Bootstrap-based Detection of the Number of Signals Correlated Across Multiple Data Sets

Tanuj Hasija[1], Yang Song[2], Peter J. Schreier[1] and David Ramírez[3, 4]
[1]Signal and System Theory Group, Universität Paderborn, Germany
[2]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[3]Signal Processing Group, Universidad Carlos III de Madrid, Leganés, Spain
[4]Gregorio Marañón Health Research Institute, Madrid, Spain
Email: {tanuj.hasija, peter.schreier}@sst.upb.de, songy@ntu.edu.sg, david.ramirez@uc3m.es

***Abstract*—We present a scheme for determining the number of signals common to or correlated across multiple data sets. Handling multiple data sets is challenging due to the different possible correlation structures. For two data sets, the signals are either correlated or uncorrelated between the data sets. For multiple data sets, however, there are numerous combinations how the signals can be correlated. Prior studies dealing with multiple data sets all assume a particular correlation structure. In this paper, we present a technique based on a series of hypothesis tests and the bootstrap, which works for arbitrary correlation structure. Numerical results show that the proposed technique correctly detects the number of correlated signals in scenarios where the competition tends to overestimate.**

***Index Terms*—** Bootstrap, correlated signals, data fusion, hypothesis testing, model-order selection, multiple data sets.

I. Introduction

Knowing the association and relationship between multiple data sets is useful for numerous applications in various fields. In biomedicine, for instance, each brain imaging modality (like EEG, fMRI and sMRI) provides a unique way of understanding the brain. EEG measures the brain activity at high temporal resolution, while the functional and structural MRI techniques acquire the brain data at high spatial resolution. Effective fusion of complementary data from these modalities helps gain insights into the brain and its associated diseases at a highly resolved spatial and temporal scale [1], [2]. Estimating the number of signals (or sources) correlated across the modalities prior to the source separation step ensures that the fusion process does not consider any unrelated signal components. As another example, in image processing, identifying objects observed by different spatially separated cameras is useful for many tracking applications [3]. Other applications include genomics, climate science, and array processing [4]–[6].

In signal processing, "model order" is the term used for the dimension of a parameter vector, i.e., the number of parameters, of the data model [7]. When dealing with two or more data sets, one particularly important model-order selection problem is to detect the dimension of the subspace common across multiple data sets. In this work, we consider the second-order moment, i.e., correlation, as a measure of commonality. In the literature, model-order selection for multiple data sets has not yet received the attention that it deserves. While the problem with two data sets has been dealt with in numerous works [8]–[10], only a few studies have addressed this for multiple data sets [6], [11]–[13]. Moreover, all of the work for multiple data sets assumes a particular correlation structure among the signals. In [6], [12], the authors assume that the signals correlated across any two data sets are also correlated across all remaining the data sets. Our work in [13] tried to slightly relax the above assumption by allowing the signals correlated between any two data sets to be uncorrelated among all remaining data sets. When these assumptions are not satisfied, all these techniques typically overestimate the number of signals correlated across all data sets.

The main challenge when dealing with multiple data sets is the number of possible correlation structures among the latent signals. For two data sets, this problem does not arise as the individual signals are either correlated or uncorrelated between the data sets. For multiple data sets, however, there are numerous combinations how the signals can be correlated. Some signals might be independent among the data sets, while some are shared only among a subset of all data sets. There can also be signals correlated across all the data sets. In this paper, we are interested in those signals that are common or correlated across all data sets and not the ones that are correlated only among a subset of data sets. We address this problem without assuming any particular correlation structure between the signals. To the best of our knowledge, no alternative approach that works for arbitrary correlation structure has yet been presented in the literature.

