MODEL-ORDER SELECTION FOR ANALYZING CORRELATION BETWEEN TWO DATA SETS USING CCA WITH PCA PREPROCESSING

Nicholas J. Roseveare and Peter J. Schreier

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

ABSTRACT

This paper is concerned with determining the number of correlated signals between two data sets using canonical correlation analysis (CCA) when a principal component analysis (PCA) preprocessing step is performed for initial rank reduction. In signal processing applications, it is commonplace in scenarios with large dimensions, insufficient samples, or knowledge of low-rank underlying signals to extract the principal components of the data before correlation is analyzed. While there exist information-theoretic criteria to either determine the number of signals in a *single* data set or the number of correlated signals between two data sets, there has yet to be a treatment of the *joint* order estimation of the number of dimensions which should be retained through the PCA preprocessing *and* the number of correlated signals. We present the likelihood and information criteria for this scenario, along with some verifying simulations.

Index Terms— Canonical correlation analysis, dimension reduction, information criteria, model-order estimation, principal component analysis

1. INTRODUCTION

Measuring and analyzing multivariate association between two sets of data is a common objective, with numerous applications in many areas of the natural and social sciences and engineering. The most widely used technique is canonical correlation analysis (CCA) [1]. In CCA, the observed data $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$ are transformed into p-dimensional internal (latent) representations $\mathbf{a} = \mathbf{S}\mathbf{x}$ and $\mathbf{b} = \mathbf{T}\mathbf{y}$, where $p \leq \min(m,n)$, using linear transformations described by the matrices $\mathbf{S} \in \mathbb{R}^{p\times n}$ and $\mathbf{T} \in \mathbb{R}^{p\times m}$. The key idea is to determine $\mathbf{S}$ and $\mathbf{T}$ so 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 $\mathbf{s}_1 \in \mathbb{R}^n$ and $\mathbf{t}_1 \in \mathbb{R}^m$ 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*. 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 $\mathbf{S} = [\mathbf{s}_1, ..., \mathbf{s}_p]^T$, $\mathbf{T} = [\mathbf{t}_1, ..., \mathbf{t}_p]^T$. CCA can be performed via the singular value decomposition [2]

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

where $\mathbf{R}_{\mathbf{xy}}$ is the cross-covariance matrix between $\mathbf{x}$ and $\mathbf{y}$, and $\mathbf{R}_{\mathbf{xx}}$ and $\mathbf{R}_{\mathbf{yy}}$ are the auto-covariance matrices of $\mathbf{x}$ and $\mathbf{y}$. The canonical correlations are the singular values, which are the diagonal elements of the diagonal matrix $\mathbf{K}$. Moreover, $\mathbf{S} = \mathbf{F}^T\mathbf{R}_{\mathbf{xx}}^{-1/2}$ and $\mathbf{T} = \mathbf{G}^T\mathbf{R}_{\mathbf{yy}}^{-1/2}$.

In practice, we do not know the covariance matrices and must estimate them from samples. Let us assume that $\mathbf{x}$ and $\mathbf{y}$ have zero mean and that we observe M sample pairs $(\mathbf{x}_i, \mathbf{y}_i)$. These are commonly assembled in matrices $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_M]$ and $\mathbf{Y} = [\mathbf{y}_1\ \mathbf{y}_2\ \cdots\ \mathbf{y}_M]$, from which the sample covariance matrices $\hat{\mathbf{R}}_{\mathbf{xx}} = \mathbf{X}\mathbf{X}^T/M$, $\hat{\mathbf{R}}_{\mathbf{xy}} = \mathbf{X}\mathbf{Y}^T/M$, and $\hat{\mathbf{R}}_{\mathbf{yy}} = \mathbf{Y}\mathbf{Y}^T/M$ are computed. If CCA is performed based on these sample covariance matrices it leads to estimated (sample) canonical correlations $\hat{k}_i$. But how can we tell from these *estimated* canonical correlations how many actual correlated components there are? This is a question of model-order selection, and may be addressed using information-theoretic criteria (ICs). The idea of these criteria is to compute a score as a function of model order (the number of free parameters). 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 number of free parameters, 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. The advantage of ICs over framing the problem as a hypothesis test is that they do not require the selection of any subjective thresholds [3]. For the two-channel problem, [4–7] have used ICs to determine the number of correlated signals.

