Complete Model Selection in Multiset Canonical Correlation Analysis

Tim Marrinan, Tanuj Hasija, Christian Lameiro, and Peter J. Schreier
Signal and System Theory Group, Universität Paderborn, Germany

***Abstract*—Traditional model-order selection for canonical correlation analysis infers latent correlations between two sets of noisy data. In this scenario it is enough to count the number of correlated signals, and thus the model order is a scalar. When the problem is generalized to a collection of three or more data sets, signals can demonstrate correlation between all sets or some subset, and one number cannot completely describe the correlation structure. We present a method for estimating multiset correlation structure that combines source extraction in the style of joint blind source separation with pairwise model order selection. The result is a general technique that describes the complete correlation structure of the collection.**

***Index Terms*—canonical correlation, hypothesis testing, joint blind source separation, MCCA, order selection**

I. Introduction

Canonical correlation analysis (CCA) and its multiset generalization (MCCA) are perhaps the most commonly used tools for describing the linear association between sets of variables [1]–[3]. The ease of implementation and interpretability of these techniques has led to applications in climate science, genomics, visual tracking, and data fusion, to name a few [4]–[8]. However, estimates of the correlation coefficients generated from observed data can be significantly distorted, to the point that separating the true correlations from spurious correlations due to noise is a nontrivial task. The problem of identifying correlated signal components is exacerbated for a collection of three or more data sets when components can be correlated across subsets of the collection. In this case, pairwise solutions for estimating the number of correlated signals are insufficient.

The goal of this work is to detect the correlation structure present across multiple data sets. We present a method for determining the number of signal components correlated across more than one set from a collection and across which data sets these components are correlated. To the best of our knowledge this is the first attempt to characterize the correlation structure of multiple data sets that distinguishes and enumerates the correlations between subsets of the total collection. Previous works in this area include methods for determining the number of correlated components between just two data sets [9], [10], or when there are more than two data sets, the total number of correlated components [6] and the number of components correlated strictly across all sets [11]–[13].

As is common in CCA, the proposed technique determines linear combinations of signal components, called canonical variates, from each data set at each stage of the algorithm [1]. The canonical variates are chosen such that they are highly correlated with those from the other sets at each stage, but uncorrelated with the canonical variates from within a set at different stages, as suggested by [2]. This extraction process has also been studied in the joint blind source separation (BSS) regime, where the canonical variates are referred to as extracted sources or endmembers. See, for example [14]. After obtaining the complete collection, inner products of same-stage canonical variates estimate each signal component's correlation between a pair of data sets. The number of significant correlations for each pair is determined using a method that applies a series of binary hypothesis tests based on the Bartlett-Lawley statistic [9]. The significant correlations are then identified with the canonical variates that have the largest inner products, and thus correlated components can be resolved at each stage whether or not they span all sets.

II. Problem statement

We consider the P-set model

$$\mathbf{x}_p = \mathbf{A}_p \mathbf{s}_p + \boldsymbol{\nu}_p \quad \text{for} \quad p = 1, \dots, P. \tag{1}$$

The mixing matrices, $\mathbf{A}_p \in \mathbb{R}^{n_p \times m_p}$, are deterministic but unknown with full column rank and unknown $m_p \leq n_p$. The signals, $\mathbf{s}_p \in \mathbb{R}^{m_p}$, are jointly Gaussian random vectors whose components have zero mean and standard deviation $\sigma_p^{(i)}$ for $i = 1, \dots, m_p$. The cross-covariance matrices are diagonal for $p \neq q$ and $q = 1, \dots, P$,

$$\mathbf{\Sigma}_{pq} = \text{diag}(\rho_{pq}^{(1)} \sigma_p^{(1)} \sigma_q^{(1)}, \dots, \rho_{pq}^{(m_{pq})} \sigma_p^{(m_{pq})} \sigma_q^{(m_{pq})}) \tag{2}$$

where $m_{pq} = \min\{m_p, m_q\}$ and $\rho_{pq}^{(i)}$ is the unknown (possibly zero) correlation coefficient between the ith components in $\mathbf{s}_p$ and $\mathbf{s}_q$. The assumption of diagonal cross-covariance matrices is a restriction but a common one in multiset BSS, because it leads to nice separability conditions for the sources [14]. We assume the auto-covariance matrices $\mathbf{\Sigma}_{pp}$ are diagonal (which can be done without loss of generality). The noise vectors $\boldsymbol{\nu}_p \in \mathbb{R}^{n_p}$ are independent of the signals and of each other. They are zero-mean Gaussian with unknown variance, but not necessarily white.

