A GLRT for estimating the number of correlated components in sample-poor mCCA

Tanuj Hasija[1] and Timothy Marrinan[2]
[1]Signal and System Theory Group, Universität Paderborn, Germany
[2]School of Electrical Engineering and Computer Science, Oregon State University, USA
Email: tanuj.hasija@sst.upb.de, marrinat@oregonstate.edu

***Abstract*— In many applications, components correlated across multiple data sets represent meaningful patterns and commonalities. Estimates of these patterns can be improved when the number of correlated components is known, but since data exploration often occurs in an unsupervised setting, the number of correlated components is generally not known. In this paper, we derive a generalized likelihood ratio test (GLRT) for estimating the number of components correlated across multiple data sets. In particular, we are concerned with the scenario where the number of available samples is small. As a result of the small sample support, correlation coefficients and other summary statistics are significantly overestimated by traditional methods. The proposed test combines linear dimensionality reduction with a GLRT based on a measure of multiset correlation referred as the generalized variance cost function (mCCA-GENVAR). By jointly estimating the rank of the dimensionality reduction and the number of correlated components, we are able to provide high-accuracy estimates in the challenging sample-poor setting. These advantages are illustrated in numerical experiments that compare and contrast the proposed method with existing techniques.**

***Index Terms*— GLRT, joint dimensionality reduction, mCCA-GENVAR, model-order selection, multiple data sets, sample poor**

I. Introduction

Multiset canonical correlation analysis (mCCA) is one of the most common tools for analyzing second-order multivariate association across multiple data sets [1]. The objective of mCCA is to extract components that are maximally correlated across two or more data sets. These extracted components have been shown to provide meaningful representations of physical processes such as brain activity patterns [2], autonomic nervous system changes [3], and gene clusters [4], and are also useful as features for image recognition in remote sensing [5] and voice activity detection [6], to name a few.

An important parameter in the mCCA pipeline is the number of components correlated across multiple sets. This number affects both the performance and the interpretation of systems that rely on mCCA. Estimating the number of correlated components can be posed as a model-order selection (MOS) problem. For two data sets, the MOS problem is well-defined. This is not the case for more than two data sets where different generalizations of model order are possible [7]. One of the most common MOS problems is to estimate the number of components correlated across all pairs of data sets [8]–[10]. However, this number ignores components that are correlated across a subset of the collection of data sets, which is also of interest in many fields [3], [11]. A more appropriate model order in multiple data sets is the number of components demonstrating correlation across all or a subset of data sets. This model order summarizes the complete set of correlations and two different methods to estimate it are proposed in [12] and [7]. The method proposed in [12] estimates the model order in two steps by first computing the number of nonzero correlations between each pair of data sets, and then jointly computing the multiset correlation to estimate the structure of these correlated components. The model order is then computed as the number of rows with at least one nonzero correlation in the estimated correlation structure. However, because the method in [12] initially relies on pairwise comparisons, it does not fully leverage the joint correlation information across all data sets. As such, the estimated model order is sensitive to inaccuracies in these pairwise comparisons when the number of data sets is large. In contrast, the method in [7] estimates the model order jointly from all available data sets, but it relies on stronger assumptions about the correlation coefficients to provably estimate the correct order. Moreover, neither [12] or [7] explicitly addresses the challenges associated with small sample support. When not enough samples are available, the correlation coefficients between the extracted components are highly overestimated [13]. This poses a major challenge in mCCA and leads to inaccurate model-order estimates from the above mentioned techniques.

In this work, we derive a generalized likelihood ratio (GLR) as a function of the model order. We show the connection between the derived GLR and a measure of multiset correlation known as the generalized variance (GENVAR) criterion in [1]. We then apply the GLR in a sequence of binary hypothesis tests to estimate the model order, and propose a novel way of estimating the unknown distribution of the test statistic under the null hypothesis. To tackle the small sample support issue, we apply linear dimensionality reduction to all data sets such that the model order and the rank of dimensionality reduction are jointly estimated. This leads to a novel joint reduced-rank mCCA (jRR-mCCA) technique where the reduced rank retains components for estimating the correct model order while excluding the undesirable uncorrelated components.

