DETECTING THE DIMENSION OF THE SUBSPACE CORRELATED ACROSS MULTIPLE DATA SETS IN THE SAMPLE POOR REGIME

Tanuj Hasija[*1]*, Yang Song*[*1]*, Peter J. Schreier*[*1] *and David Ramírez*[†2,3]

[1] Signal and System Theory Group, Universität Paderborn, Germany
[2] Signal Processing Group, University Carlos III of Madrid, Leganés, Spain
[3] Gregorio Marañón Health Research Institute, Madrid, Spain
{tanuj.hasija, yang.song, peter.schreier}@sst.upb.de, david.ramirez@uc3m.es

ABSTRACT

This paper addresses the problem of detecting the number of signals correlated across multiple data sets with small sample support. While there have been studies involving two data sets, the problem with more than two data sets has been less explored. In this work, a rank-reduced hypothesis test for more than two data sets is presented for scenarios where the number of samples is small compared to the dimensions of the data sets.

***Index Terms*—** Bartlett statistic, canonical correlation analysis, hypothesis testing, multiple data sets, small sample support

1. INTRODUCTION

The problem of detecting the number of correlated or common source signals from observations is innate to numerous signal processing applications. For example, in sensor array processing, detecting the number of sources incident on the arrays of sensors provides useful information and is also vital for algorithms estimating the direction of arrival [1]. Another example, in biomedicine, is data fusion of brain imaging modalities. Estimating the number of signals correlated among different modalities such as functional MRI (fMRI), electroencephalography (EEG), and structural MRI (sMRI) allows drawing inferences about functioning of the brain and its diseases [2]. Other applications include climate science, wireless communications, and genomics.

Canonical correlation analysis (CCA) is a widely used technique for finding relationships between two data sets [3]. The canonical correlations between two zero-mean random vectors $\mathbf{x}_1 \in \mathbb{C}^{m_1}$ and $\mathbf{x}_2 \in \mathbb{C}^{m_2}$ are given by the $p = \min(m_1, m_2)$ singular values of the coherence matrix $\mathbf{R}_{1,1}^{-1/2}\mathbf{R}_{1,2}\mathbf{R}_{2,2}^{-H/2}$. Here, $\mathbf{R}_{1,1} = E[\mathbf{x}_1\mathbf{x}_1^H]$ and $\mathbf{R}_{2,2} = E[\mathbf{x}_2\mathbf{x}_2^H]$ are the covariance matrices of $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, and $\mathbf{R}_{1,2} = E[\mathbf{x}_1\mathbf{x}_2^H]$ is their cross-covariance matrix. The superscript H denotes the Hermitian transpose, and, if necessary, the inverse is replaced by a pseudo-inverse. The number of signals correlated between $\mathbf{x}_1$ and $\mathbf{x}_2$ is given by the number of non-zero canonical correlations. However, the true covariance matrices are rarely known and have to be estimated from samples of $\mathbf{x}_1$ and $\mathbf{x}_2$. Model-order selection techniques are then used to determine the number of correlated signals from sample canonical correlations. These techniques are typically based either on information theoretic criteria (ITC) or hypothesis testing. It is shown in [4] that these two approaches are closely linked to each other. An ITC for finding the number of source signals common to two data sets was derived in [5]. Its hypothesis testing counterpart was presented in [6]. These techniques, however, work only when the sample size M is large compared to the dimensions of the data sets. Yet there are many scenarios where this assumption is not true. For such a sample poor case, an approach based on a combination of principal component analysis (PCA) and CCA was developed in [7, 8]. This approach uses a reduced-rank version of the hypothesis test, which jointly determines the dimensionality reduction for PCA and the number of correlated signals.

Only a few studies have addressed model-order selection for more than two data sets, and the few that have, did so only in the sample-rich regime. The paper [9] used an ad hoc approach for detecting the number of sources in multiple arrays. A detection technique based on ITC was derived in [10] using a similar data model as in [9]. In this work, we propose and investigate a hypothesis testing-based technique for detecting the number of correlated source signals in multiple data sets. Like [10], we assume that the number of correlated signals is the same for any pair of data sets.

