HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2018 March 15.

Published in final edited form as:

Biometrics. 2018 March ; 74(1): 300–312. doi:10.1111/biom.12715.

Integrative analysis of transcriptomic and metabolomic data via sparse canonical correlation analysis with incorporation of biological information

Sandra E. Safo, **Shuzhao Li**, and **Qi Long**

Department of Biostatistics and Bioinformatics, Department of Medicine, Division of Pulmonary, Allergy and Critical Care Medicine, Emory University, Atlanta, GA. Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA

Summary

Integrative analysis of high dimensional omics data is becoming increasingly popular. At the same time, incorporating known functional relationships among variables in analysis of omics data has been shown to help elucidate underlying mechanisms for complex diseases. In this paper our goal is to assess association between transcriptomic and metabolomic data from a Predictive Health Institute (PHI) study that includes healthy adults at a high risk of developing cardiovascular diseases. Adopting a strategy that is both data-driven and knowledge-based, we develop statistical methods for sparse canonical correlation analysis (CCA) with incorporation of known biological information. Our proposed methods use prior network structural information among genes and among metabolites to guide selection of relevant genes and metabolites in sparse CCA, providing insight on the molecular underpinning of cardiovascular disease. Our simulations demonstrate that the structured sparse CCA methods outperform several existing sparse CCA methods in selecting relevant genes and metabolites when structural information is informative and are robust to mis-specified structural information. Our analysis of the PHI study reveals that a number of gene and metabolic pathways including some known to be associated with cardiovascular diseases are enriched in the set of genes and metabolites selected by our proposed approach.

Keywords

Biological information; Canonical correlation analysis; High dimension; low sample size; Integrative analysis; Sparsity; Structural information

1. Introduction

Recent advancement in high-throughput, biomedical technologies has enabled the measurement of multiple high-dimensional omics data types in a single study, including genomics, epigenomics, transcriptomics and metabolomics. Each of these data types

Correspondence to: Qi Long.

Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 4.2 and 5 are available with this paper at the Biometrics website on Wiley Online Library. The methods are implemented in MATLAB and the code is available on Wiley Online Library

provides a different snapshot of the underlying biological system, and combining multiple data types has been shown to be very valuable in investigating complex diseases. It has been demonstrated that individual components in these data are functionally structured in networks or pathways and incorporation of such structural (or biological) information can improve analysis and lead to biologically more meaningful results (Li and Li, 2008; Pan et al., 2010; Chen et al., 2013). By the same token, it is desirable to jointly assess the association between these data types with incorporation of available structural information for each data type, enabling us to uncover drivers that individually or in combination provide better insight about the biological mechanism. In this article, we develop new canonical correlation analysis (CCA) methods for studying the overall dependency structure between transcripts and metabolites while incorporating structural information for each data type.

1.1 The PHI Study

Our work is motivated by data from the Emory University and Georgia Tech Predictive Health Institute (PHI) study. The PHI was established in 2005 with the goal of maintaining health rather than treating disease. The PHI data are collected from a longitudinal study of health measures in over 750 healthy employees of Emory University and Georgia Tech. We use data for 52 participants for whom gene expression and metabolomics data at baseline were available, and who were at a high risk of developing cardiovascular diseases defined by the Framingham risk scores (D'Agostino et al., 2008). The data consist of 32 females and 20 males with mean age of 47.35 years. The gene expression data consist of 38, 624 probes and the metabolomic data consist of 6, 009 features, where each metabolomic feature is defined by mass-to-charge ratio (*m/z*) and retention time and its relative concentration is captured by ion intensity. We exclude genes with variance and entropy expression values that are respectively less than the 90th and 20th percentile, resulting in 1, 547 genes. For the metabolomics data, we exclude features with more than 50% zeros, and use mummichog (Li et al., 2013) to annotate the *m/z* features, resulting in 252 metabolites.

Let $n = 52$ be the common samples that have both transcriptomic and metabolomic data. We denote the trancriptomic and metabolomic data by $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ ($p = 1, 547$) and $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_q)$ ($q = 252$), respectively, where $\mathbf{x}, \mathbf{y} \in \Re^n$. Structural information for genes are represented by an undirected graph $\mathcal{G}_X = (C_X, E_X, W_X)$, where C_X is the set of nodes corresponding to the p transcriptomic features, $E_X = \{i \sim j\}$ is the set of edges indicating that features i and j are associated in a biologically meaningful way, and W_X includes the weight of each node. Similarly, let $\mathcal{G}_Y = (C_Y, E_Y, W_Y)$ be the structural information for metabolites. For node i in $\mathbf{X}$, denote by d_i^X its degree i.e., the number of nodes that are directly connected to node i and by $w_i^X = f(d_i^X)$ its weight which can depend on d_i^X. Similarly, we define d_i^Y and w_i^Y. We use $w_i^X = d_i^X$ and $w_i^Y = d_i^Y$ in all our numerical studies. In our analysis of the PHI study, we obtain the gene and metabolomic network information respectively from KEGG (Kanehisa et al., 2016) and mummichog (Li et al., 2013). In the resulting gene network, there are 1, 547 genes with 479 edges in total; the distribution of d_i^X ranges from 1 to 32 with a mean of 3. In the metabolomics network, there are 252 metabolites and 190 edges in total, with each edge representing a connection between metabolites via a known metabolic reaction. The distribution of d_i^Y ranges from 1 to 13 with a mean of 3.

Our goal is to assess association between genes and metabolites with incorporation of structural information for both data types, for which, to the best of our knowledge, little work has been done in the literature. It is especially challenging when the number of features (p or q) greatly exceeds the sample size n as the case in the motivating PHI study, and in many biomedical omics studies.

1.2 Existing Methods

CCA was developed to find linear combinations of two sets of variables that have maximum correlation, which can help understand the overall dependency structure between these two sets of variables. However, it is well known that the classical CCA suffers from the singularity of sample covariance matrices when applied to high dimensional data; it also lacks biological interpretability especially when the number of variables is large. Extensions of CCA have been proposed to overcome these limitations. Some modifications deal with the singularity of sample covariance matrices by assuming sample covariance matrices are identity matrices (Witten et al., 2009; Parkhomenko et al., 2009; Chalise and Fridley, 2012), or have some structure such as sparsity, bandable or Toeplitz (Chen et al., 2013). The problem of biological interpretability has been tackled by assuming some coefficients are zero, implying that those variables do not contribute to the overall association between the two sets of variables (Parkhomenko et al., 2009; Witten et al., 2009; Chalise and Fridley, 2012; Chen et al., 2013; Safo and Ahn, 2014; Gao et al., 2015).

Despite the success of the available sparse CCA methods, their main limitation is that they do not exploit structural information among variables that is available for biological data such as transcriptomic and metabolomic data. Using available structural information, one can gain better understanding and obtain biologically more meaningful results from CCA. This has been demonstrated in the setting of sparse regression analysis (Pan et al., 2010; Li and Li, 2008). Recently, Chen et al. (2013) incorporated phylogenetic information from the bacterial taxa in CCA to study association between nutrient intake and human gut microbiome composition. We note that our work is different from the structured sparse CCA of Chen et al. (2012). In their work, they consider functional relationships among one data type and impose a group lasso penalty on the variables. Also, they do not utilize edge information among variables within pathways, which we do in the current paper.

1.3 Our Approach

We propose two structured sparse CCA methods that impose smoothness penalties on canonical correlation vectors and also allow for incorporating structural information such as gene and metabolic pathways to guide selection of important metabolites, transcripts, and pathways. Our work makes several contributions. First, the proposed methods enable us to conduct integrative analysis of transcriptomic and metabolic data that achieves variable selection and incorporates structural information for both data types, leading to biologically more meaningful results as evidenced in our data application. Second, we develop an efficient algorithm that can handle high dimensional problems. Third, our simulations demonstrate that the performance of the proposed approach is similar to or better than several existing methods even when network structure is not informative for selection of important variables. In particular, our proposed methods offer several improvements over the

recent work by Chen et al. (2013). First, our CCA formulation comes from the generalized eigenvalue problem rather than the direct CCA optimization problem. This formulation is not only simple to understand, but it also allows us to use convex objectives and constraints in the optimization problem that can be solved by most mathematical optimization softwares. Second, we use structural information from both sets of variables as opposed to only one set of variable, which is not a trivial extension. Third, their method and most sparse CCA methods assume that sample covariance matrices are identity matrices, but we relax this assumption as it can be overly restrictive in practice. Our method allows the use of sparse covariance matrices (Friedman et al., 2007) from which the underlying structural network may be inferred.

In section 2, we present the proposed structured sparse CCA after briefly reviewing sparse CCA. In Section 3, we present the algorithms for implementing the proposed sparse CCA. In Section 4, we conduct simulation studies to assess the performance of our methods in comparison with several existing methods. In Section 5, we apply our approach to the PHI study. We conclude with some discussion remarks in Section 6.

