Signal Processing 128 (2016) 449–458

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Canonical correlation analysis of high-dimensional data with very small sample support

Yang Song [a], Peter J. Schreier [a,*], David Ramírez [b,c], Tanuj Hasija [a]

[a] Signal and System Theory Group, Universität Paderborn, Paderborn, Germany
[b] Signal Processing Group, University Carlos III of Madrid, Leganés, Spain
[c] Gregorio Marañón Health Research Institute, Madrid, Spain

ARTICLE INFO

Article history:
Received 10 December 2015
Received in revised form
8 April 2016
Accepted 19 May 2016
Available online 20 May 2016

Keywords:
Bartlett-Lawley statistic
Canonical correlation analysis
Model-order selection
Principal component analysis
Small sample support

ABSTRACT

This paper is concerned with the analysis of correlation between two high-dimensional data sets when there are only few correlated signal components but the number of samples is very small, possibly much smaller than the dimensions of the data. In such a scenario, a principal component analysis (PCA) rank-reduction preprocessing step is commonly performed before applying canonical correlation analysis (CCA). We present simple, yet very effective, approaches to the *joint* model-order selection of the number of dimensions that should be retained through the PCA step *and* the number of correlated signals. These approaches are based on reduced-rank versions of the Bartlett–Lawley hypothesis test and the minimum description length information-theoretic criterion. Simulation results show that the techniques perform well for very small sample sizes even in colored noise.

1. Introduction

Correlation analysis based on only small sample support is a challenging task yet with important applications in areas as diverse as biomedicine (e.g. [1,2]), climate science (e.g. [3,4]), array processing (e.g. [5]), and others. In this paper, we look at the scenario where the data sets have large dimensions but there are only few correlated signal components. Probably the most common way of analyzing correlation between two data sets is canonical correlation analysis (CCA) [6]. In CCA, the observed data $\mathbf{x} \in \mathbb{C}^n$ and $\mathbf{y} \in \mathbb{C}^m$ are transformed into p-dimensional internal (latent) representations $\mathbf{a} = \mathbf{S}\mathbf{x}$ and $\mathbf{b} = \mathbf{T}\mathbf{y}$, where $p = \min(n, m)$, using linear transformations described by the matrices $\mathbf{S} \in \mathbb{C}^{p\times n}$ and $\mathbf{T} \in \mathbb{C}^{p\times m}$. The key idea is to determine $\mathbf{S}$ and $\mathbf{T}$ such that most of the correlation between $\mathbf{x}$ and $\mathbf{y}$ is captured in a low-dimensional subspace.

CCA proceeds as follows. First two vectors ("projectors") $\mathbf{s}_1 \in \mathbb{C}^n$ and $\mathbf{t}_1 \in \mathbb{C}^m$ are determined such that the absolute value of the scalar correlation coefficient k_1 between the internal variables $a_1 = \mathbf{s}_1^T\mathbf{x}$ and $b_1 = \mathbf{t}_1^T\mathbf{y}$ is maximized. The internal variables (a_1, b_1) constitute the first pair of *canonical variables*, and k_1 is called the first *canonical correlation (coefficient)*. The next pair of canonical variables (a_2, b_2) maximizes the absolute value of the scalar correlation coefficient k_2 (the second canonical correlation) between $a_2 = \mathbf{s}_2^T\mathbf{x}$ and $b_2 = \mathbf{t}_2^T\mathbf{y}$, subject to the constraint that they are to be uncorrelated with the first pair. A total of p correlations is determined in this manner, and $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_p]^T$, $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_p]^T$. CCA can be performed via the singular value decomposition of the coherence matrix [7]

$$\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1/2} = \mathbf{F}\mathbf{K}\mathbf{G}^H, \tag{1}$$

where $\mathbf{R}_{xy}$ is the cross-covariance matrix between $\mathbf{x}$ and $\mathbf{y}$, and $\mathbf{R}_{xx}$ and $\mathbf{R}_{yy}$ are the auto-covariance matrices of $\mathbf{x}$ and $\mathbf{y}$. The canonical correlations $0 \le k_i \le 1$ are the singular values, which are the diagonal elements of the diagonal matrix $\mathbf{K}$. The transformations that generate the latent representations $\mathbf{a}$ and $\mathbf{b}$ are then described by $\mathbf{S} = \mathbf{F}^H\mathbf{R}_{xx}^{-1/2}$ and $\mathbf{T} = \mathbf{G}^H\mathbf{R}_{yy}^{-1/2}$.

In practice, we do not know the covariance matrices and must estimate them from samples. If CCA is performed based on sample covariance matrices, it leads to sample canonical correlations $\hat{k}_i$. If the number of samples M is not significantly larger than the dimensions m and n, these $\hat{k}_i$'s can be extremely misleading as they are generally substantially overestimated. Indeed, if $M < m + n$ then $m + n - M$ sample canonical correlations are always identically one, which means that they do not carry any information at all about the true population canonical correlations [8]. In order to avoid this, we perform a dimension-reduction preprocessing step before applying CCA. The most common type of preprocessing is principal component analysis (PCA). That is, instead of applying (1) directly to the sample covariance matrices, we first extract a

[*] Corresponding author.
E-mail addresses: yang.song@sst.upb.de (Y. Song), peter.schreier@sst.upb.de (P.J. Schreier), david.ramirez@uc3m.es (D. Ramírez), tanuj.hasija@sst.upb.de (T. Hasija).

http://dx.doi.org/10.1016/j.sigpro.2016.05.020
0165-1684/© 2016 Elsevier B.V. All rights reserved.

reduced number r_x of components from $\mathbf{x}$ that account for a large fraction of the total variance in $\mathbf{x}$. Similarly, we extract r_y components from $\mathbf{y}$ that account for a large fraction of the total variance in $\mathbf{y}$. CCA is then performed on the components extracted from $\mathbf{x}$ and $\mathbf{y}$. The necessity of a PCA step preceding CCA for small sample sizes was shown in [9] using random matrix theory tools. The paper [9], however, did not answer the critical question of how to determine r_x and r_y such that the estimated $\hat{k}_i$'s best reflect the true population canonical correlations k_i.

At the same time, a key question in any correlation analysis is how many correlated signals there are. If we had access to the population canonical correlations, we could simply count the number of nonzero k_i's. Since we do not, we need to estimate the number d of correlated signals from the estimated $\hat{k}_i$'s. This is a model-order selection problem. In this paper, we present approaches to *jointly* determine, for a PCA-CCA setup, the ranks r_x and r_y of the PCA step and the number d of correlated signals based on extremely small sample support, with M possibly less or even substantially less than $m+n$. These approaches rely on the fact that, while m and n may be very large, the number of correlated signals d is often small. However, a complicating factor of the PCA-CCA setup is that PCA is designed to extract components that account for most of the variance *within one* data set, but these components are not necessarily the ones that account for most of the correlation *between two* data sets.

In the literature, most of the work on model-order selection deals with either (i) determining the number of signals in a single data set [10–12] or (ii) the number of correlated signals between two data sets, but without a PCA step [13–19]. There is only little work on the *joint* model-order selection in a PCA-CCA setup, most of which is rather ad hoc [20,21] and only [22] presents a systematic approach. However, none of these joint PCA-CCA techniques works in the sample-poor case. In the absence of any methodical approach in the sample-poor regime, it is common to use very simple rules of thumb such as “choose the PCA ranks such that a certain percentage (e.g., 70%) of the total variance/energy in each data set is retained” (see, e.g., [3]). Needless to say, such rules based on experience only work for specific scenarios.

In general, there are two main approaches to model-order selection: hypothesis tests and information-theoretic criteria. *Hypothesis tests* [13,14] are usually series of binary generalized likelihood ratio tests (GLRTs). Starting at $s=0$, they test whether the model has order s (the null hypothesis) or order greater than s (the alternative). If the null hypothesis is rejected, s is incremented and a new test is run. This proceeds until the null hypothesis is not rejected or the maximum model order is reached. The disadvantage of hypothesis tests is that they require the subjective selection of a probability of false alarm. This can be avoided by using *information-theoretic criteria* (ICs) (e.g., [10]), which compute a score as a function of model order. This score is the difference between the likelihood for the observed data, which measures how well the model fits the observed data, and a penalty function. With increasing order there is an increasing number of free parameters, and so the model fit becomes better. In order to avoid overfitting, complex models are penalized by the penalty function, which increases with model order. The best trade-off is achieved when the difference of likelihood and penalty function is maximized. It should be noted that the GLRT and IC methods for model-order selection are actually closely linked [23]—a fact that we will exploit, as well.

In this paper, we present approaches to the joint model-order selection in a PCA-CCA setup based on reduced-rank versions of both the Bartlett–Lawley hypothesis test and the minimum description length (MDL) IC [10]. As far as we know, these are currently the only techniques capable of handling the combined PCA-CCA approach in the sample-poor regime. An early version of this paper was presented at ICASSP 2015 [24].

