Received: 29 October 2017 | Accepted: 29 January 2019
DOI: 10.1111/biom.13043

BIOMETRIC METHODOLOGY

WILEY Biometrics

An iterative penalized least squares approach to sparse canonical correlation analysis

Qing Mai | **Xin Zhang**

Department of Statistics, Florida State University, Tallahassee, Florida

Correspondence
Qing Mai, Department of Statistics, Florida State University, Tallahassee, FL
Email: mai@stat.fsu.edu

Abstract

It is increasingly interesting to model the relationship between two sets of high-dimensional measurements with potentially high correlations. Canonical correlation analysis (CCA) is a classical tool that explores the dependency of two multivariate random variables and extracts canonical pairs of highly correlated linear combinations. Driven by applications in genomics, text mining, and imaging research, among others, many recent studies generalize CCA to high-dimensional settings. However, most of them either rely on strong assumptions on covariance matrices, or do not produce nested solutions. We propose a new sparse CCA (SCCA) method that recasts high-dimensional CCA as an iterative penalized least squares problem. Thanks to the new iterative penalized least squares formulation, our method directly estimates the sparse CCA directions with efficient algorithms. Therefore, in contrast to some existing methods, the new SCCA does not impose any sparsity assumptions on the covariance matrices. The proposed SCCA is also very flexible in the sense that it can be easily combined with properly chosen penalty functions to perform structured variable selection and incorporate prior information. Moreover, our proposal of SCCA produces nested solutions and thus provides great convenient in practice. Theoretical results show that SCCA can consistently estimate the true canonical pairs with an overwhelming probability in ultra-high dimensions. Numerical results also demonstrate the competitive performance of SCCA.

KEYWORDS

canonical correlation analysis (CCA), LASSO, penalized least squares, sparsity, ultra-high dimensions

1 | INTRODUCTION

Analysis of high-dimensional data has been a major research area for statisticians in the past decade. Topics such as regression, generalized linear models and principal component analysis have been intensively studied. Such research provides insights into how to generalize classical statistical methods to high dimensions. However, existing research often focuses on problems with only one high-dimensional measurement, such as the predictors in regression.

In many contemporary scientific problems, researchers have to model the relationship between two sets of high-dimensional measurements. For example, Witten et al. (2009) analyzed the relationship between gene expression levels and comparative genomic hybridization measurements of breast cancer patients; Hardoon and Shawe-Taylor (2011) investigated the relationship between paired German and Danish texts; Wang et al. (2015) investigated the interactions between groups of genes; and Fang et al. (2016) studied the association between genetic factors and brain activities of schizophrenia patients. In all these problems, we need to discover the

association of the high-dimensional measurements on each other.

Canonical correlation analysis (CCA) is a classical method to analyze the relationship between two multivariate measurements. Consider random vectors $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$. Define $\mathbf{\Sigma_{YX}} = \text{cov}(\mathbf{Y}, \mathbf{X})$, $\mathbf{\Sigma_{XX}} = \text{cov}(\mathbf{X})$ and $\mathbf{\Sigma_{YY}} = \text{cov}(\mathbf{Y})$. For a positive integer $K < \min\{p, q\}$, CCA finds canonical directions $\{\boldsymbol{\alpha}_k^*, \boldsymbol{\beta}_k^*\}_{k=1}^K$ that sequentially maximize the correlation between $(\boldsymbol{\alpha}_k^*)^{\mathrm{T}}\mathbf{Y}$ and $(\boldsymbol{\beta}_k^*)^{\mathrm{T}}\mathbf{X}$. Given the first $(k-1)$ pairs, $\{(\boldsymbol{\alpha}_l^*, \boldsymbol{\beta}_l^*)\}_{l=1}^{k-1}$, the k-th pair is $(\boldsymbol{\alpha}_k^*, \boldsymbol{\beta}_k^*) = \arg\max_{(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)} \boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{\Sigma_{YX}}\boldsymbol{\beta}_k$, s.t. $\boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{\Sigma_{YY}}\boldsymbol{\alpha}_k = 1, \boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{\Sigma_{XX}}\boldsymbol{\beta}_k = 1, \boldsymbol{\alpha}_k^{\mathrm{T}}\mathbf{\Sigma_{YY}}\boldsymbol{\alpha}_l^* = 0, \boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{\Sigma_{XX}}\boldsymbol{\beta}_l^* = 0$ for $l < k$. In practice, CCA substitutes $\mathbf{\Sigma_{YX}}$, $\mathbf{\Sigma_{XX}}$ and $\mathbf{\Sigma_{YY}}$ with their sample estimates $\widehat{\mathbf{\Sigma}}_{\mathbf{YX}}$, $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{YY}}$. When the dimensions are low, $\boldsymbol{\alpha}_k^*, \boldsymbol{\beta}_k^*$ can be estimated through singular value decomposition on $\widehat{\mathbf{\Sigma}}_{\mathbf{YY}}^{-1/2}\widehat{\mathbf{\Sigma}}_{\mathbf{YX}}\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}^{-1/2}$ (Hotelling, 1936).

However, the classical CCA approach cannot analyze high-dimensional datasets, where $p, q \gg n$, because $\widehat{\mathbf{\Sigma}}_{\mathbf{YY}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ are not invertible. On the other hand, linear combinations involving all predictors may be difficult to interpret, while sparse estimates of the canonical pairs are much more desirable. In light of these issues, many efforts have been spent on generalizing CCA to high dimensions (Parkhomenko et al., 2007; Waaijenborg et al., 2008; Le Cao et al., 2009; Witten and Tibshirani, 2009; Witten et al., 2009; Hardoon and Shawe-Taylor, 2011; Chen et al., 2017; Gao et al., 2017). All these methods assume that the canonical pairs are sparse and then employ regularization or thresholding to obtain sparse estimates.

As one of the most popular sparse CCA methods, Witten et al. (2009) proposed the penalized matrix decomposition (PMD) that replaces $\widehat{\mathbf{\Sigma}}_{\mathbf{YY}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ with identity matrices to avoid the singularity in these matrices. PMD then obtains sparse estimates of the canonical directions by penalization. To be exact, PMD estimates $(\boldsymbol{\alpha}_1^*, \boldsymbol{\beta}_1^*)$ by

$$(\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}_1) = \arg\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \boldsymbol{\alpha}^{\mathrm{T}}\widehat{\mathbf{\Sigma}}_{\mathbf{YX}}\boldsymbol{\beta},$$
$$\text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} \le 1, \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} \le 1, P_{\mathbf{Y}}(\boldsymbol{\alpha}) \le \tau_1, P_{\mathbf{X}}(\boldsymbol{\beta}) \le \tau_2,$$

where $P_{\mathbf{Y}}(\cdot), P_{\mathbf{X}}(\cdot)$ are sparsity-inducing penalty functions such as the ℓ_1 penalty. When the first $k-1$ pairs are obtained, one can obtain the k'th pair by properly modifying $\widehat{\mathbf{\Sigma}}_{\mathbf{YX}}$ to remove the variation already explained by the first $k-1$ pairs. PMD can be computed very efficiently and has good performance on many datasets. However, as noted in Chen et al. (2017), the sparse CCA directions from PMD may be inconsistent when $\mathbf{\Sigma_{XX}}$ and $\mathbf{\Sigma_{YY}}$ are far from diagonal, and thus PMD may not be suitable on such data sets. For example, biological measurements often have strong correlations among them and the true covariances may be very different from identity or diagonal matrices. In such cases, it is unclear if PMD can produce accurate estimates of the canonical pairs. Chen et al. (2017) relaxed the diagonal assumption by assuming that $\mathbf{\Sigma_{YY}}$ and $\mathbf{\Sigma_{XX}}$ (or the inverses of them) are sparse.

More recently, Gao et al. (2017) proposed a method named COLAR (standing for Convex programming with group-Lasso Refinement) that does not impose any assumption on the covariance or precision matrices. Given the number of desired pairs $K \ge 1$, define $\mathbf{A}_K = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$, $\mathbf{B}_K = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ and $\mathbf{F}_K = \mathbf{B}_K\mathbf{A}_K^{\mathrm{T}} \in \mathbb{R}^{p\times q}$. We split the data into three batches with equal sizes and compute $\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}^{(j)}, \widehat{\mathbf{\Sigma}}_{\mathbf{YY}}^{(j)}, \widehat{\mathbf{\Sigma}}_{\mathbf{XY}}^{(j)}$ as sample covariances of the jth batch. COLAR then carries out a two-stage analysis. First, it finds a sparse estimate of $\mathbf{F}_K$:

$$\widehat{\mathbf{F}}_K = \arg\max_{\mathbf{F}_K \in \mathbb{R}^{p\times q}} \left\{ \text{Tr}(\widehat{\mathbf{\Sigma}}_{\mathbf{XY}}^{(1)}, \mathbf{F}_K) - \lambda\|\mathbf{F}_K\|_1 \right\},$$
$$\text{s.t. } \|\mathbf{N}\|_* \le K, \|\mathbf{N}\|_{op} \le 1,$$

where $\lambda > 0$, $\mathbf{N} = (\widehat{\mathbf{\Sigma}}_{\mathbf{XX}}^{(1)})^{1/2}\mathbf{F}_K(\widehat{\mathbf{\Sigma}}_{\mathbf{YY}}^{(1)})^{1/2}$, and $\|\mathbf{N}\|_*$ and $\|\mathbf{N}\|_{op}$ are the summation and the maximum of the singular values of $\mathbf{N}$. Second, one decomposes $\widehat{F}_K$ on the second batch of data into $\{\widehat{\mathbf{A}}_K, \widehat{\mathbf{B}}_K\}$, which are further rescaled on the third batch of data.

