A SPARSE CCA ALGORITHM WITH APPLICATION TO MODEL-ORDER SELECTION FOR SMALL SAMPLE SUPPORT

Christian Lameiro and Peter J. Schreier

Signal and System Theory Group, Universität Paderborn, Germany, http://sst.upb.de

ABSTRACT

We address the problem of determining the number of signals correlated between two high-dimensional data sets with small sample support. In this setting, conventional techniques based on canonical correlation analysis (CCA) cannot be directly applied since the canonical correlations are significantly overestimated when computed from few samples. To overcome this problem, a principal component analysis (PCA) preprocessing step is usually performed to reduce the dimension of the data. However, PCA reduces the dimension of each data set individually without taking the correlation between the data sets into account. In this paper we propose a sparse CCA (SCCA) algorithm as an alternative to the PCA-CCA approach. This algorithm is based on ℓ_1-norm penalization, which optimizes the weight of the ℓ_1-norm to keep a prescribed number of non-zero components. The number of correlated components is then selected based on an information-theoretic criterion.

***Index Terms*—** model-order selection, small sample support, sparse CCA, ℓ_1-norm.

1. INTRODUCTION

In many applications, we are interested in assessing the correlation between two multivariate data sets. Let n and m be the dimension of the data sets $\mathbf{X}$ and $\mathbf{Y}$, respectively, and M the number of observations, so that $\mathbf{X} \in \mathbb{R}^{n\times M}$ and $\mathbf{Y} \in \mathbb{R}^{m\times M}$. The typical way to measure the correlation between $\mathbf{X}$ and $\mathbf{Y}$ is via canonical correlation analysis (CCA) [1]. CCA obtains pairs of canonical variables by means of linear projections of the original data, $\mathbf{S}^T\mathbf{X}$ and $\mathbf{T}^T\mathbf{Y}$, in such a way that the correlation between these pairs of canonical variables is maximized while they are uncorrelated with each other. A particularly important question is how many of the signal components between $\mathbf{X}$ and $\mathbf{Y}$ are correlated. This is a model-order selection problem. When CCA is used to perform model-order selection, different models with varying number of free parameters, i.e., correlated components, are usually compared by means of hypothesis testing (HT) [2, 3] or information criteria (IC) [3, 4].

In many practical situations, the data sets are high-dimensional, and only relatively few observations are available. When $M < n+m$, there are at least $n+m-M$ canonical correlations equal to one independently of the underlying model, which means that CCA cannot be directly applied to determine the number of correlated components [5]. A typical approach to overcome this issue is to perform a principal component analysis (PCA) preprocessing step before CCA [6]. By keeping only r_x and r_y components from $\mathbf{X}$ and $\mathbf{Y}$, respectively, where r_x and r_y have to be chosen wisely [6], the problem of defective canonical correlations can be alleviated [7]. However, because the PCA components are determined without regard to the correlation structure between the data sets, the extracted components might not include all correlated components.

In this paper we propose an alternative approach based on sparse CCA (SCCA). SCCA constrains the ℓ_0-norm of the projectors, i.e., they are constrained to have only few components different from zero. This can be regarded as dimensionality reduction by means of variable selection, therefore alleviating the effect of small sample support. The difference compared to the PCA preprocessing is that SCCA selects sets of variables that maximize correlation, i.e., the dimensionality reduction is performed jointly with the extraction of the canonical variables.

The main problem of SCCA is that the global optimal solution cannot be efficiently found as in regular CCA. This is due to the additional constraint on the ℓ_0-norm of the projectors, which makes the problem intractable. A conventional workaround uses the ℓ_1-norm instead [8–10] (or a variation thereof [11]), which is a convex function and hence easy to handle. However, a constraint on the ℓ_1-norm of the projectors does not permit a direct control of the sparsity level, which makes it difficult to perform the model-order selection. Other works propose suboptimal approaches directly based on the ℓ_0-norm [12, 13]. These methods use deflation to extract subsequent canonical correlations. However, deflation does not guarantee orthogonality between the canonical variables, which in turn may lead to an overestimation of the canonical correlations. This is especially critical in the model-order selection problem for small sample support, since the model-order might then easily be overestimated.