We would also like to contrast our work with so-called *sparse CCA* (e.g., [25,26]). In sparse CCA, a sparsity constraint is placed on the projectors $\mathbf{s}_i$ and $\mathbf{t}_i$, which means that each canonical variable a_i or b_i is a linear combination of only a few components in $\mathbf{x}$ and $\mathbf{y}$, respectively. While sparse CCA was not proposed to deal with the sample-poor scenario, in principle it can be used as an alternative to PCA-CCA *if* there is a priori information that the projectors are sparse. However, in many scenarios of interest (e.g., the applications in biomedicine, climate science, and array processing cited above) there is no justification to assume sparse projectors. When applied to non-sparse problems, sparse CCA will not work well.

Our program for this paper is as follows. In Section 2, we formulate the problem and illustrate the issues that arise when performing CCA based on very small sample sizes and how a combined PCA-CCA approach can address these. We present our approaches based on the hypothesis test in Section 3 and based on the MDL-IC in Section 4. Extensive simulation results are shown in Section 5.

2. Problem formulation

We observe M independent and identically distributed (i.i.d.) sample pairs $\mathbf{x}_i \in \mathbb{C}^n$, $\mathbf{y}_i \in \mathbb{C}^m$ that are drawn from the two-channel measurement model

$$\begin{aligned} \mathbf{x} &= \mathbf{A}_x\mathbf{s}_x + \mathbf{n}_x, \\ \mathbf{y} &= \mathbf{A}_y\mathbf{s}_y + \mathbf{n}_y. \end{aligned} \tag{2}$$

The signals $\mathbf{s}_x \in \mathbb{C}^{d+f_x}$ and $\mathbf{s}_y \in \mathbb{C}^{d+f_y}$ are jointly Gaussian with zero means and cross-covariance matrix

$$\mathbf{R}_{s_x s_y} = \begin{bmatrix} \mathbf{diag}(\rho_1\sigma_{x,1}\sigma_{y,1}, \ldots, \rho_d\sigma_{x,d}\sigma_{y,d}) & \mathbf{0}_{d\times f_y} \\ \mathbf{0}_{f_x\times d} & \mathbf{0}_{f_x\times f_y} \end{bmatrix},$$

where $\sigma_{x,i}$ is the unknown standard deviation of signal component $s_{x,i}$, $\sigma_{y,i}$ the unknown standard deviation of signal component $s_{y,i}$, and ρ_i the unknown correlation coefficient between $s_{x,i}$ and $s_{y,i}$. Hence, the first d components of $\mathbf{s}_x$ and $\mathbf{s}_y$ are correlated, whereas the next (f_x, f_y) components are independent between $\mathbf{s}_x$ and $\mathbf{s}_y$. The correlated components may be stronger or weaker than the independent components. Without loss of generality, we assume the auto-covariance matrices $\mathbf{R}_{s_x s_x}$ and $\mathbf{R}_{s_y s_y}$ to be diagonal. The matrices $\mathbf{A}_x \in \mathbb{C}^{n\times(d+f_x)}$ and $\mathbf{A}_y \in \mathbb{C}^{m\times(d+f_y)}$ as well as the dimensions d, f_x, and f_y are deterministic but unknown. Without loss of generality, $\mathbf{A}_x$ and $\mathbf{A}_y$ are assumed to have full column-rank. With all these assumptions, $|\rho_i|$ is the ith canonical correlation coefficient k_i between $\mathbf{s}_x$ and $\mathbf{s}_y$. The noise vectors $\mathbf{n}_x \in \mathbb{C}^n$ and $\mathbf{n}_y \in \mathbb{C}^m$ are independent of each other, independent of the signals, zero-mean Gaussian, and with *unknown* (arbitrary) covariance matrices.

Compared to the dimensions m and n (which may be very large), we assume that there are only few correlated signals and only few independent signals with variance larger than the correlated signals (but there can be many independent signals with variance smaller than the correlated signals). However, because we *do not* assume that the mixing matrices $\mathbf{A}_x$ and $\mathbf{A}_y$ in (2) are sparse, the cross-covariance matrix $\mathbf{R}_{xy}$ between the observed vectors $\mathbf{x}$ and $\mathbf{y}$ is *not sparse* and sparse CCA is generally not suitable for this scenario.

We collect the M sample pairs in data matrices $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_M]$, from which we compute the sample covariance matrices $\hat{\mathbf{R}}_{xx} = \mathbf{X}\mathbf{X}^H/M$, $\hat{\mathbf{R}}_{yy} = \mathbf{Y}\mathbf{Y}^H/M$, and $\hat{\mathbf{R}}_{xy} = \mathbf{X}\mathbf{Y}^H/M$. In the case of small sample support, the sample canonical correlations $\hat{k}_i$, $i = 1, \ldots, p$, $p = \min(n, m)$, computed from the sample covariance matrices can be extremely misleading. It has been shown in [8] that when $M < m + n$, at least $m + n - M$ sample canonical correlations will be identically one regardless of the two-channel model that generates the data samples. In such a small sample scenario, the $\hat{k}_i$'s cannot be used to infer the number of correlated signals. But even in the case with M greater (but not substantially greater) than $m+n$, the sample $\hat{k}_i$'s are generally significantly overestimated. This is shown in Fig. 1, which displays the sample canonical correlations for a model of dimension $m=n=20$, with $d=3$ correlated components and $f_x = f_y = 0$ independent components for different sample sizes M. Even for $M=200$, where the number of samples is ten times the dimension of the system, the $\hat{k}_i$'s for $i \geq 4$ are quite wrong, and it is impossible from visual inspection to determine the number of correlated components.

This motivates the use of a rank-reduction preprocessing step. The most common type of preprocessing is PCA, and a combined PCA-CCA approach is the setup that we consider in our paper. So let us investigate what effect rank reduction has on the estimated canonical correlations. The PCA step retains those r_x and r_y components in $\mathbf{X}$ and $\mathbf{Y}$, respectively, that account for most of their total variance. These components can be computed as follows. We first determine the singular value decompositions (SVDs) of the data matrices $\mathbf{X} = \mathbf{U}_x\mathbf{\Sigma}_x\mathbf{V}_x^H$ and $\mathbf{Y} = \mathbf{U}_y\mathbf{\Sigma}_y\mathbf{V}_y^H$. Then the reduced-rank PCA descriptions of $\mathbf{X}$ and $\mathbf{Y}$ are

$$\mathbf{X}_{r_x} = \mathbf{U}_x^H(:\,,1{:}\,r_x)\mathbf{X} \in \mathbb{C}^{r_x \times M},$$
$$\mathbf{Y}_{r_y} = \mathbf{U}_y^H(:\,,1{:}\,r_y)\mathbf{Y} \in \mathbb{C}^{r_y \times M}, \tag{3}$$

where $\mathbf{U}_x(:\,,1{:}\,r_x)$ denotes the $n \times r_x$ matrix containing the first r_x columns of $\mathbf{U}_x$, which are associated with the largest r_x singular values of $\mathbf{X}$, and $\mathbf{U}_y(:\,,1{:}\,r_y)$ denotes the $m \times r_y$ matrix containing the first r_y columns of $\mathbf{U}_y$, which are associated with the largest r_y singular values of $\mathbf{Y}$. Now let $\widetilde{\mathbf{R}}_{xx} = \mathbf{X}_{r_x}\mathbf{X}_{r_x}^H/M$, $\widetilde{\mathbf{R}}_{yy} = \mathbf{Y}_{r_y}\mathbf{Y}_{r_y}^H/M$, and $\widetilde{\mathbf{R}}_{xy} = \mathbf{X}_{r_x}\mathbf{Y}_{r_y}^H/M$ be the sample covariance matrices from the reduced-dimensional PCA descriptions. The corresponding estimated canonical correlations $\hat{k}_i(r_x, r_y)$ may be computed as the singular values of the reduced-dimensional sample coherence matrix, which is [8]