II. Problem Formulation

We observe M independent and identically distributed (i.i.d.) samples of P data sets, each modelled using real-valued random vectors $\mathbf{x}_p$ and generated by an unknown linear mixing of a latent component vector, $\mathbf{s}_p \in \mathbb{R}^N$ as

$$\mathbf{x}_p = \mathbf{A}_p \mathbf{s}_p, \quad p = 1, 2, \ldots, P, \tag{1}$$

where $\mathbf{A}_p \in \mathbb{R}^{N \times N}$ is an unknown but fixed mixing matrix with full rank. For simplicity, we assume each data set is of dimension N, however the results can be generalized in a natural way to data sets of differing dimensions. Each component vector, $\mathbf{s}_p \in \mathbb{R}^N$ contains N Gaussian random variables denoted by $s_p^{(n)}$, $n = 1, \ldots, N$. It is common to assume two kinds of association among the latent components:

This research was supported by the German Research Foundation (DFG) under grant SCHR 1384/3-2.

A1. Intraset independence: components within each data set are without loss of generality (w.l.o.g.) zero-mean, unit variance and uncorrelated. The component covariance matrix is

$$\mathbf{R}_{s_p s_p} = E[\mathbf{s}_p \mathbf{s}_p^\top] = \mathbf{I},$$

where $E[\cdot]$ is the expectation operator, superscript $\top$ is the transpose and $\mathbf{I}$ is the identity matrix.

A2. Interset dependence: between any two data sets p and q, components may be correlated only pairwise, i.e., component $s_p^{(n)}$ may only correlate with component $s_q^{(n)}$ for $1 \leq n \leq N$. This leads to a diagonal cross-covariance matrix between the components of data sets p and q ($p \neq q$), i.e.,

$$\mathbf{R}_{s_p s_q} = \text{diag}(\rho_{pq}^{(1)}, \rho_{pq}^{(2)}, \ldots, \rho_{pq}^{(N)}),$$

where $\rho_{pq}^{(n)}$ represents the unknown (possibly zero) correlation coefficient between their nth components.

Interset dependence (A2) is a common simplifying assumption in the literature [14], [15]. It does not represent all possible correlation structures, but it provides rather mild conditions under which the multiset correlation structure is identifiable.

Let the model order be denoted by

$$D = \left| \{n : \exists p, q \text{ for which } \rho_{pq}^{(n)} \neq 0\} \right|.$$

In other words, D is the size of the index set of components that demonstrate nonzero correlation between at least one pair of data sets. The goal of this work is summarized as follows:

Goal: *Given M i.i.d. samples jointly observed from $\mathbf{x}_1, \ldots, \mathbf{x}_P$ defined by the model in* (1)*, estimate the model order D.*

III. GLRT USING MCCA-GENVAR

When working with sample limitations, MOS techniques are well-suited for estimating the model order because they balance the trade-off between overfitting and generalizability [16]. In contrast, methods that employ heuristic user-defined thresholds on the sample correlation coefficients or treat the dimension as a hyper-parameter require careful tuning and are not appropriate in general settings. In this section, we describe the proposed method for estimating D based on a generalized likelihood ratio test (GLRT), a common MOS technique.

A. Maximum Likelihood Function

Assumptions A1 and A2 imply that the covariance matrices in model (1), $\mathbf{R}_{pp} = E[\mathbf{x}_p \mathbf{x}_p^\top]$, are nonsingular. Denote the inverse of the matrix square root by $(\cdot)^{-\frac{1}{2}}$. Then $\mathbf{y}_p = \mathbf{R}_{pp}^{-\frac{1}{2}} \mathbf{x}_p$ is a whitened version of the original data. The concatenation of these whitened data vectors, $\mathbf{y} = [\mathbf{y}_1^\top, \ldots, \mathbf{y}_P^\top]^\top$, is Gaussian distributed with zero-mean and covariance matrix $\mathbf{C} = E[\mathbf{y}\mathbf{y}^\top]$. Define $\mathbf{R}^{(n)} \in \mathbb{R}^{P \times P}$ to be the covariance matrix of the nth components of each data set. That is, if $\mathbf{s}^{(n)} = [s_1^{(n)}, \ldots, s_P^{(n)}]^\top$ then $\mathbf{R}^{(n)} = E[\mathbf{s}^{(n)} \mathbf{s}^{(n)\top}]$. Let $\widetilde{\mathbf{R}}_{ss}$ denote the block diagonal matrix,