Problem: *From M independent and identically distributed (i.i.d.) samples jointly observed from the P sets in model* (1)*, determine the number of nonzero correlation coefficients, d_{pq}, for $p \neq q$. For the jth nonzero correlation coefficients between data sets p and q, $\rho_{pq}^{(j)}$, identify all other values of p and q for which $\rho_{pq}^{(j)}$ is nonzero.*

This work was supported by the German Research Foundation (DFG), under grant SCHR 1384/3-1.

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE I: The correlation structures described in Section III and used for the performance evaluations are shown in Table Ia and Table Ib. The entries represent the correlation coefficients between different data sets for each signal component.

	$\rho_{12}^{(i)}$	$\rho_{13}^{(i)}$	$\rho_{14}^{(i)}$	$\rho_{23}^{(i)}$	$\rho_{24}^{(i)}$	$\rho_{34}^{(i)}$
$\mathbf{s}^{(1)}$	0	0	0	0.9	0.9	0.9
$\mathbf{s}^{(2)}$	0.8	0	0.8	0	0.8	0
$\mathbf{s}^{(3)}$	0	0	0	0	0.7	0
$\mathbf{s}^{(4)}$	0	0	0	0	0	0
$\mathbf{s}^{(5)}$	0	0	0	0	0	0

(a) CorrStruct1: A 4-set model with 2 signal components correlated across 3 sets, 1 correlated between a pair, and 2 independent.

	$\rho_{12}^{(i)}$	$\rho_{13}^{(i)}$	$\rho_{14}^{(i)}$	$\rho_{23}^{(i)}$	$\rho_{24}^{(i)}$	$\rho_{34}^{(i)}$
$\mathbf{s}^{(1)}$	0.9	0.9	0.9	0.9	0.9	0.9
$\mathbf{s}^{(2)}$	0	0	0	0	0.8	0
$\mathbf{s}^{(3)}$	0	0	0	0	0	0
$\mathbf{s}^{(4)}$	0	0	0	0	0	0
$\mathbf{s}^{(5)}$	0	0	0	0	0	0

(b) CorrStruct2: A 4-set model with 1 signal component correlated across all sets, 1 correlated between a pair, and 3 independent.

III. Motivating example

Suppose that we have 4 sets of data following the model in (1) with 5 signal components per set. To motivate the need for complete model selection, we describe the partial solutions provided by three existing techniques for two correlation structures defined on this model. The first structure (CorrStruct1) is defined as in Table Ia. Each element in this table shows the correlation coefficient, $\rho_{pq}^{(i)}$, for different values of p, q, and i. There are 7 nonzero correlation coefficients in CorrStruct1. The first signal component has a correlation coefficient of 0.9 in sets 2-4, so we have $\rho_{23}^{(1)} = \rho_{24}^{(1)} = \rho_{34}^{(1)} = 0.9$. The second component has $\rho_{12}^{(2)} = \rho_{14}^{(2)} = \rho_{24}^{(2)} = 0.8$. The third component of sets 2 and 4 has a correlation of $\rho_{24}^{(3)} = 0.7$. All other component pairs are uncorrelated. The second correlation structure (CorrStruct2) also has 7 nonzero correlation coefficients, as shown in Table Ib. However, the first signal component is correlated across all sets with correlation coefficients of 0.9, and the second component is correlated in sets 2 and 4, $\rho_{24}^{(2)} = 0.8$. The remaining component pairs are uncorrelated.

Considering all data sets simultaneously, [12] estimates the rank of the product of all of the pairwise coherence matrices. This rank reflects the number of signals correlated strictly across all of the data sets in a collection. For CorrStruct1 this method would estimate $d_{\text{all}} = 0$ and for CorrStruct2 it would find $d_{\text{all}} = 1$.

In a second interpretation of the model order, Informative MCCA (IMCCA) method [6] finds an eigenvalue threshold to determine how many signal components in the entire collection bear nonzero correlations. With CorrStruct1 ideally the method finds $d_{\text{total}} = 3$, and for CorrStruct2 $d_{\text{total}} = 2$.