[*]This research was supported by the Alfried Krupp von Bohlen und Halbach foundation under the program "Return of German scientists from abroad", and the German Research Foundation (DFG) under grant SCHR 1384/3-1.

[†]The work of D. Ramírez has been partly supported by Ministerio de Economía of Spain under projects: COMPREHENSION (TEC2012-38883-C02-01), OTOSIS (TEC2013-41718-R), ADVENTURE (TEC2015-69868-C2-1-R), and the COMONSENS Network (TEC2015-69648-REDC), and by the Comunidad de Madrid under project CASI-CAM-CM (S2013/ICE-2845).

2. DATA MODEL

Consider L data vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_L$, having dimensions $m_1, \ldots, m_L$ respectively. Without loss of generality, it can be assumed that $m_1 \leq m_2 \leq \ldots \leq m_L$. The generating data model is

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

Here, $\mathbf{s}_i \in \mathbb{C}^d$ is a zero-mean complex Gaussian source signal vector containing $d\ (< m_1)$ independent source signals. The matrix $\mathbf{A}_i \in \mathbb{C}^{m_i \times d}$ is an unknown but fixed mixing matrix and is assumed to have full column rank. The vector $\mathbf{n}_i \in \mathbb{C}^{m_i}$ is zero-mean complex Gaussian noise independent from the source vector. The covariance matrix $\mathbf{R}_{s_i s_i} = E[\mathbf{s}_i \mathbf{s}_i^H]$ is assumed to be the identity matrix for all sources, without loss of generality. The cross-covariance matrix between two sources is given by

$$\mathbf{R}_{s_i s_j} = E[\mathbf{s}_i \mathbf{s}_j^H] = \text{diag}(\rho_{i,j}^{(1)}, \ldots, \rho_{i,j}^{(d)}), \tag{2}$$

for any $i, j = 1, \ldots, L$ and $i \neq j$. Here, $\rho_{i,j}^{(1)}, \ldots, \rho_{i,j}^{(d)}$ denote the d unknown correlation coefficients, all assumed to be non-zero, between the elements in the i^{th} and j^{th} source vectors. Hence, we assume that the n^{th} entry in any given source vector, $s_i^{(n)}$, correlates with the n^{th} entry of any of the remaining source vectors, $s_j^{(n)}$, but not with other entries. The goal of this paper is to estimate the unknown dimension d, i.e., the number of signals that are correlated across all data sets. The noise covariance matrix $E[\mathbf{n}_i \mathbf{n}_i^H]$ is unknown and possibly colored. However, noise vectors of any two data vectors are assumed to be uncorrelated, $E[\mathbf{n}_i \mathbf{n}_j^H] = \mathbf{0}$, for $i \neq j$.

Let us define the composite data vector as $\mathbf{x} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_L^T]^T$, which has covariance matrix

$$E[\mathbf{x}\mathbf{x}^H] = \mathbf{R} = \begin{bmatrix} \mathbf{R}_{1,1} & \mathbf{R}_{1,2} & \cdots & \mathbf{R}_{1,L} \\ \mathbf{R}_{2,1} & \mathbf{R}_{2,2} & \cdots & \mathbf{R}_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{L,1} & \mathbf{R}_{L,2} & \cdots & \mathbf{R}_{L,L} \end{bmatrix},$$

where

$$\begin{aligned} &\mathbf{R}_{i,i} = \mathbf{A}_i \mathbf{R}_{s_i s_i} \mathbf{A}_i^H + E[\mathbf{n}_i \mathbf{n}_i^H] \\ &\mathbf{R}_{i,j} = \mathbf{A}_i \mathbf{R}_{s_i s_j} \mathbf{A}_j^H, \text{and} \\ &\text{rank}(\mathbf{R}_{i,j}) = d. \end{aligned}$$

3. HYPOTHESIS TESTING FOR MULTIPLE DATA SETS

A popular approach to the problem of order selection is to perform a series of binary hypothesis tests. Starting with source counter $s = 0$, we test the null hypothesis H_s : "s number of sources" against the alternative H_{s+} : "more than s sources correlated among the data sets" [6]. If H_s is rejected, s is incremented and another test of H_s vs. H_{s+} is run. This is repeated until H_s is not rejected or s reaches its maximum possible value. The obtained value of s is treated as the estimate of d. Each binary hypothesis test is a likelihood ratio test. Since the unknown parameters are replaced by their maximum likelihood estimates, this leads to a generalized likelihood ratio test (GLRT).