Our program for this paper is as follows. After defining the problem in Section II, we show in Section III that the rank of the product of coherence matrices (which are normalized cross-covariance matrices) of all possible pairs of data sets is equal to the number of signals correlated between all the data sets, provided the SNR is sufficiently large. The problem thus comes down to estimating the rank of this product of matrices. For this, we employ a standard procedure based on a series of binary hypothesis tests [14], [15], described in Section IV. Since the distribution of the utilized test statistic under the null hypothesis is difficult to derive analytically, we estimate it using the bootstrap technique. The bootstrap is a resampling technique that has been applied to solve numerous signal processing applications like bias estimation, estimation of

978-1-5386-3954-2/16/$31.00 ©2016 IEEE

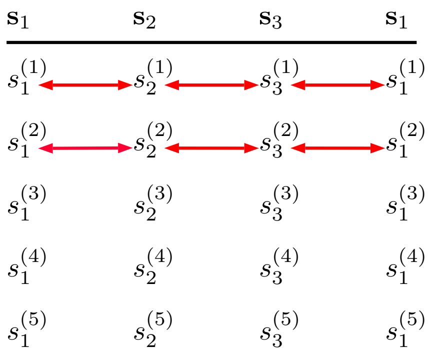

$\mathbf{s}_1$ $\mathbf{s}_2$ $\mathbf{s}_3$ $\mathbf{s}_1$
$s_1^{(1)}$ $s_2^{(1)}$ $s_3^{(1)}$ $s_1^{(1)}$
$s_1^{(2)}$ $s_2^{(2)}$ $s_3^{(2)}$ $s_1^{(2)}$
$s_1^{(3)}$ $s_2^{(3)}$ $s_3^{(3)}$ $s_1^{(3)}$
$s_1^{(4)}$ $s_2^{(4)}$ $s_3^{(4)}$ $s_1^{(4)}$
$s_1^{(5)}$ $s_2^{(5)}$ $s_3^{(5)}$ $s_1^{(5)}$

Fig. 1: Example for three data sets for the correlation structure assumed in [6], [12]. Arrows indicate correlated components of source signals. Here, $d_{12} = d_{23} = d_{31} = d = 2$.

confidence intervals, and hypothesis testing [16]–[18]. Finally, in Section V we show in simulations that our detector can reliably detect the number of signals with arbitrary correlation structure whereas other competing methods fail.

II. Problem Formulation

We consider L zero-mean complex-valued data sets (vectors), $\mathbf{x}_1, \ldots, \mathbf{x}_L$, with dimensions $m_1, \ldots, m_L$ respectively. Without loss of generality, it is assumed that $m_1 \leq m_2 \leq \ldots \leq m_L$. The generating signal-plus-noise data model is

$$\mathbf{x}_i = \mathbf{A}_i \mathbf{s}_i + \mathbf{n}_i, \quad i = 1, 2, \ldots, L, \tag{1}$$

where $\mathbf{A}_i \in \mathbb{C}^{m_i \times Q_i}$ is an unknown but fixed mixing matrix with full column rank, $\mathbf{s}_i \in \mathbb{C}^{Q_i}$ is a zero-mean complex source signal vector containing an unknown number Q_i $(< m_i)$ of independent source signals, and $\mathbf{n}_i \in \mathbb{C}^{m_i}$ is a zero-mean complex noise vector independent from all source vectors. Without loss of generality, the source covariance matrix is $\mathbf{R}_{s_i s_i} = E[\mathbf{s}_i \mathbf{s}_i^H] = \mathbf{I}_{Q_i}$, where $\mathbf{I}_{Q_i}$ is an identity matrix of size Q_i. The noise covariance matrix $\mathbf{R}_{n_i n_i} = E[\mathbf{n}_i \mathbf{n}_i^H]$ is unknown and possibly colored. However, noise vectors of any two data sets are assumed to be uncorrelated, $E[\mathbf{n}_i \mathbf{n}_j^H] = \mathbf{0}$, for $i \neq j$. The cross-covariance matrix between two source vectors, $\mathbf{R}_{s_i s_j} = E[\mathbf{s}_i \mathbf{s}_j^H]$, is the correlation-coefficient matrix whose entries on the main diagonal, $\rho_{ij}^{(q)}$, are the correlation coefficients between the q^{th} elements of source vectors $\mathbf{s}_i$ and $\mathbf{s}_j$. Entries other than on the main diagonal of $\mathbf{R}_{s_i s_j}$ are zero.