$$\widetilde{\mathbf{R}}_{ss} = \text{blkdiag}(\mathbf{R}^{(1)}, \ldots, \mathbf{R}^{(D)}, \mathbf{I}), \tag{2}$$

where $\mathbf{I} \in \mathbb{R}^{(N-D)P \times (N-D)P}$ is an identity matrix. Then there exists a permutation matrix, $\mathbf{P}$, such that $\mathbf{C}$ can be written as

$$\mathbf{C} = \mathbf{A}\mathbf{P}^\top \widetilde{\mathbf{R}}_{ss} \mathbf{P} \mathbf{A}^\top \tag{3}$$

for $\mathbf{A} = \text{blkdiag}\left((\mathbf{A}_1 \mathbf{A}_1^\top)^{-\frac{1}{2}} \mathbf{A}_1, \ldots, (\mathbf{A}_P \mathbf{A}_P^\top)^{-\frac{1}{2}} \mathbf{A}_P \right)$ [7].

Let $\mathbf{y}(m)$ denote the mth sample of $\mathbf{y}$. The log-likelihood for M i.i.d samples of $\mathbf{y}$ parametrized by $\mathbf{C}$ and D is

$$\ln f\left(\mathbf{y}(1), \ldots, \mathbf{y}(M) | \mathbf{C}, D\right) = K - \frac{M}{2} \ln \det\left(\mathbf{C}\right) - \frac{1}{2} \sum_{m=1}^{M} \left(\mathbf{y}^\top(m) \mathbf{C}^{-1} \mathbf{y}(m) \right),$$

where det$(\cdot)$ is the determinant of a matrix and the constant K is independent of the parameter space $\mathbf{C}$. If $\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(M)]$ is the matrix containing all M samples, then the log-likelihood function can be written in light of (3) as

$$\ln f\left(\mathbf{Y} | \mathbf{C}, D\right) \propto -\tfrac{M}{2} \ln \left(\frac{\det\left(\mathbf{A}\right) \det\left(\mathbf{P}^\top\right) \det\left(\widetilde{\mathbf{R}}_{ss}\right)}{\det\left(\mathbf{A}\right) \det\left(\mathbf{P}^\top\right)} \right) - \tfrac{1}{2} \sum_{m=1}^{M} \text{tr}\left(\mathbf{y}(m) \mathbf{y}^\top(m) \mathbf{C}^{-1} \right),$$

where tr$(\cdot)$ is the trace of a matrix. By further decomposing the expression with respect to (2), we can write the log-likelihood as a function of the sum of the log-determinants of covariance matrices for each of the components. That is,

$$\ln f\left(\mathbf{Y} | \mathbf{C}, D\right) \propto -\tfrac{M}{2} \Big(\ln \det\left(\mathbf{R}^{(1)}\right) + \ldots + \ln \det\left(\mathbf{R}^{(D)}\right) \Big) - \tfrac{1}{2} \sum_{m=1}^{M} \text{tr}\left(\mathbf{y}(m) \mathbf{y}^\top(m) \mathbf{C}^{-1} \right). \tag{4}$$

The log-likelihood function is then maximized when the determinant of each $\mathbf{R}^{(n)}$ is minimized and when $\mathbf{C}$ is the sample covariance matrix, $\hat{\mathbf{C}} = \frac{1}{M} \mathbf{Y}\mathbf{Y}^\top$. We use the diacritical mark $\hat{}$ to indicate when something is a sample estimate.

The GENVAR mCCA problem iteratively extracts components from each data set such that the determinant of the covariance matrix of these components is minimized [1]. The same components therefore maximize the log-likelihood function in (4). Thus, the maximum log-likelihood function is

$$\ln f\left(\mathbf{Y} | \hat{\mathbf{C}}, D\right) \propto -\frac{M}{2} \ln \left(\det\left(\hat{\mathbf{R}}^{(1)}\right) \ldots \det\left(\hat{\mathbf{R}}^{(D)}\right) \right), \tag{5}$$

where $\hat{\mathbf{R}}^{(n)}$, for $n = 1, \ldots, D$ is the estimated covariance matrix of the nth set of components obtained using mCCA-GENVAR.