Finally, three detectors based on the max min PCA-CCA paradigm are implemented in [9]. The techniques jointly estimate a reduced dimensionality for two data sets and determine the number of correlated components between the pair d_{pq}. The first two detectors use a sequence of binary hypothesis tests based on the Bartlett-Lawley statistic, while the third applies a test based on an information theoretic criterion. The max min PCA-CCA strategy also plays a role in our proposed method, where the number of pairwise correlations sets a threshold for how many jointly estimated components we identify. For CorrStruct1 max min PCA-CCA would estimate $d_{12} = 1, d_{13} = 0, d_{14} = 1, d_{23} = 1, d_{24} = 3$, and $d_{34} = 1$. It is important to note that this method alone could not identify whether the components correlated between sets 1 and 2 are the same components correlated between data sets 2 and 3. For CorrStruct2 it would find $d_{12} = d_{13} = d_{14} = d_{23} = d_{34} = 1$ and $d_{24} = 2$.

Each of these methods casts light on one aspect of the correlation structures. In contrast, we propose a solution to identify which values of $\rho_{pq}^{(i)}$ are nonzero. When successful, counting the nonzero correlations down the appropriate rows or columns of our estimates for Table Ia and Table Ib will give a proxy for the model order estimated by each of these existing techniques.

IV. Background

A. Review of Multiset Canonical Correlation Analysis

In 1971, [2] compiled a collection of new and existing criteria for analyzing the correlation of several sets of variables, which was more recently revisited by [15] to include possible constraints. Each of the criteria reduces to traditional CCA when applied to just two data sets, but provides a different perspective on the correlation structure with a larger number of data sets.

One formulation of the MCCA problem, referred to as the MAXVAR criterion, extends from traditional CCA in a natural way. Define an augmented signal vector $\mathbf{x} = [\mathbf{x}_1^T \mathbf{x}_2^T \cdots \mathbf{x}_P^T]^T$. The augmented covariance matrix between all data sets can then be constructed as

$$\mathbf{R} = \mathbb{E}[\mathbf{x}\mathbf{x}^T] = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} & \cdots & \mathbf{R}_{1P} \\ \mathbf{R}_{21} & \mathbf{R}_{22} & \cdots & \mathbf{R}_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{P1} & \mathbf{R}_{P2} & \cdots & \mathbf{R}_{PP} \end{bmatrix} \qquad (3)$$

and we define $\mathbf{R}_D = \operatorname{diag}(\mathbf{R}_{11}, \mathbf{R}_{22}, \ldots, \mathbf{R}_{PP})$.

The coherence matrix of $\mathbf{x}_p$ and $\mathbf{x}_q$ is a whitened version of their covariance matrix, and is defined by

$$\mathbb{E}[(\mathbf{R}_{pp}^{-1/2}\mathbf{x}_p)(\mathbf{R}_{qq}^{-1/2}\mathbf{x}_q)^T] = \mathbf{R}_{pp}^{-1/2}\mathbf{R}_{pq}\mathbf{R}_{qq}^{-T/2}. \qquad (4)$$

The augmented coherence matrix for the collection is then constructed similarly as $\mathbf{R}_D^{-1/2}\mathbf{R}\mathbf{R}_D^{-T/2}$. MCCA with the MAXVAR criterion searches for a vector $\mathbf{w}^{(j)} \in \mathbb{R}^{\sum_{p=1}^{P} n_p}$ that solves

$$\underset{\mathbf{w}^{(j)} \in \mathbb{R}^{\sum_{p=1}^{P} n_p}}{\arg\max} \quad \mathbf{w}^{(j)T}\mathbf{R}_D^{-1/2}\mathbf{R}\mathbf{R}_D^{-T/2}\mathbf{w}^{(j)} \tag{5}$$

at the jth stage for $j = 1, \dots, \min\{n_p\}_{p=1}^{P}$, and is often required to satisfy the constraints

$$\begin{aligned} &\mathbf{w}^{(j)T}\mathbf{w}^{(j)} = 1 \qquad\qquad \text{and} \\ &\mathbf{w}^{(j)T}\mathbf{w}^{(k)} = 0 \text{ for } k < j. \end{aligned} \tag{6}$$