In order to overcome the drawbacks of the aforementioned methods, here we propose an alternative SCCA algorithm based on the ℓ_1-norm penalization, whose weight is dynamically optimized to keep a given number of non-zero components. The difference of this SCCA algorithm compared to other SCCA techniques is that it allows us to optimize exactly r_x nonzero components for each column of $\mathbf{S}$ and r_y nonzero components for each column of $\mathbf{T}$, while satsifying the orthogonality of the canonical variables. This leads to models with different numbers of degrees of freedom, which can then be compared using either HT or IC. In this paper, we do the latter. To the best of our knowledge, SCCA has not been applied yet to model-order selection for small sample support.

2. PROBLEM FORMULATION

We consider the two-channel model

$$\mathbf{x} = \mathbf{A}_x\mathbf{s}_x + \mathbf{n}_x \ , \tag{1}$$

$$\mathbf{y} = \mathbf{A}_y\mathbf{s}_y + \mathbf{n}_y \ , \tag{2}$$

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

Matlab code is available to download from: https://github.com/SSTGroup/Correlation-Analysis-in-High-Dimensional-Data.git

where $\mathbf{x} \in \mathbb{R}^{n\times 1}$ and $\mathbf{y} \in \mathbb{R}^{m\times 1}$ are the observations of the first and second channel, respectively; $\mathbf{A}_x \in \mathbb{R}^{n\times(d+f_x)}$ and $\mathbf{A}_y \in \mathbb{R}^{m\times(d+f_y)}$ are deterministic but unknown mixing matrices; $\mathbf{s}_x \in \mathbb{R}^{(d+f_x)\times 1}$ and $\mathbf{s}_y \in \mathbb{R}^{(d+f_y)\times 1}$ are the sources in each channel, and $\mathbf{n}_x \in \mathbb{R}^{n\times 1}$ and $\mathbf{n}_y \in \mathbb{R}^{m\times 1}$ are the independent additive noises, whose covariance matrices are $\mathbf{\Psi}_x$ and $\mathbf{\Psi}_y$, respectively. We assume that the signals $\mathbf{s}_x$ and $\mathbf{s}_y$ have d correlated components with variances $\sigma_{x,j}^2$ and $\sigma_{y,j}^2$ ($j = 1,\ldots,d$), respectively, i.e., its cross-covariance matrix can be expressed as

$$\mathbf{R}_{s_x s_y} = \begin{bmatrix} \mathbf{diag}(\rho_1\sigma_{x,1}\sigma_{y,1},\ldots,\rho_d\sigma_{x,d}\sigma_{y,d}) & \mathbf{0}_{d\times f_y} \\ \mathbf{0}_{f_x\times d} & \mathbf{0}_{f_x\times f_y} \end{bmatrix} . \quad (3)$$

The number of uncorrelated components between $\mathbf{s}_x$ and $\mathbf{s}_y$ is then given by f_x and f_y, respectively. For this model, we address the following problem.

Problem: *Given M i.i.d. observations of the model described by* (1) *and* (2)*, with possibly $M < m + n$, determine the number d of correlated components.*

3. CANONICAL CORRELATION ANALYSIS

CCA is the typical approach to evaluate correlation between two data sets. The ith pair of canonical variables can be found as the solution of the following optimization problem:

$$\begin{aligned} \mathcal{P}_i: \quad & \underset{\mathbf{s}_i,\mathbf{t}_i}{\text{maximize}} && \mathbf{s}_i^T\mathbf{R}_{xy}\mathbf{t}_i \ , \\ & \text{subject to} && \mathbf{s}_i^T\mathbf{R}_{xx}\mathbf{s}_i = 1 \ , \\ & && \mathbf{t}_i^T\mathbf{R}_{yy}\mathbf{t}_i = 1 \ , \\ & && \mathbf{s}_i^T\mathbf{R}_{xx}\mathbf{s}_j = 0 \ , j = 1,\ldots,i-1 \ , \\ & && \mathbf{t}_i^T\mathbf{R}_{yy}\mathbf{t}_j = 0 \ , j = 1,\ldots,i-1 \ , \end{aligned}$$