COLAR is supported by remarkable statistical properties in that it achieves the minimax estimation risk (Gao et al., 2015). However, in the first stage of the two-stage estimation, the parameters of interest, $\mathbf{B}_K, \mathbf{A}_K$, are augmented into a much higher dimensional parameter $\mathbf{F}_K \in \mathbb{R}^{p\times q}$ and it can be computationally demanding when p, q are both large. Another potential disadvantage of COLAR is that it does not produce nested solutions. For example, if we apply COLAR for $K = 1, 2$ to obtain $(\widehat{\mathbf{A}}_1, \widehat{\mathbf{B}}_1) = (\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}_1)_{K=1}$ and $(\widehat{\mathbf{A}}_2, \widehat{\mathbf{B}}_2)$ with $\widehat{\mathbf{A}}_2 = (\widetilde{\boldsymbol{\alpha}}_1, \widetilde{\boldsymbol{\alpha}}_2)_{K=2}$ and $\widehat{\mathbf{B}}_2 = (\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_2)_{K=2}$. Then $\widehat{\mathbf{A}}_1, \widehat{\mathbf{B}}_1$ does not necessarily reside in the subspaces spanned by $\widehat{\mathbf{A}}_2, \widehat{\mathbf{B}}_2$, respectively. Hence, ambiguity arises.

We propose a new, natural yet efficient sparse CCA method in this article. Our proposal significantly differs from existing sparse CCA methods in its formulation, algorithm and analysis. Some notable advantages of the proposed SCCA are as follows. First of all, SCCA is constructed through iterative penalized least squares. The penalized least squares formulation does not involve the covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{YY}}, \widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ directly in optimization. As shown in our algorithm, $\widehat{\mathbf{\Sigma}}_{\mathbf{YY}}, \widehat{\mathbf{\Sigma}}_{\mathbf{XX}}$ appear only as a scale adjustment for the SCCA directions to ensure consistency. Hence, our SCCA approach does not involve any assumption on the covariance matrices, and it is particularly suitable to be applied to data sets with strong dependency. Second, our SCCA is a "direct" generalization of the classical CCA into high-dimensional settings, as it produces nested solutions and immediately reduces to the classical CCA solution once we set the tuning parameters of the penalty terms to zero. Consequently, we can apply it in a sequential fashion as we would perform CCA in the

low-dimensional setting. Third, our proposal is very efficient in terms of computation and storage. It can be efficiently implemented by iteratively solving penalized least squares problems. Moreover, we can use a general class of penalties to incorporate prior information and to impose structure in the final variable selection result. Finally, we provide strong theoretical justifications for our method, where the consistency of estimated SCCA directions is established in ultra-high dimensional settings under mild conditions. The theoretical development is challenging and may be of interest for future research in penalized matrix and tensor decompositions where the goal is to estimate sparse high-dimensional directions for exploring associations.

We also would like to remark that researchers in computer science and engineering also presented various algorithms that connect CCA to least squares problems. See Chu et al. (2013); Lu and Foster (2014); Ma et al. (2015); Sun et al. (2008, 2011) for example. All these works are different from our approach. Lu and Foster (2014); Ma et al. (2015) do not perform variable selection. Chu et al. (2013); Sun et al. (2008, 2011) involve finding the generalized inverse of $\hat{\mathbf{\Sigma}}_{\mathbf{YY}}, \hat{\mathbf{\Sigma}}_{\mathbf{XX}}$ when $p, q > n$, which is not needed in our proposal.

The rest of the article is organized as follows. Section 2 contains our penalized least squares formulation of the high-dimensional SCCA problem, and an efficient algorithm. Section 3 studies the theoretical properties of SCCA. Simulations and real data analysis are presented in Sections 4. Section 5 contains a discussion on SCCA. A data example, implementation details, and all technical proofs are relegated to the supporting information.

2 | METHOD

2.1 | Formulation

Assume that we have independent and identically distributed data $\{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1}^n$. We write the data matrices as $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^{\mathrm{T}}, \mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^{\mathrm{T}}$. Without loss of generality, we assume that $\mathbf{X}$ and $\mathbf{Y}$ are centered. Define the sample covariances $\hat{\mathbf{\Sigma}}_{\mathbf{XX}} = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}, \hat{\mathbf{\Sigma}}_{\mathbf{YY}} = \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}, \hat{\mathbf{\Sigma}}_{\mathbf{YX}} = \frac{1}{n}\mathbf{Y}^{\mathrm{T}}\mathbf{X}$. Our proposal is motivated by the fact that CCA can be recast as a sequential constrained quadratic problem when $n > \max\{p, q\}$. When the dimensions are low, define the solution to the classical CCA as follows:

$$
\begin{aligned}
&(\hat{\boldsymbol{\alpha}}_k^{\mathrm{CCA}}, \hat{\boldsymbol{\beta}}_k^{\mathrm{CCA}}) = \arg\max_{(\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k)} \boldsymbol{\alpha}_k^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{YX}} \boldsymbol{\beta}_k, \\
&\text{s.t. } \boldsymbol{\alpha}_k^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{YY}} \boldsymbol{\alpha}_k = 1, \boldsymbol{\beta}_k^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{XX}} \boldsymbol{\beta}_k = 1, \boldsymbol{\alpha}_k^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{YY}} \hat{\boldsymbol{\alpha}}_l = 0, \\
&\quad \boldsymbol{\beta}_k^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{XX}} \hat{\boldsymbol{\beta}}_l = 0 \text{ for any } l < k. \qquad (1)
\end{aligned}
$$

We show that (1) is equivalent to a constrained quadratic optimization problem.

Lemma 1. *When $p, q < n$, define*

$$
\begin{aligned}
(\hat{\boldsymbol{\alpha}}_k', \hat{\boldsymbol{\beta}}_k') = \arg\min_{\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k} \Bigg\{ & \frac{1}{2n} \sum_{i=1}^n (\mathbf{Y}_i^{\mathrm{T}} \boldsymbol{\alpha}_k - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k)^2 \\
& + \boldsymbol{\alpha}_k^{\mathrm{T}} \left(\sum_{l<k} \hat{\rho}_l' \hat{\mathbf{\Sigma}}_{\mathbf{YY}} \hat{\boldsymbol{\alpha}}_l' \cdot (\hat{\boldsymbol{\beta}}_l')^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{XX}} \right) \boldsymbol{\beta}_k \Bigg\}, \\
& s.t.\ \boldsymbol{\alpha}_k^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{YY}} \boldsymbol{\alpha}_k = 1,\ \boldsymbol{\beta}_k^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{XX}} \boldsymbol{\beta}_k = 1, \qquad (2)
\end{aligned}
$$

where $\hat{\rho}_l' = (\hat{\boldsymbol{\alpha}}_l')^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{YX}} \hat{\boldsymbol{\beta}}_l'$. Then we have $\hat{\boldsymbol{\alpha}}_k' = \hat{\boldsymbol{\alpha}}_k^{CCA}, \hat{\boldsymbol{\beta}}_k' = \hat{\boldsymbol{\beta}}_k^{CCA}$.

The objective function in (2) contains two terms. The first term $\frac{1}{n}\sum_{i=1}^n (\mathbf{Y}_i^{\mathrm{T}} \boldsymbol{\alpha}_k - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k)^2$ along with the two equality constraints measures the linear dependence between $\mathbf{Y}\boldsymbol{\alpha}_k$ and $\mathbf{X}\boldsymbol{\beta}_k$. Hence, a small value of this term forces $\mathbf{Y}\boldsymbol{\alpha}_k$ and $\mathbf{X}\boldsymbol{\beta}_k$ to be strongly correlated. The second term adjusts for the variability explained by the first $k-1$ pairs. The inclusion of this term removes the orthogonality constraints in (1).