There is an unknown number, d_{ij}, of sources correlated between the i^{th} and j^{th} source vectors, corresponding to the d_{ij} nonzero entries of $\mathbf{R}_{s_i s_j}$. However, only d sources are correlated across all the data sets. The goal of this work is to estimate the unknown dimension d. Prior works [6], [11], [12] have assumed $d_{ij} = d \ \forall (i,\ j) \in \{1, \ldots, L\}$, $i \neq j$. As an example for a correlation structure that satisfies this assumption, consider the scenario in Figure 1 where the source vectors of three data sets, $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$, each have five independent source signals. The source vector $\mathbf{s}_1$ is repeated again in the last column to illustrate the correlation between the sources in $\mathbf{s}_1$ and $\mathbf{s}_3$. The first two components of each source are correlated among each other. Therefore, $d_{12} = d_{23} = d_{31} = d = 2$.

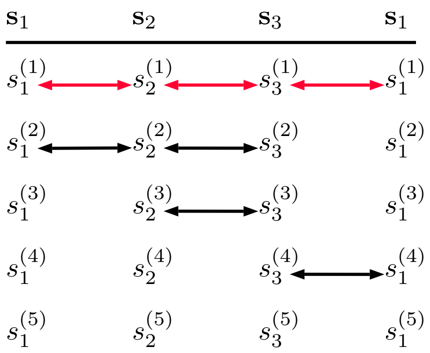

$\mathbf{s}_1$ $\mathbf{s}_2$ $\mathbf{s}_3$ $\mathbf{s}_1$
$s_1^{(1)}$ $s_2^{(1)}$ $s_3^{(1)}$ $s_1^{(1)}$
$s_1^{(2)}$ $s_2^{(2)}$ $s_3^{(2)}$ $s_1^{(2)}$
$s_1^{(3)}$ $s_2^{(3)}$ $s_3^{(3)}$ $s_1^{(3)}$
$s_1^{(4)}$ $s_2^{(4)}$ $s_3^{(4)}$ $s_1^{(4)}$
$s_1^{(5)}$ $s_2^{(5)}$ $s_3^{(5)}$ $s_1^{(5)}$

Fig. 2: Example of a correlation structure for three data sets with $d_{12} = 2$, $d_{23} = 3$, $d_{31} = 2$, and $d = 1$, not satisfying the assumptions in [6], [12], [13]. Arrows indicate correlated components, and red arrows indicate components correlated across all data sets.

In this work, we allow d_{ij} to be different for different pairs of data sets and also different than d. Our work in [13] also allows this case but still makes the simplifying assumption that the components correlated between one pair of data sets are either (i) correlated with all remaining data sets or (ii) uncorrelated with all other remaining data sets. Now consider the scenario in Figure 2, where the first components of each source, $s_1^{(1)}$, $s_2^{(1)}$ and $s_3^{(1)}$ are correlated among each other. This is the subspace of interest to us. The other sources are either only partially correlated or uncorrelated among the data sets. Therefore, the number of source signals correlated among all three data sets is just one, even though there are further source signals correlated between individual pairs of data sets. None of the works mentioned above address this case. For more than three data sets, the number of possible correlation structures increases rapidly, which makes it clear that we require a detector that works for an arbitrary correlation structure.