2. Methods

Following the notation introduced in Section 1, suppose that we have two sets of random matrices, an $n \times p$ matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, and an $n \times q$ matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_q)$, both of which, without generality, are standardized to have column mean 0 and variance 1. CCA (Hotelling, 1936) finds projections $\boldsymbol{\alpha} \in \Re^p$ and $\boldsymbol{\beta} \in \Re^q$ such that the correlation between linear combinations $\mathbf{X}\boldsymbol{\alpha}$ and $\mathbf{Y}\boldsymbol{\beta}$ is maximized. Mathematically, CCA finds vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that solve

$$\rho = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \operatorname{corr}(\mathbf{X}\boldsymbol{\alpha}, \mathbf{Y}\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \frac{\boldsymbol{\alpha}^{\mathrm{T}} \sum_{xy} \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^{\mathrm{T}} \sum_{xx} \boldsymbol{\alpha}} \sqrt{\boldsymbol{\beta}^{\mathrm{T}} \sum_{xx} \boldsymbol{\beta}}},$$

where $\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{yy}$ and $\boldsymbol{\Sigma}_{xy}$ are population covariance and cross-covariance matrices. The optimization problem is equivalent to solving

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^{\mathrm{T}} \sum_{xy} \boldsymbol{\beta} \text{ subject to } \boldsymbol{\alpha}^{\mathrm{T}} \sum_{xx} \boldsymbol{\alpha} = 1 \text{ and } \boldsymbol{\beta}^{\mathrm{T}} \sum_{yy} \boldsymbol{\beta} = 1. \quad (1)$$

Using Lagrangian multipliers and some algebra, one can show that problem (1) results in a generalized eigenvalue (GEV) problem of the form

$$\begin{bmatrix} 0 & \sum_{xy} \\ \sum_{yx} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} = \rho \begin{bmatrix} \sum_{xx} & 0 \\ 0 & \sum_{yy} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix}, \quad (2)$$

which can be solved by applying the singular value decomposition (SVD) to the matrix

$$\mathbf{K}=\sum_{xx}^{-1/2}\sum_{xy}\sum_{yy}^{-1/2}=(\mathbf{u}_1,\ldots,\mathbf{u}_k)\mathbf{D}(\mathbf{v}_1,\ldots,\mathbf{v}_k)^{\mathrm{T}}. \quad (3)$$

Here, k is the rank of the matrix $\mathbf{K}$, $\mathbf{u}_j$ and $\mathbf{v}_j$, ($j = 1, \ldots, k$) are the jth left and right singular vectors of $\mathbf{K}$, and $\mathbf{D}$ is a diagonal matrix containing singular values λ_j of $\mathbf{K}$ ordered from the largest to the smallest. It follows that the optimal coefficients in the linear combinations of $\mathbf{X}$ and $\mathbf{Y}$ are given by

$$\tilde{\boldsymbol{\alpha}}_j=\sum_{xx}^{-1/2}\mathbf{u}_j,\ \tilde{\boldsymbol{\beta}}_j=\sum_{yy}^{-1/2}\mathbf{v}_j. \quad (4)$$

The vectors $\tilde{\boldsymbol{\alpha}}_j$ and $\tilde{\boldsymbol{\beta}}_j$ are called the jth canonical correlation vectors for $\mathbf{X}$ and $\mathbf{Y}$ respectively, and are nonsparse. The random variables $\mathbf{X}\tilde{\boldsymbol{\alpha}}_j$ and $\mathbf{Y}\tilde{\boldsymbol{\beta}}_j$ are known as the jth canonical correlation variables, and $\tilde{\rho}_j = \lambda_j$ is the jth canonical correlation coefficient. Thus, the optimal coefficients in the linear combination yielding maximum correlation between $\mathbf{X}$ and $\mathbf{Y}$ is a rank one approximation of the matrix $\mathbf{K}$. When data are available, one can replace the population matrices $\sum_{xx}^{-1/2}\sum_{xy}\sum_{yy}^{-1/2}$ by the sample versions $\mathbf{S}_{xx}^{-1/2}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1/2}$, which results in consistent estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for fixed dimensions p, q, and large sample size n.

When p is greater than n, regularization is desirable in order to obtain interpretable solutions to problem (1). Despite the success of the existing regularized CCA methods, their main drawbacks, when applied to the setting of our interest, include failure to take full advantage of prior biological knowledge, and reliance on the assumption that $\mathbf{S}_{xx} = \mathbf{I}$, $\mathbf{S}_{yy} = \mathbf{I}$ which can be overly restrictive. Given the network information defined in Section 1.1, we investigate two structured sparse CCA for incorporating prior biological information.

We briefly motivate our formulation of structured sparse CCA using ideas from Dantzig Selector (DS) (Candes and Tao, 2007) for sparse estimation in regression problems. We note that the DS has been successfully used by Cai and Liu (2011) for sparse estimation in linear discriminant analysis. Given the GEV problem (2), we bound a modified version of the GEV difference with a l_∞ norm while minimizing a structured-sparsity inducing penalty of the canonical correlation loadings:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathscr{P}\left(\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix},\tau\right) \text{ subject to } \left\|\begin{bmatrix}0 & \mathbf{S}_{xy}\\ \mathbf{S}_{yx} & 0\end{bmatrix}\begin{bmatrix}\tilde{\boldsymbol{\alpha}}\\ \tilde{\boldsymbol{\beta}}\end{bmatrix}-\tilde{\rho}\begin{bmatrix}\tilde{\mathbf{S}}_{xx} & 0\\ 0 & \tilde{\mathbf{S}}_{yy}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix}\right\|_\infty \le \tau, \quad (5)$$

where $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ are the (nonsparse) solution to (2), and $\tilde{\rho}$ is the corresponding eigenvalue, where we have replaced the population covariances by their sample versions. Here, for a

random vector $\mathbf{z} \in \Re^p$, $\|\mathbf{z}\|_\infty$ is the l_∞ norm defined as $\max_{1 \le k \le p} |z_k|$, $\tau > 0$ is a tuning parameter that controls how many of the coefficients in canonical correlation loadings will be exactly zero. Let $(\tilde{\boldsymbol{\alpha}}_1, \tilde{\boldsymbol{\beta}}_1)$ be the first (nonsparse) solution to (2) and $\tilde{\rho}_1$ the corresponding eigenvalue. To obtain the structured sparse CCA loadings, we solve the problem (5) using the following two optimization problems :

$$\min_{\boldsymbol{\alpha}} \mathscr{P}(\boldsymbol{\alpha}, \tau_x) \quad \text{subject to} \quad \|\mathbf{S}_{xy}\tilde{\boldsymbol{\beta}}_1 - \tilde{\rho}_1 \tilde{\mathbf{S}}_{xx}\boldsymbol{\alpha}\|_\infty \le \tau_x, \tag{6}$$

$$\min_{\boldsymbol{\beta}} \mathscr{P}(\boldsymbol{\beta}, \tau_y) \quad \text{subject to} \quad \|\mathbf{S}_{yx}\tilde{\boldsymbol{\alpha}}_1 - \tilde{\rho}_1 \tilde{\mathbf{S}}_{yy}\boldsymbol{\beta}\|_\infty \le \tau_y. \tag{7}$$

for $\tau_x, \tau_y \ge 0$. There are at least two main advantages of this new formulation which uses ideas from the DS over existing formulation of CCA. First, the objective function $\mathscr{P}(\boldsymbol{\alpha}, \tau_x)$ and $\mathscr{P}(\boldsymbol{\beta}, \tau_y)$ can easily incorporate prior biological knowledge on both CCA loadings. Second, the optimization problems can be solved by any off-the-shelf optimization software, e.g. CVX in Matlab, given that $\mathscr{P}(\boldsymbol{\alpha}, \tau_x)$ and $\mathscr{P}(\boldsymbol{\beta}, \tau_y)$ are convex functions. In the next sections, we introduce sparse CCA methods that use network information $\mathscr{G}_\mathscr{X}$ in $\mathbf{X}$ and $\mathscr{G}_\mathscr{Y}$ in $\mathbf{Y}$.

2.1 Grouped Sparse CCA

The first approach is the Grouped sparse CCA, similar in spirit with Pan et al. (2010). Utilizing the graph structure in section 1.1, we propose the following structured sparse CCA criterion that solves the GEV problem (2) through the optimization problems (6) and (7): for the kth ($k = 1, \ldots K$) canonical correlation vector we solve iteratively until convergence the following optimization problem:

$$\begin{aligned}
&\min_{\boldsymbol{\alpha}} \left\{ (1-\eta) \sum_{i \sim j} \left(\frac{|\alpha_i|^\gamma}{w_i^X} + \frac{|\alpha_j|^\gamma}{w_j^X} \right)^{1/\gamma} + \eta \sum_{d_i^X = 0} |\alpha_i| \right\} \quad \text{subject to} \quad (A) \|\mathbf{S}_{xy}\tilde{\boldsymbol{\beta}}_k - \tilde{\rho}_k \tilde{\mathbf{S}}_{xx}\boldsymbol{\alpha}\|_\infty \le \tau_{x_1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (B) \|\tilde{\mathbf{S}}_{xx}^{-1}\mathbf{S}_{xy}\tilde{\boldsymbol{\beta}}_k - \tilde{\rho}_k \boldsymbol{\alpha}\|_\infty \le \tau_{x_2} \\
&\min_{\boldsymbol{\beta}} \left\{ (1-\eta) \sum_{i \sim j} \left(\frac{|\beta_i|^\gamma}{w_i^Y} + \frac{|\beta_j|^\gamma}{w_j^Y} \right)^{1/\gamma} + \eta \sum_{d_i^Y = 0} |\beta_i| \right\} \quad \text{subject to} \quad (A) \|\mathbf{S}_{yx}\tilde{\boldsymbol{\alpha}}_k - \tilde{\rho}_k \tilde{\mathbf{S}}_{yy}\boldsymbol{\beta}\|_\infty \le \tau_{y_1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (B) \|\tilde{\mathbf{S}}_{yy}^{-1}\mathbf{S}_{yx}\tilde{\boldsymbol{\alpha}}_k - \tilde{\rho}_k \boldsymbol{\beta}\|_\infty \le \tau_{y_2}
\end{aligned} \tag{8}$$

where for some random vector $\mathbf{x} \in \Re^p$, $\|\mathbf{x}\|_\infty$ is the l_∞ norm and is defined as $\max_i |x_i|$, $i = 1, \ldots, p$, $\tau_{x_1} > 0$ and $\tau_{y_1} > 0$ are tuning parameters, $\gamma > 1$ and $0 \le \eta < 1$ are fixed, and $\tilde{\boldsymbol{\alpha}}_k$ and $\tilde{\boldsymbol{\beta}}_k$ are the kth nonsparse canonical vectors defined in (4). As defined, (A) and (B) represent two different sets of constraints and are discussed in detail in Section 3.1 and the web supplementary material. The first term in each objective function is the weighted grouped

penalty (Pan et al., 2010), which induces grouped variable selection. It encourages both α_i and α_j (similarly both β_i and β_j) to be equal to zero or nonzero simultaneously, implying that two neighboring variables in a network are more likely to (or not to) participate in the same biological process simultaneously. In addition, the weight w_i^X encourages $|\alpha_i|/w_i^X = |\alpha_j|/w_j^X$ (similarly $|\beta_i|/w_i^Y = |\beta_j|/w_j^Y$) for two neighboring nodes *i, j*, allowing for connected features to have opposite effects. The second term in each objective function encourages variable selection of singletons that are not connected to any variable in the network. The tuning parameters τ_{x_1} or τ_{x_2} and τ_{y_1} or τ_{y_2} control the number of coefficients that are exactly zero with larger values encouraging more sparsity. The selection of τ_x and τ_y is usually data-driven, and is discussed later. We note that γ and η may also be selected using a data-driven approach. However, this will increase the complexity of the method, hence we fix them.