$$\begin{aligned}
\widetilde{\mathbf{R}}_{xx}^{-1/2}\widetilde{\mathbf{R}}_{xy}\widetilde{\mathbf{R}}_{yy}^{-1/2} &= \mathbf{U}_x^H(:\,,1{:}\,r_x)\mathbf{U}_x\left(\mathbf{\Sigma}_x\mathbf{\Sigma}_x^H\right)^{-1/2}\mathbf{U}_x^H\mathbf{U}_x(:\,,1{:}\,r_x)\mathbf{U}_x^H(:\,,1{:}\,r_x)\mathbf{U}_x\mathbf{\Sigma}_x\mathbf{V}_x^H\mathbf{V}_y\mathbf{\Sigma}_y^H\mathbf{U}_y^H\mathbf{U}_y(:\,,1{:}\,r_y)\mathbf{U}_y^H(:\,,1{:}\,r_y)\mathbf{U}_y\left(\mathbf{\Sigma}_y\mathbf{\Sigma}_y^H\right)^{-1/2}\mathbf{U}_y^H\mathbf{U}_y(:\,,1{:}\,r_y)\\
&= \left[\mathbf{I}_{r_x}, \mathbf{0}_{r_x\times(n-r_x)}\right]\left(\mathbf{\Sigma}_x\mathbf{\Sigma}_x^H\right)^{-1/2}\left[\mathbf{I}_{r_x}, \mathbf{0}_{r_x\times(n-r_x)}\right]^H\left[\mathbf{I}_{r_x}, \mathbf{0}_{r_x\times(n-r_x)}\right]\mathbf{\Sigma}_x\mathbf{V}_x^H\mathbf{V}_y\mathbf{\Sigma}_y^H\left[\mathbf{I}_{r_y}, \mathbf{0}_{r_y\times(m-r_y)}\right]^H\left[\mathbf{I}_{r_y}, \mathbf{0}_{r_y\times(m-r_y)}\right]\left(\mathbf{\Sigma}_y\mathbf{\Sigma}_y^H\right)^{-1/2}\left[\mathbf{I}_{r_y}, \mathbf{0}_{r_y\times(m-r_y)}\right]^H\\
&= \mathbf{\Sigma}_x^{-1}(1{:}\,r_x, 1{:}\,r_x)\left[\mathbf{\Sigma}_x(1{:}\,r_x, 1{:}\,r_x), \mathbf{0}_{r_x\times(M-r_x)}\right]\mathbf{V}_x^H\mathbf{V}_y\left[\mathbf{\Sigma}_y(1{:}\,r_y, 1{:}\,r_y), \mathbf{0}_{r_y\times(M-r_y)}\right]^H\mathbf{\Sigma}_y^{-1}(1{:}\,r_y, 1{:}\,r_y)\\
&= \left[\mathbf{I}_{r_x}, \mathbf{0}_{r_x\times(M-r_x)}\right]\mathbf{V}_x^H\mathbf{V}_y\left[\mathbf{I}_{r_y}, \mathbf{0}_{r_y\times(M-r_y)}\right]^H = \mathbf{V}_x^H(:\,,1{:}\,r_x)\mathbf{V}_y(:\,,1{:}\,r_y).
\end{aligned} \tag{4}$$

The thus computed canonical correlations $\hat{k}_i(r_x, r_y)$, $i = 1, \ldots, r$, $r = \min(r_x, r_y)$, depend on the ranks r_x and r_y. As seen in (4), the ith estimated canonical correlation $\hat{k}_i(r_x, r_y)$ can be found as the ith largest singular value of $\mathbf{V}_x^H(:\,,1{:}\,r_x)\mathbf{V}_y(:\,,1{:}\,r_y)$, where $\mathbf{V}_x$ and $\mathbf{V}_y$ are the matrices of right singular vectors of $\mathbf{X}$ and $\mathbf{Y}$, respectively. To avoid defective unit canonical correlations, we must choose $r_x + r_y \leq M$ and $\max(r_x, r_y) \leq p$. This, however, does not tell us what the optimum choices for r_x and r_y are such that the $\hat{k}_i(r_x, r_y)$'s are as close to the true canonical correlations as possible.

Intuitively, it seems that r_x and r_y should be chosen large enough to capture as much of the correlated signal components as possible without including too much noise. If the correlated components are weaker than some of the independent components, this will inevitably mean that the PCA preprocessing step must also keep those stronger independent components. On the other hand, if the correlated components are also the strongest components, it would be better if the PCA step got rid of the independent components. Hence, without noise, r_x would ideally be chosen between d and $d + f_x$, and r_y between d and $d + f_y$. With noise, the ranks r_x and r_y may also fall outside of these ranges, depending on the properties of the noise and the relative strengths of the signals.

It can be shown using Cauchy's interlacing theorem (the result is presented as Lemma 1 in Appendix A) that increasing the ranks of the PCA steps increases *every* estimated canonical correlation coefficient. Hence, choosing too large an r_x or r_y will lead to estimated canonical correlations that are greater, possibly significantly greater, than the true canonical correlations. On the other hand, if r_x and r_y are not large enough, then the rank-reduced representation does not contain all of the correlated components, and thus the estimated canonical correlations can be too small.

These considerations can be illustrated by the following example, where $M=30$ and $m = n = 20$. There are $d=3$ correlated signals (each with variance 1.5) and $f = f_x = f_y = 2$ independent signals (each with variance 5). Since the independent signals are stronger than the correlated signals (and the numbers f_x and f_y of independent signals in $\mathbf{s}_x$ and $\mathbf{s}_y$ are identical), we would expect $r_x = r_y = d + f = 5$ to be the optimum rank for the PCA step. Indeed, Fig. 2 shows that choosing $r = r_x = r_y$ greater than 5 leads to $\hat{k}_i$'s that are too large, whereas r less than 5 leads to $\hat{k}_i$'s that are too small. While the exact relationships depend on the variances of

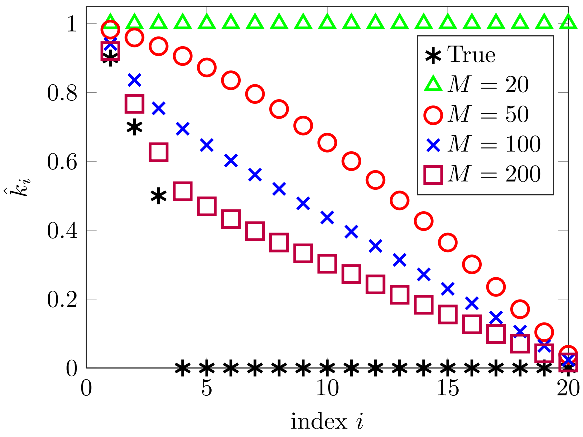


Fig. 1. Sample canonical correlation coefficients $\hat{k}_i$ for different sample sizes M, averaged over 1000 runs. There are three nonzero population canonical correlations, which are 0.9, 0.7, and 0.5, depicted as ∗. In all cases shown, the true k_i's are significantly overestimated.

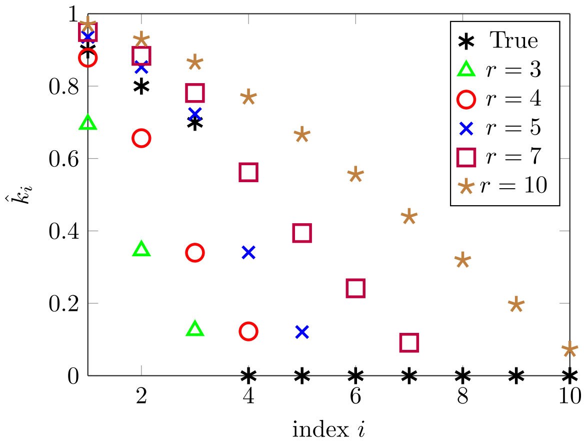


Fig. 2. Effect of rank reduction on the estimated canonical correlations $\hat{k}_i(r)$, averaged over 1000 runs. There are $d=3$ correlated signal components with population canonical correlation coefficients 0.9, 0.8, and 0.7 (depicted as ∗), and $f_x = f_y = 2$ stronger independent signal components. For $r > 5$, the canonical correlation coefficients are all overestimated. For $r < 5$, the nonzero coefficients are underestimated. The ranks of the PCA steps for $\mathbf{x}$ and $\mathbf{y}$ are the same: $r = r_x = r_y$.

signal and noise components and the correlation coefficients, the principle observed here generalizes to other settings.

3. Order selection based on hypothesis test

3.1. Traditional test

In the case of sufficient number of samples, the traditional hypothesis test [13,14] for determining the number d of correlated components between $\mathbf{x}$ and $\mathbf{y}$ is a series of binary hypothesis tests. Starting with $s=0$, it tests the null hypothesis H_0: $d=s$ versus the alternative hypothesis H_1: $d > s$. If H_0 is rejected, s is incremented and a new test is run. This proceeds until H_0 is not rejected or $s = p = \min(n, m)$ is reached.

The binary test in [13,14] is a generalized likelihood ratio test (GLRT) of H_0 vs. H_1. For a given number s of correlated signals, let Ω_s denote the parameter space of the model, which consists of the auto- and cross-covariance matrices. The maximum value of the log-likelihood function for a given number s of correlated signals, maximized over the parameter space Ω_s, is [19]

$$\ell_{\max}(\mathbf{X}, \mathbf{Y}|\Omega_s) = -M \ln \prod_{i=1}^{s}\left(1 - \hat{k}_i^2\right). \tag{5}$$

Canonical correlation coefficients close to 1 are strong evidence of correlation between $\mathbf{x}$ and $\mathbf{y}$ and thus lead to large $\ell_{\max}$. Now let $\Omega_{d>s}$ denote the parameter space of all models where the assumed number of correlated signals d is greater than s. The generalized log-likelihood ratio for testing H_0 vs. H_1 is [17]

$$\begin{aligned}\Lambda(n, m, s) &= \ell_{\max}(\mathbf{X}, \mathbf{Y}|\Omega_{d=s}) - \ell_{\max}(\mathbf{X}, \mathbf{Y}|\Omega_{d>s})\\ &= \ell_{\max}(\mathbf{X}, \mathbf{Y}|\Omega_s) - \ell_{\max}(\mathbf{X}, \mathbf{Y}|\Omega_p)\\ &= M \ln \prod_{i=s+1}^{p}\left(1 - \hat{k}_i^2\right),\end{aligned} \tag{6}$$

where the second identity follows from the fact that the maximum of the likelihood function, under the constraint $d > s$, occurs when the model has the most degrees of freedom, i.e., for $d = p = \min(n, m)$. The cross-covariance matrix has $N_\Omega(n, m, s) = 2s(m + n - s)$ degrees of freedom [19]. Wilks' theorem [27] says that $-2\Lambda(n, m, s)$ is asymptotically (as $M \to \infty$) χ^2-distributed with degrees of freedom equal to the difference of the dimensions of the parameter spaces Ω_p and Ω_s[1]:

$$\begin{aligned}N_\Lambda(n, m, s) &= N_\Omega(n, m, p) - N_\Omega(n, m, s)\\ &= 2p(m + n - p) - 2s(m + n - s)\\ &= 2(m - s)(n - s)\end{aligned} \tag{7}$$

