IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 5, 2019, accepted July 22, 2019, date of publication August 5, 2019, date of current version August 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2933255

Estimating the Number of Significant Canonical Coordinates

ABD-KRIM SEGHOUANE, (Senior Member, IEEE), AND NAVID SHOKOUHI, (Member, IEEE)
Department of Electrical and Electronic Engineering, The University of Melbourne, Melbourne, VIC 3010, Australia

Corresponding author: Abd-Krim Seghouane (abd-krim.seghouane@unimelb.edu.au)

This work was supported by the Australian Research Council under Grant FT 130101394.

ABSTRACT This study proposes a novel model selection criterion for dimensionality estimation in canonical correlation analysis (CCA), which can be used to estimate the number of correlated components between two sets of multivariate vectors, particularly in the context of canonical coordinates. The proposed method has a different form compared to existing model selection criteria, which typically use the Bartlett-Lawley measure as a goodness-of-fit term. Alternatively, we propose to use instead a goodness-of-fit term based on the sum of squares of the smallest canonical correlation coefficients. The asymptotic properties of the proposed goodness-of-fit term have been laid out and an appropriate penalty term with proof of consistency is derived based on these properties. Numerical examples are presented in the context of direction-of-arrival estimation using canonical coordinates for two channel array systems. It is shown that our proposed order selection criterion outperforms conventional criteria and is robust to various noise and interference scenarios.

INDEX TERMS Canonical coordinates, multi-view analysis, canonical correlation analysis, model selection.

I. INTRODUCTION

Canonical correlation analysis (CCA) is a popular multivariate technique used to measure linear dependencies between two sets of variables [1], where each set typically corresponds to a different view or representation of data [2]. CCA finds pairs of linear projectors to produce maximally correlated canonical variables from the original data. The pair-wise correlations, which are called canonical correlations, provide a means for quantifying the relationship between the two views/sets of variables. Due to its comparative nature, CCA has had significant impact on a wide range of applications that rely on multi-view or multi-modal analysis. Applications include biomedical signal processing and brain imaging [3]–[5], radar/sonar [6], [7], communications [8], [9], remote sensing [10], multi-view classification/clustering [11], and computer vision [12].

One of the important challenges faced when using CCA is determining the number of correlated components, which corresponds to the problem of dimensionality estimation [13], [14]. Determining the number of correlated components can help identify the linear relationships between two multivariate signals. Several methods have been proposed to estimate dimensionality in CCA. Including empirical approaches such as cross-validation [15]; hypothesis testing such as the Bartlett-Lawley test [16]–[18]; or using model selection criteria such as Akaike's information crterion (AIC) [13], [19]–[22], Bayesian information criterion (BIC) [14], [23], and Mallow's C_p [14], [24].

All of the aforementioned approaches aim to select the highest possible number of correlated components to obtain a CCA model that best fits the data, while simultaneously restricting model complexity to avoid an overfitted model. However, among all approaches, model selection criteria particularly stand out, due to their relatively simple form and broad applicability. Model selection criteria generally comprise two terms, the first is a ''goodness-of-fit'' term that quantifies information gained from the data, while the second is a ''penalty'' term that increases with the number of canonical correlation components and amount of data. The penalty term is typically obtained by exploiting information on the approximate asymptotic distribution of the goodness-of-fit term [25]. The primary contribution of our study is to propose a new model order selection criterion for estimating the number of significant components in CCA. Similar to existing criteria, it has a simple form but differs from existing criteria in both its goodness-of-fit and penalty terms.

The associate editor coordinating the review of this manuscript and approving it for publication was Ke Gu.

The present study is motivated by a recent interpretation of CCA introduced in signal processing, called the canonical coordinates perspective [8], [26]. Canonical coordinates refer to the pairs of maximally correlated variates calculated by CCA for each canonical correlation component. Canonical coordinates provide a principal subspace in which two separate views of data have the highest correlation. This concept has been employed to investigate many problems in signal processing that involve two-channel systems. Examples include multivariate Wiener filtering [8], communication systems [27], cross-spectral analysis [28], and distributed array processing [29]. For distributed array processing in particular, leveraging information from multiple sensor arrays is an important aspect of distributed sensor networks. This has led to several CCA-based approaches in distributed array processing [30], [31], where estimating CCA dimensionality is on par with estimating the dimension of the principal subspace associated with the canonical coordinates between two multivariate signals, each corresponding to an array. In this study, we will explore the role of CCA order estimation in applications related to distributed array processing. As a numerical example, we will use CCA in the context of direction of arrival (DOA) estimation using two sets of spatially separated arrays. We will show that CCA order selection can be used to estimate the number of sources in the environment, which is generally known to help improve the performance of DOA estimation in approaches such as MUltiple SIgnal Classification (MUSIC) [32].