In many cases, however, CCA is not directly applied to the raw data $\mathbf{X}$ and $\mathbf{Y}$. Especially in the cases of insufficient samples, large dimensions of $\mathbf{x}$ and $\mathbf{y}$, or known existence of low-rank underlying signals [8], it is common practice to use a principal component analysis (PCA) preprocessing step. That is, instead of applying equation (1) directly to the sample covariance matrices, we first extract a number r_1 of components from $\mathbf{x}$ that account for a large fraction of the total variance in $\mathbf{x}$, by applying an eigenvalue decomposition to the sample covariance matrix $\hat{\mathbf{R}}_{\mathbf{xx}}$. Similarly, we extract r_2 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 $\mathbf{y}$. Model-order selection now needs to determine *three* numbers: the number of components r_1 to be extracted from $\mathbf{x}$, the number of components r_2 to be extracted from $\mathbf{y}$, and the number of meaningful canonical correlations r_3 that can be extracted from the reduced-dimensional CCA. A complicating factor is that the components of $\mathbf{x}$ that account for most of the variance in $\mathbf{x}$ and the components of $\mathbf{y}$ that account for most of the variance in $\mathbf{y}$ do not have to be the components that account for most of the correlation between $\mathbf{x}$ and $\mathbf{y}$.

Model-order selection in the joint PCA-CCA approach has not

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

yet received the attention that it arguably deserves. Most previous techniques are rather heuristic (e.g., [9, 10]). In this paper, we address this deficiency by presenting ICs for jointly determining the useful PCA dimension reduction as well as the number of correlated signals in a two-channel model.

Notationally, we use $(\cdot)^{-1}$ to indicate the appropriate matrix (pseudo-)inverse (depending on the rank of the argument). Bold uppercase/lowercase symbols indicate matrices/vectors. We denote matrix trace as $\mathrm{tr}\{\cdot\}$ and determinant as $|\cdot|$.

2. PROBLEM FORMULATION

Keeping in the spirit of previous approaches to model-order determination [3–6], we use the following two-channel model:

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

In this model, $\mathbf{z}_x$ and $\mathbf{z}_y$ are q_3-dimensional signals correlated between $\mathbf{x}$ and $\mathbf{y}$, and $\mathbf{s}_x$ and $\mathbf{s}_y$ are q_1- and q_2-dimensional signals present only in $\mathbf{x}$ and $\mathbf{y}$, respectively, which are assumed to be uncorrelated with each other and uncorrelated with $\mathbf{z}_x$ and $\mathbf{z}_y$. The matrices $\mathbf{A}_x$, $\mathbf{B}_x$, $\mathbf{A}_y$, $\mathbf{B}_y$, as well as the dimensions q_1, q_2, and q_3 are fixed but unknown. Without loss of generality, we may assume that $[\mathbf{A}_x, \mathbf{B}_x]$ and $[\mathbf{A}_y, \mathbf{B}_y]$ each have full column-rank, the auto-correlation matrices of $\mathbf{s}_x$, $\mathbf{s}_y$, $\mathbf{z}_x$, $\mathbf{z}_y$ are identity matrices, and the cross-correlation matrix between $\mathbf{z}_x$ and $\mathbf{z}_y$ is diagonal. The dimension of $\mathbf{x}$ and $\mathbf{n}_x$ is n, which does not have to match the dimension m of $\mathbf{y}$ and $\mathbf{n}_y$. All signals and noise are real-valued Gaussian with zero mean, and the noise is white and uncorrelated with the signals.

While some information-theoretic model-order selection techniques do not assume white noise [5, 11], most previous work relies on this assumption [3–6]. Since we are looking for both the useful PCA dimension reduction as well as the number of correlated signals, we must assume white noise to allow distinguishing the noise space from the signal space before dimension reduction.