where $\mathbf{R}_{xy}$ is the cross-covariance matrix between $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{R}_{xx}$ and $\mathbf{R}_{yy}$ are the covariance matrices of $\mathbf{x}$ and $\mathbf{y}$, respectively, and $\mathbf{s}_i$ and $\mathbf{t}_i$ are the ith columns of $\mathbf{S}$ and $\mathbf{T}$, respectively. The covariance and cross-covariance matrices are usually unknown and must be estimated using the available observations. Therefore, $\mathbf{R}_{xy}$, $\mathbf{R}_{xx}$ and $\mathbf{R}_{yy}$ are typically replaced by their sample counterparts, namely, $\hat{\mathbf{R}}_{xy} = \frac{1}{M}\mathbf{X}\mathbf{Y}^T$, $\hat{\mathbf{R}}_{xx} = \frac{1}{M}\mathbf{X}\mathbf{X}^T$ and $\hat{\mathbf{R}}_{yy} = \frac{1}{M}\mathbf{Y}\mathbf{Y}^T$. In the case of small sample support, the canonical correlations are significantly overestimated [6] and cannot be used to infer the number d of correlated components. To this end, we consider SCCA as an alternative approach, which is described in the next section.

4. SPARSE CANONICAL CORRELATION ANALYSIS

The SCCA problem, based on the ℓ_0-norm, can generally be stated as the optimization problem

$$\begin{aligned} \mathcal{P}_i^{\ell_0}: \quad & \underset{\mathbf{s}_i,\mathbf{t}_i}{\text{maximize}} && \mathbf{s}_i^T\hat{\mathbf{R}}_{xy}\mathbf{t}_i \ , \\ & \text{subject to} && \mathbf{s}_i^T\hat{\mathbf{R}}_{xx}\mathbf{s}_i = 1 \ , \\ & && \mathbf{t}_i^T\hat{\mathbf{R}}_{yy}\mathbf{t}_i = 1 \ , \\ & && \mathbf{s}_i^T\hat{\mathbf{R}}_{xx}\mathbf{s}_j = 0 \ , j = 1,\ldots,i-1 \ , \\ & && \mathbf{t}_i^T\hat{\mathbf{R}}_{yy}\mathbf{t}_j = 0 \ , j = 1,\ldots,i-1 \ , \\ & && ||\mathbf{s}_i||_0 \le r_x \ , \\ & && ||\mathbf{t}_i||_0 \le r_y \ , \end{aligned}$$

where $||\cdot||_0$ is the ℓ_0-norm, i.e., the number of non-zero elements. Hereafter we will assume $r_x = r_y = r$. The ℓ_0-norm is non-convex and difficult to handle, and it is usually replaced by the convex ℓ_1-norm, which is defined as the sum of the absolute value of the elements. However, the ℓ_1-norm does not provide a direct measure of sparsity, and it is therefore difficult to control the number of degrees of freedom to perform the model-order selection. In the following, we propose an SCCA algorithm considering the ℓ_1-norm with an additional step that optimizes the ℓ_1-norm constraint such that a given number of non-zero components is retained. To this end, let us first consider the initial problem

$$\begin{aligned} \mathcal{P}_i^{\ell_1}: \quad & \underset{\mathbf{s}_i,\mathbf{t}_i}{\text{maximize}} && \mathbf{s}_i^T\hat{\mathbf{R}}_{xy}\mathbf{t}_i - \lambda_x||\mathbf{s}_i||_1 - \lambda_y||\mathbf{t}_i||_1 \ , \\ & \text{subject to} && \mathbf{s}_i^T\hat{\mathbf{R}}_{xx}\mathbf{s}_i = 1 \ , \\ & && \mathbf{t}_i^T\hat{\mathbf{R}}_{yy}\mathbf{t}_i = 1 \ , \\ & && \mathbf{s}_i^T\hat{\mathbf{R}}_{xx}\mathbf{s}_j = 0 \ , j = 1,\ldots,i-1 \ , \\ & && \mathbf{t}_i^T\hat{\mathbf{R}}_{yy}\mathbf{t}_j = 0 \ , j = 1,\ldots,i-1 \ , \end{aligned}$$

where λ_x and λ_y are the penalization terms for the ℓ_1-norm of $\mathbf{s}_i$ and $\mathbf{t}_i$, respectively. Since the above problem is still non-convex, we perform an alternating optimization procedure to subsequently optimize $\mathbf{s}_i$ (with fixed $\mathbf{t}_i$) and $\mathbf{t}_i$ (with fixed $\mathbf{s}_i$). For a given $\mathbf{t}_i = \mathbf{t}_i'$, the convex optimization problem for $\mathbf{s}_i$ is given by