For finite M, the closeness of the χ^2-approximation may be improved by replacing $-2\Lambda(n, m, s)$ with the Bartlett–Lawley statistic [13,14]

$$C(n, m, s) = -2\left(M - s - \frac{m + n + 1}{2} + \sum_{i=1}^{s} \hat{k}_i^{-2}\right)\ln \prod_{i=s+1}^{p}\left(1 - \hat{k}_i^2\right). \tag{8}$$

This correction makes the moments of the test statistic equal to the moments of the χ^2-distribution. As long as M is large compared to m and n, the statistic $C(n, m, s)$ is generally very close to a χ^2-distribution. Note that this is independent of the covariance matrix of the noise, since it is not used anywhere in the derivation. This allows computation of a test threshold $T(n, m, s)$ for a given probability of false alarm.

3.2. Test with PCA preprocessing

Instead of running the test directly on $\mathbf{X}$ and $\mathbf{Y}$, we would like to apply the test to the reduced-rank PCA descriptions $\mathbf{X}_{r_x}$ and $\mathbf{Y}_{r_y}$ obtained in (3). By performing PCA on $\mathbf{x}$ and $\mathbf{y}$, we create a new reduced-rank two-channel model:

$$\begin{aligned}\mathbf{x}_{r_x} &= \mathbf{U}_x^H(:, 1: r_x)\mathbf{x}\\ &= \mathbf{U}_x^H(:, 1: r_x)\mathbf{A}_x\mathbf{s}_x + \mathbf{U}_x^H(:, 1: r_x)\mathbf{n}_x\\ &= \widetilde{\mathbf{A}}_x\mathbf{s}_x + \widetilde{\mathbf{n}}_x,\\ \mathbf{y}_{r_y} &= \mathbf{U}_y^H(:, 1: r_y)\mathbf{y}\\ &= \mathbf{U}_y^H(:, 1: r_y)\mathbf{A}_y\mathbf{s}_y + \mathbf{U}_y^H(:, 1: r_y)\mathbf{n}_y\\ &= \widetilde{\mathbf{A}}_y\mathbf{s}_y + \widetilde{\mathbf{n}}_y.\end{aligned} \tag{9}$$

In this model, the new matrices $\widetilde{\mathbf{A}}_x$ and $\widetilde{\mathbf{A}}_y$ have full rank because $\mathbf{A}_x$ and $\mathbf{A}_y$ are assumed to have full rank. With the PCA preprocessing the GLRT statistic is

$$\Lambda(r_x, r_y, s) = M \ln \prod_{i=s+1}^{r}\left(1 - \hat{k}_i^2(r_x, r_y)\right), \tag{10}$$

and the Bartlett–Lawley statistic is

$$\begin{aligned}C(r_x, r_y, s) = -2\left(M - s - \frac{r_x + r_y + 1}{2} + \sum_{i=1}^{s} \hat{k}_i^{-2}(r_x, r_y)\right)\\ \ln \prod_{i=s+1}^{r}\left(1 - \hat{k}_i^2(r_x, r_y)\right)\end{aligned} \tag{11}$$

for $s = 0, \ldots, r - 1$ with $r = \min(r_x, r_y)$. The challenge in the reduced-rank version of the hypothesis test is thus to *jointly*

[1] In this difference, only the degrees of freedom associated with the cross-covariance matrix matter.

determine the best ranks r_x,r_y of the PCA steps and the number d of correlated signals. As long as the number of samples M is large compared to the minimum PCA dimension $r = \min(r_x, r_y)$ but r_x and r_y are not too small (which we will explain in the next paragraph), the new test statistic $C(r_x, r_y, d)$ under H_0: $d = s$ is still approximately χ^2-distributed with $2(r_x - d)(r_y - d)$ degrees of freedom. We denote by $r_{\max}$ the largest r for which the χ^2-distribution holds well enough. Of course, requiring M to be large with respect to r is a much more relaxed condition than requiring M to be large with respect to the dimensions n and m. This is because r_x and r_y do not have to be chosen greater (unless there are strong noise components) than $d + f_x$ and $d + f_y$, respectively, which are usually much smaller than n and m.

There is, however, a complication. By applying PCA to $\mathbf{x}$ and $\mathbf{y}$, we might eliminate some of the correlated components if the PCA ranks r_x and r_y are not chosen large enough. If this is the case, then the number of correlated components $\tilde{d}$ in the *reduced-rank descriptions* $\mathbf{x}_{r_x}$ and $\mathbf{y}_{r_y}$ will be *smaller* than the number of correlated components d between $\mathbf{x}$ and $\mathbf{y}$. As a consequence, $C(r_x, r_y, d)$ will no longer resemble a χ^2-distribution. Instead, $C(r_x, r_y, \tilde{d})$ with $\tilde{d} < d$ will now be approximately χ^2. By choosing r_x and r_y not large enough it thus becomes likely that the null hypothesis “there are $\tilde{d}$ correlated signals” is not rejected, thus deciding for a smaller number $\tilde{d}$ than the true d.

We are now getting closer to writing down a rule for jointly selecting r_x, r_y, and d. In order to motivate this rule, we summarize the preceding discussion: Provided the PCA ranks r_x and r_y are chosen sufficiently large to capture all correlated components while r is still small compared to M, i.e., $r \le r_{\max}$, the statistic $C(r_x, r_y, d)$ in (11) is approximately χ^2 (again irrespective of the noise covariance matrix). This means that in a series of binary tests of H_0: $d = s$ vs. H_1: $d > s$ (testing all values of s starting from 0 until H_0 is not rejected or the maximum $s = r_{\max}$ is reached) d would generally not be *overestimated*. It is likely, however, to be *underestimated*, if r_x and r_y are not chosen large enough. If r_x and r_y are too small, then the reduced-rank PCA descriptions do not capture all of the correlated components and thus the series of binary tests would decide for too small a d. This reasoning motivates the following decision rule.

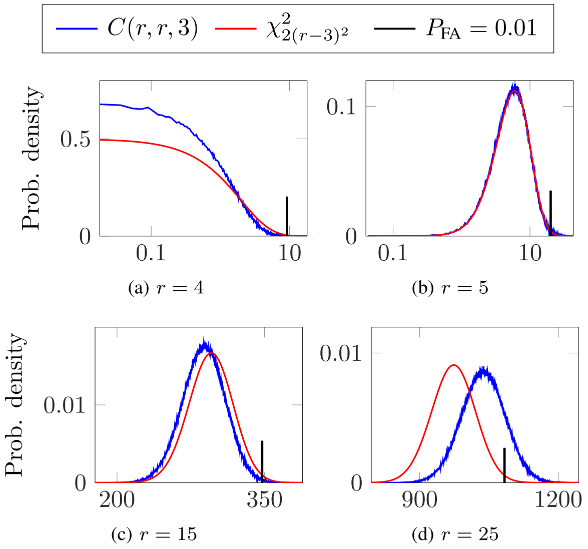
Detector 1 (“**max–min detector**”): *Choose*

$$\hat{d} = \max_{\{r_x,r_y\}=1,\ldots,r_{\max}} \ \min_{s=0,\ldots,r-1} \left\{ s\colon C(r_x, r_y, s) < T(r_x, r_y, s) \right\} \qquad (12)$$

and choose the r_x and r_y that lead to $\hat{d}$ as the PCA ranks. In (12) the min-operator chooses the smallest s such that the statistic $C(r_x, r_y, s)$ falls below the threshold $T(r_x, r_y, s)$, which ensures a given probability of false alarm. If there is no such s, then it chooses $s{=}r$. This step is similar to the traditional test, except that $T(r_x, r_y, s)$ depends on r_x and r_y. The rule (12) is based on the fact that if r_x and r_y are not chosen optimally, the min-step might return a number smaller than d. Hence, the min-step is performed for all r_x and r_y from 1 up to $r_{\max}$, and the maximum result is chosen as $\hat{d}$.

3.3. Example

We will use an example to illustrate both the closeness of the χ^2-approximation and the idea of the max–min detector. We consider a scenario with $m{=}n{=}100$, $d{=}3$ correlated signals, $f = f_x = f_y = 2$ stronger interfering signals, and $M{=}50$ samples. The noise variance is chosen small compared to the signal variances. For $s{=}d{=}3$ and $r = r_x = r_y$, Fig. 3 compares histograms of the statistic $C(r, r, 3)$ with the probability density function of a χ^2-distribution with $2(r_x - d)(r_y - d) = 2(r - 3)^2$ degrees of freedom. As long as r is large enough to capture all correlated components (which is the case for $r \ge d + f = 5$ since the independent signals are stronger than the correlated signals) but small compared to M, the statistic $C(r, r, 3)$ is very well approximated by the χ^2-distribution. This can be seen in subplots (b) $r{=}5$ and (c) $r{=}15$ (where we start to notice some divergence between the statistic and its approximation). Subplot (d) shows $r{=}25$, which is not small enough with respect to $M{=}50$. Here the χ^2-distribution is no longer a good approximation of the test statistic.



