DETERMINING THE NUMBER OF CORRELATED SIGNALS BETWEEN TWO DATA SETS USING PCA-CCA WHEN SAMPLE SUPPORT IS EXTREMELY SMALL

Yang Song, Peter J. Schreier, and Nicholas J. Roseveare

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 when the number of samples from these data sets is extremely small. In such a scenario, a principal component analysis (PCA) preprocessing step is commonly performed before applying canonical correlation analysis (CCA). We present a reduced-rank version of the hypothesis test based on the Bartlett-Lawley statistic, which allows jointly determining the required PCA dimension reduction and the number of correlated signals.

Index Terms— Bartlett-Lawley statistic, canonical correlation analysis, model-order selection, principal component analysis, small sample support.

1. INTRODUCTION

Determining the number of correlated signals between two data sets when the number of samples is extremely small is an important problem with applications in areas as diverse as biomedicine, climate science, and communications. The standard approach for determining the number of correlated signals between two zero-mean random vectors $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$ is to use canonical correlation analysis (CCA) [1]. If the (true) population auto-covariance matrices $\mathbf{R}_{xx}$ and $\mathbf{R}_{yy}$ and the cross-covariance matrix $\mathbf{R}_{xy}$ are known, then the number of correlated signals is the number of nonzero canonical correlations, which can be computed as the singular values of the coherence matrix $\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1/2}$ [2]. If the population covariances are not known, they need to be estimated from M sample pairs $(\mathbf{x}_i, \mathbf{y}_i)$, which may be arranged in data matrices $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_M]$ and $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M]$ (with the mean removed). If these samples are independent and identically distributed (i.i.d.), the most common approach is to replace the population covariances with the sample covariance matrices $\hat{\mathbf{R}}_{xx} = \mathbf{X}\mathbf{X}^T/M$, $\hat{\mathbf{R}}_{yy} = \mathbf{Y}\mathbf{Y}^T/M$, and $\hat{\mathbf{R}}_{xy} = \mathbf{X}\mathbf{Y}^T/M$, and compute sample canonical correlations as the singular values of the sample coherence matrix. Based on the sample canonical correlations, the number of correlated signals may be estimated using information-theoretic criteria [3, 4], Mallow's statistic [3], a marginal likelihood [5], or hypothesis tests [6, 7].

All of these approaches assume sufficient sample support, which often means that the number of samples M must be significantly larger than the dimensions m and n. If this assumption is violated, these approaches usually produce very misleading results. In our paper, we focus on the challenging scenario where M is very small. If $M < m+n$, some 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]. Thus, a rank-reduction preprocessing step is necessary before applying CCA. The most common way of rank reduction is principal component analysis (PCA), which extracts from $\mathbf{x}$ (and $\mathbf{y}$, respectively) those components that account for most of its variance. This raises the obvious question of how many components of $\mathbf{x}$ and $\mathbf{y}$ the PCA preprocessing step needs to keep for CCA. This is a *model-order selection* problem.

In this paper, we present a reduced-rank version of the hypothesis test based on the Bartlett-Lawley statistic [6, 9], which allows *jointly* determining the required PCA dimension reduction *and* the number of correlated signals. We are not aware of any other model-order selection technique that is designed to handle the combined PCA-CCA approach in the sample-poor regime. In this conference paper, our main aim is to provide an intuitive exposition to our technique. A forthcoming journal paper will contain a technically more rigorous version, including proofs.