When p, q exceeds n, we assume that $(\boldsymbol{\alpha}_k^*, \boldsymbol{\beta}_k^*)$ are sparse in the sense that most entries in them are zeroes. To obtain sparse estimates, we propose sparse CCA (SCCA) as follows:

$$
\begin{aligned}
(\hat{\boldsymbol{\alpha}}_k, \hat{\boldsymbol{\beta}}_k) = \arg\min_{\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k} \{ & \frac{1}{2n} \sum_{i=1}^n (\mathbf{Y}_i^{\mathrm{T}} \boldsymbol{\alpha}_k - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k)^2 \\
& + \boldsymbol{\alpha}_k^{\mathrm{T}} \left(\sum_{l<k} \hat{\rho}_l \hat{\mathbf{\Sigma}}_{\mathbf{YY}} \hat{\boldsymbol{\alpha}}_l \cdot \hat{\boldsymbol{\beta}}_l^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{XX}} \right) \boldsymbol{\beta}_k \\
& + \lambda_{\boldsymbol{\alpha}_k} \|\boldsymbol{\alpha}_k\|_1 + \lambda_{\boldsymbol{\beta}_k} \|\boldsymbol{\beta}_k\|_1 \}, \\
& \text{s.t. } \boldsymbol{\alpha}_k^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{YY}} \boldsymbol{\alpha}_k = 1, \boldsymbol{\beta}_k^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{XX}} \boldsymbol{\beta}_k = 1, \qquad (3)
\end{aligned}
$$

where $\hat{\rho}_l = \hat{\boldsymbol{\alpha}}_k^{\mathrm{T}} \hat{\mathbf{\Sigma}}_{\mathbf{YX}} \hat{\boldsymbol{\beta}}_k$ and $\lambda_{\boldsymbol{\alpha}_k}, \lambda_{\boldsymbol{\beta}_k} > 0$ are tuning parameters.

Our proposal of SCCA combines a loss function that resembles the least squares problem and a penalty that imposes sparsity. We briefly discuss these two ingredients here. First, by Lemma 1, CCA resembles the least squares problem. It will be clear in Section 2.2 that this resemblance helps us connect SCCA to an iterative penalized least squares problem, which greatly facilitates the implementation. This resemblance with least squares also distinguishes our method from PMD and COLAR. PMD is based on a connection between CCA and singular value decomposition, while COLAR shares some flavor with sparse principal component analysis (Johnstone and Lu, 2009; Cai et al., 2013; Ma, 2013; Vu et al., 2013). Second, we apply the ℓ_1 penalty (Tibshirani, 1996) to pursue the sparsity structure. We remark here that other penalty functions can be applied as well, if prior information supports a special sparsity structure. See Section 2.4 for more discussion along this line.

Our proposal of SCCA is a direct approach to performing CCA in high dimensions. First, we directly impose the sparsity assumption on the parameters of interest, while, in

contrast to PMD and Chen et al. (2017), no assumption is imposed on the nuance parameters such as the covariances or the precision matrices. We will later show that, under no sparsity assumption on the covariances, SCCA can consistently estimate the canonical pairs. Second, SCCA directly obtains estimates of the canonical pairs. We do not need to reparametrize $(\boldsymbol{\alpha}_k^*, \boldsymbol{\beta}_k^*)$ as in COLAR. Moreover, similar to CCA, SCCA yields nested solutions. There is no need to determine the number of pairs in advance. We can keep looking for canonical pairs until the canonical correlation is sufficiently small.

2.2 | Algorithm

In this section, we describe an efficient algorithm to implement SCCA. For any positive integer k, define $\widehat{\mathbf{A}}_k = (\widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_k), \widehat{\mathbf{B}}_k = (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_k), \mathbf{R}_k = \mathrm{diag}(\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{YX}}\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{YX}}\widehat{\boldsymbol{\beta}}_k), \boldsymbol{\Omega}_1 = \mathbf{I}_n, \boldsymbol{\Omega}_k = \mathbf{I}_n - \mathbf{Y}\mathbf{A}_{k-1}\mathbf{R}_{k-1}\mathbf{B}_{k-1}\mathbf{X}^{\mathrm{T}}/n$, where $\mathbf{I}_n$ denotes the identity matrix of size $n \times n$. We have the following lemma that plays a central role in our algorithm.

Lemma 2. *Suppose we have obtained the first $k-1$ pairs of solutions to* (3).

1. *If we fix $\boldsymbol{\alpha}_k$, then the minimizer to* (3), $\widehat{\boldsymbol{\beta}}_k = \{(\check{\boldsymbol{\beta}}_k)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\check{\boldsymbol{\beta}}_k\}^{-1/2} \cdot \check{\boldsymbol{\beta}}_k$, *where*

$$\check{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}_k} \left\{ \frac{1}{2n}\|\boldsymbol{\Omega}_k^{\mathrm{T}}\mathbf{Y}\boldsymbol{\alpha}_k - \mathbf{X}\boldsymbol{\beta}_k\|_2^2 + \lambda_{\beta_k}\|\boldsymbol{\beta}_k\|_1 \right\}. \quad (4)$$

2. *If we fix $\boldsymbol{\beta}_k$, then the minimizer to* (3), $\widehat{\boldsymbol{\alpha}}_k = \{(\check{\boldsymbol{\alpha}}_k)^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\check{\boldsymbol{\alpha}}_k\}^{-1/2} \cdot \check{\boldsymbol{\alpha}}_k$, *where*

$$\check{\boldsymbol{\alpha}}_k = \arg\min_{\boldsymbol{\alpha}_k} \left\{ \frac{1}{2n}\|\boldsymbol{\Omega}_k\mathbf{X}\boldsymbol{\beta}_k - \mathbf{Y}\boldsymbol{\alpha}_k\|_2^2 + \lambda_{\alpha_k}\|\boldsymbol{\alpha}_k\|_1 \right\}. \quad (5)$$

Lemma 2 further simplifies our proposed SCCA problem in (3). The problems in (4) & (5) are ℓ_1 penalized least squares problems that can be efficiently solved (Efron et al. (2004), Friedman et al. (2008, e.g)). Then Lemma 2 exhibits that if we fix one canonical direction, the other direction can be found by normalizing the solution to an ℓ_1 penalized least squares problem. Motivated by Lemma 2, we propose the following iterative algorithm:

Algorithm 1. An iterative penalized least squares algorithm for SCCA:

1. Given $\mathbf{A}_{k-1}, \mathbf{B}_{k-1}$, compute $\boldsymbol{\Omega}_k$;
2. Initialize $\{\widehat{\boldsymbol{\alpha}}_k^{(0)}, \widehat{\boldsymbol{\beta}}_k^{(0)}\}$;
3. For $m = 1, \ldots$, repeat the following two steps until convergence:
   (a) Set $\widetilde{\mathbf{Y}}_k^{(m)} = \boldsymbol{\Omega}_k^{\mathrm{T}}\mathbf{Y}\widehat{\boldsymbol{\alpha}}_k^{(m)}$. Compute

$$\check{\boldsymbol{\beta}}_k^{(m)} = \arg\min_{\boldsymbol{\beta}_k} \left\{ \frac{1}{2n}\|\widetilde{\mathbf{Y}}_k^{(m)} - \mathbf{X}\boldsymbol{\beta}_k\|_2^2 + \lambda_{\beta_k}\|\boldsymbol{\beta}_k\|_1 \right\}, \quad (6)$$

   and then set $\widehat{\boldsymbol{\beta}}_k^{(m)} = [\{\check{\boldsymbol{\beta}}_k^{(m)}\}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}}\check{\boldsymbol{\beta}}_k^{(m)}]^{-1/2} \cdot \check{\boldsymbol{\beta}}_k^{(m)}$.
   (b) Set $\widetilde{\mathbf{X}}_k^{(m)} = \boldsymbol{\Omega}_k\mathbf{X}\widehat{\boldsymbol{\beta}}_k^{(m)}$. Compute

$$\check{\boldsymbol{\alpha}}_k^{(m)} = \arg\min_{\boldsymbol{\alpha}_k} \left\{ \frac{1}{2n}\|\widetilde{\mathbf{X}}_k^{(m)} - \mathbf{Y}\boldsymbol{\alpha}_k\|_2^2 + \lambda_{\alpha_k}\|\boldsymbol{\alpha}_k\|_1 \right\}, \quad (7)$$

   and then set $\widehat{\boldsymbol{\alpha}}_k^{(m)} = [\{\check{\boldsymbol{\alpha}}_k^{(m)}\}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{YY}}\check{\boldsymbol{\alpha}}_k^{(m)}]^{-1/2} \cdot \check{\boldsymbol{\alpha}}_k^{(m)}$.
4. Output $(\widehat{\boldsymbol{\alpha}}_k, \widehat{\boldsymbol{\beta}}_k)$ at convergence.

According to Lemma 2, the objective function in Algorithm 1 monotonically decreases in each iteration. Therefore, this algorithm is guaranteed to converge. In our implementation, we solve (6)–(7) by the `R` package `glmnet`.