III. Product of Coherence Matrices

The motivation behind our proposed method is that the rank of the product of all possible source cross-covariance matrices is equal to the number d of correlated sources,

$$\operatorname{rank}\left(\prod_{i,j} \mathbf{R}_{s_i s_j}\right) = d. \tag{2}$$

Here, the indices i and j are chosen in such a way that all $L(L-1)/2$ source cross-covariance matrices are considered at least once, and the dimensions of the matrices match. This is explained in detail in the appendix. For instance, for three data sets with correlation structure shown in Figure 2, where $d = 1$,

$$\operatorname{rank}\left(\mathbf{R}_{s_1 s_2} \mathbf{R}_{s_2 s_3} \mathbf{R}_{s_3 s_1}\right) = 1.$$

However, the true sources are unobservable and d has to be estimated from the observed data. The coherence matrix of two data vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is their normalized cross-covariance matrix and is defined as $\mathbf{C}_{ij} = \mathbf{R}_{ii}^{-1/2}\mathbf{R}_{ij}\mathbf{R}_{jj}^{-H/2}$ [19]. Here and in the following, $\mathbf{R}_{ij} = E[\mathbf{x}_i\mathbf{x}_j^H]$. We will now show that the rank of the product of coherence matrices for all possible pairs of data sets is also approximately equal to d, provided that the SNR is large enough. That is,

$$\operatorname{rank}\left(\prod_{i,j}\mathbf{C}_{ij}\right) \approx d, \tag{3}$$

where the indices i and j are chosen as described by the procedure in the appendix. We now prove (3) for three data sets. For more than three data sets, the proof can be trivially extended.

Consider the product of three coherence matrices $\mathbf{C}_{12}$, $\mathbf{C}_{23}$ and $\mathbf{C}_{31}$,

$$\begin{aligned}\mathbf{C}_{123} &= \mathbf{C}_{12}\mathbf{C}_{23}\mathbf{C}_{31}\\ &= \mathbf{R}_{11}^{-1/2}\mathbf{R}_{12}\mathbf{R}_{22}^{-H/2}\mathbf{R}_{22}^{-1/2}\mathbf{R}_{23}\mathbf{R}_{33}^{-H/2}\mathbf{R}_{33}^{-1/2}\mathbf{R}_{31}\mathbf{R}_{11}^{-H/2}\\ &= \mathbf{R}_{11}^{-1/2}\mathbf{A}_1\mathbf{R}_{s_1s_2}\underbrace{\mathbf{A}_2^H\mathbf{R}_{22}^{-1}\mathbf{A}_2}_{\mathbf{P}}\mathbf{R}_{s_2s_3}\\ &\qquad\times\underbrace{\mathbf{A}_3^H\mathbf{R}_{33}^{-1}\mathbf{A}_3}_{\mathbf{Q}}\mathbf{R}_{s_3s_1}\mathbf{A}_1^H\mathbf{R}_{11}^{-H/2}.\end{aligned} \tag{4}$$

The cross-covariance matrices $\mathbf{R}_{12}$, $\mathbf{R}_{23}$, and $\mathbf{R}_{31}$ do not include any noise terms as the noise is uncorrelated with the source signals and also between any two data sets. Let us expand the expression for matrix $\mathbf{P}$ as

$$\begin{aligned}\mathbf{P} &= \mathbf{A}_2^H\mathbf{R}_{22}^{-1}\mathbf{A}_2\\ &= \mathbf{A}_2^H(\mathbf{A}_2\mathbf{A}_2^H + \mathbf{R}_{n_2n_2})^{-1}\mathbf{A}_2.\end{aligned}$$

Applying the matrix inversion lemma,

$$\begin{aligned}\mathbf{P} &= \mathbf{A}_2^H(\mathbf{R}_{n_2n_2}^{-1} - \mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2(\mathbf{I}_{Q_2} + \mathbf{A}_2^H\mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2)^{-1}\\ &\qquad\times\mathbf{A}_2^H\mathbf{R}_{n_2n_2}^{-1})\mathbf{A}_2\\ &= \mathbf{A}_2^H\mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2 - \mathbf{A}_2^H\mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2(\mathbf{I}_{Q_2} + \mathbf{A}_2^H\mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2)^{-1}\\ &\qquad\times\mathbf{A}_2^H\mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2.\end{aligned}$$