The solution vector $\mathbf{w}^{(j)}$ contains coefficients for a linear combination of the signal components in each data set. The random variable created for each set by this linear combination, $z_p^{(j)} = \mathbf{w}_p^{(j)T}\mathbf{R}_{pp}^{-1/2}\mathbf{x}_p$, is referred to as the jth stage canonical variate of the pth data set. The coefficients for each data set are concatenated in the coefficient vectors in same order as were the signals that generated the augmented covariance matrix. That is, $\mathbf{w}^{(j)} = [\mathbf{w}_1^{(j)T}\mathbf{w}_2^{(j)T} \cdots \mathbf{w}_P^{(j)T}]^T$ with $\mathbf{w}_p^{(j)} \in \mathbb{R}^{n_p}$. Rewriting this problem as a Lagrangian makes it clear that (5) is maximized for the first stage when $\mathbf{w}^{(1)}$ is the eigenvector associated with the dominant eigenvalue of the augmented coherence matrix, $\mathbf{R}_D^{-1/2}\mathbf{R}\mathbf{R}_D^{-T/2}$.

For successive stages we require canonical variates from within a particular data set to be uncorrelated, as in Equation $(9 \cdot 5)$ of [2]. In other words,

$$\mathbb{E}\left[z_p^{(j)} z_p^{(k)}\right] = 0 \tag{7}$$

for $p = 1, \dots, P$ and for all $k < j$. This is enforced with a deflationary procedure where the augmented coherence matrix is recomputed at stage j after projecting each data set onto the orthogonal complement of its existing canonical variates, $\{z_p^{(1)}, \dots, z_p^{(j-1)}\}$. Problem (5) with constraints (6) and (7) is optimized when $\mathbf{w}^{(j)}$ is the dominant eigenvector of this updated augmented coherence matrix.

The set of canonical variates for each data set forms an uncorrelated basis, and provides a solution to the BSS problem in certain scenarios. For example, [14] proposes conditions for separability when the cross-covariance matrix between pairs of data sets is diagonal with nonincreasing elements. This assumption is more limiting than our assumption of a diagonal cross-covariance matrix with no restriction on the order of the diagonal elements, but necessary conditions for separability have not yet been established in this more general context.

B. Empirical Multiset Canonical Correlation Analysis

Let $\widehat{\mathbf{X}}_p \in \mathbb{R}^{n_p \times M}$ for $p = 1, \dots, P$ be the matrices containing the M observations from each each data set. In practice $\mathbf{R}_{pp}$ and $\mathbf{R}_{pq}$ are typically unknown and must be estimated from samples, so we replace them with their maximum likelihood (ML) estimates, $\widehat{\mathbf{R}}_{pp} = \frac{1}{M}\widehat{\mathbf{X}}_p\widehat{\mathbf{X}}_p^T$ and $\widehat{\mathbf{R}}_{pq} = \frac{1}{M}\widehat{\mathbf{X}}_p\widehat{\mathbf{X}}_q^T$. Sample coherence matrices can be found without inversion via the compact singular value decomposition (SVD) following [16]. For $p = 1, \dots, P$, let $\widehat{\mathbf{X}}_p = \widehat{\mathbf{F}}_p\widehat{\mathbf{K}}_p\widehat{\mathbf{G}}_p^T$ be a compact SVD so that the pairwise coherence matrices are written as $\widehat{\mathbf{R}}_{pp}^{-1/2}\widehat{\mathbf{R}}_{pq}\widehat{\mathbf{R}}_{qq}^{-T/2} = \widehat{\mathbf{F}}_p\widehat{\mathbf{G}}_p^T\widehat{\mathbf{G}}_q\widehat{\mathbf{F}}_q^T$. If we define $\widehat{\mathbf{F}} = \operatorname{diag}(\widehat{\mathbf{F}}_1, \widehat{\mathbf{F}}_2, \dots, \widehat{\mathbf{F}}_P)$ and $\widehat{\mathbf{G}} = [\widehat{\mathbf{G}}_1\widehat{\mathbf{G}}_2 \cdots \widehat{\mathbf{G}}_P]$, the sample augmented coherence matrix is then $\widehat{\mathbf{F}}\widehat{\mathbf{G}}^T\widehat{\mathbf{G}}\widehat{\mathbf{F}}^T$.