B. Binary Hypothesis Testing Framework

The maximum-likelihood function derived in (5) is used in a hypothesis testing framework to estimate D as follows. A sequence of binary hypothesis tests are performed one at a time with a counter i starting at $i = 0$. The binary test of null hypothesis H_0 and alternative hypothesis H_1 is defined as

$$\begin{aligned} H_0 &: D = i, \\ H_1 &: D > i. \end{aligned} \tag{6}$$

If H_0 is rejected, i is incremented and the next test of H_0 vs. H_1 is run. The process is repeated until H_0 is not rejected or the maximum value of i is reached [17]. The GLR for the test in (6) is

$$\eta(i) = \frac{f(\mathbf{Y} | \hat{\mathbf{C}}, D = i)}{f(\mathbf{X} | \hat{\mathbf{C}}, D > i)},$$

where $f(\mathbf{Y} | \hat{\mathbf{C}}, D = i)$ and $f(\mathbf{X} | \hat{\mathbf{C}}, D > i)$ are the maximum likelihood functions under H_0 and H_1, respectively. Under H_1, the parameter space $i = N$ parametrizes all the possibilities

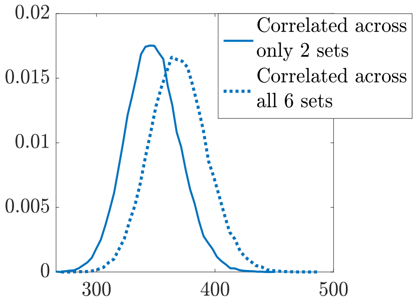

Fig. 1: Empirical estimate of $g(i)$ for $P = 6, N = 10, D = i = 3$, and $M = 5000$, computed for two different correlation structures a) components correlated across only 2 data sets (solid curve) and b) components correlated across all 6 data sets (dashed curve).

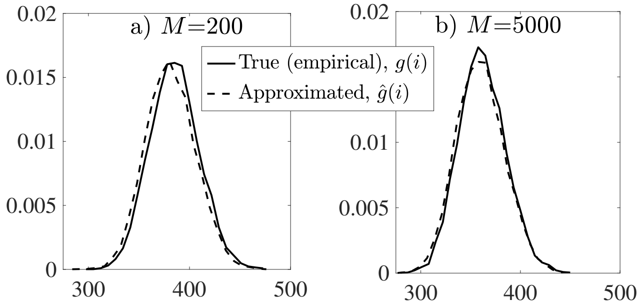

Fig. 2: True (empirical) distribution $g(i)$ and approximated distribution $\hat{g}(i)$ for $P = 6, N = 10, D = i = 3$ with a) $M = 200$ and b) $M = 5000$.

for $D > i$ and achieves the maximum likelihood. Using (5), $\eta(i)$ can be simplified as

$$\eta(i) = \left(\det(\hat{\mathbf{R}}^{(i+1)}) \dots \det(\hat{\mathbf{R}}^{(N)})\right)^{-\frac{M}{2}}. \qquad (7)$$

Consider the statistic $T(i) = -2 \ln \eta(i)$ based on the GLR. Given a threshold, $\tau(i)$, D is estimated as the smallest value i for which H_0 is not rejected, i.e.,

$$\hat{D} = \min_{i=0,\dots,N-1} \{i : T(i) < \tau(i)\}. \qquad (8)$$

If the asymptotic distribution of $T(i)$ under the true H_0 (i.e., when $D = i$) is independent of the unknown parameters, then a threshold, $\tau(i)$, can be computed to ensure a fixed probability of false alarm, P_{fa}, according to Wilks' theorem [18].