Let $\mathbf{B} = \mathbf{A}_2^H\mathbf{R}_{n_2n_2}^{-1}\mathbf{A}_2$. Therefore,

$$\begin{aligned}\mathbf{P} &= \mathbf{B} - \mathbf{B}(\mathbf{I}_{Q_2} + \mathbf{B})^{-1}\mathbf{B}\\ &= \mathbf{B} - \mathbf{B}(\mathbf{I}_{Q_2} + \mathbf{B})^{-1}(\mathbf{B} + \mathbf{I}_{Q_2} - \mathbf{I}_{Q_2})\\ &= \mathbf{B} - \mathbf{B}(\mathbf{I}_{Q_2} - (\mathbf{I}_{Q_2} + \mathbf{B})^{-1})\\ &= \mathbf{B} - \mathbf{B} + \mathbf{B}(\mathbf{I}_{Q_2} + \mathbf{B})^{-1}\\ &= \mathbf{B}(\mathbf{I}_{Q_2} + \mathbf{B})^{-1}.\end{aligned} \tag{5}$$

Typically the matrix $\mathbf{B} \gg \mathbf{I}_{Q_2}$ when the signal to noise ratio is high. Then, $(\mathbf{I}_{Q_2} + \mathbf{B})^{-1} \approx \mathbf{B}^{-1}$, hence,

$$\mathbf{P} \approx \mathbf{I}_{Q_2}. \tag{6}$$

Using the same derivation, it can be shown that

$$\mathbf{Q} \approx \mathbf{I}_{Q_3}. \tag{7}$$

Inserting the approximate values of $\mathbf{P}$ and $\mathbf{Q}$ in (4), we get

$$\mathbf{C}_{123} \approx \mathbf{R}_{11}^{-1/2}\mathbf{A}_1\mathbf{R}_{s_1s_2}\mathbf{R}_{s_2s_3}\mathbf{R}_{s_3s_1}\mathbf{A}_1^H\mathbf{R}_{11}^{-H/2}.$$

Since all other matrices are full rank, the rank of $\mathbf{C}_{123}$ will be equal to d when the approximations in (6) and (7) apply:

$$\operatorname{rank}(\mathbf{C}_{123}) = d, \quad \text{when } \mathbf{P} = \mathbf{I}_{Q_2} \text{ and } \mathbf{Q} = \mathbf{I}_{Q_3}. \tag{8}$$

Thus, the singular values of $\mathbf{C}_{123}$ are of the form,

$$\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_d > \gamma_{d+1} = \ldots = \gamma_{m_1} = 0, \tag{9}$$

where the first d largest singular values are non-zero and the rest are equal to zero. The matrices $\mathbf{P}$ and $\mathbf{Q}$ will approach identity matrices as the SNR approaches infinity. However, when these approximations are not valid, the rank of $\mathbf{C}_{123}$ will generally be greater than d.

IV. Hypothesis testing using Bootstrap

A popular approach to the problem of model-order selection is to perform a series of binary hypothesis tests. In our context, this means that, starting with source counter $s = 0$, we test the null hypothesis H_s : "s correlated sources" against the alternative H_{s+} : "more than s sources correlated among the data sets". If H_s is rejected, s is incremented and another test of H_s vs. H_{s+} is run. This is repeated until H_s is not rejected or s reaches its maximum possible value.

This approach, however, requires a statistic whose (asymptotic) distribution under the null hypothesis is known. In [12], we derived a generalized likelihood ratio test (GLRT) by assuming that the signals correlated between any two data sets are also correlated across the remaining data sets. This assumption makes the problem of maximizing the likelihood under the unknown parameters tractable [6]. However, with arbitrary correlation structure among the signals, as described in the previous sections, deriving a GLRT becomes challenging. The bootstrap is a resampling technique that can be used to estimate the distribution of a parameter of interest, particularly when the underlying distribution of the data is unknown or is too complicated to derive [17].