We can find $\hat{\boldsymbol{\alpha}}_k$ and $\hat{\boldsymbol{\beta}}_k$, $k \geq 2$ by solving (8) after projecting data onto the orthogonal complement of $[\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_{k-1}]$ and $[\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_{k-1}]$ respectively. In other words, we deflate data by obtaining $\mathbf{X}_{new} = \mathbf{X}\mathbf{P}_k^{\perp}$, where $\mathbf{P}_k^{\perp}$ is the projection matrix onto the othorgonal complement of $[\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_{k-1}]$. We obtain $\mathbf{Y}_{new}$ similarly.

In addition, in most of the existing sparse CCA methods, $\mathbf{S}_{xx}$ (and $\mathbf{S}_{yy}$) is assumed to be an identity matrix, essentially assuming that $\mathbf{X}$ (and $\mathbf{Y}$) is independent. We replace this assumption with the following variance-covariance matrices in our optimization problems

$$\tilde{\mathbf{S}}_{xx} = \mathbf{S}_{xx} + \sqrt{\log p/n}\mathbf{I}, \quad \tilde{\mathbf{S}}_{yy} = \mathbf{S}_{yy} + \sqrt{\log q/n}\mathbf{I} \qquad (9)$$

similar in spirit with Vinod (1970). The optimization problems in (8) are convex and can be solved with an off-the-shelf convex optimization package such as the CVX package in Matlab. We provide remarks on merits of constraints (A) and (B) in Section 3. The proposed method uses the nonsparse solution $(\tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k, \tilde{\rho}_k)$ as the 'initial' values; thus the performance may be dependent on the quality of initial values. To alleviate the dependence we propose to iterate the procedure by updating the $(\tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k, \tilde{\rho}_k)$ with the found $(\hat{\boldsymbol{\alpha}}_k, \hat{\boldsymbol{\beta}}_k, \hat{\rho}_k)$ until convergence, where by convergence we mean the ℓ_∞ difference between previous and current solutions $< 10^{-4}$ (i.e., $\|\hat{\boldsymbol{\alpha}}_{k(t)} - \hat{\boldsymbol{\alpha}}_{k(t-1)}\|_\infty < 10^{-4}$ or $\|\hat{\boldsymbol{\beta}}_{k(t)} - \hat{\boldsymbol{\beta}}_{k(t-1)}\|_\infty < 10^{-4}$, t and $t - 1$ are respectively the current and previous iterations). In all our empirical studies, the procedure mostly reached convergence within 4 ~ 5 iterations. Here $\hat{\rho}_k$ is the correlation coefficient between $\mathbf{X}\hat{\boldsymbol{\alpha}}_k$ and $\mathbf{Y}\hat{\boldsymbol{\beta}}_k$. Algorithm 1 below describes the procedure to obtain $\hat{\boldsymbol{\alpha}}_k$ and $\hat{\boldsymbol{\beta}}_k$, $k = 1, \ldots, K$.

2.2 Fused Sparse CCA

The second structured sparse CCA is the Fused sparse CCA, similar in spirit with Tibshirani et al. (2005). Utilizing the graph structure $\mathcal{G}$ in section 1.1, we propose the following structured sparse CCA criterion that solves the GEV problem (2) through the optimization problems (6) and (7): for the kth ($k = 1, \ldots K$) canonical correlation vector we solve iteratively until convergence the following optimization problem:

$$\min_{\boldsymbol{\alpha}} \left\{ (1-\eta)\sum_{i\sim j}\left|\frac{\alpha_i}{w_i^X}-\frac{\alpha_j}{w_j^X}\right|+\eta\sum_{d_i^X=0}|\alpha_j| \right\} \quad \text{subject to} \quad (A)\|\mathbf{S}_{xy}\tilde{\boldsymbol{\beta}}_k-\tilde{\rho}_k\tilde{\mathbf{S}}_{xx}\boldsymbol{\alpha}\|_\infty \le \tau_{x_1}$$
$$(B)\|\tilde{\mathbf{S}}_{xx}^{-1}\mathbf{S}_{xy}\tilde{\boldsymbol{\beta}}_k-\tilde{\rho}_k\boldsymbol{\alpha}\|_\infty \le \tau_{x_2}$$
$$\min_{\boldsymbol{\beta}} \left\{ (1-\eta)\sum_{i\sim j}\left|\frac{\beta_i}{w_i^Y}-\frac{\beta_j}{w_j^Y}\right|+\eta\sum_{d_i^Y=0}|\beta_j| \right\} \quad \text{subject to} \quad (A)\|\mathbf{S}_{yx}\tilde{\boldsymbol{\alpha}}_k-\tilde{\rho}_k\tilde{\mathbf{S}}_{yy}\boldsymbol{\beta}\|_\infty \le \tau_{y_1}$$
$$(B)\|\tilde{\mathbf{S}}_{yy}^{-1}\mathbf{S}_{yx}\tilde{\boldsymbol{\alpha}}_k-\tilde{\rho}_k\boldsymbol{\beta}\|_\infty \le \tau_{y_2} \quad (10)$$

where $\tau_{x1} > 0$ and $\tau_{y1} > 0$ are tuning parameters, $0 \le \eta < 1$ is assumed fixed, and $\tilde{\boldsymbol{\alpha}}_k$ and $\tilde{\boldsymbol{\beta}}_k$ are the kth nonsparse canonical vectors defined in (4). (A) and (B) are the same two sets of constraints introduced in Section 2.1 and in the web supplementary material. This penalty is a combination of fused lasso penalty on variable pairs that are connected in the network and an l_1 penalty on singletons that are not connected to any other variable in the network. This penalty is similar to the network constrained penalty of Li and Li (2008), but different in a number of ways. Their penalty $\eta_1\sum_j|\alpha_j|+\eta_2\sum_{i\sim j}\left(\frac{\alpha_i}{w_i}-\frac{\alpha_j}{w_j}\right)^2$ uses the l_2 norm and it has been shown that this does not produce sparse solutions, where sparsity refers to variables that are connected in a network. In other words, it does not encourage grouped selection of variables in the network (Pan et al., 2010). Also, the additional tuning parameter η_2 introduces more computational costs when applied to CCA as done in Chen et al. (2013); it requires solving a graph-constrained regression problem with dimension $(n+p)\times p$, incurring a high computational cost for very large p, particularly if one incorporates structural information on $\mathbf{Y}$ as well. Again, we replace $\mathbf{S}_{xx}$ and $\mathbf{S}_{yy}$ by $\tilde{\mathbf{S}}_{xx}$ and $\tilde{\mathbf{S}}_{yy}$ respectively.

3. Computation and Algorithms

3.1 Computation

Of the two constraints in optimization problems (8) and (10), constraint (B) is computationally motivated. Let $\hat{\boldsymbol{\alpha}}_F$ and $\hat{\boldsymbol{\beta}}_F$ be solution vectors from the structured sparse optimization with constraint (A) and let $\hat{\boldsymbol{\alpha}}_S$, $\hat{\boldsymbol{\beta}}_S$ be solution vectors from constraint (B). It is straightforward to show that if $\tau_{x1} = 0$, $\tau_{y1} = 0$ and $\tau_{x2} = 0$, $\tau_{y2} = 0$, then $\hat{\boldsymbol{\alpha}}_F = \hat{\boldsymbol{\alpha}}_S$ and $\hat{\boldsymbol{\beta}}_F = \hat{\boldsymbol{\beta}}_S$, that is, the solution vectors are the same. However, for $\tau_{x1} > 0$, $\tau_{y1} > 0$, $\tau_{x2} > 0$ and $\tau_{y2} > 0$, the optimization problems may yield the same objective functions but the solution vectors may not be the same, i.e., $\hat{\boldsymbol{\alpha}}_F \neq \hat{\boldsymbol{\alpha}}_S$ and $\hat{\boldsymbol{\beta}}_F \neq \hat{\boldsymbol{\beta}}_S$.

When p and q are large, problems (8) and (10) with constraint (A) are expensive to compute using the CVX package since it requires inverting $\mathbf{S}_{xx}$, a $p \times p$ matrix, and $\mathbf{S}_{yy}$, a $q \times q$ matrix, at each iteration. For constraint (B), a computationally efficient approach for very high dimensional problems is described as follows. Let $\mathbf{X}=\mathbf{U}_x\mathbf{D}_x\mathbf{V}_x^{\mathrm{T}}=\mathbf{R}_x\mathbf{V}_x^{\mathrm{T}}$ be the SVD of $\mathbf{X}$, where $\mathbf{V}_x$ is a $p \times n$ matrix of right singular vectors with orthonormal columns, $\mathbf{U}_x$ is an $n \times n$ orthogonal matrix of left singular vectors and $\mathbf{D}_x$ is a diagonal matrix of singular values. Hence $\mathbf{R}_x = \mathbf{U}_x\mathbf{D}_x$ is also $n \times n$. Also let $Y=\mathbf{U}_y\mathbf{D}_y\mathbf{V}_y^{\mathrm{T}}=\mathbf{R}_y\mathbf{V}_y^{\mathrm{T}}$ be the SVD of $\mathbf{Y}$, where $\mathbf{V}_y$ is a $q \times n$ orthonormal matrix, $\mathbf{U}_y$ is a $n \times n$ orthogonal matrix and $\mathbf{D}_y$ is a diagonal

matrix of singular values. Then $\mathbf{R}_y = \mathbf{U}_y\mathbf{D}_y$ is also $n \times n$. Plugging these into $\tilde{\mathbf{S}}_{xx}^{-1}\mathbf{S}_{xy}$, and after some careful algebra, we obtain

$\tilde{\mathbf{S}}_{xx}^{-1}\mathbf{S}_{xy}=(\mathbf{S}_{xx}+\sqrt{\log p/n}\mathbf{I})^{-1}\mathbf{S}_{xy}=\mathbf{V}_x(\mathbf{R}_x^{\mathrm{T}}\mathbf{R}_x+\sqrt{\log p/n}\mathbf{I})^{-1}\mathbf{R}_x^{\mathrm{T}}\mathbf{R}_y\mathbf{V}_y^{\mathrm{T}}$, which requires inverting a $n \times n$ matrix. Similarly,

$\tilde{\mathbf{S}}_{yy}^{-1}\mathbf{S}_{yx}=(\mathbf{S}_{yy}+\sqrt{\log q/n}\mathbf{I})^{-1}\mathbf{S}_{yx}=\mathbf{V}_y(\mathbf{R}_y^{\mathrm{T}}\mathbf{R}_y+\sqrt{\log q/n}\mathbf{I})^{-1}\mathbf{R}_y^{\mathrm{T}}\mathbf{R}_x\mathbf{V}_x^{\mathrm{T}}$. The same idea can be used in (3) and (4) for the nonsparse estimates $\tilde{\boldsymbol{\alpha}}_k$ and $\tilde{\boldsymbol{\beta}}_k$ in both constraints (A) and (B) to reduce computational cost of obtaining SVD of a $p \times q$ matrix, which is expensive as $\min(p, q)$ increases.