2.3 | Initialization

Although the sub-problems (6) & (7) are convex, the overall optimization problem in SCCA is non-convex, and different initial values may lead to different solutions. When p, q are only slightly larger than n, we observe that SCCA is reasonably insensitive to the initial value. However, when p, q are very large, we need to be careful with the initial value. To this end, we consider two possible choices of initial values.

The first initialization method sets $\{\widehat{\boldsymbol{\alpha}}_k^{(0)}, \widehat{\boldsymbol{\beta}}_k^{(0)}\}$ to be the first pair of singular vectors of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{YX}} - \sum_{l=1}^{k-1}\widehat{\rho}_l\widehat{\boldsymbol{\alpha}}_l\widehat{\boldsymbol{\beta}}_l^{\mathrm{T}}$. This was indeed the approach adopted by Witten et al. (2009). We observe that this method works well when the pairs are not overly sparse. When the canonical pairs are very sparse, the second method to be discussed in the following may produce superior results. Suppose that we have obtained $k-1$ pairs $(\widehat{\boldsymbol{\alpha}}_j, \widehat{\boldsymbol{\beta}}_j)_{j=1}^{k-1}$. Then we define $\widehat{\boldsymbol{\Sigma}}_{\mathbf{YX}}^{(k-1)} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{YX}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{YY}}(\sum_{j=1}^{k-1}\widehat{\rho}_j\widehat{\boldsymbol{\alpha}}_j\widehat{\boldsymbol{\beta}}_j^{\mathrm{T}})\widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} = (\widehat{\sigma}_{\mathbf{YX},lm}^{(k-1)})$. Define γ to be the $\sqrt{n}$'th largest entry of $|\widehat{\sigma}_{\mathbf{YX},lm}^{(k-1)}|, l = 1, \ldots, q; m = 1, \ldots, p$. We identify the sets

$$\mathbf{D}_{\mathbf{Y}}^{(k)} = \{l : \text{There exists } m \text{ s.t. } |\widehat{\sigma}_{\mathbf{YX},lm}^{(k-1)}| \geq \gamma \text{ or there exists } j \text{ s.t. } \widehat{\alpha}_{jl} \neq 0\},$$
$$\mathbf{D}_{\mathbf{X}}^{(k)} = \{m : \text{There exists } l \text{ s.t. } |\widehat{\sigma}_{\mathbf{YX},lm}^{(k-1)}| > \gamma \text{ or there exists } j \text{ s.t. } \widehat{\beta}_{jl} \neq 0\}.$$

It can be seen that $\mathbf{D}_{\mathbf{Y}}^{(k)}, \mathbf{D}_{\mathbf{X}}^{(k)}$ preserve the predictors that are marginally highly correlated or have shown correlation in the previous $k-1$ pairs. Then we perform singular value decomposition on $\{\widehat{\boldsymbol{\Sigma}}_{\mathbf{YX}}^{(k-1)}\}_{\mathbf{D}_{\mathbf{Y}}^{(k)}, \mathbf{D}_{\mathbf{X}}^{(k)}}$ and use the first pair of singular vectors as initial values in our SCCA method. We refer to this initial value as the restricted singular vectors.

As discussed, the comparison between the two initial values depends on the sparsity level of the true parameters. Hence, in practice where we have no knowledge on the truth, users are suggested to pick one from them by cross validation.

2.4 | Structured variable selection in sparse CCA

In many applications, we have additional information on the sparsity structure. If so, penalties other than the ℓ_1 penalty may be favored. For example, Chen et al. (2012) considered case where the predictors are grouped and applied group lasso (Yuan and Lin, 2006) in PMD; Witten et al. (2009) argued that in some biological problems, the coefficients should be smooth and fused lasso can be employed for this purpose (Tibshirani et al., 2005).

Our proposal of SCCA can be easily extended to such cases as well, with some proper modification on the algorithm. To see this, we consider a more general form of SCCA:

$$\begin{aligned}(\widehat{\boldsymbol{\alpha}}_k, \widehat{\boldsymbol{\beta}}_k) = \arg\min_{\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k} \{ & \frac{1}{2n} \sum_{i=1}^{n} (\mathbf{Y}_i^{\mathrm{T}} \boldsymbol{\alpha}_k - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k)^2 \\ & + \boldsymbol{\alpha}_k^{\mathrm{T}} \left(\sum_{l<k} \widehat{\rho}_l \widehat{\boldsymbol{\Sigma}}_{\mathbf{YY}} \widehat{\boldsymbol{\alpha}}_l \cdot \widehat{\boldsymbol{\beta}}_l^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} \right) \boldsymbol{\beta}_k \\ & + P_{\mathbf{Y}}(\boldsymbol{\alpha}_k) + P_{\mathbf{X}}(\boldsymbol{\beta}_k) \}, \\ & \text{s.t. } \boldsymbol{\alpha}_k^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{YY}} \boldsymbol{\alpha}_k = 1, \boldsymbol{\beta}_k^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} \boldsymbol{\beta}_k = 1, \end{aligned} \tag{8}$$

where $P_{\mathbf{Y}}(\boldsymbol{\alpha}_k)$ and $P_{\mathbf{X}}(\boldsymbol{\beta}_k)$ are properly chosen penalty functions to impose the sparsity structure. We consider the following general class of penalty functions.

(C0) For any constant $C > 0$ and $\mathbf{w} \in \mathbb{R}^q, \mathbf{v} \in \mathbb{R}^p$, the penalty functions satisfy $P_{\mathbf{Y}}(C\mathbf{w}) = CP_{\mathbf{Y}}(\mathbf{w}), P_{\mathbf{X}}(C\mathbf{v}) = CP_{\mathbf{X}}(\mathbf{v})$.

Condition (C0) is satisfied by many popular penalties, such as group lasso, fused lasso and adaptive lasso (Zou, 2006). We derive the following lemma that generalizes Lemma 2.

Lemma 3. *Suppose we have obtained the first $k-1$ pairs of solutions to (8), under Condition (C0), we have the following conclusions:*

1. *If we fix $\boldsymbol{\alpha}_k$, then the solution to (8) is $\widehat{\boldsymbol{\beta}}_k = \{(\breve{\boldsymbol{\beta}}_k)^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} \breve{\boldsymbol{\beta}}_k\}^{-1/2} \cdot \breve{\boldsymbol{\beta}}_k$, where*

$$\breve{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}_k} \left\{ \frac{1}{2n} \|\boldsymbol{\Omega}_k^{\mathrm{T}} \mathbf{Y} \boldsymbol{\alpha}_k - \mathbf{X} \boldsymbol{\beta}_k\|_2^2 + P_{\mathbf{X}}(\boldsymbol{\beta}_k) \right\}. \tag{9}$$

2. *If we fix $\boldsymbol{\beta}_k$, then the solution to (8) is $\widehat{\boldsymbol{\alpha}}_k = \{(\breve{\boldsymbol{\alpha}}_k)^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{\mathbf{XX}} \breve{\boldsymbol{\alpha}}_k\}^{-1/2} \cdot \breve{\boldsymbol{\alpha}}_k$, where*

$$\breve{\boldsymbol{\alpha}}_k = \arg\min_{\boldsymbol{\alpha}_k} \left\{ \frac{1}{2n} \|\boldsymbol{\Omega}_k \mathbf{X} \boldsymbol{\beta}_k - \mathbf{Y} \boldsymbol{\alpha}_k\|_2^2 + P_{\mathbf{Y}}(\boldsymbol{\alpha}_k) \right\}. \tag{10}$$

Lemma 3 shows that SCCA with any penalty functions satisfying Condition (C0) can be implemented by iteratively solving penalized least squares problems. For the sake of space, such an algorithm (Algorithm 2) is relegated to the supporting information.

If one wishes to apply the group lasso penalty or the fused lasso penalty, it can be easily achieved by borrowing existing algorithms to solve (9)–(10). For example, if fused lasso is applied, one can employ the algorithm in Tibshirani and Taylor (2011). If group lasso is considered appropriate, the algorithm in Yang and Zou (2015) can be borrowed to solve the penalized least squares problems.

In practice, the penalty function should be chosen based on the researchers' knowledge of the specific dataset of interest. For example, if the variables are known to contain groups that function together, the group lasso penalty should be used to perform groupwise selection in which variables in the same group are selected or excluded at the same time. If the variables are arranged in a way such that variables next to each other are expected to have similar effects, the fused lasso can be applied to encourage segments of variables to be excluded and the selected variables to have smooth coefficients. If the ℓ_1 penalized SCCA appears to be unstable in cross validation, the adaptive lasso could be a remedy.