We consider M independent and identically distributed (i.i.d.) samples of the data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_L$. Based on the result in (9), when the null hypothesis H_s is true, only the first s singular values of the product of coherence matrices are non-zero. A test statistic to check this is the difference between arithmetic and geometric mean [18], [20]

$$T_s = \left(\frac{1}{m_1 - s}\sum_{m=s+1}^{m_1}\hat{\gamma}_m\right) - \left(\prod_{m=s+1}^{m_1}\hat{\gamma}_m^{\frac{1}{m_1-s}}\right) \quad \text{for } s = 0, \ldots, m_1 - 1. \tag{10}$$

The value of this statistic under the null hypothesis will be close to zero as the sample singular values $\hat{\gamma}_m$ are close to zero. The distribution of T_s under the null hypothesis is estimated using the bootstrap as [18]

$$T_s^{\mathrm{H}}(b) = T_s^*(b) - T_s \quad \text{for } b = 1, \ldots, B,$$

where B denotes the number of bootstrap resamples, $T_s^*(b)$ is the bootstrap test statistic obtained from (10) by using the singular values $\hat{\gamma}_m^*$ estimated from the b^{th} bootstrap resampled data sets, $\mathbf{x}_1^*, \ldots, \mathbf{x}_L^*$. The significance value for the hypothesis test H_s is [18]

$$P_s = \frac{1}{B} \sum_{b=1}^{B} I\left(|T_s| \le |T_s^{\text{H}}(b)|\right)$$

where $I(\cdot)$ is the indicator function. Thus, we propose the decision rule for a given value of probability of false alarm, P_{fa},

$$\hat{d} = \min_{s=0,\ldots,m_1-1} \{s : P_s \ge P_{\text{fa}}\}. \qquad (11)$$

The min-operator choses the smallest s such that $P_s \ge P_{\text{fa}}$. If there is no such s, it chooses $s = m_1 - 1$.

V. Numerical Results

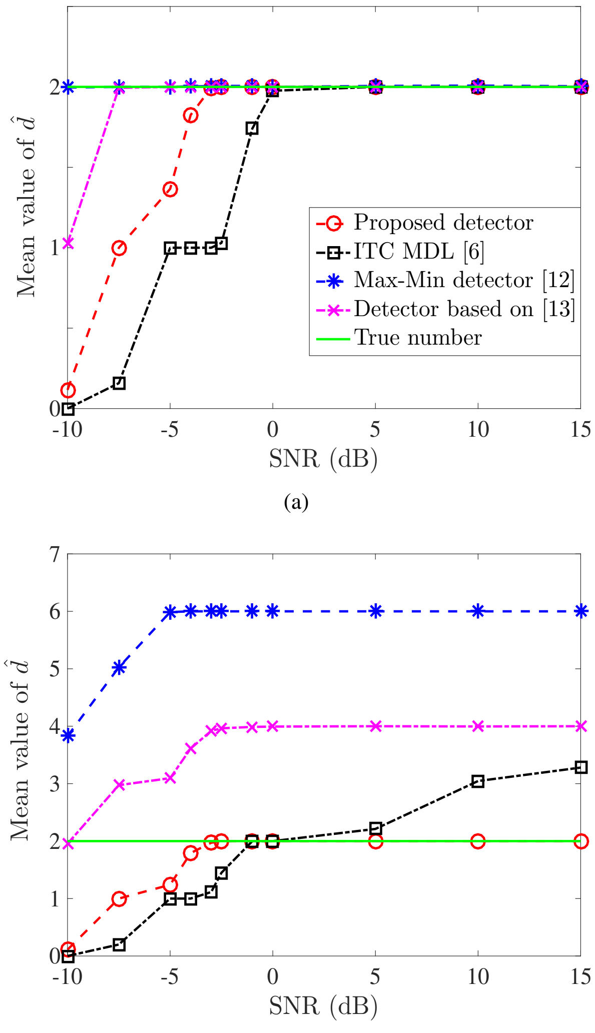
The performance of the proposed detector is evaluated using Monte Carlo simulations and compared with i) the information theoretic criterion (ITC)-based minimum description length (MDL) detector of [6], ii) Max-Min detector in [12] and iii) detector in [13]. All the competing detectors assume a particular correlation structure among the source signals. The results are shown for five data sets with dimensions $m_1 = 14$, $m_2 = 18$, $m_3 = 20$, $m_4 = 22$, $m_5 = 24$, each having $Q_i = 6$ Gaussian distributed signals with unit variance and $M = 400$ samples[1]. The mixing matrices are randomly generated unitary matrices. The added Gaussian noise is autoregressive (AR) of order 3 with coefficient vector $[2.82, 1, 1]$. The number of bootstrap resamples is $B = 500$ and $P_{\text{fa}} = 0.05$. The SNR per signal component is defined as