Algorithm 1

Optimization for obtaining the kth structured sparse CCA vector

1: **for** $k = 1, \ldots, K$ **do**

2: Initialize with nonsparse estimates: $\tilde{\boldsymbol{\alpha}}_{k0}=\tilde{\mathbf{S}}_{xx}^{-1/2}\mathbf{u}_k, \tilde{\boldsymbol{\beta}}_{k0}=\tilde{\mathbf{S}}_{yy}^{-1/2}\mathbf{v}_k$ with unity l_2 norm, and $\tilde{\rho}_{k0} = \lambda_k$. The approach discussed in Section 3.1 can be used here for problems with large p and /or q.

3: **for** t =1 until convergence or some maximum number of iterations **do**

4: Solve one of the following two optimization problems using previous estimates $\hat{\boldsymbol{\alpha}}_{k(t-1)}$ and $\hat{\boldsymbol{\beta}}_{k(t-1)}$, to obtain the kth estimates $\hat{\boldsymbol{\alpha}}_{k(t)}$ and $\hat{\boldsymbol{\beta}}_{k(t)}$:
(3i) The Grouped sparse optimization problem

$$\min_{\boldsymbol{\alpha}}\left\{(1-\eta)\sum_{i\sim j}\left(\frac{|\alpha_i|^\gamma}{w_i^X}+\frac{|\alpha_j|^\gamma}{w_j^X}\right)^{1/\gamma}+\eta\sum_{d_i^X=0}|\alpha_i|\right\}\ \text{subject to}\ (A)\|\mathbf{S}_{xy}\tilde{\boldsymbol{\beta}}_{k(t-1)}-\tilde{\rho}_{k(t-1)}\tilde{\mathbf{S}}_{xx}\boldsymbol{\alpha}\|_\infty\le\tau_{x_1}$$

$$(B)\|\tilde{\mathbf{S}}_{xx}^{-1}\mathbf{S}_{xy}\tilde{\boldsymbol{\beta}}_{k(t-1)}-\tilde{\rho}_{k(t-1)}\boldsymbol{\alpha}\|_\infty\le\tau_{x_2}$$

$$\min_{\boldsymbol{\beta}}\left\{(1-\eta)\sum_{i\sim j}\left(\frac{|\beta_i|^\gamma}{w_i^Y}+\frac{|\beta_j|^\gamma}{w_j^Y}\right)^{1/\gamma}+\eta\sum_{d_i^Y=0}|\beta_i|\right\}\ \text{subject to}\ (A)\|\mathbf{S}_{yx}\tilde{\boldsymbol{\alpha}}_{k(t-1)}-\tilde{\rho}_{k(t-1)}\tilde{\mathbf{S}}_{yy}\boldsymbol{\beta}\|_\infty\le\tau_{y_1}$$

$$(B)\|\tilde{\mathbf{S}}_{yy}^{-1}\mathbf{S}_{yx}\tilde{\boldsymbol{\alpha}}_{k(t-1)}-\tilde{\rho}_{k(t-1)}\boldsymbol{\beta}\|_\infty\le\tau_{y_2}$$

(3ii) The Fused sparse optimization problem

$$\min_{\boldsymbol{\alpha}}\left\{(1-\eta)\sum_{i\sim j}\left|\frac{\alpha_i}{w_i^X}-\frac{\alpha_j}{w_j^X}\right|+\eta\sum_{d_i^X=0}|\alpha_j|\right\}\ \text{subject to}\ (A)\|\mathbf{S}_{xy}\tilde{\boldsymbol{\beta}}_{k(t-1)}-\tilde{\rho}_{k(t-1)}\tilde{\mathbf{S}}_{xx}\boldsymbol{\alpha}\|_\infty\le\tau_{x_1}$$

$$(B)\|\tilde{\mathbf{S}}_{xx}^{-1}\mathbf{S}_{xy}\tilde{\boldsymbol{\beta}}_{k(t-1)}-\tilde{\rho}_{k(t-1)}\boldsymbol{\alpha}\|_\infty\le\tau_{x_2}$$

$$\min_{\boldsymbol{\beta}}\left\{(1-\eta)\sum_{i\sim j}\left|\frac{\beta_i}{w_i^Y}-\frac{\beta_j}{w_j^Y}\right|+\eta\sum_{d_i^Y=0}|\alpha_j|\right\}\ \text{subject to}\ (A)\|\mathbf{S}_{yx}\tilde{\boldsymbol{\alpha}}_{k(t-1)}-\tilde{\rho}_{k(t-1)}\tilde{\mathbf{S}}_{yy}\boldsymbol{\beta}\|_\infty\le\tau_{y_1}$$

$$(B)\|\tilde{\mathbf{S}}_{yy}^{-1}\mathbf{S}_{yx}\tilde{\boldsymbol{\alpha}}_{k(t-1)}-\tilde{\rho}_{k(t-1)}\boldsymbol{\beta}\|_\infty\le\tau_{y_2}$$

5: Normalize $\hat{\boldsymbol{\alpha}}_{k(\hat{\imath})}$ and $\hat{\boldsymbol{\beta}}_{k(\hat{\imath})}$ to have unity l_2 norm and obtain the canonical correlation coefficient $\hat{\rho}_{k(\hat{\imath})}$.

6: Update $(\tilde{\boldsymbol{\alpha}}_k, \tilde{\boldsymbol{\beta}}_k, \tilde{\rho}_k)$ with $(\hat{\boldsymbol{\alpha}}_k, \hat{\boldsymbol{\beta}}_k, \hat{\rho}_k)$.

7: **end for**

8: If $k \geq 2$, $k \leq \min(n-1, p, q)$, update $\mathbf{X}$ and $\mathbf{Y}$ by projecting them to the orthogonal complement of $[\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_{k-1}]$ and $[\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_{k-1}]$ respectively, and repeat steps 3 to 7.

9: **end for**

Algorithm 2

V-fold CV for tuning parameter selection

Randomly group the rows of $\mathbf{X}$ and $\mathbf{Y}$ into V roughly equal-sized groups, denoted by $\mathbf{X}^1, \ldots, \mathbf{X}^V$, and $\mathbf{Y}^1, \ldots, \mathbf{Y}^V$, respectively.

2: **for** each τ_x and a fixed τ_y **do**

(i) For $v = 1, \ldots, V$, let $\mathbf{X}^{-v}$ and $\mathbf{Y}^{-v}$ be the data matrix leaving out $\mathbf{X}^v$ and $\mathbf{Y}^v$ respectively. Apply Algorithm 1 on $\mathbf{X}^{-v}$ and $\mathbf{Y}^{-v}$ to derive the desired number of canonical correlation vectors $\hat{\boldsymbol{\alpha}}_k^{-v}(\tau_x, \tau_y)$, and $\hat{\boldsymbol{\beta}}_k^{-v}(\tau_x, \tau_y)$, $k = 1, \cdots, \min(n-1, p, q)$.

(ii) Project $\mathbf{X}^v$ and $\mathbf{Y}^v$ onto $\hat{\boldsymbol{\alpha}}_k^{-v}(\tau_x, \tau_y)$, and $\hat{\boldsymbol{\beta}}_k^{-v}(\tau_x, \tau_y)$ to obtain the testing correlation coefficients, $\hat{\rho}_{k_{test}}^{v}(\tau_x, \tau_y) = \mathrm{corr}(\mathbf{X}^v \hat{\boldsymbol{\alpha}}_k^{-v}, \mathbf{Y}^v \hat{\boldsymbol{\beta}}_k^{-v})$.

(iii) Project $\mathbf{X}^{-v}$ and $\mathbf{Y}^{-v}$ onto $\hat{\boldsymbol{\alpha}}_k^{-v}(\tau_x, \tau_y)$, and $\hat{\boldsymbol{\beta}}_k^{-v}(\tau_x, \tau_y)$ to obtain the training correlation coefficients, $\hat{\rho}_{k_{train}}^{-v}(\tau_x, \tau_y) = \mathrm{corr}(\mathbf{X}^{-v} \hat{\boldsymbol{\alpha}}_k^{-v}, \mathbf{Y}^{-v} \hat{\boldsymbol{\beta}}_k^{-v})$.

(iv) Calculate the V-fold CV score as the difference between the average training and testing absolute correlation coefficients.

$$CV(\tau_x, \tau_y) = \left| \frac{1}{V} \sum_{v=1}^{V} \left| \hat{\rho}_{k_{train}}^{v}(\tau_x, \tau_y) \right| - \frac{1}{V} \sum_{v=1}^{V} \left| \hat{\rho}_{k_{test}}^{-v}(\tau_x, \tau_y) \right| \right|. \quad (11)$$

(v) Select the optimal tuning parameter τ_x as $\tau_{x_{\text{opt}}} = \min CV(\tau_x, \tau_y)$

end for

4: **for** $\tau_{x_{\text{opt}}}$ and each τ_y **do**

(i) Repeat steps 2(*i*) to 2(*iv*)

(ii) Select the optimal tuning parameter $\tau_{y_{\text{opt}}}$ as $\tau_{y_{\text{opt}}} = \min\{CV(\tau_{x_{\text{opt}}}, \tau_y)\}$

end for

6: Apply $\tau_{\text{opt}} = (\tau_{x_{\text{opt}}}, \tau_{x_{\text{opt}}})$ on the whole training data $\mathbf{X}$, $\mathbf{Y}$ using Algorithm 1 to obtain the optimal CCA vectors $\hat{\boldsymbol{\alpha}}_k$, $\hat{\boldsymbol{\beta}}_k$, and coefficients ρ_k at each iteration until convergence.

3.2 Algorithms

We describe two algorithms for the proposed structured sparse CCA methods. The first obtains the kth canonical correlation vector for fixed tuning parameters τ_x and τ_y. The second algorithm provides a data driven approach for selecting the optimal tuning parameters.