2. PROBLEM FORMULATION

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

$$\begin{aligned} \mathbf{x} &= \mathbf{A}_x\mathbf{s}_x + \mathbf{n}_x, \\ \mathbf{y} &= \mathbf{A}_y\mathbf{s}_y + \mathbf{n}_y. \end{aligned} \tag{1}$$

The signals $\mathbf{s}_x \in \mathbb{R}^{d+f}$ and $\mathbf{s}_y \in \mathbb{R}^{d+f}$ are jointly Gaussian with zero mean and cross-covariance matrix

$$E\{\mathbf{s}_x\mathbf{s}_y^T\} = \begin{bmatrix} \mathbf{diag}(\rho_1, \ldots, \rho_d) & \mathbf{0}_{d\times f} \\ \mathbf{0}_{f\times d} & \mathbf{0}_{f\times f} \end{bmatrix},$$

where ρ_i is the unknown correlation coefficient between $s_{x,i}$ and $s_{y,i}$ for $i = 1, \ldots, d$. Hence, the first d components of $\mathbf{s}_x$

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."

and $\mathbf{s}_y$ are correlated,[1] whereas the next f components are independent between $\mathbf{s}_x$ and $\mathbf{s}_y$. Without loss of generality, we may assume the auto-covariance matrices $E\{\mathbf{s}_x\mathbf{s}_x^T\}$ and $E\{\mathbf{s}_y\mathbf{s}_y^T\}$ to be diagonal. The correlated components may be stronger or weaker than the independent components. The matrices $\mathbf{A}_x \in \mathbb{R}^{n\times(d+f)}$ and $\mathbf{A}_y \in \mathbb{R}^{m\times(d+f)}$ as well as the dimensions d and f are deterministic but unknown. Without loss of generality, $\mathbf{A}_x$ and $\mathbf{A}_y$ may be assumed to have full column-rank. The noise $\mathbf{n}_x \in \mathbb{R}^n$ and $\mathbf{n}_y \in \mathbb{R}^m$ is independent of the signals, zero-mean Gaussian, white spatio-temporally, with power σ_n^2 per component.

From the data matrices $\mathbf{X}$ and $\mathbf{Y}$ we compute the sample covariance matrices $\hat{\mathbf{R}}_{xx}$, $\hat{\mathbf{R}}_{yy}$, and $\hat{\mathbf{R}}_{xy}$. In the case of small sample support, $\hat{\mathbf{R}}_{xx}$ and $\hat{\mathbf{R}}_{yy}$ may be singular, which means that we cannot determine the sample canonical correlations $\hat{k}_i$, $i = 1,...,p$, $p = \min(m,n)$, as the singular values of the sample coherence matrix $\hat{\mathbf{R}}_{xx}^{-1/2}\hat{\mathbf{R}}_{xy}\hat{\mathbf{R}}_{yy}^{-1/2}$, because this would require the computation of the matrix inverses $\hat{\mathbf{R}}_{xx}^{-1/2}$ and $\hat{\mathbf{R}}_{yy}^{-1/2}$. An easy workaround, following [8], is to first compute the compact (or economy) singular value decompositions (SVDs) of the data matrices $\mathbf{X} = \mathbf{U}_x\mathbf{\Sigma}_x\mathbf{V}_x^T$ and $\mathbf{Y} = \mathbf{U}_y\mathbf{\Sigma}_y\mathbf{V}_y^T$ (i.e., $\mathbf{V}_x \in \mathbb{R}^{n\times p}$ and $\mathbf{V}_y \in \mathbb{R}^{m\times p}$) and then to determine the sample canonical correlations as the singular values of $\mathbf{V}_x^T\mathbf{V}_y$.

However, the fact that we may thus always compute the sample canonical correlations $\hat{k}_i$, even in the case of low sample support, does not mean that the so determined $\hat{k}_i$'s are actually meaningful. 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. Therefore, a rank-reduction preprocessing step is required, the most common way of which is PCA. A combined PCA-CCA approach is the setup that we consider in our paper. For such a setup, we would like to *jointly* determine the required PCA dimension reduction and the number of correlated signals d, when the sample support is extremely small, possibly $M < m+n$.