In this section, we first introduce a GLRT for multiple data sets in the sample rich regime, and then propose its reduced-rank version for small sample support.

3.1. Sample rich regime

Consider M independent and identically distributed (i.i.d.) samples of the composite data vector $\mathbf{x}$, arranged as the M columns of the data matrix $\mathbf{X} = \big[\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(M)\big]$. The generalized likelihood ratio for the hypothesis test is defined as

$$\eta = \frac{\max_{\boldsymbol{\theta}_s} f(\mathbf{X}|\boldsymbol{\theta}_s)}{\max_{\boldsymbol{\theta}_{s+}} f(\mathbf{X}|\boldsymbol{\theta}_{s+})}, \tag{3}$$

where $f(\mathbf{X}|\boldsymbol{\theta}_s)$ is the likelihood function of $\mathbf{X}$, and $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_{s+}$ are the parameter spaces under the hypotheses H_s and H_{s+}, respectively.

The maximization of the likelihood $f(\mathbf{X}|\boldsymbol{\theta}_s)$ for $L = 2$ data sets under H_s yields [5]

$$\max_{\boldsymbol{\theta}_s} f(\mathbf{X}|\boldsymbol{\theta}_s) \propto \left\{ \prod_{m=1}^{s} (1 - \hat{k}_m^2) \right\}^{-M}.$$

This maximization is performed under the constraint that $\text{rank}(\mathbf{R}_{1,2}) = s$. Here, $\hat{k}_m$ is the m^{th} largest sample canonical correlation between data sets $\mathbf{X}_1$ and $\mathbf{X}_2$.

For $L > 2$ data sets, the maximized likelihood $f(\mathbf{X}|\boldsymbol{\theta}_s)$ is the product of $L-1$ likelihoods for two data sets and is given by [10]

$$\max_{\boldsymbol{\theta}_s} f(\mathbf{X}|\boldsymbol{\theta}_s) \propto \left\{ \prod_{i=1}^{L-1} \prod_{m=1}^{s} (1 - \hat{k}_m^2(i)) \right\}^{-M}. \tag{4}$$

The term with $i = 1$ uses the sample canonical correlations $\hat{k}_m(1)$ between the first data set $\mathbf{X}_1$ and the remaining data sets $\mathbf{Y}_1 = [\mathbf{X}_2^T, \ldots, \mathbf{X}_L^T]^T$. The term with $i = 2$ uses the sample canonical correlations $\hat{k}_m(2)$ between the second data set $\mathbf{X}_2$ and $\mathbf{Y}_2 = [\mathbf{X}_3^T, \ldots, \mathbf{X}_L^T]^T$, and so on. The maximization of the first likelihood function is performed under the constraint that $\text{rank}(E[\mathbf{x}_1 \mathbf{y}_1^H]) = \text{rank}(\mathbf{R}_{1,2}, \ldots, \mathbf{R}_{1,L}) = s$, and similarly for the subsequent likelihood functions.

