the $l_1$ regularization, the population version of PMD is to maximize $u^T \Sigma_{12} v$ subject to $||u|| \lor ||v|| \leq 1$, which gives the maximizer in the direction of $(\Sigma_1 \theta, \Sigma_2 \eta)$ instead of $(\theta, \eta)$. When the covariance matrices $\Sigma_1$ and $\Sigma_2$ are sufficiently sparse, $(\Sigma_1 \theta, \Sigma_2 \eta)$ and $(\theta, \eta)$ are close. This explains that in Scenario I, the PMD method performs well. However, in Scenario II, we assume the precision matrices $\Omega_1$ and $\Omega_2$ are sparse. In this case, the corresponding $\Sigma_1$ and $\Sigma_2$ are not necessarily sparse, implying that $(\Sigma_1 \theta, \Sigma_2 \eta)$ could be far away from $(\theta, \eta)$. The PMD method is not consistent in this case, as is illustrated in Figure 1. In contrast, our method takes advantage of the sparsity of $\Omega_1$ and $\Omega_2$, and accurately recovers the canonical directions.
Fig. 2. Left: Visualization of the genomic coordinates of detected methylation probes. Genes that represented by more than one probes are highlighted by square symbols. Right: Canonical correlations of disease associated genes and methylation probes on eight chromosomes in the training set and the test set.

Comparison with CoLaR. The CoLaR estimator proposed by Gao et al. (2014) is a follow-up work of this paper during the review process. They extended the rank-one setting of ours and proposed a two-step estimator that does not require structural assumptions on the marginal covariance matrices. The two-step CoLaR procedure involves a $p_1 \times p_2$ dimensional convex optimization step and a $p_1 + p_2$ dimensional penalized linear regression step. The computational cost is much greater than the CAPIT estimator. Table 1 and Table 2 show that the extra computational cost helps produce very accurate results. However, in the setting of the current paper where structural assumptions on the marginal covariance matrices are given, the CAPIT estimator plus an OLS refinement gives comparable (or even better) performances with much lower computational cost. Therefore, we would recommend CoLaR when little information on the covariance structures is available given sufficient computational resources. When the marginal covariance matrices are structured, we would recommend the proposed CAPIT+OLS in this paper, which is both accurate and scalable especially in a very high-dimensional setting.

6. Real Data Analysis

DNA methylation plays an essential role in the transcriptional regulation (VanderKraats et al., 2013). In tumor, DNA methylation patterns are frequently altered. However, how these alterations contribute to the tumorigenesis and how they affect gene expression and patient survival remain poorly characterized. Thus it is of great interest to investigate the relationship between methylation and gene expression and their interplay with survival status of cancer patients. We applied the proposed method to a breast cancer dataset
Table 3. Sparse CCA results for methylation sites and gene expression that are associated with disease free status for TCGA breast cancer data. In the analysis, methylation and gene expression data are assumed to have sparse precision matrix structure. Sparse CCA were performed on the same set of genes with methylation probes on different chromosomes. Chromosomes with canonical correlation greater than 0.5 on both the training and test set are listed with the number of probes associated with disease free status and the probes form the support of canonical directions.

| Chromosome | Number of probes | Genes | Methylation probes |
|------------|------------------|-------|--------------------|
| 2          | 269              | RRM2, ILF2, ORC6L, SUSD3, SHCBP1 |
|            |                  | cg04779988, cg08022717, cg10414274, cg13052887, cg16297938, cg24011073, cg26364080, cg27066555 |
| 4          | 143              | SLC26A9, C15orf52, NPM2, DNAH11, RAB6B, LIN28, STC2 |
|            |                  | cg04612351, cg14505741, cg1556751, cg15763121, cg17232091 |
| 9          | 89               | RGS6, ORC6L, PTPRH, GPX2, QSOX2, NPM2, SCG3, RAB6B, LICAM, STC2, REG1A |
|            |                  | cg18129806, cg21237984, cg35933628, cg34133594, cg34866237, cg3847987, cg4094457, cg14443051, cg21123355 |
| 10         | 116              | RRM2, SLC26A9, ORC6L, PTPRH, DNAH11, SCG3, LIN28, UMMDL1, C11orf9 |
|            |                  | cg00827318, cg01162610, cg01520297, cg0182629, cg14522796, cg14999031, cg19982462 |
| 12         | 175              | QSOX2, SCG3 |
|            |                  | cg00417147, cg13074783, cg11883138 |
| 15         | 92               | C15orf52, NPM2, DNAH11, SELE, RAB6B |
|            |                  | cg11465484, cg8581777, cg21735516 |
| 18         | 37               | ILF2, SPRK2D, ADCY4, RAB6B, C11orf9, REG1A, SHCBP1 |
|            |                  | cg07740836, cg15531009, cg18953516, cg19363888 |
| 19         | 162              | ORC6L, NPM2, GPHM6 |
|            |                  | cg16392429, cg00555246 |