$$\begin{aligned} \mathcal{P}_i^{\ell_1}(\mathbf{t}_i'): \quad & \underset{\mathbf{s}_i}{\text{maximize}} && \mathbf{s}_i^T\hat{\mathbf{R}}_{xy}\mathbf{t}_i' - \lambda_x||\mathbf{s}_i||_1 \ , \\ & \text{subject to} && \mathbf{s}_i^T\hat{\mathbf{R}}_{xx}\mathbf{s}_i \le 1 \ , \\ & && \mathbf{s}_i^T\hat{\mathbf{R}}_{xx}\mathbf{s}_j = 0 \ , j = 1,\ldots,i-1 \ . \end{aligned}$$

The optimal solution of this problem is characterized in the following lemma.

Lemma 1. *Let the columns of $\mathbf{N}_s$ and $\mathbf{N}_{xx}$ span the null space of $\hat{\mathbf{R}}_{xx}[\mathbf{s}_1,\ldots,\mathbf{s}_{i-1}]$ and $\mathbf{N}_s^T\hat{\mathbf{R}}_{xx}\mathbf{N}_s$, respectively. The optimal solution of $\mathcal{P}_i^{\ell_1}(\mathbf{t}_i')$ is then given by*

$$\begin{aligned} \mathbf{s}_i^\star =& \frac{\tilde{\mathbf{s}}_i}{\sqrt{\tilde{\mathbf{s}}_i^T\hat{\mathbf{R}}_{xx}\tilde{\mathbf{s}}_i}} \ , \\ \tilde{\mathbf{s}}_i =& \mathbf{N}_s\left(\mathbf{N}_s^T\hat{\mathbf{R}}_{xx}\mathbf{N}_s\right)^\dagger\left(\mathbf{N}_s^T\hat{\mathbf{R}}_{xy}\mathbf{t}_i' + \mathbf{N}_s^T\boldsymbol{\mu}_x\right) \\ & + \mathbf{N}_s\mathbf{N}_{xx}\boldsymbol{\phi}_x \ , \end{aligned} \quad (4)$$

where $(\cdot)^\dagger$ is the pseudo-inverse and $\boldsymbol{\mu}_x$ satisfies $|\boldsymbol{\mu}_x| \le \lambda_x\mathbf{1}$, with $|\cdot|$ and $\le$ applied element-wise, and $\mathbf{N}_{xx}^T\boldsymbol{\mu}_x = \mathbf{0}$. Furthermore, let $\mathbf{q} = [\mathbf{q}^+;\mathbf{q}^-]$ be the vector containing the indexes of the non-zero elements of $\mathbf{s}_i^\star$, where $\mathbf{q}^+$ and $\mathbf{q}^-$ are the indexes of the positive and negative elements of $\mathbf{s}_i^\star$, respectively. Then $(\boldsymbol{\mu}_x)_j = \lambda_x$, where $(\cdot)_j$ denotes the jth element of its vector argument, for all elements of $\boldsymbol{\mu}_x$ with indexes in $\mathbf{q}^-$, $(\boldsymbol{\mu}_x)_j = -\lambda_x$ for all elements of $\boldsymbol{\mu}_x$ with indexes in $\mathbf{q}^+$, and $|(\boldsymbol{\mu}_x)_j| < \lambda_x$ otherwise.

Proof. $\mathcal{P}_i^{\ell_1}(\mathbf{t}_i')$ can equivalently be written as

$$\begin{aligned} \tilde{\mathcal{P}}_i^{\ell_1}(\mathbf{t}_i'): \quad & \underset{\mathbf{s}_i,\mathbf{u}}{\text{maximize}} && \mathbf{s}_i^T\hat{\mathbf{R}}_{xy}\mathbf{t}_i' - \lambda_x\mathbf{1}^T\mathbf{u} \ , \\ & \text{subject to} && \mathbf{s}_i^T\hat{\mathbf{R}}_{xx}\mathbf{s}_i \le 1 \ , \\ & && \mathbf{s}_i^T\hat{\mathbf{R}}_{xx}\mathbf{s}_j = 0 \ , j = 1,\ldots,i-1 \ , \\ & && \mathbf{u} \ge \mathbf{s}_i \ , \ \mathbf{u} \ge -\mathbf{s}_i \ . \end{aligned}$$