Since the parameter space $\boldsymbol{\theta}_{m_i}$ is sufficient to parametrize all the possibilities in $\boldsymbol{\theta}_{s+}$,

$$\max_{\boldsymbol{\theta}_{s+}} f(\mathbf{X}|\boldsymbol{\theta}_{s+}) \propto \left\{ \prod_{i=1}^{L-1} \prod_{m=1}^{m_i} (1 - \hat{k}_m^2(i)) \right\}^{-M}. \tag{5}$$

Using (4) and (5), the generalized likelihood ratio η defined in (3) can be simplified to

$$\eta = \left\{ \prod_{i=1}^{L-1} \prod_{m=s+1}^{m_i} (1 - \hat{k}_m^2(i)) \right\}^M. \quad (6)$$

Bartlett statistic - According to Wilks' theorem, the statistic $-2\ln\eta$ is asymptotically χ^2 distributed with the degrees of freedom (d.f.) equal to the difference between the number of free parameters corresponding to $\boldsymbol{\theta}_{s+}$ and $\boldsymbol{\theta}_s$ when H_s is true, i.e. $s = d$ [11]. Based on the results in [5, 10], it is not difficult to derive that the d.f. are $\sum_{i=1}^{L-1} 2(a_1(i) - s)(a_2(i) - s)$. Here, $a_1(i)$ and $a_2(i)$ depend on the dimensions of the data sets forming the i^{th} likelihood term. For instance, for $i = 1$, $a_1(1) = m_1$ and $a_2(1) = m_2 + m_3 + \ldots + m_L$. Similarly, for $i = 2$, $a_1(2) = m_2$ and $a_2(2) = m_3 + m_4 + \ldots + m_L$, etc. For small sample size, Bartlett's statistic [12] provides a better approximation of the χ^2 distribution. After applying Bartlett's correction to each of the $L - 1$ likelihoods in (6), we obtain the Bartlett statistic for multiple data sets

$$C(s) = -\sum_{i=1}^{L-1} [2M - (a_1(i) + a_2(i) + 1)] \times \times \ln \prod_{m=s+1}^{m_i} (1 - \hat{k}_m^2(i)), \quad (7)$$

which replaces the statistic $-2\ln\eta$.

3.2. Sample poor regime

For any two data sets $\mathbf{X}_1$ and $\mathbf{X}_2$, the sample canonical correlations can be computed as the singular values of $\mathbf{V}_1^H(:, 1 : m_1)\mathbf{V}_2(:, 1 : m_2)$, where $\mathbf{V}_1$ and $\mathbf{V}_2$ are calculated from the singular value decomposition of $\mathbf{X}_1 = \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{V}_1^H$ and $\mathbf{X}_2 = \mathbf{U}_2\mathbf{\Lambda}_2\mathbf{V}_2^H$, respectively, and $\mathbf{V}_1(:, 1 : m_1)$ denotes the matrix containing the first m_1 columns of $\mathbf{V}_1$ [13]. When the number of samples is smaller than the sum of ranks of $E[\mathbf{x}_1\mathbf{x}_1^H]$ and $E[\mathbf{x}_2\mathbf{x}_2^H]$, i.e., $M < m_1 + m_2$, it is proved in [13] that at least $(m_1 + m_2) - M$ sample canonical correlations will be equal to one, irrespective of the model from which $\mathbf{X}_1$ and $\mathbf{X}_2$ are generated. Moreover, even if $M > m_1 + m_2$, but not significantly greater, the sample canonical correlations significantly overestimate the true canonical correlations [13, 8]. This calls for rank reduction either before or alongside the detection of the number of correlated signals.

The reduced-rank version of Bartlett's statistic for multiple data sets in (7) is

$$C(s, r) = -\sum_{i=1}^{L-1} [2M - (2r + 1)] \ln \prod_{m=s+1}^{r} (1 - \hat{k}_m^2(r, i)),$$

where r denotes the rank of the PCA applied in the i^{th} term to the data sets $\mathbf{X}_i$ and $\mathbf{Y}_i = [\mathbf{X}_{i+1}^T, \ldots, \mathbf{X}_L^T]^T$, and the source

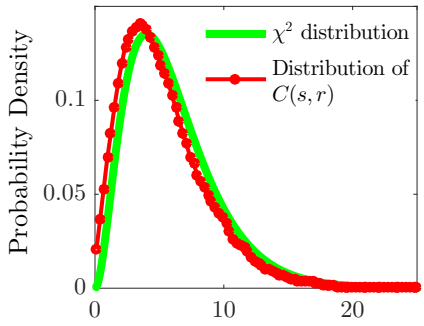

Fig. 1: Empirical distribution of $C(s, r)$ for $s = d = 4$ and $r = 5$.