C. Distribution of $T(i)$ under H_0

Let $g(i)$ denote the asymptotic distribution of $T(i)$ under the true H_0. For $P = 2$, $g(i)$ is independent of unknown parameters such as the correlation coefficients and the variance of components. It depends only on the known parameters N and i [17]. For $P > 2$, however, $g(i)$ also depends on the correlation *structure* of the correlated components. Fig. 1 illustrates the dependence of $g(i)$ on the correlation structure for an example where $P = 6$, $N = 10$, and $D = i = 3$. The two curves in Fig. 1 are empirical estimates of $g(i)$ computed from 10^4 independent Monte Carlo trials. For each trial, $T(i)$ was estimated from $M = 5000$ samples, (which is much larger than N) to approximate the asymptotic distribution of $T(i)$. The difference in the two plots is that the solid curve is obtained from data where components are correlated across only 2 of the 6 sets, whereas the dashed curve is obtained from components that are correlated across all 6 data sets.

Unless the correlation structure is known *a priori*, $g(i)$ cannot be accurately characterized as a function of only n, P, and i. However, except in special cases [8], assuming a correlation structure *a priori* is prohibitively restrictive in applications where the data sets are heterogeneous and contain components correlated with arbitrary correlation structure [3], [4], [19]. For the example shown in Fig. 1, when the components are correlated with arbitrary correlation structure, $g(i)$ will lie in between the two extreme versions shown by the solid and dashed curves. In this case, if the solid curve is used as an approximation for $g(i)$, the estimator in (8) tends to overestimate the value of D. On the other hand, if the dashed curve is used, (8) would tend to underestimate.

Obtaining an approximate distribution for $g(i)$: To provide a balance between over- and under- estimation, we randomize the number of data sets across which each set of components is correlated when estimating the distribution of $T(i)$. The parameters N and P are fixed for any given problem. Thus, for i specified by the current null hypothesis, an approximate distribution, $\hat{g}(i)$, is estimated for $T(i)$ via Monte Carlo trials. For a single trial, we generate P data sets, each with M samples, N dimensions, and i correlated components. The ιth set of components ($\iota = 1, \dots, i$) is correlated across $q^{(\iota)}$ data sets where $q^{(\iota)}$ is selected uniformly at random from $\{2, 3, \dots, P\}$. The variance of all components is 1, as is each nonzero correlation coefficient. The mixing matrices, $\mathbf{A}_p$, are orthogonal, and randomly sampled according to [20]. Thus, each trial results in a sample of $T(i)$, and $\hat{g}(i)$ is the empirical distribution obtained from these samples over multiple trials.

To illustrate the effectiveness of this approach, consider a second example. The parameters remain the same as those used for Fig. 1, except now the components are correlated across 6, 4, and 2 data sets, respectively, with associated correlation coefficients of 0.9, 0.8, and 0.7. The variance of the correlated components is 2 and that of uncorrelated components is 1. The empirical version of $g(i)$ along with its approximated distribution $\hat{g}(i)$ are shown in Fig. 2 for two sample sizes of a) $M = 200$ and b) $M = 5000$. We see that $\hat{g}(i)$ closely approximates $g(i)$ in both cases irrespective of the unknown parameters.

IV. Joint reduced-rank mCCA estimator

Both the proposed estimator in (8) and the competing MOS techniques assume M to be large relative to N. When M is of the same order as, or smaller than, N, the sample correlation coefficients are highly overestimated [13]. This overestimation is especially pronounced for correlation coefficients that are supposed to be zero, which leads to an inaccurate estimate of $\eta(i)$ in (7), and consequently an unreliable estimate of D. This hurdle is particularly relevant because sample-poor scenarios are common in many fields, often due to resource availability or the cost of gathering samples. For example, in biomedicine samples may refer to human subjects, which are limited due to the number of participants in a given study [2]. In oceanography, the samples might correspond to measurements of a physical property like sea surface temperature which can only be observed a few times per year or in a few locations due to the cost of measurement collection [21].

When working with small M, a common practice is to assume a low-rank generative model such that each data set in (1) is generated from the product of a tall, skinny mixing matrix $\mathbf{A}_p$ and low-rank component vector $\mathbf{s}_p$ [22]. This means that even though the dimension N is comparable to, or even larger than, M, the number of components and D are both still small relative to M. One reasonable procedure to tackle small-sample support is to apply dimensionality reduction so that the reduced dimension, r, is small compared to M. In [13], a principal component analysis (PCA) pre-processing is proposed as a dimensionality reduction step before performing mCCA. However, PCA retains the components with most variance within a data set and these components are not necessarily the ones that are correlated across multiple data sets. If the components are retained only on the basis of their variance, then the PCA step before mCCA will likely retain undesirable uncorrelated components.