We take M independent and identically distributed (i.i.d.) samples from the model (2), from which we compute the sample covariance matrices $\hat{\mathbf{R}}_{\mathbf{xx}}$, $\hat{\mathbf{R}}_{\mathbf{xy}}$, and $\hat{\mathbf{R}}_{\mathbf{yy}}$. Our task is now to jointly determine, from these sample covariance matrices and based on ICs, how many dimensions r_1 the PCA of $\mathbf{x}$ should keep, how many dimensions r_2 the PCA of $\mathbf{y}$ should keep, and how many correlated components r_3 there are. The goal is to choose the dimensions r_1 and r_2 such that the PCA preprocessing step preserves the correlated components. However, we are *not* interested in the components $\mathbf{s}_x$ and $\mathbf{s}_y$ that are present only in $\mathbf{x}$ and $\mathbf{y}$, respectively. Hence, it would actually be beneficial if the PCA preprocessing step were to eliminate those components (which is not generally possible).

3. INFORMATION-THEORETIC CRITERIA

ICs are computed as the likelihood for observing the data given a particular model and model order, minus a penalty term that increases with model order [12–14]. In our case, the likelihood is computed *assuming* a PCA-CCA model. That is, the coordinate system for the initial rank reduction is given by PCA (i.e., the eigenvectors of $\hat{\mathbf{R}}_{\mathbf{xx}}$ and $\hat{\mathbf{R}}_{\mathbf{yy}}$), and the coordinate system for analyzing correlation is determined by CCA. Therefore, our likelihood is a function of three parameters only: r_1, r_2, and r_3. In general, PCA would *not* be optimum for the initial rank reduction. However, in this work we do not attempt to find such an optimum rank reduction scheme.

It is necessary to make distinct the model versus the sample auto-covariances. The model auto-covariance for the $\mathbf{x}$-channel is [3]

$$\begin{aligned}\mathbf{R}_{\mathbf{xx}} &= \mathbf{U}_{r_1}(\mathbf{\Lambda}_{r_1} - \mathbf{I}_{r_1}\sigma_{\mathbf{x}}^2)\mathbf{U}_{r_1}^T + \mathbf{I}_n\sigma_{\mathbf{x}}^2 \\ &= \mathbf{U}_{\mathbf{x}}\begin{bmatrix}\mathbf{\Lambda}_{r_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\cdot\sigma_{\mathbf{x}}^2\end{bmatrix}\mathbf{U}_{\mathbf{x}}^T,\end{aligned} \tag{3}$$

where $\mathbf{U}_{\mathbf{x}}$ is composed of the eigenvectors, and $\mathbf{U}_{r_1}$ of the first r_1 eigenvectors, of $\hat{\mathbf{R}}_{\mathbf{xx}}$, $\mathbf{\Lambda}_{r_1}$ is an $r_1 \times r_1$ matrix with the largest r_1 eigenvalues on the diagonal, and $\sigma_{\mathbf{x}}^2 = \frac{1}{n-r_1}\sum_{i=r_1+1}^{n}\lambda_{\mathbf{x},i}$, following [3], represents the approximation of the white noise power (within $\mathbf{x}$) assumed present in the smallest eigenvalues. The model for $\mathbf{R}_{\mathbf{xx}}$ in (3) depends on r_1, which is the number of signal components kept. A similar form with r_2 components constitutes the model for $\mathbf{R}_{\mathbf{yy}}$ with noise power approximation $\sigma_{\mathbf{y}}^2$.

In order to define the model cross-covariance, we start with the singular value decomposition (SVD)

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

where $k_i(r_1, r_2),\ i = 1, \ldots, r_3, \ldots, \min\{n, m\}$ are the diagonal elements of $\mathbf{K}$, which are the model canonical correlations. A rank-r_3 model is now obtained by considering only the r_3 largest canonical correlations in the SVD. We write this as

$$\mathbf{C}_{\mathbf{xy}} = \mathbf{F}_{r_3}\mathbf{K}_{r_3}\mathbf{G}_{r_3}^T,$$

where $\mathbf{F}_{r_3}$ is $n \times r_3$, $\mathbf{K}_{r_3}$ is $r_3 \times r_3$, and $\mathbf{G}_{r_3}$ is $m \times r_3$. We call $\mathbf{C}_{\mathbf{xy}}$ the model coherence matrix. Hence, the model cross-covariance is obtained as

$$\mathbf{R}_{\mathbf{xy}} = \mathbf{R}_{\mathbf{xx}}^{1/2}\mathbf{C}_{\mathbf{xy}}\mathbf{R}_{\mathbf{yy}}^{1/2}.$$