(a) $r = 4$

(b) $r = 5$

(c) $r = 15$

(d) $r = 25$

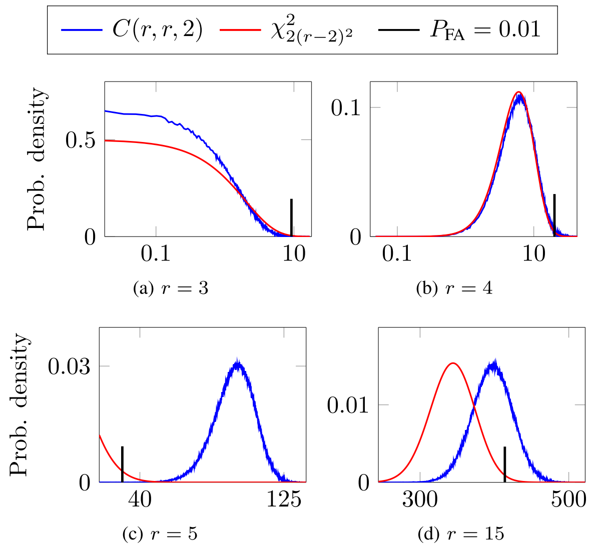
Fig. 3. Histogram of the test statistic $C(r, r, 3)$ (in blue) and the probability density function of a χ^2-distribution with $2(r - 3)^2$ degrees of freedom (in red), for $s = d = 3$ and different PCA ranks $r = r_x = r_y$. Histograms are computed from 10^6 independent trials. Also shown as vertical lines are the thresholds $T(r, r, 3)$ for a probability of false alarm $P_{\text{FA}} = 0.01$. The horizontal axis uses a logarithmic scale. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

On the other hand, if $r < 5$ then the PCA step eliminates some correlated components. This can be observed in subplot (a) for $r{=}4$, where the histogram of $C(4, 4, 3)$ does not approximate a χ^2-distribution. Because the PCA steps with $r_x = r_y = 4$ keep the two stronger independent signals and only two of the three weaker correlated signals, the reduced-rank PCA descriptions $\mathbf{x}_{r_x}$ and $\mathbf{y}_{r_y}$ only have $\tilde{d} = 2$ correlated signals rather than $d{=}3$. It can be observed in Fig. 4 (b) that $C(4, 4, 2)$ indeed well approximates a χ^2-distribution with $2(r - \tilde{d})^2 = 2(4 - 2)^2 = 8$ degrees of freedom.

So let us look at how the max–min detector would proceed in this example. To illustrate this, we again consider Figs. 3 and 4, which compare histograms of $C(r, r, s)$ with χ^2-distributions with $2(r - s)^2$ degrees of freedom for $s = d = 3$ (Fig. 3) and $s{=}2$ (Fig. 4). Also shown in these figures are the thresholds $T(r, r, s)$ for a probability of false alarm $P_{\text{FA}} = 0.01$. According to (12), for given r_x and r_y, the detector needs to find the minimum s (between 0 and r) such that the statistic C falls below the threshold T. Consider first $r_x = r_y = 4$, which is too small because the PCA steps eliminate one correlated component. From Fig. 4(b), we see that it is likely that $C(4, 4, 2)$ falls below $T(4, 4, 2)$, which means that for $r_x = r_y = 4$, the min-step of the detector would likely return too small a number of correlated signals ($s{=}2$).[2]

Now consider $r_x = r_y = 5$, which is large enough so that the PCA steps capture all correlated components. It can now be observed in

[2] If we plotted the test statistics and thresholds also for $s{=}0$ and $s{=}1$ we would see that it is unlikely that a value $s < 2$ would be chosen.



(a) $r=3$

(b) $r=4$

(c) $r=5$

(d) $r=15$

Fig. 4. Histogram of the test statistic $C(r, r, 2)$ (in blue) and the probability density function of a χ^2-distribution with $2(r-2)^2$ degrees of freedom (in red), for $d=3$ but $s=2$ and different PCA ranks $r = r_x = r_y$. The vertical lines are the thresholds $T(r, r, 2)$ for a probability of false alarm $P_{FA} = 0.01$. The horizontal axis uses a logarithmic scale. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Fig. 4(c) that for $s=2$, $C(5, 5, 2)$ will likely not fall below $T(5, 5, 2)$.[3] On the other hand, Fig. 3(b) shows that it is likely that $C(5, 5, 3)$ falls below $T(5, 5, 3)$, hence returning $s=3$ in the min-step of the detector.

Finally, consider $r_x = r_y = 15$, which is larger than needed to capture all correlated components. If r_x and r_y are too large then it becomes increasingly difficult, as can be observed in Fig. 2, to distinguish between the sample correlation coefficients that are associated with the correlated signals and those that are not. The min-step of the detector would still not generally overestimate d (because the χ^2-approximation remains valid under H_0) but it might *underestimate* it. This becomes clear from looking at Fig. 4(d), which shows that there is a rather high chance that the min-step would select $s=2$. However, an underestimating min-step is not a problem for the max-min detector because it selects the maximum of all min-step results.

4. Order selection based on information-theoretic criterion

A disadvantage of the hypothesis testing approach to order selection is the requirement of selecting a probability of false alarm P_{FA}. Setting P_{FA} too high will lead to a detector that tends to overfit, setting it too low will generally underfit. Achieving the best performance thus requires the right trade-off. In this section, we present two alternative approaches that do not require the manual selection of a threshold and are based on the minimum description length (MDL)-IC. The first approach will remain a hypothesis test but with automatic P_{FA} – selection exploiting a link between the GLRT and the IC for model-order selection. The second approach will be a max–min detector based directly on the MDL-IC.

4.1. Setting the threshold based on the MDL-IC

The MDL-IC for selecting the number of correlated signals in two data sets (without PCA steps) is [19]

$$I_{MDL}(n, m, s) = -\ell_{\max}(\mathbf{X}, \mathbf{Y}|\Omega_s) + \frac{1}{2}\ln(M)N_\Omega(n, m, s)$$

$$= M\ln\prod_{i=1}^{s}\left(1-\hat{k}_i^2\right) + \ln(M)s(m+n-s). \tag{13}$$

In this expression, the second term is the penalty term that depends on the degrees of freedom of the model[4] and thus penalizes overly complex models. The model order chosen is the value of s for which $I_{MDL}(n, m, s)$ is minimized. The reduced-rank version of (13), which accounts for the PCA steps, is

$$I_{MDL}(r_x, r_y, s)$$

$$= M\ln\prod_{i=1}^{s}\left(1-\hat{k}_i^2(r_x, r_y)\right) + \ln(M)s(r_x+r_y-s). \tag{14}$$

As has been noted in [23], there is the following connection between the MDL-IC and the log-likelihood ratio of the reduced-rank GLRT H_0: $d = s$ vs. H_1: $d > s$:

$$I_{MDL}(r_x, r_y, r) - I_{MDL}(r_x, r_y, s)$$

$$= \Lambda(r_x, r_y, s) + \ln(M)N_\Lambda(r_x, r_y, s) \tag{15}$$

with $N_\Lambda(r_x, r_y, s) = (r_x - s)(r_y - s)$. When choosing between model orders s and r based on the MDL-IC, we decide for model order s if $I_{MDL}(r_x, r_y, r) > I_{MDL}(r_x, r_y, s)$. Because of (15) we can implement this decision rule also based on the GLRT. We decide for model order s rather than a model order greater than s if

$$\Lambda(r_x, r_y, s) > \underbrace{-\ln(M)(r_x-s)(r_y-s)}_{T_{MDL}(r_x, r_y, s)}. \tag{16}$$

The term on the right-hand side of this inequality is thus the threshold for the GLRT, which is determined based on the MDL-IC. Note that it is unnecessary to apply the Bartlett–Lawley correction because this would amount to multiplying both sides of the inequality (16) with the same factor. Thus, we obtain the following max-min decision rule in terms of $\Lambda(r_x, r_y, s)$ rather than $C(r_x, r_y, s)$.

Detector 2 (max–min detector with threshold set by MDL-IC): *Choose*

$$\hat{d} = \max_{\{r_x, r_y\}=1,\ldots,r_{\max}} \min_{s=0,\ldots,r-1} \{s: \Lambda(r_x, r_y, s) > T_{MDL}(r_x, r_y, s)\}, \tag{17}$$

where $\Lambda(r_x, r_y, s)$ is given in (10) and $T_{MDL}(r_x, r_y, s)$ is given in (16), and choose the r_x and r_y that lead to $\hat{d}$ as the PCA ranks.