In the first stage, $\widehat{\mathbf{w}}^{(1)}$ is the dominant eigenvector of $\widehat{\mathbf{F}}\widehat{\mathbf{G}}^T\widehat{\mathbf{G}}\widehat{\mathbf{F}}^T$, and

$$\widehat{\mathbf{z}}_p^{(1)} = \frac{\sqrt{M}(\widehat{\mathbf{w}}_p^{(1)T}\mathbf{F}_p\mathbf{G}_p^T)}{\|\widehat{\mathbf{w}}_p^{(1)T}\mathbf{F}_p\mathbf{G}_p^T\|} \in \mathbb{R}^M \tag{8}$$

is a vector containing the sample canonical variates, normalized to have sample variance 1. For $j > 1$, a deflationary procedure is used to enforce (7). Define a matrix

$$\widehat{\mathbf{Z}}_p^{(j-1)} = \left[\frac{\widehat{\mathbf{z}}_p^{(1)T}}{\|\widehat{\mathbf{z}}_p^{(1)}\|}\, \frac{\widehat{\mathbf{z}}_p^{(2)T}}{\|\widehat{\mathbf{z}}_p^{(2)}\|} \cdots \frac{\widehat{\mathbf{z}}_p^{(j-1)T}}{\|\widehat{\mathbf{z}}_p^{(j-1)}\|}\right] \in \mathbb{R}^{M \times (j-1)} \tag{9}$$

for $p = 1, \dots, P$, so that $\widehat{\mathbf{X}}_p^{(j)} = \widehat{\mathbf{X}}_p(\mathbf{I} - \widehat{\mathbf{Z}}_p^{(j-1)}\widehat{\mathbf{Z}}_p^{(j-1)T})$ is the deflated pth data set. Each compact SVD is then recomputed, $\widehat{\mathbf{X}}_p^{(j)} = \widehat{\mathbf{F}}_p^{(j)}\widehat{\mathbf{K}}_p^{(j)}\widehat{\mathbf{G}}_p^{(j)T}$ for $p = 1, \dots, P$, and $\widehat{\mathbf{w}}^{(j)}$ is the dominant eigenvector of the jth stage augmented sample coherence matrix. Finally, the vector containing samples of the jth canonical variate of data set p is

$$\widehat{\mathbf{z}}_p^{(j)} = \frac{\sqrt{M}(\widehat{\mathbf{w}}_p^{(j)T}\mathbf{F}_p^{(j)}\mathbf{G}_p^{(j)T})}{\|\widehat{\mathbf{w}}_p^{(j)T}\mathbf{F}_p^{(j)}\mathbf{G}_p^{(j)T}\|} \in \mathbb{R}^M. \tag{10}$$

V. Pairwise Model Order for Multiset Data

Empirical MCCA determines sets sample canonical variates, $\{\widehat{\mathbf{z}}_1^{(j)}, \dots, \widehat{\mathbf{z}}_P^{(j)}\}$ for $j = 1, \dots, \min\{n_p\}_{p=1}^{P}$. When these sets solve the BSS problem, the jth set of sample canonical variates will approximate the ith set of signal components (with some ambiguity about which i). Thus the inner products of same-stage sample canonical variates provide the sample canonical correlations, $\widehat{\rho}_{pq}^{(j)} = |\frac{1}{M}\widehat{\mathbf{z}}_p^{(j)}\widehat{\mathbf{z}}_q^{(j)T}|$, and approximate the true correlations $\rho_{pq}^{(i)}$. The accuracy of this approximation is affected by numerous parameters of the observed data including the power of the noise, the number of data sets in the collection, the number of sets across which a particular signal component is correlated, and the number of observed samples. To bring this technique into the model-selection paradigm, we determine which values of $\widehat{\rho}_{pq}^{(j)}$ are significant for $j = 1, \dots, n_{pq}$ with $n_{pq} = \min\{n_p, n_q\}$ for $p \neq q$.