3. HYPOTHESIS TEST

3.1. Effect of rank reduction on sample canonical correlations

Consider again the SVDs of the data matrices $\mathbf{X} = \mathbf{U}_x\mathbf{\Sigma}_x\mathbf{V}_x^T$ and $\mathbf{Y} = \mathbf{U}_y\mathbf{\Sigma}_y\mathbf{V}_y^T$. With a PCA preprocessing step, only r, instead of m or n, column vectors are kept in $\mathbf{V}_x$ and $\mathbf{V}_y$, which is denoted by $\mathbf{V}_x(:,1:r)$ and $\mathbf{V}_y(:,1:r)$. Thus, the ith sample canonical correlation $\hat{k}_i(r)$, which now depends on the rank r of the PCA preprocessing step, can be found as the ith largest singular value of $\mathbf{V}_x^T(:,1:r)\mathbf{V}_y(:,1:r)$. In order to avoid defective unit sample canonical correlations, we must choose $2r \le M$ and $r \le p$. This of course does not answer the question of what the optimum choice for r would be.

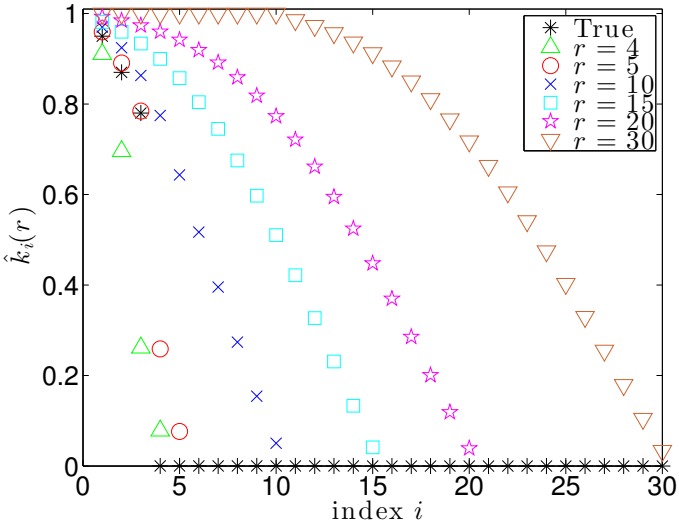
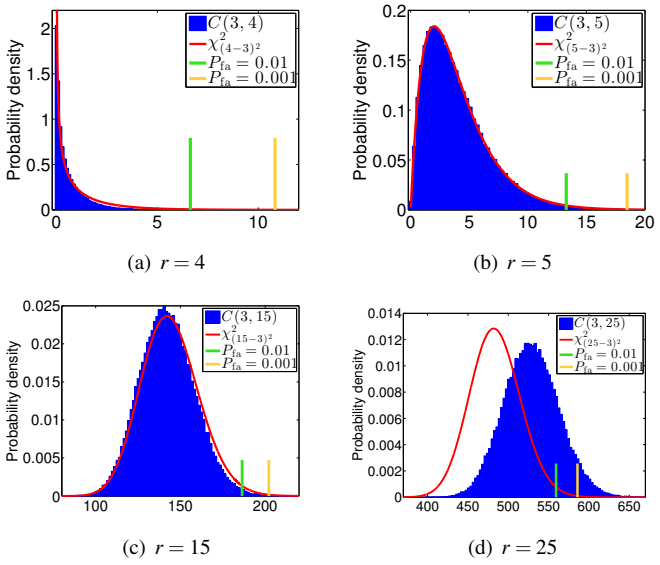


Fig. 1. Effect of rank reduction on the sample canonical correlations $\hat{k}_i(r)$, averaged over 10^3 runs. The true population canonical correlations are denoted by $*$.

Intuitively, it seems that r should be chosen large enough to capture as much of the cross-correlated signal components as possible without including too much noise. If the cross-correlated components are weaker than some of the independent components, this will inevitably mean that the PCA preprocessing step also keeps those stronger independent components. Hence, without noise, r would always be chosen as a number between d and $d+f$.