3.1. Likelihood

We would like to obtain model-order estimates that account only for the correlated terms between $\mathbf{x}$ and $\mathbf{y}$. Hence, we are not interested in how well the model fits the observations $\mathbf{X}$ and $\mathbf{Y}$. Rather, we would like to know how well the model fits the correlated parts of $\mathbf{X}$ and $\mathbf{Y}$, in other words, those parts of $\mathbf{X}$ that tell us something about $\mathbf{Y}$, and vice versa. One might first think of tackling this issue by looking at how well $\mathbf{X}$ can be estimated from $\mathbf{Y}$. However, optimal (i.e., minimum mean-squared error) estimation is not symmetric: Estimating $\mathbf{x}$ from $\mathbf{y}$ leads to a different mean-squared error than estimating $\mathbf{y}$ from $\mathbf{x}$. Since CCA is a symmetric correlation analysis technique (i.e., the roles of $\mathbf{x}$ and $\mathbf{y}$ may be interchanged), we instead look at how well whitened $\mathbf{u} = \mathbf{R}_{\mathbf{xx}}^{-1/2}\mathbf{x}$ can be estimated from whitened $\mathbf{v} = \mathbf{R}_{\mathbf{yy}}^{-1/2}\mathbf{y}$. For this setup, interchanging the roles of $\mathbf{u}$ and $\mathbf{v}$ leaves the mean-squared error unchanged.

Thus, we consider the estimation error for estimating $\mathbf{u}$ from $\mathbf{v}$, which is

$$\mathbf{e} = \mathbf{u} - \mathbf{C}_{\mathbf{xy}}\mathbf{v},$$

and look at how well our model explains the sample error matrix

$$\mathbf{E} = \mathbf{R}_{\mathbf{xx}}^{-1/2}\mathbf{X} - \mathbf{C}_{\mathbf{xy}}\mathbf{R}_{\mathbf{yy}}^{-1/2}\mathbf{Y}.$$

The model error covariance is

$$\mathbf{Q} = E\{\mathbf{e}\mathbf{e}^T\} = \mathbf{F}\left(\mathbf{I} - \begin{bmatrix}\mathbf{K}_{r_3}\mathbf{K}_{r_3}^T & \\ & \mathbf{0}\end{bmatrix}\right)^{-1}\mathbf{F}^T, \tag{4}$$

and the sample error covariance matrix is

$$\hat{\mathbf{Q}} = \frac{1}{M}\mathbf{E}\mathbf{E}^T = \mathbf{R}_{\mathbf{xx}}^{-1/2}\hat{\mathbf{R}}_{\mathbf{xx}}\mathbf{R}_{\mathbf{xx}}^{-1/2} - \mathbf{C}_{\mathbf{xy}}\mathbf{R}_{\mathbf{yy}}^{-1/2}\hat{\mathbf{R}}_{\mathbf{xy}}^T\mathbf{R}_{\mathbf{xx}}^{-1/2} - \mathbf{R}_{\mathbf{xx}}^{-1/2}\hat{\mathbf{R}}_{\mathbf{xy}}\mathbf{R}_{\mathbf{yy}}^{-1/2}\mathbf{C}_{\mathbf{xy}}^T + \mathbf{C}_{\mathbf{xy}}\mathbf{R}_{\mathbf{yy}}^{-1/2}\hat{\mathbf{R}}_{\mathbf{yy}}\mathbf{R}_{\mathbf{yy}}^{-1/2}\mathbf{C}_{\mathbf{xy}}^T.$$

Since we have i.i.d. Gaussian samples, the likelihood for $\mathbf{E}$ given our model is [15]

$$f(\mathbf{E};\mathbf{Q}) = \text{const.} \times |\mathbf{Q}|^{-M/2} \exp\left\{-\frac{1}{2}\text{tr}\left\{\mathbf{Q}^{-1}\hat{\mathbf{Q}}\right\}\right\}. \quad (5)$$