The remainder of this study is organized in the following manner. In Section II, we present some preliminary knowledge regarding canonical correlation analysis, canonical coordinates, and dimensionality estimation using model selection criteria. In Section III, we propose an alternative model selection criterion for CCA and present the main results of our study, including proof of consistency for our proposed method. In Section IV, we present some background knowledge on DOA estimation using CCA and compare the performance of our proposed method with conventional model selection criteria. Finally, in Section V, we present some concluding remarks.

II. BACKGROUND

This section presents some background on canonical correlation analysis, canonical coordinates, and order selection criteria for CCA.

A. CCA AND CANONICAL COORDINATES

Assume we have an $n \times (p+q)$ data matrix $(X^\top, Y^\top)$, where X is $p \times n$ and Y is $q \times n$ with $n > p \geq q$. As in [33] we assume that the columns $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}^q$, $i = 1, \ldots, n$, are represented by multivariate zero-mean Gaussian random vectors x and y, respectively, with densities $\mathcal{N}(0, \mathbf{\Sigma}_{xx})$ and $\mathcal{N}(0, \mathbf{\Sigma}_{yy})$, where:

$$\mathbf{\Sigma}_{xx} = E[xx^\top] \quad \mathbf{\Sigma}_{yy} = E[yy^\top].$$

The $p+q$ Gaussian vector $(x^\top, y^\top)^\top$ is zero-mean and has the following covariance matrix:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_{yy} \end{pmatrix} \tag{1}$$

where $\mathbf{\Sigma}_{xy} = \mathbf{\Sigma}_{yx}^\top = E[xy^\top]$ are $p \times q$ cross-covariance matrices. The coherence matrix of x and y is defined as [33], [34]:

$$C = \mathbf{\Sigma}_{xx}^{-1/2} \mathbf{\Sigma}_{xy} \mathbf{\Sigma}_{yy}^{-\top/2}, \tag{2}$$

where it is assumed that both $\mathbf{\Sigma}_{xx}$ and $\mathbf{\Sigma}_{yy}$ are invertible and the superscript $^{1/2}$ denotes matrix square-root. The singular value decomposition of $C = FKG^\top$ results in two column orthogonal matrices F and G of size $p \times s$ and $q \times s$, respectively, where $s = min(p, q)$. Since F and G are by definition orthonormal (i.e., $F^\top F = I_s$ and $G^\top G = I_s$), pre-multiplying C by $F^\top$ and post-multiplying it by G gives the rectangular diagonal matrix K, which contains the real-valued singular values of C, with non-zero entries only in the descending sequence $k_1 \geq k_2 \geq \ldots k_s \geq 0$. The non-zero elements of K are referred to as canonical correlations and correspond to the directions of maximal correlation between the random variables $u_i = f_i^\top \mathbf{\Sigma}_{xx}^{-1/2} x$ and $v_i = g_i^\top \mathbf{\Sigma}_{yy}^{-1/2} y$ [33], [34], where f_i and g_i are columns of F and G, for $i = 1, \ldots, s$. To illustrate this, consider the definition of k_1, which is:

$$\begin{aligned} k_1 &= \max_{f,g} f^\top C g = \max_{f,g} f^\top \mathbf{\Sigma}_{xx}^{-1/2} E\left[xy^\top\right] \mathbf{\Sigma}_{yy}^{-\top/2} g \\ &= \max_{f,g} E\left[f^\top \mathbf{\Sigma}_{xx}^{-1/2} xy^\top \mathbf{\Sigma}_{yy}^{-\top/2} g\right] \\ &= \max E\left[u_1 v_1^\top\right] = \max corr(u_1, v_1). \end{aligned} \tag{3}$$

The last line of (3) uses the fact that $Cov(u_1) = Cov(v_1) = 1$ in the definition of correlation $corr(u, v)$. This is sequentially performed for $k_2, \ldots, k_s$ under the conditions of orthogonality for the columns of F and G. The i^{th} pair of columns in F and G, namely f_i and g_i, provide a linear combination of the whitened variables $\mathbf{\Sigma}_{xx}^{-1/2} x$ and $\mathbf{\Sigma}_{yy}^{-1/2} y$ that are maximally correlated with each other and uncorrelated with all previous variables corresponding to $1, \ldots, i-1$.

In the context of two-channel signal processing, the vectors $u = F^\top \mathbf{\Sigma}_{xx}^{-1/2} x$ and $v = G^\top \mathbf{\Sigma}_{yy}^{-1/2} y$ are called canonical coordinates for x and y, respectively [8]. Therefore, the canonical correlations, which were defined as the singular values of the coherence matrix, can also be viewed as the correlations between the canonical coordinates [8]. Since correlation coefficients are bounded by 1, we have $1 \geq k_1 \geq k_2 \geq \cdots \geq k_s \geq 0$.