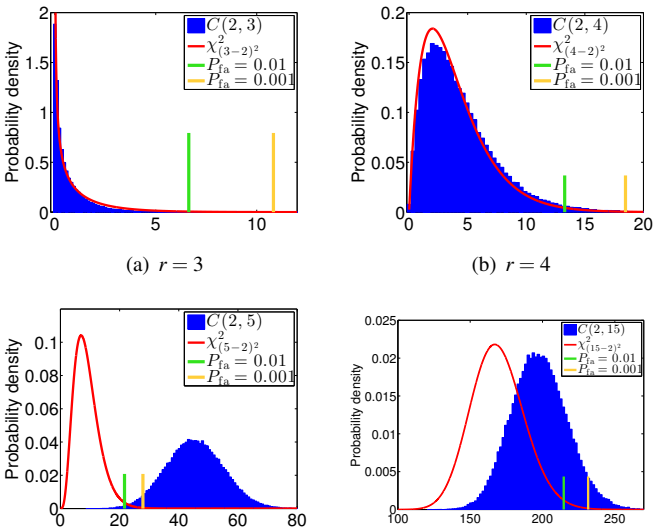
The optimum choice for r is closely linked with the effect that it has on the sample canonical correlations. It can be shown using Cauchy's interlacing theorem (the proof will be presented in the forthcoming journal paper) that $\hat{k}_i(r+1) > \hat{k}_i(r)$. Choosing too large an r will thus lead to sample canonical correlations that are greater, possibly significantly greater, than the true canonical correlations. On the other hand, if r is not large enough, then the rank-reduced representation does not contain all of the cross-correlated components, and thus the sample canonical correlations are too small. Figure 1 shows these effects for $M = 50$, $m = n = 40$, $d = 3$ correlated signals, and $f = 2$ independent signals, in the noise-free case. In our shown scenario, the independent signals are 3 dB stronger than the correlated signals, so the optimum choice for r should be $r = d+f = 5$. Indeed, we observe that choosing $r > 5$ leads to $\hat{k}_i$'s that are (significantly) too large, whereas choosing $r < 5$ leads to $\hat{k}_i$'s, for $i = 1,2,3$, that are too small.

[1] It is not difficult to generalize the results presented in this paper to the complex case and the case where the number of independent signals in $\mathbf{x}$ and $\mathbf{y}$ are different.



(a) $r=4$ (b) $r=5$ (c) $r=15$ (d) $r=25$

Fig. 2. The histogram of the statistic $C(s,r)$ (in blue) and the probability density function of a χ^2 distribution with $(r-s)^2$ degrees of freedom (in red), which is the distribution of $C(s,r)$ under the null hypothesis $s=d$, for different PCA ranks r. In all plots, $M=50$, $m=n=40$, $s=3$, $d=3$, and $f=2$. Histograms are computed from 10^5 statistically independent trials. Also shown as vertical lines are the thresholds $T(s,r)$ for two different probabilities of false alarm.



(a) $r=3$ (b) $r=4$ (c) $r=5$ (d) $r=15$

Fig. 3. Same settings as in Fig. 2, except for $s=2$.

3.2. Traditional hypothesis test

In the case of sufficient samples, the traditional hypothesis test for determining d 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 of H_0 vs. H_1 is run. This proceeds until H_0 is not rejected or $s=p$ is reached. The binary test is based on the Bartlett-Lawley statistic [6, 9]

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

When the data is Gaussian, the *asymptotic* distribution (as $M\to\infty$) of the Bartlett-Lawley statistic $C(s)$ under H_0 is $\chi^2_{(m-s)(n-s)}$ (with $(m-s)(n-s)$ degrees of freedom). This allows computation of the test threshold $T(s)$ for a given probability of false alarm.

3.3. A rank-reduced version of the hypothesis test

The Bartlett-Lawley statistic may be modified to account for the PCA preprocessing as

$$C(s,r) = -\left(M-s-r-\frac{1}{2}+\sum_{i=1}^{s}\hat{k}_i^{-2}\right)\ln\prod_{i=s+1}^{r}\left(1-\hat{k}_i^2(r)\right)$$

for $s=0,\ldots,r-1$. It can be shown that, as long as the PCA rank r is small relative to the number of samples M, the reduced-rank statistic $C(s,r)$ under H_0 is still approximately $\chi^2_{(r-d)^2}$. This can be observed in Figs. 2(a)–2(c), which use the same settings as before ($d=3$ correlated signals and $f=2$ stronger interfering signals). However, for too large an r, the statistic $C(s,r)$ is no longer approximately $\chi^2_{(r-d)^2}$, as is evident from Fig. 2(d) where $r=25=M/2$ (the largest r that does not result in defective unit sample canonical correlations). Thus, in a small sample scenario, r must be chosen sufficiently smaller than $M/2$. This maximum allowable r we will denote by $r_{\max}$.