In the case of two data sets with M samples, the sample canonical correlations are ML estimates of the true correlations. Thus the number of correlated components, $d_{pq} \leq n_{pq}$, between the random vectors $\mathbf{x}_p \in \mathbb{R}^{n_p}$ and $\mathbf{x}_q \in \mathbb{R}^{n_q}$ can be estimated through a series of binary hypothesis tests based on the Bartlett-Lawley statistic [17], [18]. Starting with $s = 0$ the test compares the null hypothesis $H_0 : d_{pq} = s$ with the alternative hypothesis $H_1 : d_{pq} > s$. If H_0 is rejected, the value of s is incremented, and the test is repeated until H_0 is not rejected or the maximum dimension is reached. The Bartlett-Lawley statistic,

$$C(s) = -\left(M - s - \frac{n_p + n_q + 1}{2} + \sum_{j=1}^{s} \frac{1}{(\widehat{\rho}_{pq}^{(j)})^2} \right) \times \ln \prod_{j=s+1}^{n_{pq}} \left(1 - (\widehat{\rho}_{pq}^{(j)})^2 \right), \tag{11}$$

is a function of the number of correlated components, s, being tested. However, parameters that include the sample canonical correlations, $\{\widehat{\rho}_{pq}^{(j)}\}_{j=1}^{n_{pq}}$, and the dimensions of the data sets, n_p and n_q, are also required to define the statistic. The asymptotic distribution of $C(s)$ under H_0 is χ^2 with $(n_p - s)(n_q - s)$ degrees of freedom. In the finite sample regime, this Bartlett-Lawley statistic contains a correction term to match the moments of the χ^2-distribution. Using this statistic allows the computation of a test threshold, $T(s)$, for each hypothesis test with a given probability of false alarm.

In the multiset scenario, $\widehat{\rho}_{pq}^{(j)} = |\frac{1}{M}\widehat{\mathbf{z}}_p^{(j)}\widehat{\mathbf{z}}_q^{(j)T}|$ is not an ML estimate of the pairwise correlation coefficient. This means that $C(s)$ for $s = d_{pq}$ is not guaranteed to follow the χ^2 distribution, and a test based on the Bartlett-Lawley statistic will underestimate the number of correlated components. This hurdle is overcome by pairwise estimating the number of correlated components between each combination of data sets, and then identifying which of the multiset sample canonical variates for those two data sets have the highest correlations.

VI. Proposed model selection framework

Although the multiset sample canonical correlations are smaller than their pairwise counterparts, the jointly estimated canonical variates with the largest inner products are certainly the ones that demonstrate the greatest correlation. Thus, signal components correlated across any number of data sets can be identified using the following three-step approach.

1) Compute sample canonical variates $\{\widehat{\mathbf{z}}_1^{(j)}, \dots, \widehat{\mathbf{z}}_P^{(j)}\}$ for stages $j = 1, \dots, \min\{n_p\}_{p=1}^{P}$ using the MCCA method with the MAXVAR criterion and constraints (6) and (7).
2) Estimate the pairwise model order for each combination of data sets. This is done by computing ML estimates of the canonical correlations between $\mathbf{x}_p$ and $\mathbf{x}_q$, and applying the sequence of hypothesis tests based on the Bartlett-Lawley statistic to find d_{pq}.
3) Identify which sample canonical variates have the largest magnitude, same-stage inner products for each pair of data sets. If $|\frac{1}{M}\widehat{\mathbf{z}}_p^{(j)}\widehat{\mathbf{z}}_q^{(j)T}|$ is one of the d_{pq} largest inner products for data set p and data set q, a nonzero correlation is identified between the ith signal components of $\mathbf{x}_p$ and $\mathbf{x}_q$. Note that these will not necessarily correspond to stages $1, \dots, d_{pq}$ of the MCCA method.

In practical situations, estimates of correlation are affected by the number of observed samples. When there are few observations relative to the number of signal components, canonical correlations are overestimated. Ideally a method for multiset model selection would jointly estimate a reduced dimension for all data sets, but this is beyond the scope of the current work. However, pairwise correlation estimates can be improved in small sample scenarios with detectors based on a max min PCA-CCA paradigm. Detector 1 in [9] is one such technique, and the proposed algorithm employs this method in Step 2 to provide a better estimate of pairwise model order over the Bartlett-Lawley statistic without rank reduction.