It can be shown that the trace term in this expression may be simplified to yield the following log-likelihood (omitting any terms that do not depend on r_1, r_2, and r_3):

$$\ell(r_1,r_2,r_3) = -\frac{M}{2}\cdot\log\left(\prod_{i=1}^{r_3}(1-k_i^2(r_1,r_2))\right) - \frac{1}{2}\sum_{i=1}^{r_3}\frac{k_i^2(r_1,r_2)}{1-k_i^2(r_1,r_2)}\left(\gamma_{r_1,i}+\gamma_{r_2,i}-2\right). \quad (6)$$

As before, we include the arguments (r_1,r_2) of the canonical correlations to emphasize the dependence on the rank of the whitening matrices. Furthermore, we call $\gamma_{r_1,i}$ and $\gamma_{r_2,i}$ the alignment coefficients, which are defined as

$$\gamma_{r_1,i} = \sum_{k=1}^{r_1}|\mathbf{f}_i^T\mathbf{u}_{\mathbf{x},k}|^2 + \sum_{k=r_1+1}^{n}\frac{\lambda_{\mathbf{x},k}}{\sigma_{\mathbf{x}}^2}|\mathbf{f}_i^T\mathbf{u}_{\mathbf{x},k}|^2$$
$$\gamma_{r_2,i} = \sum_{k=1}^{r_2}|\mathbf{g}_i^T\mathbf{u}_{\mathbf{y},k}|^2 + \sum_{k=r_2+1}^{m}\frac{\lambda_{\mathbf{y},k}}{\sigma_{\mathbf{y}}^2}|\mathbf{g}_i^T\mathbf{u}_{\mathbf{y},k}|^2. \quad (7)$$

Here, $\mathbf{f}_i$, $\mathbf{g}_i$, $\mathbf{u}_{\mathbf{x},k}$, and $\mathbf{u}_{\mathbf{y},k}$ denote the respective columns of $\mathbf{F}$, $\mathbf{G}$, $\mathbf{U}_{\mathbf{x}}$, and $\mathbf{U}_{\mathbf{y}}$. The alignment coefficients represent the degree to which the mismatch between the model covariance $\mathbf{R}_{\mathbf{xx}}$ and the sample covariance $\hat{\mathbf{R}}_{\mathbf{xx}}$ (resp., $\mathbf{R}_{\mathbf{yy}}$ and $\hat{\mathbf{R}}_{\mathbf{yy}}$) affects the canonical representations of correlated data (i.e., the left and right singular vectors of the coherence matrix). Thus, extreme mismatches in the sample eigenvalues and the estimated white noise level cause non-unit alignment coefficients to act as an internal penalty to the likelihood.

If there is no initial rank-reduction (i.e., no PCA pre-processing step), then $r_1 = n$ and $r_2 = m$ and the alignment coefficients are $\gamma_{r_1,i} = 1$ and $\gamma_{r_2,i} = 1$, $\forall\ i$. In this case, the nonconstant terms in the log-likelihood can be shown to be

$$\ell(r_3) = \frac{M}{2}\log\left(\prod_{i=r_3+1}^{\min\{n,m\}}(1-k_i^2)\right)$$

which is the solution given in [5,6]. Note that in this case the canonical correlation coefficients k_i are obviously no longer functions of r_1 and r_2.

3.2. Akaike and Bayesian information criteria

The number of model terms (degrees of freedom) utilized in the model covariance matrices can be determined similar to [3, 5] and results in

$$d = \frac{r_1}{2}(2n-r_1+1) + \frac{r_2}{2}(2m-r_2+1) + r_3(n+m-r_3) + 2. \quad (8)$$

The first summand in d accounts for the degrees of freedom in modeling the signal part of $\mathbf{R}_{\mathbf{xx}}$, the second summand for the degrees of freedom in the signal part of $\mathbf{R}_{\mathbf{yy}}$, the third for the correlated parts in $\mathbf{R}_{\mathbf{xy}}$, and the final summand accounts for the two model parameters $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{y}}$.

Utilizing the log-likelihood (6) along with the number of free parameters d, we can write down the Akaike IC (AIC) and the Bayesian IC (BIC), which are to be maximized over r_1, r_2, and r_3, as

$$\text{AIC} = 2\ell(r_1,r_2,r_3) - 2\cdot d \quad (9)$$
$$\text{BIC} = \ell(r_1,r_2,r_3) - \frac{1}{2}\log(M)\cdot d. \quad (10)$$

It is well-known that the AIC tends to overfit the data, i.e., results in too large a model order, as the number of samples M increases.