The challenge in the reduced-rank version of the hypothesis test is to *jointly* determine r and d. We propose to select

$$\hat{d} = \max_{r=1,\ldots,r_{\max}}\ \min_{s=0,\ldots,r}\ \{s : C(s,r) < T(s,r)\}. \qquad (2)$$

The r that leads to $\hat{d}$ is the optimum rank for the PCA step. In (2) the min operator chooses the smallest s such that the statistic $C(s,r)$ falls below the threshold $T(s,r)$, which guarantees 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 the threshold $T(s,r)$ depends on both s and r. The rule (2) is based on the fact that if r is not chosen optimally, then it is likely that the min-step returns a number smaller than d. Hence, the min-step is performed for all r from 1 up to $r_{\max}$, and the maximum result is chosen as $\hat{d}$.

Let us explain the motivation for this test based on our example. Since the interfering independent signals are stronger than the correlated signals, $r=d+f=5$ should be the optimum choice for the PCA rank reduction. Figure 3(c) shows that for this choice it is very likely that we would reject the

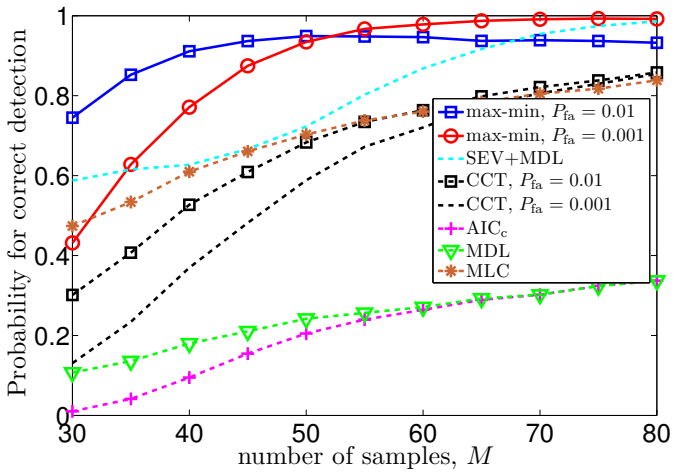


Fig. 4. Comparison of proposed "max-min" approach with canonical correlation test (CCT) [7], Akaike information criterion (AIC$_{\rm c}$) corrected for small sample sizes [10], minimum description length (MDL) [4, 5], and the marginal likelihood criterion (MLC) [5], for different sample sizes M. For all competing techniques, the PCA preprocessing step keeps approximately 70% of the total energy in $\mathbf{x}$ and $\mathbf{y}$ each. The only exception is SEV+MDL, where the approach in [11] is used to select r.

(incorrect) null hypothesis $d = 2$ since the test statistic $C(2,5)$ is unlikely to fall below the threshold $T(2,5)$ computed from a χ^2_3 distribution (unless the false alarm probability has been set extremely small). The choices $d < 2$ are even more likely to be rejected. On the other hand, Fig. 2(b) shows that the (correct) null hypothesis $d = 3$ is likely not to be rejected.

Now what happens if r is not chosen to be the optimum rank? If $r < 5$, then we do not capture all of the correlated signal components, and hence the test would decide for too small a d. Indeed, Figs. 3(a) and 3(b) show that for $r = 3$ and $r = 4$, it is likely that when testing $H_0 : d = 2$ versus $H_1 : d > 2$, the test would not reject H_0. On the other hand, if $r > 5$, then it becomes more difficult to distinguish between the canonical correlations that are associated with the correlated signals and those that are not, as was shown in Fig. 1. The effect can be observed in Fig. 3(d), which for $r = 15$ shows the distribution under $H_0 : d = 2$ and the corresponding histogram for $C(2,15)$. The red line and the histogram overlap much more than they do in Fig. 3(c), which means that it is rather likely that H_0 would not be rejected.