4.2. Min-MDL detector

Another approach that does not require the selection of P_{FA} applies the max–min idea directly to the MDL-IC. Let us first write down the decision rule and interpret it afterwards.

Detector 3 ("max-min MDL-IC detector"): *Choose*

$$\hat{d} = \max_{\{r_x, r_y\}=1,\ldots,r_{\max}} \operatorname*{argmin}_{s=0,\ldots,r-1} I_{MDL}(r_x, r_y, s) \tag{18}$$

and choose the r_x and r_y that lead to $\hat{d}$ as the PCA ranks.

In order to understand this detector, we note, based on the discussion in the preceding subsection, that

[3] As before, if we plotted the test statistics also for $s=0$ and $s=1$, we would see that it is even less likely that $C(5, 5, s)$ falls below $T(5, 5, s)$ if $s < 2$.

[4] As before, only the degrees of freedom associated with the cross-covariance matrix are considered because the degrees of freedom associated with the auto-covariance matrices do not depend on s. Hence, they do not matter in the following optimization problems.

$$
\begin{aligned}
&\underset{s=0,\ldots,r-1}{\operatorname{argmin}}\, I_{\mathrm{MDL}}(r_x, r_y, s)\\
&= \underset{s=0,\ldots,r-1}{\operatorname{argmax}}\,[-I_{\mathrm{MDL}}(r_x, r_y, s)]\\
&= \underset{s=0,\ldots,r-1}{\operatorname{argmax}}\,[I_{\mathrm{MDL}}(r_x, r_y, r) - I_{\mathrm{MDL}}(r_x, r_y, s)]\\
&= \underset{s=0,\ldots,r-1}{\operatorname{argmax}}\,[\Lambda(r_x, r_y, s) - T_{\mathrm{MDL}}(r_x, r_y, s)]. \qquad (19)
\end{aligned}
$$

The min-step in Detector 3 thus chooses the value of s that *maximizes* the difference between the GLRT statistic and the MDL test threshold. This is different from the min-step in Detector 2, which picks the *smallest* s for which the test statistic exceeds the threshold. Therefore, Detector 3 will never pick a $\hat{d}$ smaller, but possibly larger, than Detector 2.

5. Performance evaluation

In this section, we compare the performance of our three model-order selection schemes among each other and with competing approaches. In the absence of a competing systematic approach to the *joint* model-order selection in PCA-CCA, we determined the PCA ranks r_x and r_y separately from the number of correlated signals d. We used the sample eigenvalue-based (SEV) technique [11] for selecting r_x and r_y because it is one of the few techniques that can handle the sample-poor case for a single channel. For the selection of d we used the canonical correlation test (CCT) [17] with $P_{\mathrm{FA}} = 0.005$, the Akaike information criterion (AIC) [19], and the MDL criterion [19].

Figs. 5–12 show the probability of selecting the correct d for different setups. In the first setup, shown in Figs. 5–7, we consider a system with $d=2$ correlated signals (each with variance 5 and correlation coefficients 0.8 and 0.7), and $f_x=3$ and $f_y=4$ independent signals (each with variance 1.5). The matrices $\mathbf{A}_x$ and $\mathbf{A}_y$ are randomly generated unitary matrices. For each data point, we ran 1000 independent Monte Carlo trials.

We first consider a system with fixed dimension $m=n=40$. In Fig. 5, we show the performance as a function of the number of samples M when the noise is white and each noise component has unit variance. We see that the performance of Detector 1 depends on the choice of P_{FA}: For smaller M, P_{FA} should be chosen larger, whereas for larger M, a smaller P_{FA} performs better. Detector 2 does this trade-off automatically and performs very well even for very small sample sizes. All other approaches (including Detector 3) still perform quite well but require larger sample support.

The picture completely changes when we have colored rather than white noise. We now generate the noise from a spatially varying moving average (MA) process of order 3 with coefficients $[\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}]$. Before the spatial averaging, the noise components have variance 1/3. It can be seen in Fig. 6 that methods that select r_x and r_y separately from d completely fail. This is because a single-channel technique such as SEV cannot distinguish between signal and noise eigenvalues if the noise is colored. The performance of our detectors, on the other hand, is actually improved particularly for very small sample sizes.

In Fig. 7, we reconsider the white noise case but with varying dimensions $m=n$ and fixed sample size $M=100$. When the noise is independent of space and time, increasing the ratio of the data dimensions m, n to the number of samples M shrinks the signal-subspace [11], which worsens the detection performance. We note, however, that the decrease in performance affects the SEV+X techniques much more than our detectors. Indeed, Detector 2 again shows a very reliable performance even for large dimensions. The main reason behind this effect is that the SEV technique is designed to keep all the signal components (i.e., correlated *and* independent components) whereas our detectors aim to eliminate weaker independent components in the PCA step. The presence of independent components deteriorates the detection performance of the subsequent CCA step.

So far, we have looked at a case where the correlated signals are *stronger* than the independent signals. We now investigate what

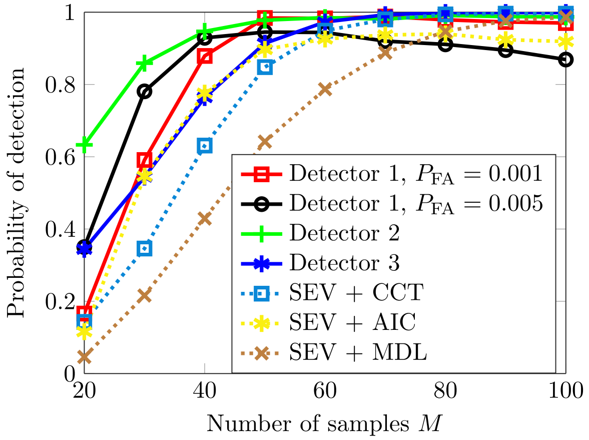


Fig. 5. Performance of our Detectors 1, 2, 3 and competing approaches for *white* noise. System dimensions are $m=n=40$.

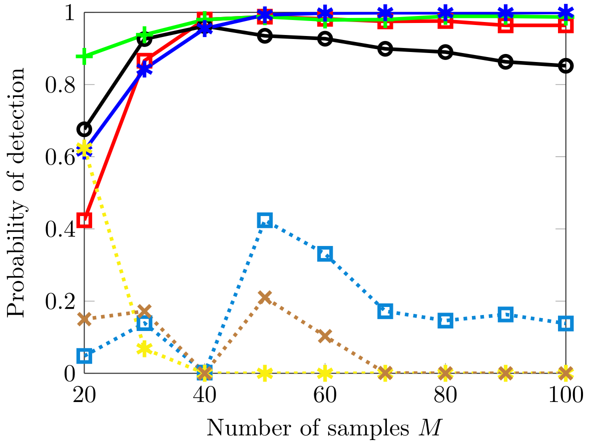


Fig. 6. Same setup as in Fig. 5 but with *colored* MA noise. For the meaning of the colored markers, refer to the legend of Fig. 5.

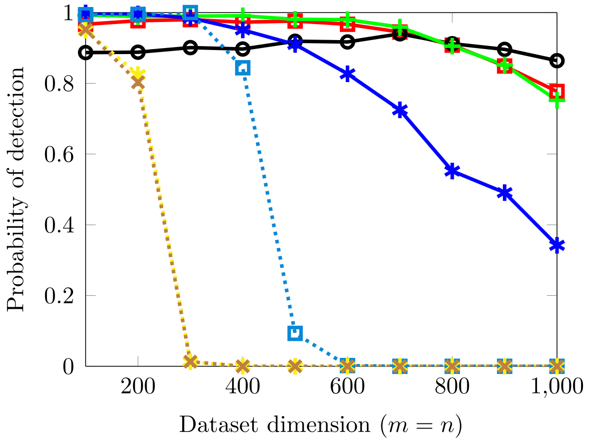


Fig. 7. Same setup as in Fig. 5 but with varying dimensions $m=n$ and fixed sample size $M=100$. For the meaning of the colored markers, refer to the legend of Fig. 5.

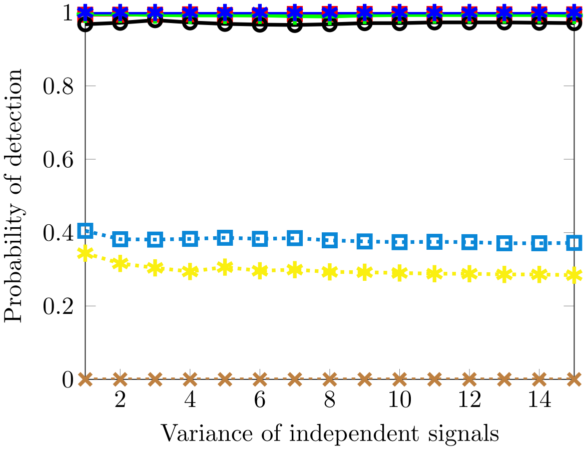


Fig. 8. Effect of the independent signals' variance on performance. Settings: $d=7$ correlated signals with variance 10 and correlation coefficients (0.92, 0.9, 0.88, 0.85, 0.83, 0.8, 0.75), $f_x = f_y = 2$ independent signals of varying variance, $m=n=80$, $M=150$, colored AR(1) noise with coefficient 0.65. For the meaning of the colored markers, refer to the legend of Fig. 5.