3 | THEORIES

In this section, we present theoretical results of SCCA. Although many penalty functions can be applied in SCCA, we consider the ℓ_1 penalized SCCA in (3) to fix ideas. Recall that $(\boldsymbol{\alpha}_k^*, \boldsymbol{\beta}_k^*)_{k=1}^K$ are the canonical pairs we aim to estimate. Our proofs repeatedly make use of the fact (Chen et al., 2017) that for $\{\boldsymbol{\alpha}_k^*, \boldsymbol{\beta}_k^*\}_{k=1}^K$, we must have $\boldsymbol{\Sigma}_{\mathbf{YX}} = \boldsymbol{\Sigma}_{\mathbf{YY}} \left\{ \sum_{k=1}^K \rho_k^* \boldsymbol{\alpha}_k^* (\boldsymbol{\beta}_k^*)^{\mathrm{T}} \right\} \boldsymbol{\Sigma}_{\mathbf{XX}}$. We remark here that our proposal of SCCA does not require the knowledge of K, in the sense that if user wishes to only obtain $K_1 \le K$ pairs, it can be shown that $\{\widehat{\boldsymbol{\alpha}}_k, \widehat{\boldsymbol{\beta}}_k\}_{k=1}^{K_1}$ are consistent estimates of $\{\boldsymbol{\alpha}_k^*, \boldsymbol{\beta}_k^*\}_{k=1}^{K_1}$. However, for ease of presentation, we consider the estimation of K pairs.

Now we introduce some notations. Define $\tau_0 = \max_k \{\|\boldsymbol{\alpha}_k^*\|_1, \|\boldsymbol{\beta}_k^*\|_1\}$. We use $C > 0$ to denote a generic positive constant that could vary from line to line. For simplicity, we assume that all tuning parameters are equal to $\lambda > 0$, but our conclusion still holds if all tuning parameters

decay to zero at the same order. For the sake of space, technical conditions (C1)–(C4) are listed in the supporting information, where readers can also find a discussion of their implications. With these mild conditions, we have the following result.

Theorem 1. *Under Conditions (C1)–(C3), for any* $\epsilon > 0$, *there exist* $0 < C_L < C_U$ *and* $\epsilon_0 > 0$ *such that if* $0 < \epsilon < \epsilon_0$, *and* $C_L\tau_0\epsilon < \lambda < C_U\tau_0\epsilon$, *with a probability greater than* $1 - C(p+q)^2\exp(-Cn\epsilon^2)$, *there exists a local minimizer to* (3), $(\widehat{\boldsymbol{\alpha}}_k, \widehat{\boldsymbol{\beta}}_k)_{k=1}^K$, *such that* $\cos^2\{\angle(\widehat{\boldsymbol{\alpha}}_k, \boldsymbol{\alpha}_k^*)\} \geq 1 - C\tau_0^2\epsilon, \cos^2\{\angle(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k^*)\} \geq 1 - C\tau_0^2\epsilon$ *for* $k = 1, \ldots, K$.

To better see the implication of Theorem 1, we translate it to asymptotic results.

Corollary 1. *Let* $b_n = \tau_0\dfrac{\log(p+q)}{n}$. *Under Conditions (C1)–(C4), there exist* $0 < C_L < C_U$ *such that if* $C_L b_n < \lambda < C_U b_n$ *then there exists a local minimizer to* (3), $(\widehat{\boldsymbol{\alpha}}_k, \widehat{\boldsymbol{\beta}}_k)_{k=1}^K$, *such that* $\cos\{\angle(\widehat{\boldsymbol{\alpha}}_k, \boldsymbol{\alpha}_k^*)\} \to 1, \cos\{\angle(\widehat{\boldsymbol{\beta}}_k, \boldsymbol{\beta}_k^*)\} \to 1$ *for* $k = 1, \ldots, K$ *with a probability tending to 1.*

Corollary 1 shows that, SCCA consistently estimates the canonical pairs even when the dimension grows at an exponential rate of the sample size. This result does not require any sparsity assumption on the covariance matrices. Such a result supports the application of SCCA on high-dimensional datasets, especially when strong dependence is present.

4 | NUMERICAL STUDIES

4.1 | Simulations

In this section we demonstrate the application of sparse CCA on simulated datasets. All of our simulations are based on 200 replicates. In all the simulations, we set $n = 500$, but p, q range from 300 to 1500. Recall that, by Chen et al. (2017), we must have $\boldsymbol{\Sigma}_{\mathbf{YX}} = \boldsymbol{\Sigma}_{\mathbf{YY}}\left\{\sum_{k=1}^K \rho_k^*\boldsymbol{\alpha}_k^*(\boldsymbol{\beta}_k^*)^{\mathrm{T}}\right\}\boldsymbol{\Sigma}_{\mathbf{XX}}$. Therefore, in each simulation setting, we specify $\boldsymbol{\Sigma}_{\mathbf{XX}}, \boldsymbol{\Sigma}_{\mathbf{YY}}$ and $\{\boldsymbol{\alpha}_k^*, \boldsymbol{\beta}_k^*\}_{k=1}^K$. Then we simulate $(\mathbf{Y}, \mathbf{X}) \sim N(0, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{YY}} & \boldsymbol{\Sigma}_{\mathbf{YY}}\{\sum_{k=1}^K \rho_k^*\boldsymbol{\alpha}_k^*(\boldsymbol{\beta}_k^*)^{\mathrm{T}}\}\boldsymbol{\Sigma}_{\mathbf{XX}} \\ \boldsymbol{\Sigma}_{\mathbf{XX}}\{\sum_{k=1}^K \rho_k^*\boldsymbol{\beta}_k^*(\boldsymbol{\alpha}_k^*)^{\mathrm{T}}\}\boldsymbol{\Sigma}_{\mathbf{YY}} & \boldsymbol{\Sigma}_{\mathbf{XX}} \end{pmatrix}$$

To evaluate the methods, write $\mathbf{A}^* = (\boldsymbol{\alpha}_1^*, \ldots, \boldsymbol{\alpha}_K^*)$ and $\mathbf{B}^* = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_K^*)$. For any matrix $\mathbf{M}$, write $\mathbf{P}_{\mathbf{M}}$ as the projection matrix onto the column space of $\mathbf{M}$. For all the methods with estimates $\widehat{\mathbf{A}} = (\widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_K), \widehat{\mathbf{B}} = (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)$, we compare $\|\mathbf{P}_{\mathbf{A}^*} - \mathbf{P}_{\widehat{\mathbf{A}}}\|_F$ and $\|\mathbf{P}_{\mathbf{B}^*} - \mathbf{P}_{\widehat{\mathbf{B}}}\|_F$.

For easy computation, we generated an independent validation set $(\mathbf{X}^{val}, \mathbf{Y}^{val})$ with the same size of the training set to choose the tuning parameters for all the methods included for comparison. In SCCA and PMD, we let $\lambda_{\boldsymbol{\alpha}_k} = \lambda_{\boldsymbol{\beta}_k} = \lambda_k, k = 1, \ldots, K$. Given $\{\widehat{\boldsymbol{\alpha}}_l, \widehat{\boldsymbol{\beta}}_l\}, l < k$, λ_k is chosen as follows. For a candidate set of tuning parameters $\lambda_k^{(1)}, \ldots, \lambda_k^{(M)}$, we obtain the corresponding $\{\widehat{\boldsymbol{\alpha}}_k^{(m)}, \widehat{\boldsymbol{\beta}}_k^{(m)} \mid \lambda_k^{(m)}\}_{m=1}^M$ and evaluate the correlations $\widehat{\rho}_m^{val} = \widehat{\mathrm{cor}}\{\mathbf{Y}^{val}\widehat{\boldsymbol{\alpha}}_k, \mathbf{X}^{val}\widehat{\boldsymbol{\beta}}_k \mid \lambda_k^{(m)}\}$ on the validation set. The $\lambda_k^{(m)}$ with the largest $\widehat{\rho}_m^{val}$ is used in the final estimation. In COLAR, there are two tuning parameters $\lambda_{(1)}, \lambda_{(2)}$ that generate all the K pairs simultaneously. For each candidate pair of tuning parameters, we compute the canonical pairs $\{\widehat{\boldsymbol{\alpha}}_k, \widehat{\boldsymbol{\beta}}_k \mid \lambda_{(1)}, \lambda_{(2)}\}_{k=1}^K$ and evaluate $\widehat{\phi}^{val}\{\lambda_{(1)}, \lambda_{(2)}\} = \sum_{k=1}^K \widehat{\mathrm{cor}}^2\{\mathbf{Y}^{val}\widehat{\boldsymbol{\alpha}}_k, \mathbf{X}^{val}\widehat{\boldsymbol{\beta}}_k \mid \lambda_{(1)}, \lambda_{(2)}\}$ on the validation set. The maximizer of $\widehat{\phi}^{val}\{\lambda_{(1)}, \lambda_{(2)}\}$ is used in the final estimate.

We consider two scenarios: moderately high dimensions and very high dimensions. A comparison of computation time is also given at the end of this section.