Inspired by the strategy for two data sets in [23], we propose a joint reduced-rank mCCA method (jRR-mCCA) where the reduced dimension, r, and the model order, D, are jointly estimated. In what follows, we use PCA as the dimensionality reduction technique to show a fair comparison with the competing PCA-based methods from [12], [13], however, the PCA step in our test can be replaced by another linear dimensionality reduction step as long as the reduced-rank data can be described by (1) and its assumptions.

Let the rank-r PCA descriptions be $\widetilde{\mathbf{X}}_p = \mathbf{U}_p^\top(r)\mathbf{X}_p$, where $\mathbf{X}_p = [\mathbf{X}_p(1), \dots, \mathbf{X}_p(M)]$ is the sample data matrix and the columns of $\mathbf{U}_p(r)$ are the first r dominant eigenvectors of $\hat{\mathbf{R}}_{pp}$. The rank-reduced statistic, $T(i,r)$, is computed using the covariance matrices of the components extracted from the rank-reduced descriptions $\widetilde{\mathbf{X}}_1, \dots, \widetilde{\mathbf{X}}_P$ via mCCA-GENVAR, and the threshold $\tau(i,r)$ is computed from the approximate distribution $\hat{g}(i,r)$ using the known parameters r, P, and i. The jRR-mCCA estimator for D is

$$\hat{D} = \max_{r=1,\dots,r_{\text{max}}} \min_{i=0,\dots,r-1} \{i : T(i,r) < \tau(i,r)\}, \qquad (9)$$

The decision rule in (9) is motivated by the fact that the min-step will generally not overestimate d, while the max-step ensures that the rank of the dimensionality reduction, r, is chosen large enough to capture all of the correlated components [23]. Here, r_{max} is a user-defined upper limit, typically chosen to be smaller than M/P to avoid ill-conditioning of $\hat{\mathbf{C}}$. Thus, the dimensionality reduction retains r components from each data set that are the most informative for estimating D, and excludes weaker uncorrelated components.

V. Numerical Results

In this section, we compare the performance of the proposed technique with two existing multiset MOS methods for estimating D, the pairwise mCCA-HT of [12] and the joint eigenvalue decomposition (jointEVD) technique of [7]. The jointEVD technique is not well-suited to small-sample scenarios. We therefore create a hybrid approach that first reduces the rank of the data following the proposal in [13] to do PCA-preprocessing and second applies the jointEVD technique to the low-rank data. The rank-reduction in [13] takes the desired rank as a parameter, which we obtain from [24]. We refer to this combination of [7] + [13] + [24] as the "hybrid jointEVD" approach.

The simulation setup is as follows[1]. We observe $M = 200$ samples from each of $P = 10$ data sets with $N = 50$ dimensions per data set. The collection of data sets has $D = 5$ components that are correlated across 10, 9, 8, 7, and 6 data sets, respectively, with associated correlation coefficients of $0.85, 0.8, 0.75, 0.7$ and 0.6. The variance is $\sigma_{\text{c}}^2 = 1$ for each correlated component. In addition, there are two stronger uncorrelated components with variance of 2 and 5 weaker uncorrelated components with variance of 0.5 in each data set. All data sets are corrupted with an additive white noise whose variance, σ_{n}^2, is chosen according to the signal-to-noise ratio (SNR) defined as $10\log_{10}\{\frac{\sigma_c^2}{\sigma_n^2}\}$.

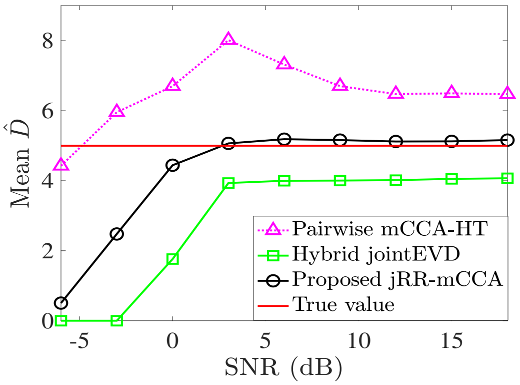

Fig. 3: Mean $\hat{D}$ for proposed and competing techniques for $P = 10, N = 50$ and $M = 200$.