counter s can take values from 0 to $r - 1$. Note that the PCA rank applied to all the data sets is r because we assume that the number of correlated signals is the same for any pair of data sets. Similar to $C(s)$, $C(s, r)$ is also a sum of $L - 1$ statistics for two data sets. The first statistic for $i = 1$ involves the sample canonical correlations of the rank-reduced versions of $\mathbf{X}_1$ and $\mathbf{Y}_1$. Hence, the r sample canonical correlations $\hat{k}_m(r, 1)$ are obtained as the singular values of $\mathbf{V}_1^H(:, 1 : r)\tilde{\mathbf{V}}_1(:, 1 : r)$, where $\tilde{\mathbf{V}}_1$ contains the right singular vectors of $\mathbf{Y}_1$. Similarly, the second statistic for $i = 2$ in $C(s, r)$ involves the singular values of $\mathbf{V}_2^H(:, 1 : r)\tilde{\mathbf{V}}_2(:, 1 : r)$, where $\tilde{\mathbf{V}}_2$ contains the right singular vectors of $\mathbf{Y}_2$, and so forth.

It was shown in [8] that under the null hypothesis $s = d$, each of these $L - 1$ statistics is approximately χ^2 distributed with $2(r - s)^2$ d.f., as long as the PCA rank r is large enough to capture all correlated components, yet sufficiently smaller than the number of samples M (this is typically the case when $r < M/3$). As long as M is large with respect to r, these statistics are close to the χ^2 distribution. Under the same conditions, $C(s, r)$ is also approximately χ^2-distributed with $2(L - 1)(r - s)^2$ d.f. This is demonstrated in Figure 1 for an example of four data sets with $m_1 = 40$, $m_2 = 50$, $m_3 = 55$, and $m_4 = 60$, and $M = 80$ samples. The number of correlated sources is $d = 4$, all of which have equal signal power. The correlation coefficients $\rho_{i,j}^{(1)}, \rho_{i,j}^{(2)}, \rho_{i,j}^{(3)}, \rho_{i,j}^{(4)}$ are chosen as 0.9, 0.9, 0.8 and 0.7, respectively, for all data sets. The noise is white with small power compared to the signal power. The empirical distribution of $C(s, r)$ is shown in Figure 1 along with the χ^2 distribution for $s = 4$ and $r = 5$. It can be seen that when H_s is true, i.e. $s = d = 4$, $C(s, r)$ closely follows the χ^2 distribution.

This means that in a series of binary tests of H_s vs H_{s+} based on $C(s, r)$, d is generally not overestimated. It is likely, however, to be underestimated if r is not chosen large enough. If r is too small, then the reduced-rank PCA descriptions do not capture all of the correlated components and thus the se-

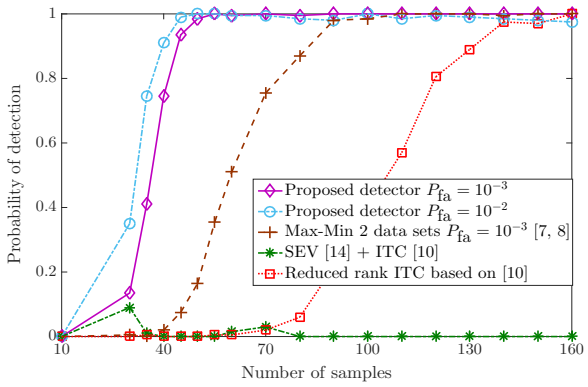

Fig. 2: Performance comparison of the proposed detector with competing detectors for four data sets having dimensions, $m_1 = 40$, $m_2 = 50$, $m_3 = 55$ and $m_4 = 60$.