B. ESTIMATION OF DIMENSIONALITY FOR CCA

The number of correlated components between x and y is determined by the number of non-zero canonical correlations in K. When only $r \leq s$ components are correlated, the diagonal matrix K_r can be defined with the r non-zero canonical correlation coefficients $k_1 \geq k_2 \geq \ldots k_r > 0$. The value r is called CCA dimensionality [13]. When $r < s$, the coherence

matrix takes the form:

$$C = F \begin{pmatrix} K_r & 0 \\ 0 & 0 \end{pmatrix} G^\top \tag{4}$$

where the $p \times p$ orthogonal matrix $F = (F_r, F_{p-r})$ is constructed from F_r of size $p \times r$ and $F_{(p-r)}$ of size $p \times (p-r)$, and the orthonormal columns of F_r and F_{p-r} are orthogonal to each other. Similarly, the $q \times q$ orthogonal matrix $G = (G_r, G_{q-r})$ contains G_r of size $q \times r$ and G_{q-r} of size $q \times (q-r)$. The matrix $K_r = \text{diag}(k_1, \ldots, k_r)$ is diagonal and contains the non-zero singular values of C in descending order, $k_1 \geq \cdots \geq k_r > 0$. The number of non-zero singular values is the rank of C, thus the singular value decomposition of C is equivalent to

$$C = F_r K_r G_r^\top \tag{5}$$

The r left singular vectors F_r corresponding to the r largest singular values of K span the range or column space of C (i.e., $\mathcal{R}(C)$), whereas the associated right singular vectors G_r span the row space of C or $\mathcal{R}(C^\top)$.

The objective of CCA dimensionality estimation can be viewed as that of finding the rank r. This is trivial when $\mathbf{\Sigma}_{xx}$, $\mathbf{\Sigma}_{yy}$, and $\mathbf{\Sigma}_{xy}$ are known. However, in practice only the sample covariance matrix is available:

$$S = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} \tag{6}$$

where

$$S_{xx} = \frac{1}{n-1} XX^\top, \quad S_{yy} = \frac{1}{n-1} YY^\top, \text{ and}$$
$$S_{xy} = \frac{1}{n-1} XY^\top$$

Resulting in the sample coherence matrix:

$$\hat{C} = S_{xx}^{-1/2} S_{xy} S_{yy}^{-\top/2} \tag{7}$$

It is assumed that both S_{xx} and S_{yy} are positive definite. The corresponding SVD of the sample coherence matrix $\hat{C}$ is:

$$\hat{C} = \hat{F} \hat{K} \hat{G}^\top \tag{8}$$

where $\hat{K} = \text{diag}(\hat{k}_1, \ldots, \hat{k}_s) = (\hat{K}_r, \hat{K}_{s-r})$ is an $s \times s$ diagonal matrix with positive decreasing singular values and $s = \min(p, q)$. We are interested in finding r so that the matrix

$$\hat{C} = \hat{F} \begin{pmatrix} \hat{K}_r & 0 \\ 0 & \hat{K}_{s-r} \end{pmatrix} \hat{G}^\top \tag{9}$$

can be associated to (4) or equivalently to (5).

One method to estimate $rank(C)$ is to use an information criterion (IC) based on the choice of models

$$C = F_m K_m G_m^\top, \tag{10}$$

for different $m = 1, \ldots, s$. In such methods, the IC is defined in a way that it reaches a minimum point when $m = r$. Historically, ICs for CCA can be derived in the context of regression analysis [35], which addresses the conditional distribution of y given x (or vise versa). Under the problem conditions described at the beginning of this section, the conditional distribution can be fully determined by the regression matrix $\mathbf{\Sigma}_{yx} \mathbf{\Sigma}_{xx}^{-1}$, which has the same rank as C, since $\mathbf{\Sigma}_{xx}$ and $\mathbf{\Sigma}_{yy}$ are positive definite [13]. Some of the popular ICs derived through this method [14] are the Akaike's information criterion (AIC):

$$AIC(m) = -n \sum_{i=m+1}^{s} \log(1 - \hat{k}_i^2) - 2(p-m)(q-m), \tag{11}$$

the Bayesian information criterion:

$$BIC(m) = -n \sum_{i=m+1}^{s} \log(1-\hat{k}_i^2) - \log(n-1)\left[(p-m)(q-m)\right], \tag{12}$$

and Mallow's C_p:

$$C_p(m) = n \sum_{i=k+1}^{s} \left(\frac{\hat{k}_i^2}{1 - \hat{k}_i^2} \right) - 2(p-m)(q-m). \tag{13}$$

A common interpretation of these criteria is to consider the first component as the goodness-of-fit term (e.g., $-n \sum_{i=m+1}^{s} \log(1 - \hat{k}_i^2)$ in AIC), which determines how well the model (10) corresponding to m fits the data matrix. The second term is a penalizing term (e.g., $2(p-m)(q-m)$ in AIC), which limits the complexity of the model and is derived based on the degree-of-freedom of the first term [14]. The goodness-of-fit terms derived for *AIC* and *BIC* are based on the log-likelihood function of the conditional probability $P(y|x)$, while the goodness-of-fit of C_p is based on the mean squared prediction error of y given x [24].