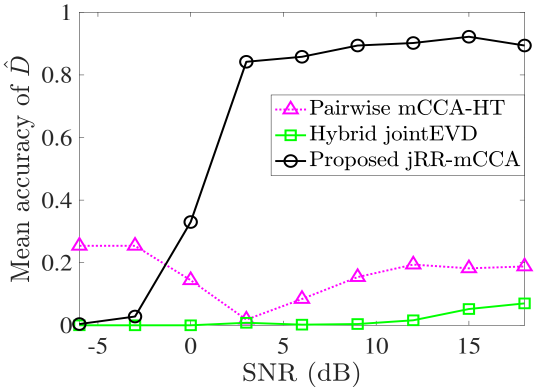

Fig. 4: Mean accuracy of $\hat{D}$ for proposed and competing techniques for the same parameter setting as Fig. 3.

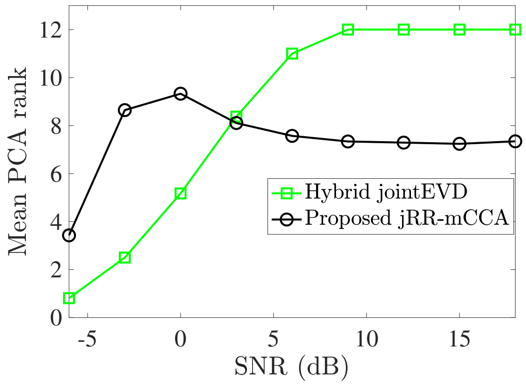

Fig. 5: Mean PCA rank estimated by the proposed and competing technique for the same parameter setting as Fig. 3.

Fig. 3 and Fig. 4 plot the mean value and the mean accuracy (the number of correct estimates divided by number of trials) of $\hat{D}$, respectively, estimated by each method from 500 Monte Carlo trials as a function of the SNR. The pairwise mCCA-HT

[1] Code for the techniques and experiments is available at https://github.com/SSTGroup/Correlation-Analysis-in-High-Dimensional-Data

technique combines the model-order estimates from each pair of data sets (45 pairs for $P = 10$) to compute $\hat{D}$. Thus, an error in any of the pairwise model orders leads to an inaccurate estimate of D. This can be seen in Fig. 3, where the pairwise mCCA-HT technique overestimates D even for large SNR values. On the other hand, the hybrid jointEVD technique tends to underestimate D and approaches the true value only for large SNR. The proposed jRR-mCCA technique estimates D with high accuracy even for small SNR values as seen in Fig. 4 and outperforms both of the competing techniques.

In Fig. 5, we illustrate the benefit of the proposed joint approach for estimating D and r over the hybrid jointEVD technique. Ideally the reduced-rank data would contain the correlated components and as little else as possible. We can see from Fig. 5 that the mean estimated rank from the proposed method (black curve with circle markers) is close to 7. This makes sense because there are 5 correlated components with variance 1 and two uncorrelated components with variance 2. This means that the proposed joint method retains all high variance components necessary to estimate the true number of correlations, and no unnecessary uncorrelated dimensions. The hybrid approach (green curve with square markers), on the other hand, estimates the rank r independently of the number of correlated components D, and as such overestimates the rank to be 12. This represents the total number of components in each data set, but contains weak uncorrelated components and means that the number of available samples is smaller relative to the number of retained dimensions, further reducing estimation accuracy. In this context, there is a significant advantage to jointly estimating the rank of dimensionality reduction with the model order.

VI. Conclusion and Future Work

We derived a GLRT-based hypothesis testing framework for determining the number of components correlated across more than two data sets in the sample-poor regime. To the best of our knowledge, this is the first method for joint dimensionality reduction and model-order selection under the broad umbrella of multiset correlation analysis. The GLRT is achieved by the maximally correlated components that minimize the GENVAR cost function of mCCA. To estimate the distribution of the test statistic, which depends on the unknown correlation structure, we propose a novel approximation method using Monte-Carlo trials and random correlation structures. The GLRT is combined with a dimensionality reduction step to improve estimation accuracy when the number of available samples is relatively small. The final end-to-end pipeline demonstrates superior performance to the existing techniques in challenging sample-poor settings. The numerical experiments in this manuscript are limited to synthetic data to clearly demonstrate the scenarios in which the proposed method shines. In a forthcoming sequel, the proposed method is used in a practical application of brain imaging data fusion, in which accurately estimating the number of correlated components is crucial for the interpretability of brain activity patterns.