An additional comment regarding the implementation of these ICs is in order: In the presence of perfectly correlated signals (i.e., there is $k_i = 1$ for some i), we must assume the correlated signal a part of the final model and evaluate the remainder of the likelihood terms. Moreover, a mismatch in the terms used for the white noise approximation can cause a canonical correlation greater than one. The test indices for such a whitening model are left out from consideration as the correct model.

4. SIMULATION RESULTS

In this section we show simulation results to demonstrate the effectiveness of the presented ICs. The data are generated according to the model in (2). To construct the $\mathbf{A}_x$ and $\mathbf{B}_x$ matrices, the eigendecomposition of the outer product of an $n\times M$ matrix with random normal entries with itself is taken, from which q_1+q_3 eigenvectors are randomly selected (an analogous procedure is applied to construct $\mathbf{A}_y$ and $\mathbf{B}_y$). In our examples, the signals $\mathbf{z}_x$ and $\mathbf{z}_y$ are perfectly correlated, i.e., $\mathbf{z}_x = \mathbf{z}_y$.

In the first example, there is one independent $\mathbf{x}$-channel signal ($q_1 = 1$) with variance 3, two independent $\mathbf{y}$-channel signals ($q_2 = 2$) also with variance 3, and one correlated signal ($q_3 = 1$) with variance 5. All noise components have variance 1. The dimension of $\mathbf{x}$ is $n = 20$ and the dimension of $\mathbf{y}$ is $m = 18$. Figure 1 illustrates the behavior of the model-order estimates for different sample sizes M. Here, $\square$'s represent the estimate of a the model rank r_1 for the PCA of $\mathbf{x}$, $\circ$'s the estimate of the model rank r_2 for the PCA of $\mathbf{y}$, and $\times$'s the estimate of the model rank r_3 for the number of correlated terms. As the interfering independent signals are weaker than the correlated signal, we would expect $r_1 = r_2 = r_3 = 1$, which is the choice of both AIC and BIC for a sufficiently large number of samples. The bottom plot in the figure shows the corresponding probability of error in estimating the model orders. A close look reveals that AIC starts to overestimate the model order for large number of samples.

In the second example, we consider the case where the independent (interfering) signals are stronger than the correlated signal. Here, the variance of each independent signal component is 10, whereas the correlated signal has variance 7. Otherwise the settings are the same as before. We would expect $r_1 = q_1+q_3 = 2$, $r_2 = q_2+q_3 = 3$, and $r_3 = q_3 = 1$, which again is the choice of both AIC and BIC for a sufficiently large number of samples. As shown in Figure 2, BIC performs better than AIC, especially when determining the number of correlated signals r_3 for smaller number of samples.

5. CONCLUSIONS

This paper presents an extension of the very well treated and ubiquitously applied information criteria for model-order selection to the