We first normalize the columns of $\mathbf{X}$ and $\mathbf{Y}$ to have mean zero and unit variance. Let $\mathbf{u}_k$ and $\mathbf{v}_k$ be the kth left and right singular vectors of $\tilde{\mathbf{S}}_{xx}^{-1/2}\mathbf{S}_{xy}\tilde{\mathbf{S}}_{yy}^{-1/2}$, and let λ_k be the kth singular value. The approach discussed in Section 3.1 can be used here for problems with large p and/or q. For fixed positive tuning parameters $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_y$, use Algorithm 1 for the kth sparse CCA vectors, $\hat{\boldsymbol{\alpha}}_k$ and $\hat{\boldsymbol{\beta}}_k$, and correlation $\hat{\rho}_k$. Of note $k \leq \min(n-1, p, q)$; usually few correlation vectors are sufficient to explain the overall dependency structure between the data types. In practice, one could use the k that explains some desired correlation using cumulative contribution ratio (Fujikoshi et al. (2010)) given as $\sum_{j=1}^{k}\hat{\rho}_j / \sum_{j=1}^{\min(n-1,p,q)}\lambda_j$.

The tuning parameters $\boldsymbol{\tau} = (\boldsymbol{\tau}_x, \boldsymbol{\tau}_y)$ control the model complexity and their optimal values need to be selected. We use V-fold cross validation (CV) to select $\boldsymbol{\tau}$ at each iteration of Algorithm 1. The optimal tuning parameter pair is chosen by performing a grid search over the entire pre-specified set of parameter values. To further reduce computational costs, we cross search over the pre-specified set of parameters. For a fixed value in the $\boldsymbol{\tau}_y$ set of values (we fix $\boldsymbol{\tau}_y$ as the middle value of the set of values), we search over the entire space of $\boldsymbol{\tau}_x$ values and select $\boldsymbol{\tau}_{x_{\text{opt}}}$ that minimizes criterion (11) given $\boldsymbol{\tau}_y$. Using $\boldsymbol{\tau}_{x_{\text{opt}}}$, we search the entire $\boldsymbol{\tau}_y$ space and choose $\boldsymbol{\tau}_{y_{\text{opt}}}$ that also minimizes criterion (11). We choose $\boldsymbol{\tau}_{\text{opt}} = (\boldsymbol{\tau}_{x_{\text{opt}}}, \boldsymbol{\tau}_{y_{\text{opt}}})$ at each iteration in Algorithm 1 since the selected optimal pair from previous iterations may be too large and may result in a trivial solution at the subsequent iteration.

4. Simulations

4.1 Simulation Set-up

We conduct simulations to assess the performance of the proposed methods in comparison with several existing sparse CCA methods, and one structured sparse CCA method. In each simulation experiement, 200 Monte Carlo (MC) datasets are generated as follows. The first data type $\mathbf{X}$ have p variables and the second data type $\mathbf{Y}$ have q variables, all drawn on the same samples with size $n = 80$. $(\mathbf{X}, \mathbf{Y})$ are simulated from MVN$(\mathbf{0}, \boldsymbol{\Sigma})$ with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$ partitioned as $\sum = \begin{pmatrix} \sum_{xx} & \sum_{xy} \\ \sum_{yx} & \sum_{yy} \end{pmatrix}$, where $\boldsymbol{\Sigma}_{xy}$ is the covariance between $\mathbf{X}$ and $\mathbf{Y}$, and $\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{yy}$ are respectively the covariance of $\mathbf{X}$ and $\mathbf{Y}$ that describe the network structure in each data type. Without loss of generality, we let the first 36 variables form the networks in $\mathbf{X}$ and $\mathbf{Y}$, where within each data type there are 6 main variables, each connected to 5 variables. The resulting network has 36 variables and edges with a maximum degree of 5, and $p - 36$ and $q - 36$ singletons in $\mathbf{X}$ and $\mathbf{Y}$ respectively. Using the notation in Section 1.1, the graph structure is given by $\mathcal{G}_X = \mathcal{G}_Y = \{C, E, W\}$, where $C = \{i, j \in p, q\}$, $E = \{i \sim j | i, j = 1, \cdots, 36\}$, and $W = \{w_i | w_i = \text{degree of node } i, i = 1, \cdots, 36\}$. The network structure in each data type is captured by the covariance matrices $\sum_{xx} = \begin{pmatrix} \bar{\sum}_{36\times 36} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-36} \end{pmatrix}$, $\sum_{yy} = \begin{pmatrix} \bar{\sum}_{36\times 36} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{q-36} \end{pmatrix}$, where $\bar{\boldsymbol{\Sigma}}$ is block diagonal with 6 blocks of size 6, between-block correlation 0 and within each block there is a 5×5 compound symmetric submatrix with correlation 0.49 describing the correlation structure of the connected variables. The correlation between a main and a connecting variable is 0.7.

The covariance between $\mathbf{X}$ and $\mathbf{Y}$ is $\boldsymbol{\Sigma}_{xy} = \rho\boldsymbol{\Sigma}_{xx}\boldsymbol{\alpha}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{yy}$, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the true canonical correlation vectors and ρ is the canonical correlation coefficient. We consider four simulation scenarios.

(1) Scenario one: *All networks in* X *are correlated with all networks in* Y—In the first scenario, all 6 networks in $\mathbf{X}$ and $\mathbf{Y}$ are associated and contribute to the correlation between the sets of variables, while the remaining singletons do not contribute to the correlation and thus have zero coefficients. We generate the true canonical correlation vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as follows

$$\left(-20, \tfrac{-20}{\sqrt{5}}, \ldots, \tfrac{-20}{\sqrt{5}}, 20, \tfrac{20}{\sqrt{5}}, \ldots, \tfrac{20}{\sqrt{5}}, -17, \tfrac{-17}{\sqrt{5}}, \ldots, \tfrac{-17}{\sqrt{5}}, 17, \tfrac{17}{\sqrt{5}}, \ldots, \tfrac{17}{\sqrt{5}}, -10, \tfrac{-10}{\sqrt{5}}, \ldots, \tfrac{-10}{\sqrt{5}}, 10, \tfrac{10}{\sqrt{5}}, \ldots, \tfrac{10}{\sqrt{5}}, 0, \ldots, 0\right)$$

and normalize such that $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}_{xx}\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{yy}\boldsymbol{\beta} = 1$. We set $\rho = 0.9$ and 0.5.

(2) Scenario two: *Two networks in* X *and* Y *are correlated—*In the second scenario, only the first 2 networks in $\mathbf{X}$ and $\mathbf{Y}$ contribute to the correlation structure between the sets of variables. The remaining networks and singletons do not contribute to the correlation between the two data types, even though within each data type, each network exhibit strong association between variables. The true canonical correlation vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are generated as $\left(-20, \tfrac{-20}{\sqrt{5}}, \ldots, \tfrac{-20}{\sqrt{5}}, 20, \tfrac{20}{\sqrt{5}}, \ldots, \tfrac{20}{\sqrt{5}}, 0, \ldots, 0\right)$ and we normalize each to have $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}_{xx}\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{yy}\boldsymbol{\beta} = 1$. We set $\rho = 0.9$ and 0.5.

(3) Scenario three: *Two orthogonal CCA vectors in* X *and* Y—In the third scenario, there are two orthogonal canonical correlation vectors $\mathbf{A} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ and $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ in $\mathbf{X}$ and $\mathbf{Y}$ respectively that induce the correlation between $\mathbf{X}$ and $\mathbf{Y}$. Specifically, there are four networks in $\boldsymbol{\alpha}_1$, which are the first 24 variables with nonzero loadings, and these are associated with the first 18 variables (or 3 networks) in $\boldsymbol{\beta}_1$. The next 12 variables, forming the remaining two networks are found in $\boldsymbol{\alpha}_2$, and these are correlated with the next 3 networks in $\boldsymbol{\beta}_2$. Then, the covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$ is $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_{xx}\mathbf{ADB}^{\mathrm{T}}\boldsymbol{\Sigma}_{yy}$, where $\mathbf{D} = \mathrm{diag}(0.9, 0.6)$ (and $\mathbf{D} = \mathrm{diag}(0.5, 0.3)$) is a diagonal matrix with diagonal values being the first and second canonical correlation coefficients. We normalize the vectors

$$\begin{aligned}
\boldsymbol{\alpha}_1 &= \left(-20, \tfrac{-20}{\sqrt{5}}, \ldots, \tfrac{-20}{\sqrt{5}}, 20, \tfrac{20}{\sqrt{5}}, \ldots, \tfrac{20}{\sqrt{5}}, -17, \tfrac{-17}{\sqrt{5}}, \ldots, \tfrac{-17}{\sqrt{5}}, 17, \tfrac{17}{\sqrt{5}}, \ldots, \tfrac{17}{\sqrt{5}}, 0, \ldots, 0\right) \\
\boldsymbol{\alpha}_2 &= \left(0, \ldots, 0, 17, \tfrac{17}{\sqrt{5}}, \ldots, \tfrac{17}{\sqrt{5}} - 17, \tfrac{-17}{\sqrt{5}}, \ldots, \tfrac{-17}{\sqrt{5}}, 0, \ldots, 0\right) \\
\boldsymbol{\beta}_1 &= \left(-20, \tfrac{-20}{\sqrt{5}}, \ldots, \tfrac{-20}{\sqrt{5}}, 20, \tfrac{20}{\sqrt{5}}, \ldots, \tfrac{20}{\sqrt{5}}, -17, \tfrac{-17}{\sqrt{5}}, \ldots, \tfrac{-17}{\sqrt{5}}, 0, \ldots, 0\right) \\
\boldsymbol{\beta}_2 &= \left(0, \ldots, 0, 17, \tfrac{17}{\sqrt{5}}, \ldots, \tfrac{17}{\sqrt{5}}, -10, \tfrac{-10}{\sqrt{5}}, \ldots, \tfrac{-10}{\sqrt{5}}, 10, \tfrac{10}{\sqrt{5}}, \ldots, \tfrac{10}{\sqrt{5}}, 0, \ldots, 0\right)
\end{aligned}$$

to have $\boldsymbol{\alpha}_i^{\mathrm{T}}\sum_{xx}\boldsymbol{\alpha}_i = 1, \boldsymbol{\beta}_i^{\mathrm{T}}\sum_{yy}\boldsymbol{\beta}_i = 1$, $i = 1, 2$, $\boldsymbol{\alpha}_1^{\mathrm{T}}\sum_{xx}\boldsymbol{\alpha}_2 = 0$, and $\boldsymbol{\beta}_1^{\mathrm{T}}\sum_{yy}\boldsymbol{\beta}_2 = 0$.