Choose the desired number r of non-zero elements, initialize the empty matrices $\mathbf{T}$ and $\mathbf{S}$.
for $i = 1$ **to** r **do**
  Choose an starting point $\mathbf{t}_i = \mathbf{t}_i'$ satisfying the corresponding constraints in $\mathcal{P}_i^{\ell_1}$.
  repeat
    1. Initialize the vector of indexes $\mathbf{q}$ as an empty vector.
    2. Solve $\tilde{\mathcal{P}}_i^{\lambda_x}(\mathbf{t}_i')$ to obtain the indexes of the non-zero elements and denote them as $\tilde{\mathbf{q}}$.
      if the length of $\tilde{\mathbf{q}}$ is smaller than $r+1$ **then**
        Take $\mathbf{q} = \tilde{\mathbf{q}}$ and repeat step 2
      else
        Continue to step 3
      end if
    3. Obtain $\mathbf{s}_i^\star$ using (4) and repeat steps 1 and 2 solving the analog problems for $\mathbf{t}_i$ while keeping $\mathbf{s}_i = \mathbf{s}_i^\star$ fixed.
  until convergence criterion is met.
  Update $\mathbf{S} = [\mathbf{S}\ ,\ \mathbf{s}_i^\star]$ and $\mathbf{T} = [\mathbf{T}\ ,\ \mathbf{t}_i^\star]$.
end for

Algorithm 1: Proposed SCCA algorithm.

Since this problem is convex and satisfies Slater's condition, its optimal solution fulfills the Karush-Kuhn-Tucker (KKT) conditions [14], which yield (4). Through the KKT conditions we also obtain that $\boldsymbol{\mu}_x = \boldsymbol{\mu}_x^- - \boldsymbol{\mu}_x^+$, where $\boldsymbol{\mu}_x^-$ and $\boldsymbol{\mu}_x^+$ are the Lagrange multipliers of $\mathbf{u} \geq -\mathbf{s}_i$ and $\mathbf{u} \geq \mathbf{s}_i$, respectively, and $\boldsymbol{\mu}_x^- + \boldsymbol{\mu}_x^+ = \lambda_x \mathbf{1}$. Furthermore, let q_1 be an index such that $(\mathbf{s}_i^\star)_{q_1} > 0$. Then it is clear that $\mathbf{u} \geq -\mathbf{s}_i$ is not active, hence $(\boldsymbol{\mu}_x^-)_{q_1} = 0$, $(\boldsymbol{\mu}_x^+)_{q_1} = \lambda_x$ and consequently $(\boldsymbol{\mu}_x)_{q_1} = -\lambda_x$. Similarly, let q_2 be an index such that $(\mathbf{s}_i^\star)_{q_2} < 0$. Then $\mathbf{u} \geq \mathbf{s}_i$ is not active, yielding $(\boldsymbol{\mu}_x^+)_{q_2} = 0$, $(\boldsymbol{\mu}_x^-)_{q_2} = \lambda_x$ and $(\boldsymbol{\mu}_x)_{q_2} = \lambda_x$. Constraints $\mathbf{u} \geq \mathbf{s}_i$ and $\mathbf{u} \geq -\mathbf{s}_i$ are simultaneously active for the index q_3 satisfying $(\mathbf{s}_i^\star)_{q_3} = 0$, which yields $0 < (\boldsymbol{\mu}_x^+)_{q_3} < \lambda_x$, $(\boldsymbol{\mu}_x^-)_{q_3} = \lambda_x - (\boldsymbol{\mu}_x^+)_{q_3}$ and thus $|(\boldsymbol{\mu}_x)_{q_3}| < \lambda_x$, which concludes the proof. □

Lemma 1 permits determining the optimal λ_x by subsequently identifying the elements of $\tilde{\mathbf{s}}_i$ that must be non-zero. It is clear that, by increasing λ_x, we decrease the number of non-zero elements. Furthermore, if λ_x is too high, the optimal solution of $\mathcal{P}_i^{\ell_1}(\mathbf{t}_i')$ will be an all-zero vector. The minimum value of λ_x such that $\tilde{\mathbf{s}}_i = \mathbf{0}$ can be obtained as the solution of