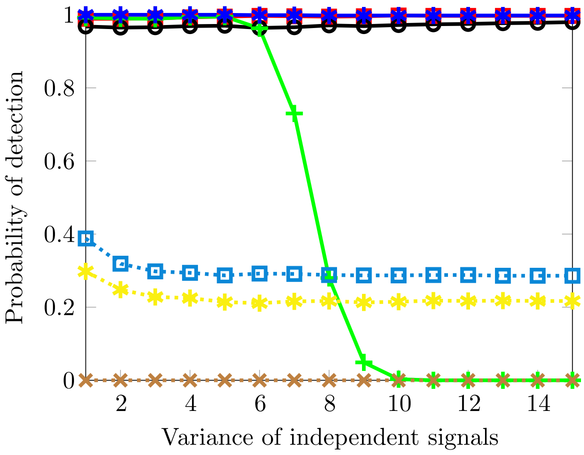


Fig. 9. Same setting as in Fig. 8, except that now there are $f_x = f_y = 4$ independent signals of varying variance. For the meaning of the colored markers, refer to the legend of Fig. 5.

happens when the correlated signals are *weaker* than some or all of the independent signals. First consider a scenario with 2 independent signals of varying variance and 7 correlated signals of variance 10. Fig. 8 shows the probability of detection as a function of the independent signals' variance. We see that the variance has only little effect on the performance of all techniques.

Now we increase the number of independent signals to 4, leaving all other settings unchanged. The most dramatic effect that can be observed in Fig. 9 is the failure of Detector 2 once the independent signals reach a variance close to the correlated signals' variance. This may be explained as follows. Detector 2 sets its threshold based on MDL, which generally does not overestimate the number of correlated signals, but may *underestimate* it if the sample size is not sufficiently large compared to the system dimension (i.e., the PCA rank). In the case shown in Fig. 9, there are 7 correlated signals and 4 independent signals. Once the independent signals become as strong as or stronger than the correlated signals, this leads to an optimum PCA rank of 11. As the number of samples $M=150$ is not significantly larger than 11, MDL starts to underestimate the model order. This affects Detector 2 more severely than Detector 3 because Detector 3 will always return a model as large as, but possibly larger than, Detector 2 (see the discussion in Section 4.2).

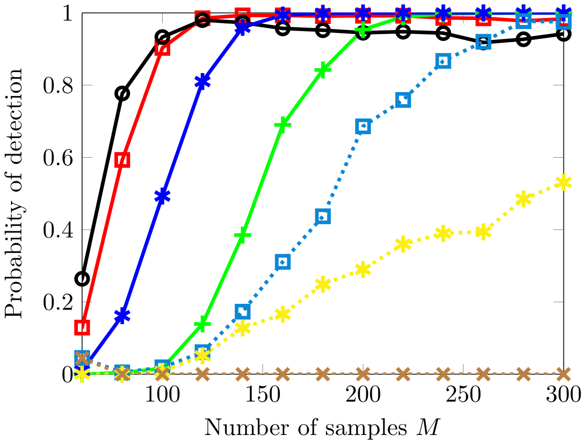


Fig. 10. Same setting as in Fig. 8, except that there are $d=5$ correlated signals with variance 8, and $f_x = f_y = 7$ independent signals, two of which have variance 12 and 5 of which have variance 3. Performance as a function of number of samples M. For the meaning of the colored markers, refer to the legend of Fig. 5.

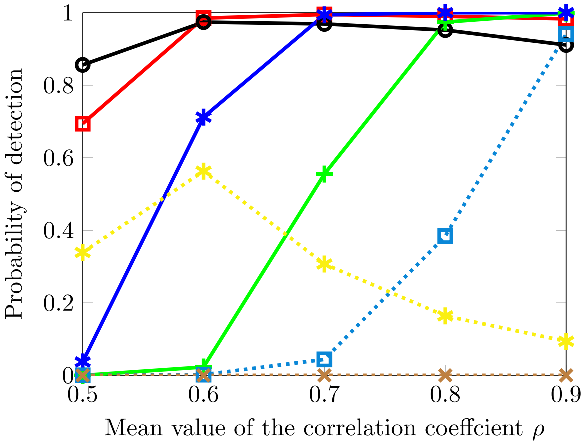


Fig. 11. Performance as a function of the mean correlation coefficient ρ. Settings: $m=n=100$, $M=180$, $d=5$ correlated signals with correlation coefficients drawn from a uniform distribution between $[\rho - 0.05, \rho + 0.05]$, $f_x = f_y = 2$ stronger independent signals, AR(1) noise with coefficient 0.65. For the meaning of the colored markers, refer to the legend of Fig. 5.

This explanation can be validated by investigating the effect of the number of samples in a scenario where there are strong independent signals. We now consider a case with $d=5$ correlated signals with variance 8, and $f_x = f_y = 7$ independent signals, two of which have variance 12 and 5 of which have variance 3. In Fig. 10, we look at the performance as a function of the number of samples *M*. It can be observed that, among our three detectors, Detector 2 needs the largest number of samples for satisfactory performance, followed by Detector 3. The lesson that can be learned here is that in the presence of strong independent signals, Detector 1 should be preferred if only a very small number of samples are available.

Let us now investigate the effect that the value of the correlation coefficients among the correlated signals have. Here we consider a scenario with $d=5$ correlated signals with variance 8, and $f_x = f_y = 2$ stronger independent signals of variance 10. In Fig. 11, we plot the performance as function of ρ. The correlation coefficients for the 5 correlated signals are drawn from a uniform distribution between $[\rho - 0.05, \rho + 0.05]$. As expected, stronger correlation leads to better performance. Since the independent signals are stronger than the correlated signals, Detector

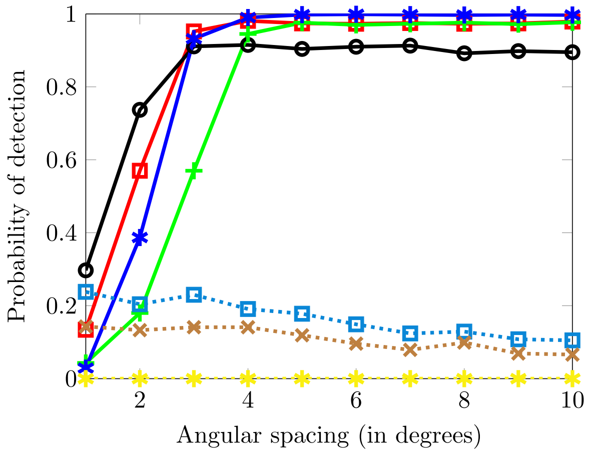


Fig. 12. Array processing toy example to illustrate the effect of ill-conditioned mixing matrices. For the meaning of the colored markers, refer to the legend of Fig. 5.

1 outperforms Detector 3, which in turn outperforms Detector 2. All of our detectors outperform the competition.

In our last example we examine an array processing toy application to see what happens if the mixing matrices $\mathbf{A}_x$ and $\mathbf{A}_y$ become ill-conditioned. We consider two spatially separated uniform linear arrays (ULAs) with 40 sensors (i.e., $m=n=40$) and inter-sensor spacing of $\lambda/2$, which take $M=60$ samples. There are 5 fixed point-sources in the far-field emitting narrow-band Gaussian signals at wavelength λ, which impinge upon ULA 1 at angles $[\theta_{x,1}, \theta_{x,2}, \dots, \theta_{x,5}] = [20°, 20° + \delta, \dots, 20° + 4\delta]$. Similarly, 6 such signals impinge upon ULA 2 at angles $[\theta_{y,1}, \theta_{y,2}, \dots, \theta_{y,6}] = [50°, 50° + \delta, \dots, 50° + 5\delta]$. Two of these signals are correlated between ULAs 1 and 2 (i.e., $d=2$, $f_x=3$, $f_y=4$) with correlation coefficients 0.8 and 0.7. The correlated signals each have variance 5 and the independent signals each have variance 1.5. The noise is colored and generated as in the setup for Fig. 6.

With these assumptions, the ith column of $\mathbf{A}_x$ is $[1, e^{j\frac{\pi}{2}\sin\theta_{x,i}}, \dots, e^{j\frac{\pi}{2}(n-1)\sin\theta_{x,i}}]^T$, $i = 1, \dots, 5$, and the ith column of $\mathbf{A}_y$ is $[1, e^{j\frac{\pi}{2}\sin\theta_{y,i}}, \dots, e^{j\frac{\pi}{2}(m-1)\sin\theta_{y,i}}]^T$, $i = 1, \dots, 6$. As the angular spacing δ decreases, the mixing matrices $\mathbf{A}_x$ and $\mathbf{A}_y$ become more ill-conditioned. Fig. 12 shows the performance of all detectors for angular spacing δ ranging from 1° to 10°. We can see that due to the presence of colored noise, all SEV+X methods fail irrespective of δ. Our detectors, on the other hand, are able to provide very good detection rates from $\delta=4°$ onward. Detectors 1 and 3 provide the best results for small δ.