Moderately high dimensions. We start with four models where the dimensions are only moderately high at $p = q = 300$, but the covariance matrices can be very different from identity matrices. Similar simulations with higher dimensions $p = q = 600$ are included in Web Appendix C in the supporting information. In all these four models, $\boldsymbol{\alpha}^* = \boldsymbol{\beta}^* \in \mathbb{R}^{p\times 2}$ are only nonzero at rows $(1, 6, 11, 16, 21)$. In particular, we first choose $\boldsymbol{\eta}^* \in \mathbb{R}^{p\times 2}$ with $\boldsymbol{\eta}^*_{(1,6,11,16,21),1} \propto (-2, -1, -1, 2, 2)$, $\boldsymbol{\eta}^*_{(1,6,11,16,21),2} \propto (0, 0, 0, 1, 1)$ and all other entries of $\boldsymbol{\eta}^*$ are zeros. In each model setting, we normalize $\boldsymbol{\eta}^*$ with respect to the chosen covariance matrices to obtain $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$. The canonical correlations are $(0.9, 0.8)$.

Model 1 (Identity covariances): $p = q = 300$. $\boldsymbol{\Sigma}_{\mathbf{YY}} = \boldsymbol{\Sigma}_{\mathbf{XX}} = \mathbf{I}$.

Model 2 (Moderate correlation): $p = q = 300$. $\boldsymbol{\Sigma}_{\mathbf{YY}} = \boldsymbol{\Sigma}_{\mathbf{XX}} = AR(0.3)$.

Model 3 (High correlation): $p = q = 300$. $\boldsymbol{\Sigma}_{\mathbf{YY}} = \boldsymbol{\Sigma}_{\mathbf{XX}} = AR(0.8)$.

Model 4 (Sparse precision matrices): $p = q = 300$. $\boldsymbol{\Sigma}_{\mathbf{XX}} = \boldsymbol{\Sigma}_{\mathbf{YY}} = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}_0\boldsymbol{\Lambda}$, $\boldsymbol{\Sigma}_0 = \boldsymbol{\Omega}^{-1}$, $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\Sigma}_0)^{-1/2}$ where $\omega_{jj} = 1, \omega_{jk} = 0.5$ if $|j-k| = 1$ and $\omega_{jk} = 0.4$ if $|j-k| = 2$.

Note that Models 1, 2 & 4 have been used to demonstrate COLAR in Gao et al. (2017) as well. We compare our SCCA proposal with COLAR (Gao et al., 2017) and PMD (Witten et al., 2009). Gao et al. (2017) suggested that in practice COLAR should be conducted without sample splitting to achieve higher estimation accuracy. We follow this suggestion. As suggested by a referee, we also include the (unpenalized) CCA in our comparison.

The results are reported in Table 1. CCA has very poor performance in all the models, as it does not enforce sparsity. This fact further confirms the importance of developing

TABLE 1 Simulation results for Models 1–4. The reported numbers are the medians and standard errors (in parentheses) of $\text{Err}(\widehat{\mathbf{A}}) = \|\mathbf{P}_{\widehat{\mathbf{A}}} - \mathbf{P}_{\mathbf{A}}\|_F$ and $\text{Err}(\widehat{\mathbf{B}}) = \|\mathbf{P}_{\widehat{\mathbf{B}}} - \mathbf{P}_{\mathbf{B}}\|_F$, over 200 replicates.

	Model 1		Model 2		Model 3		Model 4	
$p, q = 300$	$\text{Err}(\widehat{\mathbf{A}})$	$\text{Err}(\widehat{\mathbf{B}})$	$\text{Err}(\widehat{\mathbf{A}})$	$\text{Err}(\widehat{\mathbf{B}})$	$\text{Err}(\widehat{\mathbf{A}})$	$\text{Err}(\widehat{\mathbf{B}})$	$\text{Err}(\widehat{\mathbf{A}})$	$\text{Err}(\widehat{\mathbf{B}})$
SCCA	0.1149	0.1155	0.1129	0.1158	0.2156	0.2274	0.1510	0.1594
	(0.0031)	(0.0032)	(0.0050)	(0.0029)	(0.0076)	(0.0087)	(0.0038)	(0.0043)
Colar	0.1371	0.1430	0.1386	0.1420	0.2193	0.2242	0.1849	0.1864
	(0.0026)	(0.0026)	(0.0027)	(0.0026)	(0.0048)	(0.0045)	(0.0034)	(0.0030)
PMD	0.3018	0.3146	0.3848	0.4198	1.0737	1.0961	1.5931	1.5888
	(0.0191)	(0.0201)	(0.0187)	(0.0156)	(0.0137)	(0.0168)	(0.0073)	(0.0060)
CCA	1.9862	1.9867	1.9918	1.9910	1.9945	1.9940	1.9945	1.9941
	(0.0015)	(0.0015)	(0.0005)	(0.0006)	(0.0003)	(0.0004)	(0.0004)	(0.0004)

sparse methods in the presence of high dimensions. As for the sparse methods, SCCA uniformly outperforms the two competitors. The SCCA estimator is significantly closer to the true parameters than the estimates given by COLAR and PMD. It is also worth noting that from Model 1 to Model 4, the covariance structure gradually deviates from identity matrices and the performance of PMD deteriorates in the meantime. On the other hand, SCCA and COLAR produce accurate estimates across all the covariance structures. This confirms that methods with no assumption on the covariance matrices have broader applicability.

Very high dimensions. Now we consider four models where the dimensions are very high. The dimensions $p = q$ are set to $1000, 1200, 1500, 2000$. We denote $\mathbf{1}_r$ as an r-dimensional vectors with all entries being 1, and $\mathbf{0}_r$ as an r-dimensional vectors with all entries being 0. The canonical correlation of the first pair is 0.9, and that of the second pair (if exists) is 0.8.

Model 5: $\mathbf{\Sigma_{XX}} = \mathbf{\Sigma_{YY}} = \mathbf{I}$. There is one canonical pair $\boldsymbol{\alpha}_1^* = \boldsymbol{\beta}_1^* \propto (\mathbf{1}_4, \mathbf{0}_{p-4})$.

Model 6: $\mathbf{\Sigma_{XX}} = \mathbf{\Sigma_{YY}} = AR(0.5)$. There is one canonical pair $\boldsymbol{\alpha}_1^* = \boldsymbol{\beta}_1^* \propto (\mathbf{1}_8, \mathbf{0}_{p-8})$.

Model 7: $\mathbf{\Sigma_{XX}} = \mathbf{\Sigma_{YY}} = AR(0.5)$. There are two canonical pairs. For $\boldsymbol{\eta} \in \mathbb{R}^{p\times 2}$, set $\boldsymbol{\eta}_{(1,2,3,4),1} = \boldsymbol{\eta}_{(51,52,53,54),2} = \mathbf{1}_4$. Then $\boldsymbol{\eta}$ is normalized with respect to $\mathbf{\Sigma_{XX}}$ to obtain $\boldsymbol{\alpha}^* = \boldsymbol{\beta}^*$.

Model 8: $\mathbf{\Sigma_{XX}} = \mathbf{\Sigma_{YY}} = CS(0.5)$. There are two canonical pairs. For $\boldsymbol{\eta} \in \mathbb{R}^{p\times 2}$, set $\boldsymbol{\eta}_{(1,2,3,4),1} = \boldsymbol{\eta}_{(51,52,53,54),2} = \mathbf{1}_4$. Then $\boldsymbol{\eta}$ is normalized with respect to $\mathbf{\Sigma_{XX}}$ to obtain $\boldsymbol{\alpha}^* = \boldsymbol{\beta}^*$.

When the dimensions exceed 1000, it takes a long time to run COLAR. Therefore, we only consider SCCA and PMD on these very high-dimensional settings. The original PMD uses the singular vectors as initial values, but SCCA considers both the singular vectors and the restricted singular vectors as initials (See Section 2.3 for details). As suggested by a referee, we further consider PMD with the restricted singular values in Models 5 & 6. We denote this method as PMD*. We do not apply PMD* in Models 7 & 8 because there are two canonical pairs in these two models. Since the initial value for the second pair depends on the first pair, the two methods will have different initial values even if both of them use restricted singular vectors. The simulation results are reported in Table 2. It can be seen that again SCCA uniformly outperforms PMD. This observation suggests that SCCA continues to produce accurate estimates even when the dimensions are very high. Moreover, SCCA is more resistant to both the increase of dimensions and the change of the correlation structure. On the other hand, PMD* is much better than PMD in Model 5, where the covariance matrices are identities. This fact suggests that, when the identity covariance assumptions is met, the performance of PMD may be improved by carefully chosen initial values, but a thorough study along this line is out of the scope of this article. In Model 6, PMD* is only slightly better than PMD. We suspect that this is because the identity covariance matrix assumption is violated in this model, and the resulting bias cannot be corrected by different choices of initial values. Nevertheless, SCCA is still better than PMD* in both models.