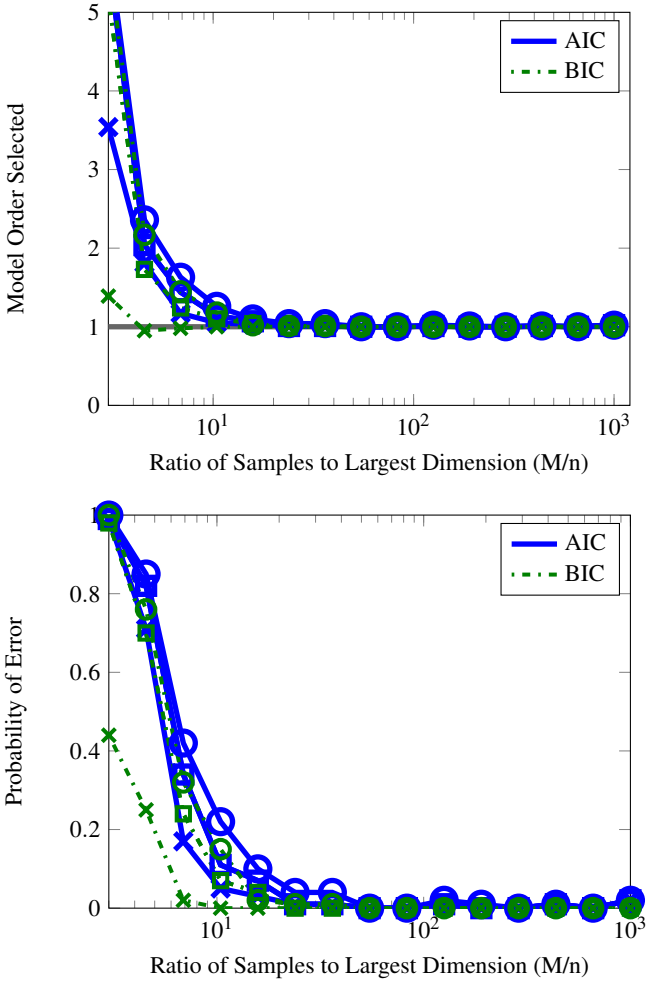

Fig. 1. Varying the number of samples M. □'s represent the estimate of r_1, ○'s the estimate of r_2, and ×'s the estimate r_3, averaged over 100 Monte Carlo runs. In this scenario, the correlated signal is stronger than the independent signals.

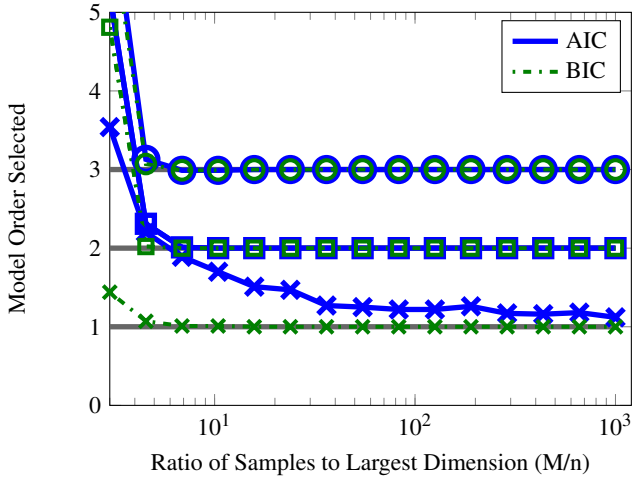

Fig. 2. □'s represent the estimate of r_1, ○'s the estimate of r_2, and ×'s the estimate r_3. The independent signal is stronger than the correlated signal.

scenario where the goal is to determine the number of correlated signals between two data sets in a combined PCA-CCA approach. Most previous approaches have been rather heuristic. To the best of our knowledge, this is the first paper to provide a systematic way of *jointly* choosing the optimum PCA dimension reduction *and* the number of correlated signals in CCA using information criteria.

6. REFERENCES

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

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

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

[4] J. J. Fuchs, "Estimation of the number of signals in the presence of unknown correlated sensor noise," *IEEE Trans. Signal Processing*, vol. 40, no. 5, pp. 1053–1061, May 1992.

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

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

[7] B. K. Gunderson and R. J. Muirhead, "On estimating the dimensionality in canonical correlation analysis," *J. Multivariate Analysis*, vol. 62, no. 1, pp. 121–136, July 1997.

[8] I. T. Jolliffe, *Principle Component Analysis*, John Wiley & Sons, 2005.

[9] W. R. Zwick and W. F. Velicer, "Comparison of five rules for determining the number of components to retain," *Psych. Bull.*, vol. 99, no. 3, pp. 432–442, 1986.

[10] H. Hwang, K. Jung, Y. Takane, and T. S. Woodward, "A unified approach to multiple-set canonical correlation analysis and principal components analysis.," *British J. Math. Stat. Psych.*, vol. 66, no. 2, pp. 308–21, May 2013.

[11] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On detection of the number of signals when the noise covariance matrix is arbitrary," *J. Multivariate Analysis*, vol. 20, pp. 26–49, Oct. 1986.

[12] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, vol. 19, no. 6, 1974.

[13] G. Schwartz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, 1978.

[14] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[15] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection theory*, Prentice Hall, 1998.