$$
\begin{aligned}
&\mathcal{P}_i^{\lambda_x}(\mathbf{t}_i'): \\
&\underset{\lambda_x, \boldsymbol{\phi}_x, \boldsymbol{\mu}_x}{\text{minimize}} \quad \lambda_x\ , \\
&\text{subject to} \quad \mathbf{N}_s \left(\mathbf{N}_s^T \hat{\mathbf{R}}_{xx} \mathbf{N}_s\right)^\dagger \left(\mathbf{N}_s^T \hat{\mathbf{R}}_{xy} \mathbf{t}_i' + \mathbf{N}_s^T \boldsymbol{\mu}_x\right) \\
&\qquad + \mathbf{N}_s \mathbf{N}_{xx} \boldsymbol{\phi}_x = \mathbf{0}\ , \\
&\qquad |\boldsymbol{\mu}_x| \leq \lambda_x \mathbf{1}\ , \\
&\qquad \mathbf{N}_{xx}^T \boldsymbol{\mu}_x = \mathbf{0}\ .
\end{aligned}
$$

Let $\lambda_x^\star$ be the optimal solution of $\mathcal{P}_i^{\lambda_x}(\mathbf{t}_i')$. Then $\lambda_x = \lambda_x^\star - \epsilon$, with $\epsilon > 0$, will yield a solution with at least one non-zero component. These non-zero components can be identified using Lemma 1.

Table 1. Probability of detection for varying d. The correlation coefficient of the second and third components is 0.85 and 0.75, respectively.

	Proposed SCCA	Rank-1 SCCA	PCA-CCA
$d = 0$	0.92	0.39	0.74
$d = 1$	0.91	0.41	0.76
$d = 2$	0.65	0.63	0.71
$d = 3$	0.30	0.60	0.66

Specifically, the indexes $\mathbf{q}$ of $\tilde{\mathbf{s}}_i$ corresponding to the non-zero components are those satisfying $|(\boldsymbol{\mu}_x)_{\mathbf{q}}| = \lambda_x$. This permits obtaining the sparsest non-zero projection $\tilde{\mathbf{s}}_i$. If we want a higher number of non-zero elements, we have to modify problem $\mathcal{P}_i^{\ell_1}(\mathbf{t}_i')$ to remove the indexes $\mathbf{q}$ of the previously obtained non-zero elements and enforce $|(\boldsymbol{\mu}_x)_{\mathbf{q}}| = \lambda_x$. In other words, $\mathcal{P}_i^{\lambda_x}(\mathbf{t}_i')$ is modified as

$$
\begin{aligned}
&\tilde{\mathcal{P}}_i^{\lambda_x}(\mathbf{t}_i'): \\
&\underset{\lambda_x, \boldsymbol{\phi}_x, \boldsymbol{\mu}_x}{\text{minimize}} \quad \lambda_x\ , \\
&\text{subject to} \quad \tilde{\mathbf{N}}_s \left(\mathbf{N}_s^T \hat{\mathbf{R}}_{xx} \mathbf{N}_s\right)^\dagger \left(\mathbf{N}_s^T \hat{\mathbf{R}}_{xy} \mathbf{t}_i' + \mathbf{N}_s^T \boldsymbol{\mu}_x\right) \\
&\qquad + \tilde{\mathbf{N}}_s \mathbf{N}_{xx} \boldsymbol{\phi}_x = \mathbf{0}\ , \\
&\qquad |\boldsymbol{\mu}_x| \leq \lambda_x\ , \\
&\qquad (\boldsymbol{\mu}_x)_{\mathbf{q}^+} = -\lambda_x\ ,\ (\boldsymbol{\mu}_x)_{\mathbf{q}^-} = \lambda_x\ , \\
&\qquad \mathbf{N}_{xx}^T \boldsymbol{\mu}_x = \mathbf{0}\ ,
\end{aligned}
$$

where $\tilde{\mathbf{N}}_s$ is obtained by removing the rows of $\mathbf{N}_s$ with indexes in $\mathbf{q}$. The above problem can be shown to admit a closed-form solution, which significantly reduces the computational complexity. Finally, the proposed SCCA algorithm is summarized in Algorithm 1.