(4) Scenario four: *Randomly selected features in* X *and* Y *are correlated—*In the fourth scenario, there are two networks and 12 variables in $\mathbf{X}$ and $\mathbf{Y}$ that are correlated. These two networks are the same as those in scenario two, but are randomly dispersed and are not necessarily the first 12 variables in $\mathbf{X}$ and $\mathbf{Y}$. We normalize the vectors to have $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}_{xx}\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{yy}\boldsymbol{\beta} = 1$. We set $\rho = 0.9$ and 0.5. This setting assesses performance in

cases where the structural information is mis-specified or uninformative and sheds light on robustness of the proposed methods.

In the analysis of each MC dataset, we fix $\eta = .5$ to give equal weights to the selection of networks and singletons. We set $\gamma = 2$ and 8 in the L_γ-norm penalty of the Grouped structured sparse CCA method since $\gamma = 2$ and $\gamma = 8$ gave better results in the original paper (Pan et al., 2010). We consider the dimensions $(p, q) = (500, 500)$ for all scenarios. We use 5-fold cross validation to select the optimal tuning parameters from criterion 11, and then obtain $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ using the entire training set. We evaluate the proposed methods based on their ability to select relevant features while maximizing correlation between $\mathbf{X}$ and $\mathbf{Y}$. The results are summarized in terms of sensitivity, specificity, and Matthew's correlation coefficient (MCC) (Matthews, 1975) which are defined as follows:

$\text{Sensitivity} = \frac{TP}{TP+FN}$, $\text{Specificity} = \frac{TN}{TN+FP}$, $\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}$, where TP, FP, TN, and FN are true positives, false positives, true negatives, and false negatives, respectively. Of note, MCC lies in the interval $[-1, 1]$, with a value of 1 corresponding to selection of all signal variables and no noise variables, a perfect selection. A value of -1 implies that $FP = 1$, $FN = 1$ and $TP = 0$, $TN = 0$, and a value of 0 implies $TP = TN = FP = FN = 0.5$.

4.2 Simulation results

We denote the proposed methods, Grouped and Fused structured sparse CCA as Grouped$_A$, Grouped$_B$, and Fused$_A$, Fused$_B$ with subscripts A and B respectively indicating constraints A and B in (8) and (10). We compare with the following sparse methods: sparse CCA (SCCA) (Parkhomenko et al., 2009), penalized matrix decomposition CCA (PMD) (Witten et al., 2009), sparse CCA with SCAD penalty (SCAD) (Chalise and Fridley, 2012) and sparse estimation via linear programming for CCA (SELP)(Safo and Ahn, 2014). We also compare the proposed methods with Chen et al. (2012) (SGCCA). SGCCA imposes a group-lasso penalty on CCA vector in the second data type to select variables in a group, but do not consider the interactions among variables within a group, unlike the proposed methods. We implement SGCCA using Matlab software provided on the author's website. In the simulations, we set the number of groups as the number of networks in the simulated data, and choose the regularization parameters following ideas in Chen et al. (2012).

Figures 1 and 2 show the sensitivity, specificity and MCC for the methods compared with sparse CCA methods when $\rho = 0.9$ and 0.5 respectively. We observe a competitive performance of the proposed methods, in particular Fused$_A$ and Fused$_B$, in selecting the true signals in all but scenario four. Fused$_A$ and Fused$_B$ perform well in scenarios one, three and four while Grouped$_A$ performs better in scenario four. The other sparse methods especially SCCA and SCAD tend to select a large number of noise variables, evident by the low specificity and MCC proportions in Figure 1. In general, we expect the performance of the sparse and structured sparse CCA methods to deteriorate for low canonical correlation coefficients, and this is what we observe when $\rho = 0.5$ (Figure 2). However, the findings for the proposed methods compared to the sparse CCA methods still hold true.

For the proposed methods, it is noticeable from the sensitivity and MCC proportions that Grouped$_A$ and Grouped$_B$ have a suboptimal performance in scenarios one, two and three, yet these are better than the sparse methods. In scenario two, Fused$_A$ and Fused$_B$ select more FP than Grouped$_A$ and Grouped$_B$ as evidenced by the low specificity, yet they are comparable to the other sparse methods. Recall that in scenario two, only 2 networks in **X** and **Y** contribute to the overall correlation between **X** and **Y**. However, within each network, there is high correlation, causing the Fused methods to read these as signals and therefore select them, though they do not contribute to the association between **X** and **Y**. In scenario four, the performance of all the methods deteriorates from scenarios one to three, yet the proposed methods still outperform the other sparse methods, suggesting that the proposed methods are robust to uninformative network information.

When we compare constraints A and B for Grouped and Fused methods, we notice similar performances in terms of variable selection and MCC for both Fused$_A$ and Fused$_B$, but the latter is computationally more efficient and can be used for very high dimensional problems. Grouped$_A$ has high specificity and high MCC values (Figures 1 and 2), but Grouped$_B$ has better sensitivity. In general, Grouped$_A$ outperforms Grouped$_B$ at higher computational cost.

Comparing Fused and Grouped methods, we notice that in general, the performance of Fused is better than Grouped. However, Grouped (specifically Grouped$_A$) tends to achieve better performance in terms of specificity. We also compare Grouped$_A$ and Grouped$_B$ when $\gamma = 2$ and $\gamma = 8$ (Tables 1 and 2 of supplementary material). We observe comparable results for Grouped$_A$ but mixed results for Grouped$_B$. We find that Grouped$_B$ yields slightly higher sensitivity and lower specificity when $\gamma = 8$, and $\rho = 0.9$ or $\rho = 0.5$, but higher or comparable specificity when $\gamma = 2$. These findings suggest that higher γ may not always lead to improved performance, which is consistent with the findings from Pan et al. (2010).

We compare our proposed methods with structured sparse CCA, SGCCA (Chen et al., 2012). We note that SGCCA imposes structural information when estimating CCA vector for **Y**. Tables 1 and 2 report sensitivities and specificities, and MCC estimates. SGCCA has suboptimal performance in selecting true signals in **X** across all scenarios for both $\rho = 0.9$ and $\rho = 0.5$, but comparable or better specificities when $\rho = 0.5$. In estimating the CCA vector for **Y**, we observe the proposed methods show better performance in selecting true signals while ignoring noise variables, with the exception of Grouped, which has suboptimal performance compared to SGCCA in terms of specificity. The MCC estimates for **Y** under SGCCA are below 0 across all scenarios; this suggests that the overall selectivity for SGCCA is poor, and it is not able to distinguish the true signals (networks) from the noise variables.

We also compare our proposed methods with SGCCA in terms of computational time (see web supplementary material for more details). Figure 1 (web supplementary material) shows the average CPU time (in seconds) for estimating CCA vectors. SGCCA is faster than the proposed methods, but has suboptimal performance in terms of sensitivity, specificity and MCC. Of the proposed methods, Fused$_B$ is faster than Fused$_A$, Grouped$_A$, and Grouped$_B$.

The results in Figures 1 and 2, and Tables 1 and 2 demonstrate that the structured sparse CCA methods we propose exhibit superior performance over the other sparse and structured sparse methods that are considered, evidenced by their high sensitivity, high specificity and high MCC proportions. The performance of the other sparse methods is worse in scenarios one and three than in scenario two. This shows that if the features in each set of variables are interconnected in the form of networks, and if most of these networks contribute to the association between **X** and **Y**, the existing sparse methods encounter difficulty in selecting the important networks. On the other hand, the proposed structured sparse methods can exploit the prior biological knowledge to increase sensitivity, specificity, and MCC.

5. Analysis of the PHI Study Data

We apply the proposed methods to integrative analysis of the transcriptomic **X** and metabolomic **Y** data in the PHI study. We log 10 transform the metabolomics data and normalize both the transcriptomic and metabolomics data to have mean 0 and variance 1 for each transcriptomic or metabolomic feature. Our goal is to identify a subset of transcriptomic and metabolomic features that capture the overall association between transcripts and metabolites.

We analyze the PHI data using the proposed methods as well as the existing structured and sparse CCA methods considered in our simulations. We use 5-fold cross validation to select optimal tuning parameters in our proposed methods, and then apply the selected tuning parameters to the whole data to estimate the maximal canonical correlation coefficient and vectors. For the Grouped method, we fix $\gamma = 2$. For SGCCA (Chen et al., 2012), we set the number of groups to 50. Table 3 shows the number of genes and metabolites from the first canonical correlation vectors. From Table 3, we observe that the proposed methods, especially Grouped$_B$ and Fused$_B$ have high estimated canonical correlation coefficients compared to SELP even though all select a similar number of genes and metabolites. Grouped$_A$ is more sparse, which is consistent with the simulation results as observed by the low sensitivity and high specificity (Figures 1 and 2) when compared with Fused$_A$, Grouped$_A$, and Grouped$_B$. In addition, the genes and metabolites identified by Fused$_A$ are subsets of those identified by Fused$_B$. It is noticeable in Table 4 that there is considerable overlap of the genes and metabolites identified by the proposed methods and the existing methods considered.

We also investigate the biological relationships between the selected genes and metabolites using ToppGene Suite (Chen et al., 2009) and MetaboAnalyst 3.0 (Xia et al., 2015) respectively. These genes and metabolites are taken as input in ToppGene and MetaboAnalyst 3.0 online tools to identify pathways that are significantly enriched. The pathways that are significantly enriched in the genes selected by Fused$_B$ include mitochondrial ATP synthesis coupled proton transport and Oxidative phosphorylation. For the metabolites, the pathways identified in Fused$_B$ include purine and histidine metabolism. These pathways play essential roles in some important biological processes including orderly cell division and survival. For instance, cardiovascular research suggests that oxidative phosphorylation is implicated in mitochondrial dysfunction, a major factor in heart failure (Rosca1 et al., 2008). Also, epidemiological research suggest that uric acid, the final

end product of purine metabolism (Maiuolo et al., 2015), is an important independent risk factor for cardiovascular diseases (Alderman and Aiyer, 2004). Grouped$_A$ selected only 9 genes and 4 metabolites. It identified gene VLDLR in the reeling signaling pathway, but we regard this finding with caution.