6. Conclusions

PCA-CCA is a common approach to the analysis of correlation between two data sets when there is only small sample support. In the past, selecting the ranks of the PCA steps and identifying the number of correlated signals was often done by ad hoc rules or based on experience. In this paper, we have presented a systematic approach to the joint order selection of PCA ranks and number of correlated signals, based on a GLRT and information-theoretic criteria. Simulation results have shown that the techniques perform very well for extremely sample-poor scenarios in particular in the presence of colored noise. Of course, it is important to remember that there is no free lunch. While we do not need many samples compared to the dimensions of the data sets, the techniques do require the number of samples to be sufficiently greater than the sum of the numbers of correlated signals and stronger independent signals (i.e., variance larger than the correlated signals).

Acknowledgments

This research was supported by the German Research Foundation (DFG) under grant SCHR 1384/3-1, and the Alfried Krupp von Bohlen und Halbach foundation under its program “Return of German scientists from abroad.” The work of D. Ramírez has been partly supported by Ministerio de Economía of Spain under projects: COMPREHENSION (TEC2012-38883-C02-01), 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).

Appendix A. Effect of PCA on estimated canonical correlations

Lemma 1. *The estimated canonical correlation coefficients increase with increasing PCA ranks r_x and r_y: $\hat{k}_i(\tilde{r}_1, \tilde{r}_2) \ge \hat{k}_i(r_x, r_y)$, $i = 1, \dots, \min(r_x, r_y)$, for $1 \le r_x < \tilde{r}_1$ and $1 \le r_y < \tilde{r}_2$.*

Proof. Define the following matrices:

$$\mathbf{G} = \mathbf{V}_x^H(:,1:\tilde{r}_1)\mathbf{V}_y(:,1:\tilde{r}_2) = \begin{bmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \end{bmatrix},$$

$$\mathbf{G}_1 = \mathbf{V}_x^H(:,1:r_x)\mathbf{V}_y(:,1:\tilde{r}_2) = \begin{bmatrix} \mathbf{G}_{1,1} & \mathbf{G}_{1,2} \end{bmatrix},$$

$$\mathbf{G}_2 = \mathbf{V}_x^H(:,r_x+1:\tilde{r}_1)\mathbf{V}_y(:,1:\tilde{r}_2),$$

$$\mathbf{G}_{1,1} = \mathbf{V}_x^H(:,1:r_x)\mathbf{V}_y(:,1:r_y),$$

$$\mathbf{G}_{1,2} = \mathbf{V}_x^H(:,1:r_x)\mathbf{V}_y(:,r_y+1:\tilde{r}_2).$$

According to the Cauchy interlacing theorem, we have

$$\lambda_i\left(\mathbf{G}\mathbf{G}^H\right) = \lambda_i\left(\begin{bmatrix} \mathbf{G}_1\mathbf{G}_1^H & \mathbf{G}_1\mathbf{G}_2^H \\ \mathbf{G}_2\mathbf{G}_1^H & \mathbf{G}_2\mathbf{G}_2^H \end{bmatrix}\right) \ge \lambda_i\left(\mathbf{G}_1\mathbf{G}_1^H\right)$$

for $i = 1, \dots, r_x$, where $\lambda_i(\cdot)$ represents the ith largest eigenvalue. Furthermore, as a result of the Weyl inequality, we also have $\lambda_i\left(\mathbf{G}_1\mathbf{G}_1^H\right) = \lambda_i\left(\mathbf{G}_{1,1}\mathbf{G}_{1,1}^H + \mathbf{G}_{1,2}\mathbf{G}_{1,2}^H\right) \ge \lambda_i\left(\mathbf{G}_{1,1}\mathbf{G}_{1,1}^H\right)$. Together with the first inequality, this yields $\lambda_i\left(\mathbf{G}\mathbf{G}^H\right) \ge \lambda_i\left(\mathbf{G}_{1,1}\mathbf{G}_{1,1}^H\right)$. As $\mathbf{V}_x$ and $\mathbf{V}_y$ represent the matrices of right-singular vectors of $\mathbf{X}$ and $\mathbf{Y}$, respectively, it follows that the squared sample canonical correlation coefficient $\hat{k}_i^2(\tilde{r}_1, \tilde{r}_2) = \lambda_i\left(\mathbf{G}\mathbf{G}^H\right)$ is greater than or equal to the squared sample canonical correlation coefficient $\hat{k}_i^2(r_x, r_y) = \lambda_i\left(\mathbf{G}_{1,1}\mathbf{G}_{1,1}^H\right)$. □

References

[1] Z. Lin, C. Zhang, W. Wu, X. Gao, Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs, IEEE Trans. Biomed. Eng. 53 (12) (2006) 2610–2614.

[2] N.M. Correa, T. Adali, Y.-O. Li, V.D. Calhoun, Canonical correlation analysis for data fusion and group inferences: examining applications of medical imaging data, IEEE Signal Process. Mag. 27 (4) (2010) 39–50, http://dx.doi.org/10.1109/MSP.2010.936725.

[3] J.M. Wallace, C. Smith, C.S. Bretherton, Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies, J. Clim. 5 (6) (1992) 561–576.

[4] A. Shabbar, W. Skinner, Summer drought patterns in Canada and the relationship to global sea surface temperatures, J. Clim. 17 (2004) 2866–2880.

[5] H.Y. Ge, I.P. Kirsteins, X.L. Wang, Does canonical correlation analysis provide reliable information on data correlation in array processing?, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2009, pp. 2113–2116.

[6] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3–4) (1936) 321.

[7] L.L. Scharf, C.T. Mullis, Canonical coordinates and the geometry of inference, rate, and capacity, IEEE Trans. Signal Process. 48 (3) (2000) 824–831.

[8] A. Pezeshki, L.L. Scharf, M.R. Azimi-Sadjadi, M. Lundberg, Empirical canonical correlation analysis in subspaces, in: Proceedings of the Asilomar Conference Signals, Systems, and Computers, vol. 1, 2004, pp. 7–10.

[9] R.R. Nadakuditi, Fundamental finite-sample limit of canonical correlation analysis based detection of correlated high-dimensional signals in white noise, in: Proceedings of the IEEE Statistical Signal Processing Workshop (SSP), 2011, pp. 397–400.

[10] M. Wax, T. Kailath, Detection of signals by information theoretic criteria, IEEE Trans. Acoust., Speech Signal Process. 33 (2) (1985) 387–392.

[11] R.R. Nadakuditi, A. Edelman, Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples, IEEE Trans. Signal Process. 56 (7) (2008) 2625–2638.

[12] Z. Lu, A.M. Zoubir, Source enumeration in array processing using a two-step test, IEEE Trans. Signal Process. 63 (10) (2015) 2718–2727.

[13] M.S. Bartlett, The statistical significance of canonical correlations, Biometrika 32 (1) (1941) 29–37.

[14] D.N. Lawley, Tests of significance in canonical analysis, Biometrika 46 (1–2) (1959) 59–66.

[15] Y. Fujikoshi, L.G. Veitch, Estimation of dimensionality in canonical correlation analysis, Biometrika 66 (2) (1979) 345–351.

[16] Q.T. Zhang, K.M. Wong, Information theoretic criteria for the determination of the number of signals in spatially correlated noise, IEEE Trans. Signal Process. 41 (4) (1993) 1652–1663.

[17] W. Chen, J.P. Reilly, K.M. Wong, Detection of the number of signals in noise with banded covariance matrices, IEE Proc. - Radar, Sonar, Navig. 143 (5) (1996) 289–294.

[18] B.K. Gunderson, R.J. Muirhead, On estimating the dimensionality in canonical correlation analysis, J. Multivar. Anal. 62 (1997) 121–136.

[19] P. Stoica, K.M. Wong, Q. Wu, On a nonparametric detection method for array signal processing in correlated noise fields, IEEE Trans. Signal Process. 44 (4) (1996) 1030–1032.

[20] W.R. Zwick, W.F. Velicer, Comparison of five rules for determining the number of components to retain, Psychol. Bull. 99 (1986) 432–442.

[21] H. Hwang, K. Jung, Y. Takane, T.S. Woodward, A unified approach to multiple-set canonical correlation analysis and principal components analysis, Br. J. Math. Stat. Psychol. 66 (2013) 308–321.

[22] N.J. Roseveare, P. J. Schreier, Model-order selection for analyzing correlation between two data sets using CCA with PCA preprocessing, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2015.

[23] P. Stoica, Y. Selén, J. Li, On information criteria and the generalized likelihood ratio test of model order selection, IEEE Signal Proc. Lett. 11 (10) (2004) 794–797.

[24] Y. Song, P. Schreier, N. J. Roseveare, Determining the number of correlated signals between two data sets using PCA-CCA when sample support is extremely small, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2015.

[25] X. Chen, H. Liu, J.G. Carbonell, Structured sparse canonical correlation analysis, in: Proceedings of the 15th International Conference on Artificial Intelligence, Statistics, 2012, pp. 199–207.

[26] D.R. Hardoon, J. Shawe-Taylor, Sparse canonical correlation analysis, Mach. Learn. 83 (2011) 331–353.

[27] S.S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Stat. 9 (1938) 60–62.