Computation costs. Now we illustrate the computational efficiency and scalability of our algorithm. We consider Model 6 in our simulations with $n = 500$ and dimension $p = q$ varying from 200 to 1000. For each dimension, we generated 20 datasets and ran each method with tuning parameter selected from (different) candidate pools of five tuning parameters, except for COLAR with $p = q = 1000$ we only generated 5 datasets because of its high computation costs. More details on the comparison can be found in supporting information.

The averaged CPU time elapsed on the log scale for fitting one data set with each methods are reported in Figure 1 (left panel), which shows drastically different computational costs for the three methods. To illustrate the scalability, Figure 1 also summarizes the computing time ratios of COLAR versus SCCA and PMD versus SCCA. SCCA is not only much faster but also more scalable than COLAR: it is over 200 times faster than COLAR at the moderately high-dimension, $p = q = 200$, and is even about 5000 times faster than COLAR at the very high dimension, $p = q = 1000$. In the meantime, because SCCA does not make any assumption on the covariance, it is slower than PMD. However, the ratio between these two methods decreases slowly as the dimensions increase. In our

TABLE 2 Simulation results for Models 5–8. The reported numbers are the medians and standard errors (in parentheses) of $\text{Err}(\widehat{\mathbf{A}}) = \|\mathbf{P}_{\widehat{\mathbf{A}}} - \mathbf{P}_{\mathbf{A}}\|_F$ and $\text{Err}(\widehat{\mathbf{B}}) = \|\mathbf{P}_{\widehat{\mathbf{B}}} - \mathbf{P}_{\mathbf{B}}\|_F$, over 200 replicates.

	(a) $p = q = 1000$		(b) $p = q = 1200$		(c) $p = q = 1500$		(d) $p = q = 2000$	
	$\text{Err}(\widehat{\mathbf{A}})$	$\text{Err}(\widehat{\mathbf{B}})$	$\text{Err}(\widehat{\mathbf{A}})$	$\text{Err}(\widehat{\mathbf{B}})$	$\text{Err}(\widehat{\mathbf{A}})$	$\text{Err}(\widehat{\mathbf{B}})$	$\text{Err}(\widehat{\mathbf{A}})$	$\text{Err}(\widehat{\mathbf{B}})$
	Model 5							
SCCA	0.0597	0.0624	0.0635	0.0621	0.0662	0.0667	0.0655	0.0622
	(0.0026)	(0.0030)	(0.0026)	(0.0016)	(0.0032)	(0.0026)	(0.0021)	(0.0034)
PMD	1.4142	1.4142	1.4142	1.4142	1.4142	1.4142	1.4142	1.4142
	(0.0771)	(0.0770)	($< 10^{-4}$)	($< 10^{-4}$)	($< 10^{-4}$)	($< 10^{-4}$)	($< 10^{-4}$)	($< 10^{-4}$)
PMD*	0.1344	0.1423	0.1470	0.1405	0.1445	0.1465	0.1971	0.1892
	(0.0048)	(0.0052)	(0.0050)	(0.0054)	(0.0057)	(0.0063)	(0.0033)	(0.0038)
	Model 6							
SCCA	0.2101	0.2053	0.2060	0.2125	0.2052	0.2105	0.2067	0.2150
	(0.0070)	(0.0044)	(0.0048)	(0.0051)	(0.0050)	(0.0047)	(0.0057)	(0.0047)
PMD	0.2885	0.2927	0.3107	0.3134	0.3152	0.3187	0.3527	0.3510
	(0.0056)	(0.0045)	(0.0037)	(0.0042)	(0.0053)	(0.0073)	(0.0151)	(0.0200)
PMD*	0.2873	0.2914	0.2996	0.3046	0.2932	0.2931	0.3035	0.2986
	(0.0048)	(0.0046)	(0.0038)	(0.0049)	(0.0040)	(0.0024)	(0.0058)	(0.0035)
	Model 7							
SCCA	0.2271	0.2430	0.2461	0.2430	0.2500	0.2500	0.2347	0.2482
	(0.0057)	(0.0068)	(0.0062)	(0.0072)	(0.0070)	(0.0067)	(0.0062)	(0.0093)
PMD	0.5440	0.5769	0.7592	0.7519	0.8571	0.8623	1.1574	1.1855
	(0.0476)	(0.0431)	(0.1012)	(0.0913)	(0.0924)	(0.0891)	(0.1408)	(0.1415)
	Model 8							
SCCA	0.4097	0.4064	0.4022	0.3987	0.4178	0.4231	0.4085	0.4257
	(0.0103)	(0.0104)	(0.0094)	(0.0116)	(0.0083)	(0.0114)	(0.0097)	(0.0127)
PMD	0.6366	0.6084	0.6334	0.6275	0.7104	0.6852	0.7582	0.8458
	(0.0112)	(0.0111)	(0.0168)	(0.0162)	(0.0182)	(0.0198)	(0.0253)	(0.0392)

experiment, when we varied the dimensions from $p = q = 200$ to $p = q = 1000$, the ratio slightly decreases from 0.24 to 0.18. It is clear that SCCA scales well with the dimension. Also, note that SCCA has consistently better estimation accuracy than PMD when the covariance matrices deviate from identities.

4.2 | A real data example

We demonstrate the application of SCCA on the breast cancer data (Chin et al., 2006), which is included in the R package PMA. This dataset includes 89 samples, with 19672 gene expression measurements and 2149 comparative genomic

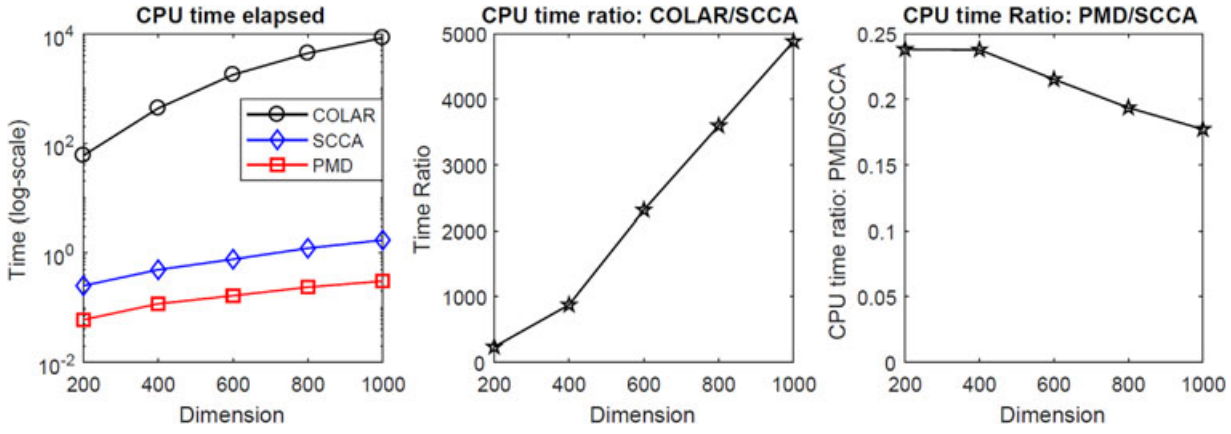

FIGURE 1 Elapsed CPU time for fitting each method. Left panel summarizes the averaged time (in log-scale) for each method to fit one data set with tuning based on five candidate tuning parameters. The averaged CPU time elapsed for each methods are (in seconds): 59.06 (COLAR), 0.25 (SCCA) and 0.06 (PMD) when $p = q = 200$; and 8419.3 (COLAR), 1.72 (SCCA) and 0.31 (PMD) when $p = q = 1000$. The middle and right panels summarize the computing time ratios of COLAR versus SCCA and PMD versus SCCA. This figure appears in color in the electronic version of this article.

hybridization (CGH) measurements. The genes and CGH spots locate on 23 chromosomes. For demonstration, we conduct the same analysis on the first two chromosomes, focusing on one chromosome at a time. The analysis of the other chromosomes can be carried out similarly. We let $\mathbf{X}$ (with dimension p) be the CGH measurements, and $\mathbf{Y}$ (with dimension q) be the gene expression measurements. The dimensions are $p = 136, q = 1942$ and $p = 72, q = 1329$ for the two chromosomes, respectively.

We apply SCCA and PMD on these two chromosomes. COLAR was not included because of its computational cost. In both SCCA and PMD, we use the ℓ_1 penalty. Following Witten et al. (2009), we restrict our attention to the first canonical pair and use the (unpenalized) SVD solution as the initial value. For simplicity, we use equal tuning parameters $\lambda_{\alpha_1} = \lambda_{\beta_1}$. For SCCA, the tuning parameter is chosen by cross validation, while for PMD, the tuning parameter is chosen by the function `CCA.permute` in their `R` package `PMA`.