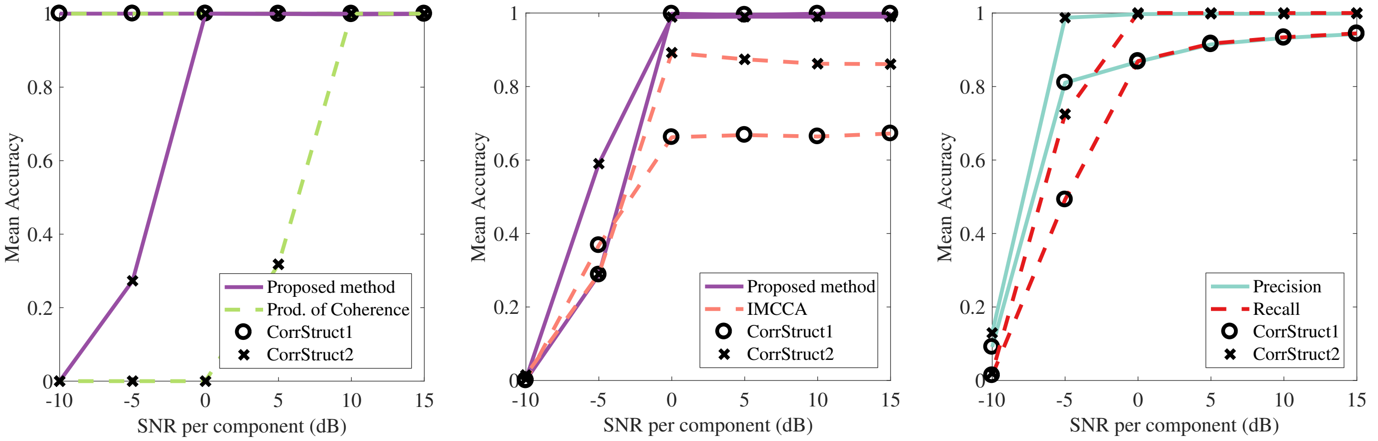
VII. Numerical Evaluations

The result of the proposed method is an estimate of which correlation coefficients are nonzero, and its performance can be evaluated against the techniques in Section III by counting the number the estimated nonzero correlations in the appropriate manner. To compare with the product of coherence matrices method [12] we determine d_{all} by counting the number of rows that have a nonzero correlation in every column of the estimated versions of Table Ia and Ib. To evaluate against IMCCA [6] we estimate d_{total} by counting how many rows have at least one nonzero correlation. We do not evaluate against max min PCA-CCA [9], because it is used in the proposed method and results would be identical.

These comparisons have been simulated for data sets with $M = 500$ samples using CorrStruct1 (Table Ia) and CorrStruct2 (Table Ib). The mixing matrices have $n_p = 5$ orthogonal rows. Correlated signals have a variance of 10, independent signals have a variance of 3, and zero-mean white noise is added to create SNRs of -10 dB to 15 dB. 1000 bootstrap resamplings were run for the product of coherence matrices method. The number of 'informative' dimensions for IMCCA was chosen to preserve 90% of the variance, and a threshold for eigenvalues associated with correlated components was estimated from empirical distributions of 1000 randomly generated data sets. For all methods, the probability of false alarm was set at 0.001, and 1000 Monte Carlo simulations were run for each experiment. Fig. 1a and Fig. 1b plot the mean accuracy for each task, that is, the number of times the estimated value was equal to the true value divided by the number of simulations. Additionally, Fig. 1c shows the precision (the number of true positives divided by the sum of the true positives and false positives) and recall (the number of true positives divided by the sum of the true positives and false negatives) for nonzero correlation coefficients using the proposed method. For all plots, the '∘' marker denotes CorrStruct1 and the '×' marker indicates CorrStruct2. Solid purple lines show the results of the proposed method.

In Fig. 1a we see high accuracy from both the proposed method and the product of coherence matrices method (green dashed lines) in identifying the components correlated across all data sets for high SNR. With CorrStruct2 both methods eventually fail as the noise variance increases, however detection persists for lower SNR with the proposed method. The lowered accuracy is associated with both methods incorrectly estimating $d_{\text{all}} = 1$ correlation across all data sets.

Fig. 1b shows that the proposed method outperforms IMCCA (pink dashed lines) in estimating the total number of correlated components for nearly all SNRs. The inconsistency



(a) Correlations across all data sets, d_{all} (b) Total number of correlated signals, d_{total} (c) Proposed method precision and recall

Fig. 1: Model (order) selection evaluation of the proposed method for two different correlation structures with 4 data sets and 7 nonzero correlations in each case. CorrStruct1 is summarized in Table Ia and CorrStruct2 is described in Table Ib.

in the accuracy of the IMCCA method between CorrStruct1 and CorrStruct2 is due to the total number of correlated signals being underestimated in CorrStruct1, with the method often reporting $d_{\text{total}} = 2$ instead of 3.