from The Cancer Genome Atlas project (TCGA, 2012). This dataset consists both DNA methylation and gene expression data for 193 breast cancer patients. The DNA methylation was measured from Illumina Human methylation 450 BeadChip, which contains 482,431 CpG sites that cover 96% of the genome-wide CpG islands. Since no batch effect has either been reported from previous studies or been observed from our analysis, we do not further process the data. For methylation data, there are two popular metrics used to measure methylation levels, β-value and M-value statistics. β-value is defined as the proportion of methylated probes at a CpG site. M-value is defined as the log 2 ratio of the intensities of methylated probe versus un-methylated probe, which is reported as approximately homoscedastic in a previous study Du et al. (2010). We choose to use M-value for methylation data in our analysis.

To investigate the relationship of methylation and gene expression and their interplay with clinical outcomes, we follow the supervised sparse CCA procedure suggested in Witten and Tibshirani (2009). More specifically, we first select methylation probes and genes that are marginally associated with the disease free status by performing a screening on methylation and gene expression data, respectively. There are 135 genes and 4907 methylation probes marginally associated with disease free status with a P-value less than 0.01. We further reduce the number of methylation probes to 3206 by selecting the ones with sample variance greater than 0.5. Compared to the sample size, the number of methylation probes is still too large. To control the dimension of input data, we apply our
Chen et al.

Table 4. Detected methylation probes and their corresponding genes on chromosome 9.

| Probe          | Gene       | Function                                                                 | \( r \) |
|---------------|------------|--------------------------------------------------------------------------|--------|
| cg1728066     | MIR600     | microRNA regulating estrogen factors                                      | 0.282  |
| cg14004457    | MIR455     | microRNA regulating estrogen factors                                      | 0.347  |
| cg2127980, cg13413384 | RXRA | retinoic X receptors                                                     | 0.269, 0.242 |
| cg83603999    | CEB1       | fat catalyzation and vitamin absorption                                   | 0.286  |
| cg13486627    | RGS6TD3    | RNA (guanine-9-) methyltransferase                                        | 0.479  |
| cg13479978    | ABL1       | a protein tyrosine kinase functioned in cell differentiation and stress response | 0.334  |
| cg21123355    | VAV2       | a member of the VAV guanine nucleotide exchange factor family of oncoproteins | 0.312  |
| cg14443041    | Intergenic region |                                                                | 0.384  |

methods to 135 genes with the methylation probes on each chromosome separately. Since it is widely believed that genes operate in biological pathways, the graph for gene expression data is expected to be sparse. We apply the proposed procedure under the sparse precision matrix setting (Section 3.3.1). As we have discussed in the simulation studies, the canonical correlation structure under the sparse precision matrix setting cannot be estimated by the current methods in the literature, such as PMD.

For the purpose of interpretation, the tuning parameters are selected such that a sparse representation of \((\hat{\theta}, \hat{\eta})\) is obtained while the canonical correlation is high. More specifically, we require the number of non-zero genes or probes is less than 10 for each chromosome. We split the data into two halves as a test set and a training set. Then we applied the proposed procedure on the training set. To remove false discoveries, we required the canonical correlation on the test set is greater than 0.5. In total, there are eight chromosomes that have methylation probes satisfying the above criteria (shown in Figure 2). In Table 3, we list genes and methylation probes on each chromosome that form the support of detected canonical directions. A further examination of the genomic coordinates of detected methylation probes reveal the physical closeness of some probes. Some detected probes correspond to the same gene. LPIN1 on chromosome 2, RXRA on chromosome 9, DIP2C on chromosome 10, AACS on chromosome 12, and NFATC1 on chromosome 18 are represented by more than one methylation probes (shown in Figure 2). Moreover, 16 of the 25 genes listed in Table 3 are detected more than once as candidate genes associated with methylation probes. ORC6L, RRM2, RAB6B are independently detected from four chromosomes. All these genes have been proposed as prognosis signature for the metastasis of breast cancer (Weigelt et al., 2005; Ma et al., 2003; van’t Veer et al., 2002). Our results suggest the interplay of their expression with detected methylation sites. We list the functional annotation of probes detected on Chromosome 9 in Table 4 †.