One of the short-comings of the goodness-of-fit terms in (11), (12), and (13) is their sensitivity to slight variations in the canonical correlation coefficients $\hat{k}_i$. This is problematic, since realistic conditions such as limited number of observations n can significantly deviate the sample canonical correlations from their expected values [36]. In the next section, we will propose a new criterion that is proven to be a consistent estimator of the rank of C.

III. PROPOSED METHOD

This section proposes a new criterion for estimating the number of canonical coordinates based on the sum of squared canonical correlation coefficients $\hat{k}_i$. The proposed criterion takes the form:

$$IC(m) = n \sum_{i=m+1}^{s} \hat{k}_i^2 - c_n(p-m)(q-m) \tag{14}$$

where $\hat{k}_i$ are the sample canonical correlations or the singular values of $\hat{C}$. The second term in (14) is the so called penalty term, for which we will derive an appropriate c_n to obtain a consistent estimate of the dimension. In the remainder of this section we will show that the order m that minimizes (14) is a consistent estimator of dimensionality for CCA. Goodness-of-fit terms of the form presented in the

first term in (14) have been used before as statistical tests of dimensionality [37]–[40].

We start by deriving some properties of the sample coherence matrix $\hat{C}$, the estimator of the coherence matrix C. Using the definitions of the sample covariance, cross-covariance, and coherence matrices in (6) and (7), we can state:

Proposition 1: The sample coherence matrix $\hat{\mathbf{C}}$ *defined in (7) is a* $\sqrt{n}$*-consistent estimator of* $\mathbf{C}$.

In the proof of Proposition 1, presented in Appendix A, it is shown that $\sqrt{n}\text{vec}\left(\hat{C}-C\right) \xrightarrow{d} \mathcal{N}(0,\mathbf{\Omega})$ where $\mathbf{\Omega}$ is a $pq \times pq$ covariance matrix and $\xrightarrow{d}$ means convergence in distribution. From the $\sqrt{n}$-consistency of $\hat{C}$, we can address the consistency of $\hat{F}_r$ and $\hat{G}_r$ by adopting a random matrix approach on the matrix $\hat{C}$ which is a function of x_i and y_i, $i = 1, \ldots, n$.

Proposition 2: Let $r = rank(\mathbf{C})$ *and assume that there exists a small positive value* ϵ *such that* $k_r > \epsilon > 0$*, where* k_r *is the smallest non-zero singular value of* $\mathbf{C}$*. Then as* $n \to \infty$

$$\hat{\mathbf{F}}_r \xrightarrow{p} \mathbf{F}_r \quad \text{and} \quad \hat{\mathbf{G}}_r \xrightarrow{p} \mathbf{G}_r$$

where $\xrightarrow{p}$ *denotes convergence in probability.*

The proof of Proposition 2 is presented in Appendix B, which shows that $\hat{F}_r$ and $\hat{G}_r$ are $\sqrt{n}$-consistent estimators of F_r and G_r, respectively. From Proposition 2, the following result is immediately derived for the orthogonal projectors $\hat{Q}_F$ and $\hat{Q}_G$ calculated from $\hat{C}$.

Corollary 1: The orthogonal projectors on the null spaces of $\hat{\mathbf{C}}^\top$ *and* $\hat{\mathbf{C}}$ *satisfy*

$$\hat{\mathbf{Q}}_F = \hat{\mathbf{F}}_{p-r}\hat{\mathbf{F}}_{p-r}^\top = \mathbf{F}_{p-r}\mathbf{F}_{p-r}^\top + O_p\left(\frac{1}{\sqrt{n}}\right) \quad (15)$$

$$\hat{\mathbf{Q}}_G = \hat{\mathbf{G}}_{q-r}\hat{\mathbf{G}}_{q-r}^\top = \mathbf{G}_{q-r}\mathbf{G}_{q-r}^\top + O_p\left(\frac{1}{\sqrt{n}}\right) \quad (16)$$

Proof of Corollary 1 is presented in Appendix C. Based on the assumption in (4), which states that only the first r canonical correlations are non-zero, the following quantity

$$\hat{K}_{s-r} = \sqrt{n}\hat{F}_{p-r}^\top \hat{C}\hat{G}_{q-r} \quad (17)$$

approaches zero as the sample size n increases. $\hat{K}_{s-r}$ can be calculated using the projection matrices $\hat{Q}_F$ and $\hat{Q}_G$ defined in (15) and (16). Proposition 3 below shows the relation between these orthogonal projections and the proposed goodness-of-fit of (14) as well as the asymptotic properties of the final $s-r$ correlation coefficients.