The precision and recall plot in Fig. 1c shows that the proposed method does better with data from CorrStruct2 for high SNR, and that both precision and recall are high when the noise variance is less than or equal to that of the correlated components. When the noise is stronger than the correlated components, it is harder to identify nonzero correlation coefficients and the method tends to underestimate the number of correlated components.

VIII. Conclusion

We have presented a method for estimating which signals are correlated in multiple sets of data, whether between all data sets or a subset. The method provides better results than existing techniques for estimating model order, while providing more information about the overall model. Future work is required to identify necessary conditions for blind source separation, and ways of extracting sample canonical variates when those conditions are not met. Additionally, model order selection is most difficult when sample support is small, and it remains to find a method for jointly reducing the dimension of all data sets. Code for the proposed method will be made publicly available at `https://github.com/SSTGroup`.

References

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[2] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, pp. 433–451, 1971.

[3] J. D. Carroll, "Generalization of canonical correlation analysis to three or more sets of variables," in *Proc. Annu. Conv., Am. Psych. Assoc.*, vol. 3, 1968, pp. 227–228.

[5] Y. Yamanishi, J.-P. Vert, A. Nakaya, and M. Kanehisa, "Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis," *Bioinformatics*, vol. 19, pp. 323–330, 2003.

[4] M. K. Tippett, T. DelSole, S. J. Mason, and A. G. Barnston, "Regression-based methods for finding coupled patterns," *J. Climate*, vol. 21, pp. 4384–4398, 2008.

[6] N. Asendorf and R. R. Nadakuditi, "Improving multiset canonical correlation analysis in high dimensional sample deficient settings," in *Proc. IEEE Asilomar Conf. on Signals, Systems & Computers*, 2015, pp. 112–116.

[7] J. Sui, T. Adali, G. Pearlson, H. Yang, S. R. Sponheim, T. White, and V. D. Calhoun, "A CCA + ICA based model for multi-task brain imaging data fusion and its application to schizophrenia," *NeuroImage*, vol. 51, pp. 123–134, 2010.

[8] Y. Levin-Schwartz, Y. Song, P. J. Schreier, V. D. Calhoun, and T. Adalı, "Sample-poor estimation of order and common signal subspace with application to fusion of medical imaging data," *NeuroImage*, vol. 134, pp. 486–493, 2016.

[9] Y. Song, P. J. Schreier, D. Ramírez, and T. Hasija, "Canonical correlation analysis of high-dimensional data with very small sample support," *Signal Processing*, vol. 128, pp. 449–458, 2016.

[10] R. R. Nadakuditi and A. Edelman, "Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples," *IEEE Trans. Signal Process.*, vol. 56, pp. 2625–2638, 2008.

[11] Y. Song, T. Hasija, P. J. Schreier, and D. Ramírez, "Determining the number of signals correlated across multiple data sets for small sample support," in *Proc. European Signal Proc. Conf. (EUSIPCO)*, 2016, pp. 1528–1532.

[12] T. Hasija, Y. Song, P. J. Schreier, and D. Ramírez, "Bootstrap-based detection of the number of signals correlated across multiple data sets," in *Proc. IEEE Asilomar Conf. on Signals, Systems & Computers*, 2016, pp. 610–614.

[13] T. Hasija, Y. Song, P. J. Schreier, and D. Ramírez, "Detecting the dimension of the subspace correlated across multiple data sets in the sample poor regime," in *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, 2016, pp. 1–5.

[14] Y.-O. Li, T. Adali, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Trans. Signal Process.*, vol. 57, pp. 3918–3929, 2009.

[15] A. A. Nielsen, "Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data," *IEEE Trans. Image Process.*, vol. 11, pp. 293–305, 2002.

[16] A. Pezeshki, L. L. Scharf, M. R. Azimi-Sadjadi, and M. Lundberg, "Empirical canonical correlation analysis in subspaces," in *Proc. IEEE Asilomar Conf. on Signals, Systems & Computers*, 2004, pp. 994–997.

[17] M. Bartlett, "The statistical significance of canonical correlations," *Biometrika*, pp. 29–37, 1941.

[18] D. Lawley, "Tests of significance in canonical analysis," *Biometrika*, pp. 59–66, 1959.