$$\text{SNR} = 10 \log_{10}\left(\frac{1}{\sigma_n^2}\right),$$

where σ_n^2 is the variance of the white noise component before applying AR filtering. The SNR is the same for all data sets. For each data point, we ran 500 independent trials.

Figure 3a shows the mean of the estimated value $\hat{d}$ as a function of the per-component SNR when the correlation structure between the source signals is chosen such that $d_{ij} = d = 2$ $\forall (i,j) \in \{1, \ldots, L\}, i \neq j$, i.e., the assumption of [6], [12] and [13] applies. In this case, all the detectors approach the true number of correlated signals as the SNR increases. Since the detectors of [12], [13] are designed specifically for the sample-poor regime and incorporate a principal component analysis (PCA) pre-processing step, they approach the true value faster compared to the proposed and the MDL detectors.

Next, Figure 3b shows the performance for a correlation structure where the assumption in [6], [12] and [13] does not apply. In this setup, $d = 2$, but there are signals that are correlated across some data sets but not all i.e. $d_{ij} > d$ $\forall (i,j) \in \{1, \ldots, L\}, i \neq j$. It can be observed that the competing detectors overestimate the number of correlated signals, d, as the SNR increases, whereas the proposed technique continues to give the right answer.



(a)

(b)

Fig. 3: Mean of the estimated number of correlated signals for the proposed detector and competing detectors in [6], [12], [13] for five data sets with $d = 2$ correlated signals. (a) The number of pairwise correlated signals, $d_{ij} = d = 2$ $\forall (i,j) \in \{1, \ldots, L\}, i \neq j$, i.e. assumption in [6], [12], [13] is satisfied. (b) Correlation structure does not satisfy assumption in [6], [12], [13] (Refer to the legend of Fig. 3a for the meaning of the colored markers). For all cases and all detectors, the variance of $\hat{d}$ is small.

VI. Conclusion

We have presented a bootstrap-based hypothesis testing technique for detecting the dimension of the subspace common across multiple data sets. Compared to previous works, the proposed technique does not make any assumption on the correlation structure among the latent signals present in the data sets.

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

Appendix
Procedure for Generating the Product of Coherence Matrices

The indices i and j for generating the product of coherence matrices are chosen using the following procedure:

- If L is odd, $N = (L-1)/2$ groups of data pairs are formed.
  1) For $n = 1, 2, \ldots, N$, repeat the following steps to generate the n^{th} group:
     a) Set $i = 1$.
     b) Compute
     $$j = \begin{cases} \mathrm{mod}(i+n, L) & \text{if } i+n \neq L, \\ L & \text{if } i+n = L. \end{cases}$$
     c) Include $\mathbf{C}_{ij}$ in the product of coherence matrices.
     d) If $j = 1$, then this group is complete. If $j \neq 1$, then let $i = j$ and go back to Step 1b.
  2) Include the $\mathbf{C}_{ij}$'s determined for all groups in the product of coherence matrices.
- If L is even, perform this procedure for $L+1$ data sets with the following modification: For each group n, there will be two pairs $\{i, L+1\}$ and $\{L+1, j\}$. Remove these and instead include $\{i, j\}$ in the product.

For instance, for $L = 5$, the following $N = 2$ groups of data pairs can be formed.