5. MODEL-ORDER SELECTION

We use IC to perform model-order selection based on the estimated canonical correlations obtained with Algorithm 1. Therefore, the number d of correlated components is estimated as

$$
\hat{d} = \underset{d \in [0,\ldots,r]}{\arg\max} \max_{r \in [1,\ldots,r_{\max}]} \mathrm{IC}(d, r)\ , \tag{5}
$$

where $r_{\max}$ is the maximum number of estimated canonical correlations, which needs to be sufficiently smaller than the number of samples M. In practice, $r_{\max} < M/3$ seems to be working well [6]. In the above expression, $\mathrm{IC}(d, r)$ is the IC score. Assuming that the sources and noise are Gaussian distributed, the IC score is given by

$$
\mathrm{IC}(d, r) = -\frac{M}{2} \log \prod_{i=1}^{d} \left(1 - \hat{k}_i^2(r)\right) - p(d, r)\ . \tag{6}
$$

The first term of the IC score corresponds to the log-likelihood of the observations, where $\hat{k}_i(r)$ is the ith estimated canonical correlation, and the second term penalizes the complexity of the model. Following the minimum description length (MDL) approach [15], we use the penalization term

$$
p(d, r) = \frac{1}{2} n_f(d, r) \log M\ , \tag{7}
$$

Table 2. Average of the estimated canonical correlations for $d = 3$ by the proposed SCCA.

	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	true
$\hat{k}_1$	0.77	0.88	0.92	0.95	0.96	0.97	0.95
$\hat{k}_2$	0	0.75	0.83	0.87	0.90	0.93	0.85
$\hat{k}_3$	0	0	0.65	0.75	0.75	0.86	0.75
$\hat{k}_4$	0	0	0	0.50	0.55	0.70	0
$\hat{k}_5$	0	0	0	0	0.39	0.56	0
$\hat{k}_6$	0	0	0	0	0	0.40	0

Table 3. Average of the estimated canonical correlations for $d = 3$ by the SCCA algorithm from [13].

	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	true
$\hat{k}_1$	0.81	0.92	0.95	0.96	0.96	0.96	0.95
$\hat{k}_2$	0	0.79	0.90	0.92	0.93	0.93	0.85
$\hat{k}_3$	0	0	0.73	0.84	0.88	0.86	0.75
$\hat{k}_4$	0	0	0	0.62	0.75	0.80	0
$\hat{k}_5$	0	0	0	0	0.55	0.70	0
$\hat{k}_6$	0	0	0	0	0	0.47	0

Table 4. Probability of detection for $d = 3$ and varying M.

	Proposed SCCA	SCCA [13]	PCA-CCA [6]
$M = 20$	0.30	0.60	0.66
$M = 30$	0.43	0.72	0.73
$M = 40$	0.71	0.84	0.92
$M = 50$	0.84	0.90	0.90

Table 5. Probability of detection for varying noise variance σ^2.

	Proposed SCCA	SCCA [13]	PCA-CCA [6]
$\sigma^2 = 1$	0.92	0.41	0.76
$\sigma^2 = 5$	0.84	0.09	0.78
$\sigma^2 = 10$	0.76	0	0.73
$\sigma^2 = 20$	0.62	0	0.72

where $n_f(d, r)$ is the number of free parameters of the model. In our problem, this corresponds to the number of free adjusted parameters to model the covariance matrices $\mathbf{R}_{xy}$, $\mathbf{R}_{xx}$ and $\mathbf{R}_{yy}$. Following the lines of [4, 15], the number of free adjusted parameters can be obtained as follows. As we project the observations onto an r-dimensional subspace, each auto-covariance matrix has a total of $(r + 1)d$ parameters, but the normality and mutual orthogonality of the eigenvectors impose $\frac{1}{2}d(d + 1)$ constraints. Similarly, the cross-covariance matrix has $(2r+1)d$ parameters with $d(d+1)$ additional constraints, thus the total number of free adjusted parameters is