We investigate stability of the variables identified using bootstrap resampling (50 bootstrap datasets). Table 3 (web supplementary material) shows number of genes and metabolites selected all 50 times. The sparse methods did not identify any gene or metabolite. Of the structured sparse methods, Fused$_B$ identified 132 genes and 145 metabolites, and SGCCA identified 241 metabolites but no gene. Of note, the significantly enriched pathways identified by Fused$_B$ in all 50 bootstrap datasets are those described in the previous paragraph. Figure 2 (web supplementary material) is box plots of the number of genes and metabolites, and the estimated canonical correlations. The proposed methods have less variability compared to the sparse methods. Compared to SGCCA, SGCCA is less variable.

Our analyses show that the proposed structured sparse CCA methods lead to biologically meaningful results that may shed light on the etiology of cardiovascular diseases.

6. Discussion

To address challenges associated with integrative analysis of transcriptomic and metabolomic data, we adopt an analysis strategy that is both data-driven and knowledge-based and our proposed structured sparse CCA approach allows us to not only characterize associations between two data types using a subset of relevant genes and metabolites, but also incorporate structural information from each data type. Simulation studies demonstrate that our methods achieve better performance and are robust to mis-specified and uninformative network information. Analysis of the PHI study shows that a number of gene and metabolic pathways including some known to be associated with cardiovascular diseases are enriched in the set of genes and metabolites selected by our approach.

Of the two methods proposed, our numerical studies show that Fused sparse CCA performs better than Grouped sparse CCA in terms of MCC and sensitivity, while Grouped sparse CCA outperforms Fused sparse CCA in terms of specificity. Our recommendation is to use the Fused sparse CCA (particularly, Fused$_B$) for $p \gg n$ problems, and the Grouped sparse CCA (particularly, Grouped$_A$) for small to moderate dimensional problems. The methods are implemented in MATLAB and are available at the authors website. If the graph information is not available, one can estimate network structures from observed data using approaches for sparse estimation of precision matrices (Friedman et al., 2007).

While our current work has focused on continuous data, it is of interest to develop similar methods for discrete data such as SNP data. When data are not continuous, CCA cannot be directly applied. To tackle this difficulty, one approach is to assume that there is a latent continuous variable for each discrete variable and use these latent variables to model the discrete variables where correlation among the latent variables is assessed using CCA. It is also of interest to extend our methods to conduct integrative analysis of more than two data types and assess nonlinear associations between multiple omics data types.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Emory Predictive Health Institute for providing us with the gene expression, metabolomics, and clinical data. We also thank the Associate Editor and two reviewers for their suggestions which greatly improved the manuscript. Sandra Safo's work is supported by NIH grant K12HD085850, and Qi Long's work by NIH grants R03CA173770, R03CA183006, R21NS091630 and P30CA016520. The content is the responsibility of the authors and does not represent the views of NIH.

References

Alderman M, Aiyer KJ. Uric acid: role in cardiovascular disease and effects of losartan. Current Medical Research Opinion. 2004

Cai T, Liu W. A direct estimation approach to sparse linear discriminant analysis. Journal of the American Statistical Association. 2011; 106:1566–1577.

Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics. 2007; 35:2313–2351.

Chalise P, Fridley BL. Comparison of penalty functions for sparse canonical correlation analysis. Computational Statistics and Data Analysis. 2012; 56:245–254. [PubMed: 21984855]

Chen J, Bardes EE, Aronow BJ, Jegga AG. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009; 37:W305–11. [PubMed: 19465376]

Chen J, Bushman FD, Lewis JD, Wu GD, Li H. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. Biostatistics. 2013; 14:244–258. [PubMed: 23074263]

Chen, X., Liu, H., Carbonell, JG. Structured sparse canonical correlation analysis. Proceedings of the 15th International Conference on Artificial Intelligence and Statistics; 2012.

D'Agostino RBS, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the framingham heart study. Circulation. 2008; 117:743–753. [PubMed: 18212285]

Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with graphical lasso. Biostatistics. 2007; 0:1–10.

Fujikoshi, Y., Ulyanov, VV., Shimizu, R. Wiley Series in Probability and Mathematical Statistics. 2010. Multivariate Statistics: High-Dimensional and Large-Sample Approximations.

Gao, C., Ma, Z., Zhou, HH. Sparse cca: Adaptive estimation and computational barriers. 2015. http://arxiv.org/pdf/1409.8565.pdf

Hotelling H. Relations between two sets of variables. Biometrika pages. 1936:312–377.

Kanehisa M, Sato Y, Kawashima M, MMF, Tanabe M. Kegg as a reference resource for gene and protein annotation. Nucleic Acids Research. 2016

Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008; 24:1175–1182. [PubMed: 18310618]

Li S, Park Y, Duraisingham S, Strobel HF, Khan N, Soltow AQ, Jones PD. Predicting network activity from high throughput metabolomics. PLoS Comput Biol. 2013

Maiuolo J, Oppedisano F, Gratteri S, Muscoli C, Mollace V. Regulation of uric acid metabolism and excretion. International Journal of Cardiology. 2015

Matthews B. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 1975

Pan W, Xie B, Shen X. Incorporating predictor network in penalized regression. Biometrics. 2010; 66:474–484. [PubMed: 19645699]

Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. Statistical Applications in Genetics and Molecular Biology. 2009

Rosca1 MG, Vazquez EJ, Kerner J, Parland W, Chandler MP, Stanley W, Sabbah HN, Hoppel CL. Cardiac mitochondria in heart failure: decrease in respirasomes and oxidative phosphorylation. Cardiovascular Research. 2008; 80:30–39. [PubMed: 18710878]

Safo, S., Ahn, J. PhD thesis. University of Georgia; 2014. Sparse Analysis for High Dimensional Data.

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67:91–108.

Vinod HD. Canonical ridge and econometrics of joint production. Journal of Econometrics. 1970:147–166.

Witten DM, Tibshirani RJ, Hastie T. A penalized matrix decomposition, with applications to sparse prinicial components and canonical correlation analysis. Biostatistics. 2009; 10:515–534. [PubMed: 19377034]

Xia J, Sinelnikov IV, Han B, Wishart DS. Metaboanalyst 3.0 making metabolomics more meaningful. Nucleic Acids Research. 2015

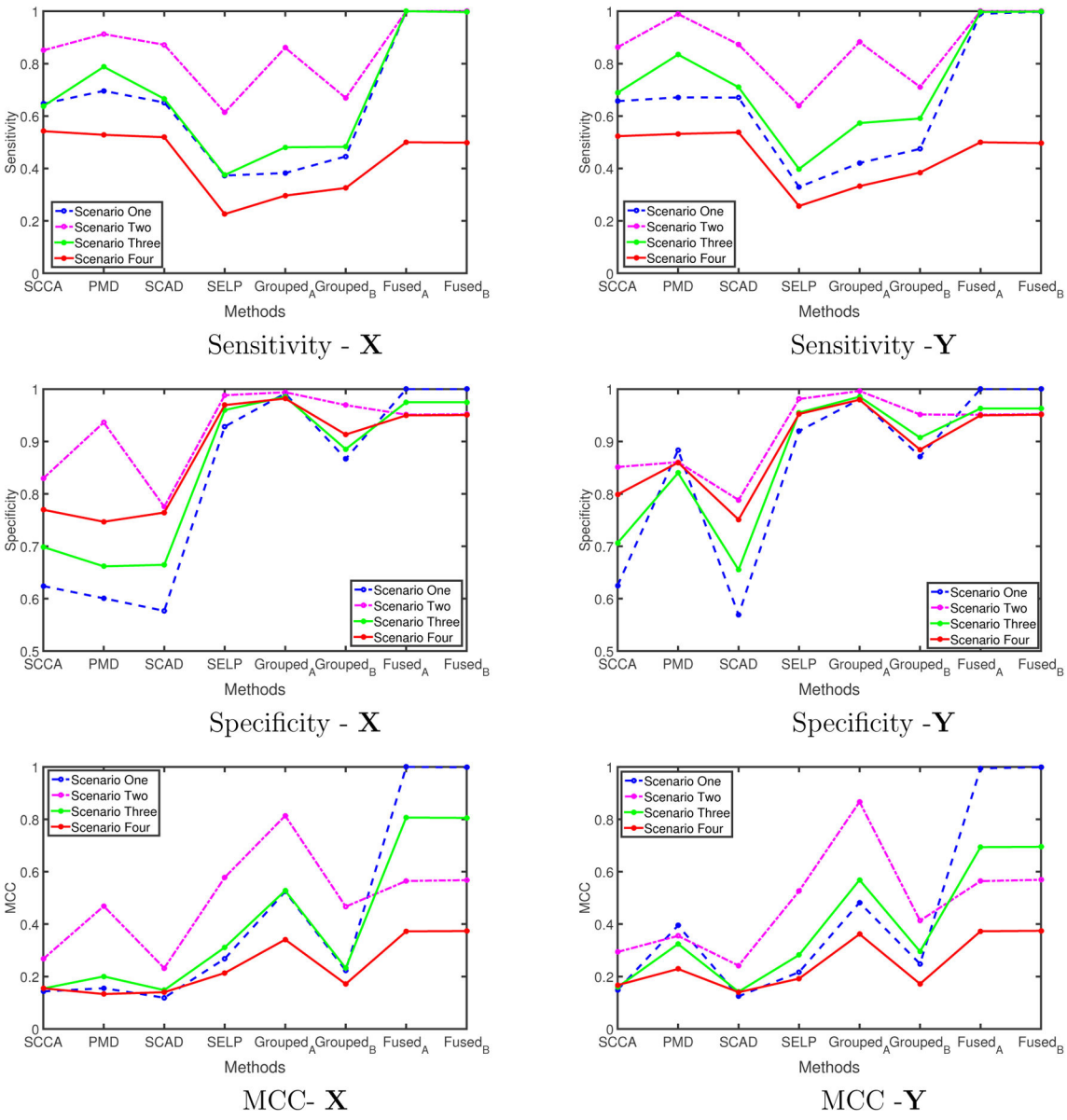

Figure 1.

Comparison of structured sparse CCA with existing sparse CCA methods under scenarios one to four with $\rho = 0.9$. $\gamma = 2$ for Grouped$_A$ and Grouped$_B$. MCC, Matthew's correlation coefficient.