ries of binary tests decides for too small a dimension d. This reasoning leads to the decision rule [7, 8]

$$\hat{d} = \max_{r=1,\ldots,r_{\max}} \min_{s=0,\ldots,r-1} \{s : C(s,r) < T(s,r)\}, \quad (8)$$

and the r that leads to $\hat{d}$ is the PCA rank. In (8), $r_{\max}$ should be chosen sufficiently smaller than $M/2$ (typically $M/3$) and $T(s,r)$ is the threshold value chosen to maintain a specific value of probability of false alarm, P_{fa}. The min-operator choses the smallest s such that $C(s,r) < T(s,r)$. If there is no such s, it chooses $s = r$. The rule (8) is motivated by the fact that if r is not chosen optimally, the min-step might return a number smaller than d. Because the min-step will not overfit, we can take the maximum result for all r from 1 up to $r_{\max}$. More details about the detector in (8) for the case of two data sets are provided by [7, 8].

4. SIMULATION RESULTS

The performance of the proposed "max-min" technique for multiple data sets is evaluated using Monte Carlo simulations. As before, we use four data sets with same dimensions as in the previous section. There are $d = 5$ correlated signals, with correlation coefficients 0.9, 0.8, 0.7, 0.7, and 0.6 between all data sets, and the noise is autoregressive (AR) with coefficients [1 0.2 0.5]. The signal components have unit variance in all the data sets. The variance of the white noise component before applying the AR filtering is 0.1. The mixing matrix $\mathbf{A}$ is randomly generated unitary matrix for all four data sets. Figure 2 shows the probability of correctly detecting the true number of correlated signals as a function of M, for the proposed method with two different probabilities of false alarm. For each point, we ran 500 independent trials. Comparisons with the competing techniques based on ITC [10] and "max-min" for two data sets [7] are also shown. Since the number of correlated signals d is the same for any pair of data sets, the "max-min" detector for two data sets can also be used. Here, we use the first two data sets for the "max-min" approach with two data sets. Since the technique based on ITC works only in the sample rich regime, the sample eigenvalue-based (SEV) technique [14] is used to determine the rank r for PCA, before determining d. In addition, we also show the performance of the rank-reduced version of the ITC based detector in [10] which selects r and d that jointly minimize the ITC function. While this modified reduced-rank ITC detector performs much better than its original version in [10] with the SEV preprocessing step, it still lags behind in performance compared to the detector proposed in this work. It can be observed that the proposed "max-min" detector for multiple data sets performs better than all competing approaches.

5. CONCLUSION

In this paper, a hypothesis testing based technique for detecting the number of signals correlated across multiple data sets has been presented. The technique shows promising results and an improvement in performance over the state of the art in scenarios where the number of samples is extremely small. Following the ideas in [7, 8], ITC-based versions of our approach can also be developed.

6. REFERENCES

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996.

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

[3] H. Hotelling, "Relations between two sets of variates," *Biometrika*, pp. 321–377, 1936.

[4] P. Stoica, Y. Selén, and J. Li, "On information criteria and the generalized likelihood ratio test of model order selection," *IEEE Signal Processing Letters*, vol. 11, no. 10, pp. 794–797, 2004.

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

[6] W. Chen, J. P. Reilly, and K. M. Wong, "Detection of the number of signals in noise with banded covariance ma-

trices," *IEE Proceedings-Radar, Sonar and Navigation*, vol. 143, no. 5, pp. 289–294, 1996.

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

[8] Y. Song, P. J. Schreier, D. Ramírez, and T. Hasija, "Canonical correlation analysis of high-dimensional data with very small sample support," 2016, arXiv:1604.02047.

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

[10] Y. Wu, K. W. Tam, and F. Li, "Determination of number of sources with multiple arrays in correlated noise fields," *IEEE Transactions on Signal Processing*, vol. 50, no. 6, pp. 1257–1260, 2002.

[11] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[12] M. S. Bartlett, "The statistical significance of canonical correlations," *Biometrika*, pp. 29–37, 1941.

[13] A. Pezeshki, L. L. Scharf, M. R. Azimi-Sadjadi, and M. Lundberg, "Empirical canonical correlation analysis in subspaces," *Proceedings of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 994–997, 2004.

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