$$n_f(d, r) = (4r - 2d + 1)d \, . \tag{8}$$

6. NUMERICAL EXAMPLES

We generate our observations according to (1) and (2), where each entry of the mixing matrices $\mathbf{A}_x$ and $\mathbf{A}_y$ is drawn independently from a Gaussian distribution with zero mean and unit variance, and the noise covariance matrices are $\mathbf{\Psi}_x = \mathbf{\Psi}_y = \sigma^2\mathbf{I}$. Unless otherwise stated, we consider the scenario: $n = m = 50$, $M = 20$, $d = 1$ with variance 10 and correlation coefficient 0.95, $f_x = f_y = 4$ with variance 3, and $\sigma^2 = 1$. We compare the proposed SCCA approach with the following existing techniques: the joint PCA-CCA technique for small sample support [6] (we use Detector 2 from [6]) and the SCCA algorithm proposed in [13], which directly constrains the ℓ_0-norm of the projections and can thus also be used with the proposed model-order selection. Table 1 shows the probability of correct detection for $d = \{0, 1, 2, 3\}$. The proposed approach significantly outperforms the benchmarks for $d = 0$ and $d = 1$, but the detection rate substantially decreases for $d = 2$ and $d = 3$. On the other hand, the existing SCCA approach exhibits an abnormal behavior, as the probability of detection increases with d. This is because it provides estimates of the canonical correlations that are generally greater than those obtained by the proposed SCCA algorithm (see Tables 2 and 3). Therefore, the SCCA algorithm in [13] is more likely to overfit the data. Nevertheless, Table 4 shows that the detection probability of the proposed SCCA algorithm for $d = 3$ significantly increases with the number of observations and approaches that of the benchmark techniques. In Table 5 we provide the results for increasing noise power. We observe that the existing SCCA is much more sensitive to the noise level than the proposed SCCA, whereas the performance of PCA-CCA is almost unaffected by the noise. This is due to the PCA step, which eliminates most of the noise. However, SCCA provides higher estimates of the canonical correlations increasing the probability of overfitting. The opposite behavior is observed when we increase the variance of the uncorrelated components, as shown in Table 6. The estimates of the canonical correlations by SCCA decrease, increasing the probability of underfitting.

Table 6. Probability of detection for varying variance of uncorrelated components $\tilde{\sigma}^2$.

	Proposed SCCA	SCCA [13]	PCA-CCA [6]
$\tilde{\sigma}^2 = 3$	0.92	0.41	0.76
$\tilde{\sigma}^2 = 20$	0.45	0.40	0.72
$\tilde{\sigma}^2 = 100$	0.28	0.38	0.74

7. CONCLUSIONS

In this paper we have applied SCCA for model-order selection with small sample support. To this end, we have proposed a new algorithm based on ℓ_1-norm penalization, which dynamically adjusts the penalization weight to keep a given number of non-zero components. We have shown that SCCA can be applied to determine the number of signals correlated between two data sets when there are only few samples available. Although we have focused on model-order selection, the proposed SCCA algorithm can be used for other applications as an alternative to existing SCCA methods, e.g., to ease the interpretation of the canonical variables or when the mixing matrices are known to be sparse.

8. REFERENCES

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, Dec. 1936.

[2] M. S. Bartlett, "The statistical significance of canonical correlations," *Biometrika*, vol. 32, no. 1, pp. 29–37, Jan. 1941.

[3] Y. Fujikoshi and L. G. Veitch, "Estimation of dimensionality in canonical correlation analysis," *Biometrika*, vol. 66, no. 2, pp. 345–351, Aug. 1979.

[4] Q. T. Zhang and K. M. Wong, "Information theoretic criteria for the determination of the number of signals in spatially correlated noise," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1652–1663, Apr. 1993.

[5] A. Pezeshki, L. L. Scharf, M. R. Azimi-Sadjadi, and M. Lundberg, "Empirical canonical correlation analysis in subspaces," in *Conference Records of the Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, 2004, pp. 994–997.

[6] 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, Nov. 2016.

[7] 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)*, Nice, France, June 2011, pp. 397–400.

[8] D. R. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis," *Machine Learning*, vol. 83, no. 3, pp. 331–353, Nov. 2011.

[9] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, July 2009.

[10] I. Wilms and C. Croux, "Sparse canonical correlation analysis from a predictive point of view," *Biometrical Journal*, vol. 57, no. 5, pp. 834–851, Sept. 2015.

[11] D. Torres, D. Turnbull, L. Barrington, and G. Lanckriet, "Identifying words that are musically meaningful," in *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, Sept. 2007, pp. 405–410.

[12] A. Wiesel, M. Klieger, and A. O. Hero III, "A greedy approach to sparse canonical correlation analysis," *ArXiv e-prints*, Jan. 2008.

[13] A. Aïssa-El-Bey and A. K. Seghouane, "Sparse canonical correlation analysis based on rank-1 matrix approximation and its application for fMRI signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016, pp. 4678–4682.

[14] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[15] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, Apr. 1985.