References

[1] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[2] T. Adali, Y. Levin-Schwartz, and V. D. Calhoun, "Multimodal data fusion using source separation: Two effective models based on ICA and IVA and their properties," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1478–1493, 2015.

[3] S. Vieluf, T. Hasija, R. Jakobsmeyer, P. J. Schreier, and C. Reinsberger, "Exercise-induced changes of multimodal interactions within the autonomic nervous network," *Frontiers in Physiology*, vol. 10, p. 240, 2019.

[4] 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, no. suppl 1, pp. 323–330, 2003.

[5] X. Yang, W. Liu, D. Tao, J. Cheng, and S. Li, "Multiview canonical correlation analysis networks for remote sensing image recognition," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1855–1859, 2017.

[6] T. Hasija, M. Gölz, M. Muma, P. J. Schreier, and A. M. Zoubir, "Source enumeration and robust voice activity detection in wireless acoustic sensor networks," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 1257–1261.

[7] T. Hasija, T. Marrinan, C. Lameiro, and P. J. Schreier, "Determining the dimension and structure of the subspace correlated across multiple data sets," *Signal Processing*, p. 107613, 2020.

[8] Y. Wu, K. W. Tam, and F. Li, "Determination of number of sources with multiple arrays in correlated noise fields," *IEEE Transactions on Signal Processing*, vol. 50, no. 6, pp. 1257–1260, 2002.

[9] T. Hasija, Y. Song, P. J. Schreier, and D. Ramírez, "Bootstrap-based detection of the number of signals correlated across multiple data sets," in *2016 50th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2016, pp. 610–614.

[10] S. Bhinge, Y. Levin-Schwartz, and T. Adalı, "Estimation of common subspace order across multiple datasets: Application to multi-subject fMRI data," in *Proceedings of the 51st Annual Conference on Information Sciences and Systems*, 2017.

[11] I. Lehmann, E. Acar, T. Hasija, M. Akhonda, V. D. Calhoun, P. J. Schreier, and T. Adali, "Multi-task fMRI data fusion using IVA and PARAFAC2," *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, 2022.

[12] T. Marrinan, T. Hasija, C. Lameiro, and P. J. Schreier, "Complete model selection in multiset canonical correlation analysis," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1082–1086.

[13] N. Asendorf and R. R. Nadakuditi, "Improving multiset canonical correlation analysis in high dimensional sample deficient settings," in *2015 49th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2015, pp. 112–116.

[14] Y.-O. Li, T. Adalı, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3918–3929, 2009.

[15] X.-F. Gong, Q.-H. Lin, F.-Y. Cong, and L. De Lathauwer, "Double coupled canonical polyadic decomposition for joint blind source separation," *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3475–3490, 2018.

[16] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, 2004.

[17] W. Chen, J. P. Reilly, and K. M. Wong, "Detection of the number of signals in noise with banded covariance matrices," *IEE Proceedings-Radar, Sonar and Navigation*, vol. 143, no. 5, pp. 289–294, 1996.

[18] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[19] 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.

[20] F. Mezzadri, "How to generate random matrices from the classical compact groups," *arXiv preprint math-ph/0609050*, 2006.

[21] M. K. Tippett, T. DelSole, S. J. Mason, and A. G. Barnston, "Regression-based methods for finding coupled patterns," *Journal of Climate*, vol. 21, no. 17, pp. 4384–4398, 2008.

[22] N. Asendorf and R. R. Nadakuditi, "Improved detection of correlated signals in low-rank-plus-noise type data sets using informative canonical correlation analysis (ICCA)," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3451–3467, 2017.

[23] 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.

[24] R. R. Nadakuditi and A. Edelman, "Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2625–2638, 2008.