Proposition 3: The asymptotic distribution of $\sqrt{n}vec\left(\hat{\mathbf{Q}}_F\hat{\mathbf{C}}\hat{\mathbf{Q}}_G\right)$ *is given by*

$$\sqrt{n}vec\left(\hat{\mathbf{Q}}_F\hat{\mathbf{C}}\hat{\mathbf{Q}}_G\right) \xrightarrow{d} \mathcal{N}\left(0, (\mathbf{Q}_G \otimes \mathbf{Q}_F)\mathbf{\Omega}(\mathbf{Q}_G \otimes \mathbf{Q}_F)\right) \quad (18)$$

where $\mathbf{\Omega}$ *is the* $pq \times pq$ *covariance matrix of* $\text{vec}\left(\hat{\mathbf{C}}-\mathbf{C}\right)$ *as derived in Appendix A.*

The proof of Proposition 3 is provided in Appendix D. Since according to (18), $\sqrt{n}\text{vec}\left(\hat{Q}_F\hat{C}\hat{Q}_G\right)$ is multivariate normal, the following can be said regarding the squared sum of its entries.

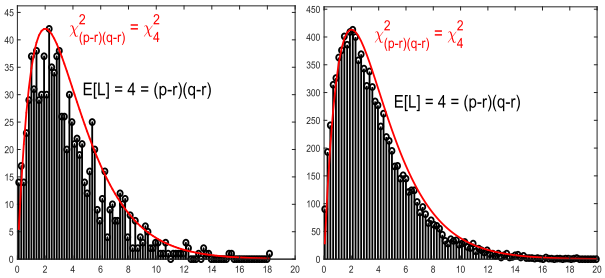

FIGURE 1. Illustration of the asymptotic behavior of the proposed goodness-of-fit term $L(\hat{C}, r)$. Left: Histogram of $L(\hat{C}, r)$ in black for 1000 runs (i.e., empirical results) vs. theoretical distribution in red. Right: For 10000 runs. In both small sample and large sample cases, the mean of $L(\hat{C}, r)$ matches the degrees-of-freedom of the data.

Corollary 2: Let $L(\hat{\mathbf{C}}, r)$ *be defined by*

$$L(\hat{\mathbf{C}}, r) = n\sum_{i=r+1}^{s}\hat{k}_i^2.$$

Then $L(\hat{\mathbf{C}}, r)$ *follows a* $(p-r)(q-r)$ *mixture of weighted* χ_1^2 *distributions where the weights are the eigenvalues of* $\mathbf{\Gamma} = (\mathbf{Q}_G \otimes \mathbf{Q}_F)\mathbf{\Omega}(\mathbf{Q}_G \otimes \mathbf{Q}_F)$*. Furthermore,* χ_γ^2 *with* $\gamma = (p-r)(q-r)$ *degrees-of-freedom can be used as an approximation to the distribution of* $L(\hat{C}, r)$.

Proof of Corollary 2 is presented in Appendix E. This result shows that the term $n\sum_{i=r+1}^{s}\hat{k}_i^2$ can be used as a goodness-of-fit term for a CCA model order selection criterion. For $1 \leq m \leq s$, the asymptotic bias of $n\sum_{i=m+1}^{s}\hat{k}_i^2$ can be corrected by $(p-m)(q-m)$, resulting in the criterion presented in (14) and the order estimation $\hat{r} = arg\,min\; mIC(m)$.

According to Corollary 2, the goodness-of-fit is asymptotically χ^2-distributed. This can be shown by considering a small example where two multivariate vectors are constructed according to the model: $x = W_x s + n_x$, and $y = W_y s + n_y$, where W_x and W_y are two random matrices of dimension $p \times r$ and $q \times r$. The latent vector s is the common hidden variable, which is of dimension r. The vectors n_x and n_y correspond to white Gaussian noise $\mathcal{N}(0, I)$. Figure 1 shows the histogram of $L(\hat{C}, r)$ calculated in two scenarios corresponding to 1000 and 10000 runs, with $n = 1000$ samples of x and y, $p = q = 5$, and $r = 3$. According to Corollary 2 the distribution of $L(\hat{C}, r)$ is asymptotically χ^2 with $(p-r)(q-r) = 4$ degrees-of-freedom, as it is empirically evident from Figure 1. In our example, the j^{th} entry of s for the i^{th} sample was set to $\sin(w(j)i + \phi(j))$, for $i = 1, \ldots, n$, where w and ϕ were randomly generated for each sinusoid.

The asymptotic distribution of $L(\hat{C}, r)$ together with its degrees-of-freedom $(p-r)(q-r)$ provide a guideline for model order selection. However, as a final step we must introduce an adjustment factor between the two terms to ensure consistency of estimation [41], [42]. Proposition 4 below shows the conditions required for the factor c_n in (14) that ensures this consistency.