Both methods yield high canonical correlations on the two chromosomes (see the row "Correlation on the full dataset" in Table 3). To validate these results, we adopt the bootstrap method proposed in Witten et al. (2009) to produce p-values. We permute the dataset N times to refit SCCA/PMD and obtain $\hat{\rho}^i, i = 1, \ldots, N$. Then we compute p-value as $\frac{1}{N}\sum_{i=1}^{N} 1(\hat{\rho}^i > \hat{\rho})$, where $\hat{\rho}$ is the canonical correlation obtained on the original dataset with the corresponding method. All the p-values are much smaller than 0.05 (see the row "p-value" in Table 3), indicating that our discovery is significant. Hence, the correlations are significant between gene expressions and CGH measurements within each chromosome.

Further, we repeatedly split the data into training sets $(\mathbf{Y}^{\text{train}}, \mathbf{X}^{\text{train}})$ and testing sets $(\mathbf{Y}^{\text{test}}, \mathbf{X}^{\text{test}})$ with a 3:1 ratio for 200 times. In each replicate, we find $(\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}_1)$ on $(\mathbf{Y}^{\text{train}}, \mathbf{X}^{\text{train}})$, and evaluate the correlations $\widehat{cor}(\mathbf{Y}^{\text{test}}\widehat{\boldsymbol{\alpha}}_1, \mathbf{X}^{\text{test}}\widehat{\boldsymbol{\beta}}_1)$ on the testing set. We compare the average correlations and the number of selected variables in Table 3. The correlations on the testing sets continue to be far away from zero, which confirms their statistical significance. Both methods suggest that the correlations are stronger on the first chromosome than the second chromosome. Also, both methods select subsets of variables. Such results can be used as guidance for biologists to investigate the relationship between genes and CGH spots.

On average SCCA produces higher correlations than PMD. Paired t-tests show that the improvements are significant on both chromosomes. Moreover, SCCA gives much sparser estimates. In other words, SCCA identifies smaller sets of variables that account for more association. We further compare the variables selected by the two methods. We say that a variable is "frequently selected" if it is selected in more than half of the replicates. All the variables frequently selected by SCCA were also frequently selected by PMD, with the only exception of the gene `FLJ20225` on Chromosome 1. This gene was selected by PMD 35% of the time. In general, SCCA is more aggressive than PMD in excluding variables.

Finally, to see why SCCA has better performance, we examine the covariance matrices of the frequently selected CGH measurements by SCCA. The covariance matrices all appear to be far away from identity matrices. For example, on Chromosome 2, ten genes were frequently selected, including: `HDLBP`, `LRRFIP1`, `CYP27A1`, `TNRC15`, `RAB1A`, `GGCX`, `PDE6D`, `FLJ20752`, `HRB`, `SNTG2`. Their covariance matrix has a condition number (i.e, the ratio between the largest and the smallest eigenvalues) over 15.7. Hence, the identity covariance assumption imposed by PMD is severely violated, and SCCA has better performance on this dataset.

5 | DISCUSSION

We propose SCCA based on iterative penalized least squares. Our proposal provides nested solutions and can be combined with different penalty functions to perform structured variable selection. SCCA is implemented in an efficient algorithm. We show that SCCA consistently estimates the canonical pairs with an overwhelming probability even when the dimension grows at an exponential rate of the sample size. Numerical studies on simulated and real datasets also show that SCCA compares favorably to existing methods.

TABLE 3 Results on Chromosomes 1 & 2. For the last three rows, the numbers are means based on 200 replicates. In the parentheses are the standard errors.

	Chromosome 1		Chromosome 2	
	$p = 136, q = 1942$		$p = 72, q = 1329$	
	SCCA	**PMD**	**SCCA**	**PMD**
Correlation on the full dataset	0.9907	0.8192	0.9926	0.7232
p-value	0.005	$< 10^{-3}$	0.012	0.018
Correlation on the testing set	0.7940	0.7417	0.5832	0.3638
	(0.0060)	(0.0077)	(0.0169)	(0.0130)
# Selected CGH spots	12.8	50.1	10.7	28.8
	(0.28)	(2.37)	(0.45)	(1.29)
# Selected genes	46.8	836.5	41.6	639.3
	(0.62)	(40.67)	(1.33)	(30.52)

As pointed out by a referee, SCCA has some noticeable benefits in real data analysis as well. Its low computational costs allow researchers to conduct more in-depth analysis. For example, we demonstrated in Section 5 that SCCA can be combined with the bootstrap method proposed by Witten et al. (2009) to produce *p*-values. This is made possible by the low computation cost of SCCA, since the bootstrap method involves estimating the canonical pair many times. Hence, SCCA provides a approach to perform inference without assumptions on the covariance matrices, which was not feasible previously.

SCCA is designed for continuous data. In many real life problems, we may need to model relationships between discrete or even binary data, with SNP data being an important example. The extension of SCCA to the analysis of such data sets is left for future research.

ACKNOWLEDGEMENTS

The authors are grateful to the editor, the associate editor, and two referees for insightful suggestions that greatly improved the quality of this article. Mai is partially supported by the grant CCF-1617691 from National Science Foundation, and Zhang is partially supported by the grants CCF-1617691 and DMS-1613154 from National Science Foundation.

ORCID

Qing Mai http://orcid.org/0000-0003-0563-5622

REFERENCES

Cai, T. T., Ma., Z., and Wu, Y. (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics* 41, 3074–3110.

Chen, M., Gao, C., Ren, Z., and Zhou, H. (2017). Sparse cca via precision adjusted iterative thresholding. *Proceedings of International Congress of Chinese Mathematicians 2016*.

Chen, X., Liu, H., and Carbonell, J. G. (2012). Structured sparse canonical correlation analysis. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W. L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B. M., Esserman, L., Albertson, D. G., Waldman, F. M., and Gray, J. W. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* 10, 529–541.

Chu, D., Liao, L., Ng, M., and Zhang, X. (2013). Sparse kernel canonical correlation analysis. In *Proceedings of International Multiconference of Engineers and Computer Scientists*.

Efron, B., Hastie, B., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* 32, 407–499.

Fang, J., Lin, D., Schulz, S. C., Xu, Z., Calhoun, V. D., and Wang, Y.-P. (2016). Joint sparse canonical correlation analysis for detecting differential imaging genetics modules. *Bioinformatics* 32, 3480–3488.

Friedman, J. H., Hastie, T. J., and Tibshirani, R. J. (2008). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Software* 33, 1–22.

Gao, C., Ma, Z., Ren, Z., and Zhou, H. (2015). Minimax estimation in sparse canonical correlation analysis. *Annals of Statistics* 43, 2168–2197.

Gao, C., Ma, Z., Zhou, H. H., et al. (2017). Sparse cca: Adaptive estimation and computational barriers. *The Annals of Statistics* 45, 2074–2101.

Hardoon, D. R. and Shawe-Taylor, J. (2011). Sparse canonical correlation analysis. *Machine Learning Journal* 83, 331–353.

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

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Statist. Assoc.* 104, 682–693.

Le Cao, K., Pascal, M., C., R.-G., and Philippe, B. (2009). Sparse canonical methods for biological data integration: Application to a crossplatform study. *BMC Bioinfo.* 10, 34.

Lu, Y. and Foster, D. P. (2014). Large scale canonical correlation analysis with iterative least squares. In *Advances in Neural Information Processing Systems*, pages 91–99.

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

Ma, Z., Lu, Y., and Foster, D. (2015). Finding linear structure in large datasets with scalable canonical correlation analysis. In *International Conference on Machine Learning*, pages 169–178.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC proceedings* 1, S119.

Sun, L., Ji, S., and Ye, J. (2008). A least squares formulation for canonical correlation analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 1024–1031. ACM.

Sun, L., Ji, S., and Ye, J. (2011). Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 194–200.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* 58, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B.* 67, 91–108.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *Annals of Statistics* 39, 1335–1371.

Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse pca. *In Advances in Neural Information Processing Systems* pages 2670–2678.

Waaijenborg, S., Verselewel de Witt Hamer, P. C., and Zwinderman, A. H. (2008). Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. *Stat Appl Genet Mol Biol.* 7, Article 3.

Wang, Y. X. R., Jiang, K., Feldman, L. J., Bickel, P., and Huang, H. (2015). Inferring gene-gene interactions and functional modules using sparse canonical correlation analysis. *The Annals of Applied Statistics* 9, 300–323.

Witten, D. M. and Tibshirani, R. (2009). Extensions of sparse canonical correlation analysis, with applications to genomic data. *Stat Appl Genet Mol Biol* 8, Article 28.

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

Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing* 25, 1129–1141.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B* 68, 49–67.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Am. Statist. Assoc.* 101, 1418–1429.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article, including all the proofs and an example of applying SCCA with R.

How to cite this article: Mai Q, Zhang X. An iterative penalized least squares approach to sparse canonical correlation analysis. *Biometrics.* 2019;75:734–744. https://doi.org/10.1111/biom.13043