- $n = 1 : \{1,2\}, \{2,3\}, \{3,4\}, \{4,5\}, \{5,1\}$.
- $n = 2 : \{1,3\}, \{3,5\}, \{5,2\}, \{2,4\}, \{4,1\}$.

The two groups are combined to generate the indices for the product of coherence matrices. For $L = 4$, again the data pairs for $L = 5$ are generated. In the first group, $\{4,5\}$ and $\{5,1\}$ are replaced with $\{4,1\}$, and in the second group, $\{3,5\}$ and $\{5,2\}$ are replaced with $\{3,2\}$. Hence:

- $n = 1 : \{1,2\}, \{2,3\}, \{3,4\}, \{\mathbf{4},\mathbf{1}\}$.
- $n = 2 : \{1,3\}, \{\mathbf{3},\mathbf{2}\}, \{2,4\}, \{4,1\}$.

For an odd number of data sets, following this procedure will mean that each pair of coherence matrices appears in the product exactly once. However, for an even number of data sets, $\lceil (L-1)/2 \rceil$ number of pairs will be repeated, where $\lceil\rceil$ denotes the ceiling function.

Acknowledgment

This research was supported by the Alfried Krupp von Bohlen und Halbach foundation under the program "Return of German scientists from abroad", and the German Research Foundation (DFG) under grant SCHR 1384/3-1. The work of D. Ramírez has been partly supported by Ministerio de Economía of Spain under projects: OTOSIS (TEC2013-41718-R) and the COMONSENS Network (TEC2015-69648-REDC), by the Ministerio de Economía of Spain jointly with the European Commission (ERDF) under project ADVENTURE (TEC2015-69868-C2-1-R), and by the Comunidad de Madrid under project CASI-CAM-CM (S2013/ICE-2845).

References

[1] T. Adali, Y. Levin-Schwartz, and V. D. Calhoun, "Multimodal data fusion using source separation: Application to medical imaging," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1494–1506, 2015.

[2] 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.

[3] N. Asendorf and R. R. Nadakuditi, "Improving multiset canonical correlation analysis in high dimensional sample deficient settings," in *Proceedings of the 49th Asilomar Conference on Signals, Systems and Computers*.

[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. i323–i330, 2003.

[5] 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.

[6] 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.

[7] 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.

[8] P. Stoica, K. M. Wong, and Q. Wu, "On a nonparametric detection method for array signal processing in correlated noise fields," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 1030–1032, 1996.

[9] 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.

[10] 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.

[11] M. Bhandary, "Detection of the number of signals in the presence of white noise in decentralized processing," *IEEE Transactions on Signal Processing*, vol. 46, no. 3, pp. 800–803, 1998.

[12] 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 *Proceedings of the IEEE Workshop on Statistical Signal Processing, Palma de Mallorca, Spain*, 2016.

[13] 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 *Proceedings of the European Signal Processing Conference, Budapest, Hungary*, 2016.

[14] M. S. Bartlett, "A note on the multiplying factors for various χ^2 approximations," *Journal of the Royal Statistical Society*, pp. 296–298, 1954.

[15] D. Lawley, "Tests of significance for the latent roots of covariance and correlation matrices," *Biometrika*, vol. 43, pp. 128–136, 1956.

[16] B. Efron and R. J. Tibshirani, *An Introduction To The Bootstrap*. CRC press, 1994.

[17] A. M. Zoubir and D. R. Iskander, *Bootstrap Techniques For Signal Processing*. Cambridge University Press, 2004.

[18] R. F. Brcich, A. M. Zoubir, and P. Pelin, "Detection of sources using bootstrap techniques," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 206–215, 2002.

[19] L. L. Scharf and C. T. Mullis, "Canonical coordinates and the geometry of inference, rate, and capacity," *IEEE Transactions on Signal Processing*, vol. 48, no. 3, pp. 824–831, 2000.

[20] S. Aouada, D. Traskov, N. d'Heureuse, and A. M. Zoubir, "Application of the bootstrap to source detection in nonuniform noise," *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing, Philadelphia PA, USA*, 2005.