Proposition 4: The criterion defined by

$$IC(m) = n \sum_{i=m+1}^{s} \hat{k}_i^2 - c_n \gamma \tag{19}$$

where $\gamma = (p-m)(q-m)$ *can be used to obtain a strongly consistent estimator of r (i.e.,* $lim_{n\to\infty} m_n = r$ *a.s.). Here* $m_n = \arg\min_m IC(m)$ *and* c_n *is taken such that* $lim_{n\to\infty} \frac{c_n}{n} = 0$ *and* $lim_{n\to\infty} \frac{c_n}{\log\log n} = \infty$.

Proof of Proposition 4 is presented in Appendix F. This concludes our derivation of (14) as a model selection criterion for the rank of the coherence matrix used to calculate CCA. In the next section, we provide some numerical results in the context of a practical application of CCA order selection for distributed array processing.

IV. NUMERICAL RESULTS

A recent application of CCA has been to address direction-of-arrival (DOA) estimation using a pair of spatially distributed arrays [29], [43]. The simulations adopted in this section are based on the method proposed in [29], which introduced a model based on the canonical coordinates between two spatially separated sensor arrays to derive an objective function for DOA estimation. This objective can be viewed as an extension of conventional single-array processing algorithms such as MUSIC [32]. In this section, we will use the cross-spectral analysis [30] technique proposed in [29] to demonstrate the performance of our proposed order estimation criterion in the context of DOA estimation and compare it with conventional criteria as described in Section II-B. We start with a brief overview of the DOA estimation technique proposed in [29], which from here on after will be referred to as CCA-DOA.

A. CCA-DOA

Consider r uncorrelated far-field narrow-band sources impinging on two arrays denoted by x and y that are spatially separated. In general, x and y have different dimensions (i.e., number of sensors per array) p and q, respectively. The source signals received at the arrays are denoted by the vectors s_x and s_y. The dimension of the vector s_x is r_x (where $r \leq r_x \leq p$) and the dimension of s_y is r_y (where $r \leq r_y \leq q$). The additional $r_x - r$ signals received at array x correspond to local interferences and therefore are considered to be nuisances, which we would like to ignore (similarly for y). The objective is to find $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_r]^\top$, which is the DOA of the r common sources between x and y with respect to a reference point. We are given a total of n snapshots ($i = 1, \ldots, n$, with $n \geq p + q + 1$):

$$\begin{aligned} x_i &= A_x s_{xi} + w_{xi} \\ y_i &= A_y s_{yi} + w_{yi}, \quad \text{for } i = 1, \ldots, n \end{aligned} \tag{20}$$

The vectors w_{xi} and w_{yi} represent white Gaussian measurement noise at each array. The arrays are separated far enough in space so that the noise sources w_x and w_y can be assumed uncorrelated. The $p \times r_x$ and $q \times r_y$ dimensional matrices

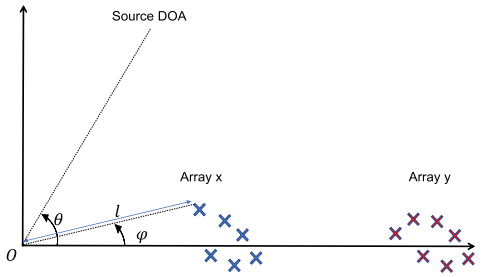

FIGURE 2. DOA estimation problem setup. Two spatially separate arrays x and y with known positions in polar coordinates, for example (l, ϕ) depicted in figure above. The goal is to find θ using n snapshots. In the diagram we have only shown one source. The effective distance with respect to the origin O is $d = l\cos(\theta - \phi)$.

$A_x = [a_{x1}(\boldsymbol{\theta}), \ldots, a_{xr_x}(\boldsymbol{\theta})]$ and $A_y = [a_{y1}(\boldsymbol{\theta}), \ldots, a_{yr_y}(\boldsymbol{\theta})]$ contain the steering vectors of x and y. Figure 2 illustrates the DOA schematic in a 2-dimensional coordinate system from which A_x and A_y can be derived. Note that all vectors defined in this section are complex valued.[1] Based on this, the covariance/cross-covariance matrices between x and y are [29]:

$$\begin{aligned} \boldsymbol{\Sigma}_{xx} &= A_x E[s_x s_x^H] A_x^H + E[w_x w_x^H], \\ \boldsymbol{\Sigma}_{yy} &= A_y E[s_y s_y^H] A_y^H + E[w_y w_y^H], \\ \boldsymbol{\Sigma}_{xy} &= A_x E[s_x s_y^H] A_y^H. \end{aligned} \tag{21}$$

In the last line of (21), we have $E[w_x w_y^H] = 0$.