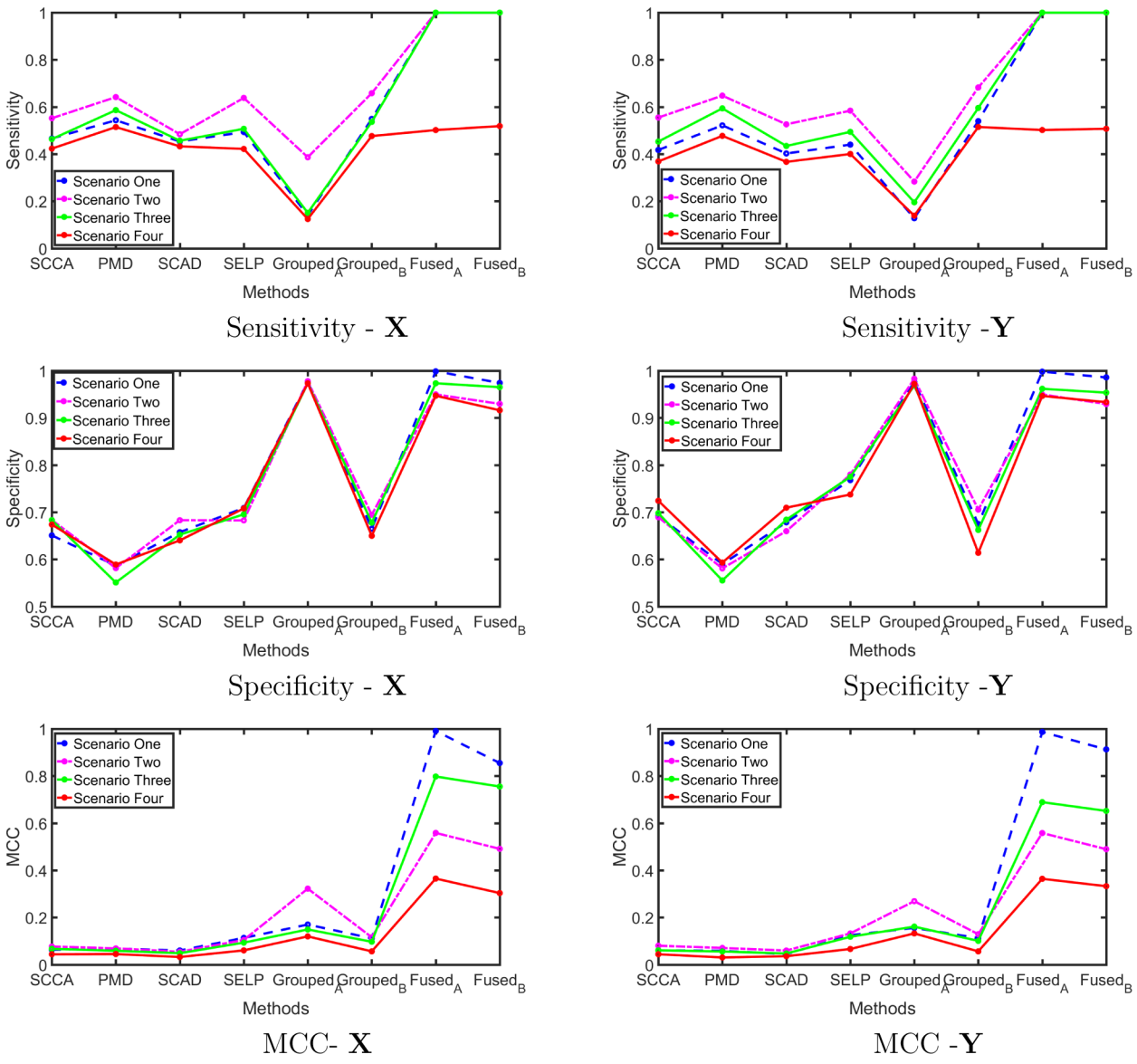

Figure 2.

Comparison of structured sparse CCA with existing sparse CCA methods under scenarios one to four with $\rho = 0.5$. $\gamma = 2$ for Grouped$_A$ and Grouped$_B$. MCC, Matthew's correlation coefficient.

Table 1

Comparison of proposed methods with existing structured sparse CCA methods under scenarios one to four with $\rho = 0.9$, $\gamma = 2$ for Grouped$_A$ and Grouped$_B$. MCC, Matthew's correlation coefficient.

	Sensitivity-X	Specificity-X	MCC-X	Sensitivity-Y	Specificity-Y	MCC -Y
Scenario One						
SGCCA	0.2771	0.8698	0.1092	0.8769	0.0187	−0.1689
Grouped$_A$	0.3825	0.9917	0.5242	0.4208	0.9803	0.4823
Grouped$_B$	0.4460	0.8668	0.2229	0.4747	0.8713	0.2481
Fused$_A$	1.0000	1.0000	1.0000	0.9903	1.0000	0.9946
Fused$_B$	0.9972	1.0000	0.9985	0.9982	1.0000	0.9990
Scenario Two						
SGCCA	0.6296	0.7437	0.1294	0.9213	0.0187	−0.0655
Grouped$_A$	0.8608	0.9940	0.8139	0.8825	0.9964	0.8670
Grouped$_B$	0.6692	0.9694	0.4668	0.7113	0.9515	0.4133
Fused$_A$	1.0000	0.9512	0.5647	1.0000	0.9511	0.5642
Fused$_B$	1.0000	0.9520	0.5680	1.0000	0.9523	0.5693
Scenario Three						
SGCCA	0.3781	0.8192	0.1074	0.9094	0.0187	−0.0927
Grouped$_A$	0.4808	0.9854	0.5278	0.5733	0.9854	0.5687
Grouped$_B$	0.4827	0.8854	0.2323	0.5906	0.9073	0.2956
Fused$_A$	1.0000	0.9749	0.8067	0.9972	0.9630	0.6939
Fused$_B$	0.9969	0.9750	0.8055	0.9986	0.9631	0.6952
Scenario Four						
SGCCA	0.2160	0.8666	0.0513	0.9148	0.0190	−0.0962
Grouped$_A$	0.2963	0.9820	0.3410	0.3325	0.9797	0.3618
Grouped$_B$	0.3254	0.9133	0.1715	0.3846	0.8845	0.1719
Fused$_A$	0.5000	0.9498	0.3724	0.5000	0.9499	0.3728
Fused$_B$	0.4985	0.9505	0.3737	0.4967	0.9512	0.3746

Biometrics. Author manuscript; available in PMC 2018 March 15.

Table 2

Comparison of proposed methods with existing structured sparse CCA methods under scenarios one to four with $\rho = 0.5$, $\gamma = 2$ for Grouped_A and Grouped_B. MCC, Matthew's correlation coefficient.

	Sensitivity-X	Specificity-X	MCC-X	Sensitivity-Y	Specificity-Y	MCC -Y
Scenario One						
SGCCA	0.0624	0.9682	0.0435	0.7931	0.0179	−0.2795
Grouped_A	0.1399	0.9763	0.1702	0.1281	0.9765	0.1560
Grouped_B	0.5492	0.6649	0.1116	0.5398	0.6751	0.1127
Fused_A	0.9993	0.9988	0.9913	1.0000	0.9981	0.9869
Fused_B	1.0000	0.9742	0.8552	1.0000	0.9859	0.9131
Scenario Two						
SGCCA	0.0525	0.9716	0.0220	0.7729	0.0175	−0.2159
Grouped_A	0.3867	0.9779	0.3225	0.2833	0.9830	0.2694
Grouped_B	0.6583	0.6945	0.1163	0.6825	0.7056	0.1292
Fused_A	1.0000	0.9498	0.5588	1.0000	0.9497	0.5583
Fused_B	1.0000	0.9298	0.4913	1.0000	0.9293	0.4898
Scenario Three						
SGCCA	0.0740	0.9629	0.0407	0.8417	0.0181	−0.1736
Grouped_A	0.1510	0.9733	0.1494	0.1956	0.9686	0.1614
Grouped_B	0.5365	0.6778	0.0972	0.5956	0.6624	0.1010
Fused_A	1.0000	0.9734	0.7981	1.0000	0.9618	0.6893
Fused_B	1.0000	0.9652	0.7555	1.0000	0.9536	0.6521
Scenario Four						
SGCCA	0.0615	0.9566	0.0188	0.8925	0.0187	−0.1268
Grouped_A	0.1246	0.9733	0.1199	0.1390	0.9722	0.1326
Grouped_B	0.4765	0.6500	0.0565	0.5148	0.6139	0.0564
Fused_A	0.5019	0.9470	0.3652	0.5021	0.9466	0.3642
Fused_B	0.5188	0.9162	0.3037	0.5073	0.9334	0.3329

Table 3

Number of genes and metabolites selected in the first canonical correlation vector in the PHI study.

	Genes Selected $\hat{a}_1$	Metabolites selected $\hat{\beta}_1$	Correlation Coefficient $\hat{\rho}_1$
SCCA	86	154	0.7248
PMD	654	36	0.8745
SCAD	31	252	0.7036
SELP	508	152	0.8982
SGCCA	755	252	0.8632
Grouped$_A$	9	4	0.8168
Grouped$_B$	535	137	0.9871
Fused$_A$	297	146	0.8658
Fused$_B$	536	168	0.9814

Table 4

Overlapping genes and metabolites selected in the first canonical correlation vectors in the PHI study. (·, ·) represents number of genes and metabolites common for each pair of method compared.

	SCCA	**PMD**	**SCAD**	**SELP**	**SGCCA**	**Grouped$_A$**	**Grouped$_B$**	**Fused$_A$**	**Fused$_B$**
SCCA	(86,154)								
PMD	(14,20)	(654,36)							
SCAD	(31,154)	(10,36)	(31,252)						
SELP	(16,92)	(503,33)	(11,152)	(508,152)					
SGCCA	(24,154)	(577,36)	(11,252)	(482,252)	(755,252)				
Grouped$_A$	(0,2)	(9,3)	(0,4)	(9,3)	(9,4)	(9,4)			
Grouped$_B$	(17,82)	(427,30)	(9,137)	(383,101)	(445,137)	(9,3)	(535,137)		
Fused$_A$	(13,89)	(124,24)	(4,146)	(91,91)	(132,146)	(2,4)	(92,77)	(297,146)	
Fused$_B$	(21,104)	(342,31)	(8,168)	(297,108)	(349,168)	(9,4)	(323,98)	(297,146)	(536,168)