In this analysis, we assume there is one pair of canonical directions between methylation and gene expression. We note that when the underlying canonical correlation structure is

† The corresponding canonical vector on genes RGS6, ORC6L, PTPRHI, GPX2, QSOX2, NPM2, SCG3, RAB6B, L1CAM, STC2, REG1A is \((-0.252, -0.27, -0.286, -0.244, -0.35, -0.282, -0.256, -0.367, -0.256, 0.357, -0.358)\).
low-rank, the pair of canonical directions obtained from the proposed method lie in the subspace of true canonical directions. The extracted canonical directions can still be used to identify sets of methylate sites that are correlated with gene expression.

References

Avants, B. B., P. A. Cook, L. Ungar, J. C. Gee, and M. Grossman (2010). Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with sparse canonical correlation analysis. *Neuroimage* 50(3), 1004–1016.

Bickel, P. J. and E. Levina (2008a). Regularized estimation of large covariance matrices. *The Annals of Statistics* 36(1), 199–227.

Bickel, P. J. and E. Levina (2008b). Covariance regularization by thresholding. *The Annals of Statistics* 36(6), 2577–2604.

Birnbaum, A., I. M. Johnstone, B. Nadler, and D. Paul (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Annals of statistics* 41(3), 1055.

Cai, T. T., W. Liu, and X. Luo (2011). A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* 106(494), 594–607.

Cai, T. T., Z. Ren, and H. H. Zhou (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probability Theory and Related Fields* 156(1-2), 101–143.

Cai, T. T. and M. Yuan (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics* 40(4), 2014–2042.

Cai, T. T., C.-H. Zhang, and H. H. Zhou (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* 38(4), 2118–2144.

Cai, T. T. and H. H. Zhou (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics* 40(5), 2389–2420.

Du, P., X. Zhang, C.-C. Huang, N. Jafari, W. Kibbe, L. Hou, and S. Lin (2010). Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 11(1), 587.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456), 1348–1360.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3), 432–441.

Gao, C., Z. Ma, and H. H. Zhou (2014). Sparse CCA: Adaptive estimation and computational barriers. arXiv preprint arXiv:1409.8365.

Golub, G. H. and C. F. Van Loan (1996). *Matrix computations*. The Johns Hopkins University Press.

Hardoon, D. R. and J. Shawe-Taylor (2011). Sparse canonical correlation analysis. *Machine Learning* 82(3), 331–353.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* 28(3/4), 321–377.

Johnstone, I. (2001). Thresholding for weighted Chi-squared. *Statistica Sinica* 11(3), 691–704.

Johnstone, I. M. and A. Y. Lu (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* 104(486).
Lê Cao, K.-A., P. G. Martin, C. Robert-Granié, and P. Besse (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* **10**(1), 34.

Ma, X.-J., R. Salunga, J. T. Tuggle, J. Gaudet, E. Enright, P. McQuary, T. Payette, M. Pistone, K. Stecker, B. M. Zhang, et al. (2003). Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences* **100**(10), 5974–5979.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics* **41**(2), 772–801.

Massart, P. (2007). *Concentration inequalities and model selection*. Springer.

Meinshausen, N. and P. Bühlmann (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**(3), 1436–1462.

Parkhomenko, E., D. Tritchler, and J. Beyene (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology* **8**(1), 1–34.

Ren, Z., T. Sun, C.-H. Zhang, and H. H. Zhou (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical model. *The Annals of Statistics* **43**(3), 991–1026.

Stewart, G. W. and J.-G. Sun (1990). *Matrix perturbation theory*. Academic Press.

TCGA (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(3), 611–622.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer.

VanderKraats, N. D., J. F. Hiken, K. F. Decker, and J. R. Edwards (2013). Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. *Nucleic Acids Research*.

van’t Veer, L. J., H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**(6871), 530–536.

Waaljenborg, S. and A. H. Zwinderman (2009). Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC Bioinformatics* **10**(1), 315.

Weigelt, B., Z. Hu, X. He, C. Livasy, L. A. Carey, M. G. Ewend, A. M. Glas, C. M. Perou, and L. J. van’t Veer (2005). Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer Research* **65**(20), 9155–9158.

Wiesel, A., M. Kliger, and A. O. Hero III (2008). A greedy approach to sparse canonical correlation analysis. *arXiv preprint arXiv:0801.2748*.

Witten, D., R. Tibshirani, S. Gross, B. Narasimhan, and M. D. Witten (2013). Package “PMA”. *Genetics and Molecular Biology* **8**(1), 28.

Witten, D. M., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**(3), 515–534.

Witten, D. M. and R. J. Tibshirani (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology* **8**(1), 1–27.

Yang, D., Z. Ma, and A. Buja (2013). A sparse SVD method for high-dimensional data. *arXiv preprint arXiv:1112.2433*. 

---

*Chen et al.*