In practice, we use the $p \times n$ and $q \times n$ data matrices X and Y by horizontally stacking x_i and y_i for all n snapshots to calculate the sample covariance/cross-covariances. The sample coherence matrix is $\hat{C} = S_{xx}^{-1/2} S_{xy} S_{yy}^{-H/2} = \hat{F}\hat{K}\hat{G}^H$, where S_{xx}, S_{yy}, and S_{xy} are as defined in (6), except using the Hermitian operator. As in conventional DOA methods [32], CCA-DOA finds an objective function that depends on a candidate angle-of-arrival θ. Given the number of sources r, [29] proposes to use the sum of the first r canonical correlation coefficients to derive the objective. Since only the first r singular values correspond to the signal component, we use the maximum corresponding to the truncated matrices $\hat{F}_r$ and $\hat{G}_r$. Defining P as the matrix of source cross-correlations corresponding to $E[s_x s_y^H]$ (not to be mistaken with K) and using the trace operator $tr(.)$, we have:

$$\begin{aligned} \sum_{j=1}^{r} \hat{k}_j &= tr\left(\hat{F}_r^H \hat{C} \hat{G}_r\right) = tr\left(\hat{F}_r^H S_{xx}^{-1/2} S_{xy} S_{yy}^{-H/2} \hat{G}_r\right) \\ &= tr\left(\hat{F}_r^H S_{xx}^{-1/2} A_x \mathrm{P} A_y^H S_{yy}^{-H/2} \hat{G}_r\right) \\ &= \sum_{j=1}^{r} \mathrm{P}(j,j) \left| A_y(:,j)^H S_{yy}^{-H/2} \hat{G}_r \hat{F}_r^H S_{xx}^{-1/2} A_x(:,j) \right| \end{aligned} \tag{22}$$

[1] Therefore, when calculating the covariance, cross-covariance, and coherence matrices in this section, we must use the matrix Hermitian operator ^{H} instead of transpose $^\top$. The singular values of the coherence matrix of x and y are still real and the proposed order estimation technique is the same as that of (14). Chapter 10 of [44] shows that the asymptotic behavior of the canonical correlation is similar for real and complex scenarios.

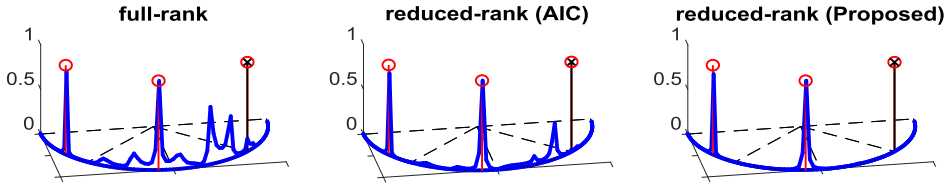

FIGURE 3. Example of DOA objective function $J(\theta)$ (in blue) using (22) with 2 impinging sources (in) and one interfering source that only affects one of the arrays (in black marked by ×). These plots show the effect of rank-reduction. In the left plot, with no rank-reduction, the objective function has peaks in directions not corresponding to the signals that are caused by noise. In the middle plot, where rank-reduction accuracy using AIC is 81%, some unwanted peaks can still be observed. In the right plot, the only peaks that are observed in the objective function come from the two desired source signals, resulting in better DOA estimation.

where $A_x(:,j) = a_{xj}(\boldsymbol{\theta})$ and $A_y(:,j) = a_{yj}(\boldsymbol{\theta})$ are the j^{th} columns of A_x and A_y, respectively. Since the term $S_{yy}^{-H/2}\hat{G}_r\hat{F}_r^H S_{xx}^{-1/2}$ is fully determined by the data, the maximization can be performed by sweeping the steering vectors $a_x(\theta)$ and $a_y(\theta)$ over $(-\pi/2, \pi/2)$ to determine the maximizing θ that satisfies (22). The third equality in (22) uses (21) and the last equality uses the fact that the sources are mutually uncorrelated. The objective of CCA-DOA is therefore:

$$J(\theta) = \frac{\left|a_x(\theta)^H S_{xx}^{-H/2} F_r G_r^H R_{yy}^{-1/2} a_y(\theta)\right|}{\left|a_x(\theta)^H S_{xx}^{-1} a_x(\theta) a_y(\theta)^H S_{yy}^{-1} a_y(\theta)\right|} \tag{23}$$

where the denominator is added as a standard normalizing term that helps remove the effect of signal power. As in standard MUSIC-like algorithms, the local maxima of $J(\theta)$ correspond to the DOA of the common sources between x and y. Ideally, we would like there to be exactly r peaks in $J(\theta)$. It is observed that CCA-DOA provides additional power compared to standard single-array DOA, as it is able to account for unwanted local interferences when r_x or r_y are greater than r. For more information on CCA-DOA, please see [29], [43].