In summary, in all those cases where r is not chosen optimally, it is likely that the null hypothesis $H_0 : s = d$ would not be rejected for some value smaller than the true d. If r is chosen optimally, then the fact that H_0 still is approximately χ^2 guarantees that we will not pick too large a number of correlated components. This justifies the decision rule (2).

4. NUMERICAL RESULTS

We performed Monte Carlo simulations to evaluate the performance of our approach. The settings for our simulations were $n = m = 80$, $d = 3$ correlated signals with correlation coefficients 0.93, 0.85, and 0.78, $f = 4$ independent signals, and $r_{\max} = \min(\lfloor 0.3M \rfloor, p)$. The variance of each correlated signal is 2, the variance of each independent signal is 0.5, and the variance of each noise component is 1. Figure 4 shows the probability of correctly choosing $\hat{d} = 3$ for different number of samples M for our max-min-approach for two different probabilities of false alarm $P_{\rm fa} = 0.01$ and 0.001. We see that the max-min approach shows good performance from a very small sample size onward. However, the performance does depend on choosing the best $P_{\rm fa}$, an issue which will be addressed in the journal paper.

Our technique is compared with several competing approaches in Fig. 4. None of these approaches works at all without a PCA preprocessing step, which raises the question of how to choose r. We have done this in two different ways. First, we use the commonly employed rule of thumb "keep approximately P% of the total energy in $\mathbf{x}$ and $\mathbf{y}$ in the PCA preprocessing step." Because in practice there is no simple way of optimizing the performance with respect to P, we have not attempted to do so and have chosen a typical number for P (e.g., [12] suggests 70). Of course, the performance depends crucially on choosing P, and the percentage P that leads to the best performance depends very much on the scenario (SNR, relative powers of correlated and interfering signals, and strength of correlation). As an alternative, we use the approach [11], which is based on sample eigenvalues (denoted SEV in Fig. 4) and works with small sample support. We note that this approach aims to identify the number of signals in *one* dataset, so it is not designed for a PCA-CCA setup. Nevertheless, a combined SEV-MDL approach works quite well, even though it is still outperformed by the max-min technique by a significant margin.

5. CONCLUSION

In this paper, we have presented a technique that jointly determines the dimension of the rank reduction and the number of correlated signals using a combined PCA-CCA approach in the sample-poor regime. Of course there is no free lunch. Our technique only works if the number of signals is small compared to the number of samples. This matches intuition: We would not expect to be able to identify, say, 20 signals based on 30 samples. Another issue concerns the selection of the probability of false alarm. Setting it too high will lead to an estimator that tends to overfit, setting it too low will generally underfit. Achieving the best probability of detection requires the right tradeoff. Our focus in this paper was an intuitive exposition; a more rigorous treatment, including proofs, will be presented in a forthcoming journal paper.

6. REFERENCES

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

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

[3] Y. Fujikoshi and L. G. Veitch, "Estimation of dimensionality in canonical correlation analysis," *Biometrika*, vol. 66, no. 2, pp. 345–351, 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, April 1993.

[5] B. K. Gunderson and R. J. Muirhead, "On estimating the dimensionality in canonical correlation analysis," *Journal of Multivariate Analysis*, vol. 62, pp. 121–136, 1997.

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

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

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

[9] D. N. Lawley, "Tests of significance in canonical analysis," *Biometrika*, vol. 46, no. 1/2, pp. 59–66, June 1959.

[10] C. M. Hurvich and C. L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, pp. 297–307, 1976.

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

[12] J. M. Wallace, C. Smith, and C. S. Bretherton, "Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies," *J. Climate*, vol. 5, pp. 561–576, 1992.