B. SIMULATIONS

Using the full rank $\hat{F}$ and $\hat{G}$ matrices invalidates the equality of (22), therefore rank-reduction is necessary to limit the objective function to only contain local maxima at the directions of arrival. As in most DOA estimation techniques, the performance of CCA-DOA heavily relies on the number of sources and its accurate estimation [45], [46]. Figure 3 shows an example where $r = 2$ sources, with impinging wavefronts from angles $-60°$ and $-7.5°$, are measured at $n = 200$ snapshots by two uniform linear arrays (ULAs). Additionally, another equal power interfering source is received only by one of the arrays from angle 45° (marked in Fig. 3 with a black ×). We would like to be able to ignore the interference source, as it is "local" to one of the arrays. The $0dB$ noise sources w_x and w_y are white Gaussian and uncorrelated with each other. The two arrays are placed 10 meters apart with $p = q = 16$ sources that are randomly placed in a plane with known positions (the steering matrices are determined using the description provided in Fig. 2). The distance between sensors in each array is $\lambda/2$, where λ is the wavelength. The curves in Fig. 3 show the objective $J(\theta)$ calculated for different $\theta \in (-90°, 90°)$ with θ swept at 0.6° precision. The importance of accurate rank-reduction can be viewed by comparing the three plots in Fig. 3. In the left plot, we see that without truncating $\hat{F}$ and $\hat{G}$, the objective has several unwanted peaks. These peaks are caused by the noise and interference subspaces corresponding to $\hat{k}_i$ for $i > r$. The middle plot in Fig. 3 shows inaccurate rank-reduction where sometimes r is overestimated. These ranks were estimated using AIC, for which r is correctly estimated approximately 85% of times. In the right plot, we can see that when using our proposed method, which under these conditions always correctly estimates r with 100% accuracy, the objective function $J(\theta)$ only has local maxima at the DOAs. While these plots are for demonstration purposes only and all three approaches on average converge to the correct DOAs, using the proposed order estimation technique significantly improves consistency by correctly estimating the DOA every single time.

An interesting observation that can be made with regards to DOA estimation accuracy is the mean-squared-error (MSE) as a function of the truncation of $\hat{F}$ and $\hat{G}$. To show this, we repeat the previous experiment 100 times, except with $r = 10$ equally spaced sources located at angles ranging from $-60°$ to 45°. The SNR for this experiment is set to $0dB$ and the number of sensors per array is set to 32, so as to obtain higher accuracy. We estimate the DOA by finding the peaks of $J(\theta)$ calculated by truncating $\hat{F}$ and $\hat{G}$ at values ranging between 1 to 25. Calculating the mean-squared-error of DOA estimation is not straight-forward, since $J(\theta)$ may not contain exactly 10 peaks. To find the most optimistic error, we find the closest estimated value for each true DOA. After adding the corresponding squared error, we remove both the DOA and its suspected estimation and repeat the process. This iteration continues until either all true DOAs or all estimated DOAs have been considered. Figure 4 shows the MSE as a function of truncation and the average MSE for 100 runs. From this figure we can observe the different effects of under-estimation (truncation $<$ 10) and over-estimation (truncation $>$ 10). For

TABLE 1. Order selection accuracy (%) vs. number of sources for two ULAs.

	r	1	2	3	4	5	6	7	8	9	10
wide table	AIC	90.8	91.6	90.1	89.5	92.4	89.9	92.1	91.2	91.3	93
	BIC	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	32.5
	C_p	81.1	83.1	83.7	84.6	87	84.6	88.7	87.5	88.4	91.7
	Proposed	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**99.8**

TABLE 2. Order selection accuracy (%) vs. number of sources with randomly positioned sensors.

r	1	2	3	4	5	6	7	8	9	10
AIC	90.2	90.5	90.3	89.9	89.2	90.7	91.2	93	90.6	90.3
BIC	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	99.8	90.2
C_p	79.9	82.1	82.7	83.3	84.2	86.1	87.3	90.5	89.4	88.6
Proposed	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**95.4**

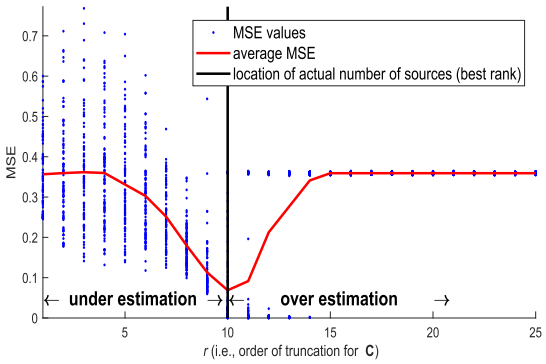

FIGURE 4. Investigating the effect of order estimation on MSE values for 100 runs. The black vertical line shows the number of sources. Truncating the rank of $\hat{C}$ with values on the left of the black line equates to underestimation, which increases the variance of MSE. Truncating on the right of the back line results on overestimation, which reduces the MSE variance but has high bias. The trade-off, as determined by the average MSE over 100 runs, coincides with the number of sources (here 10).

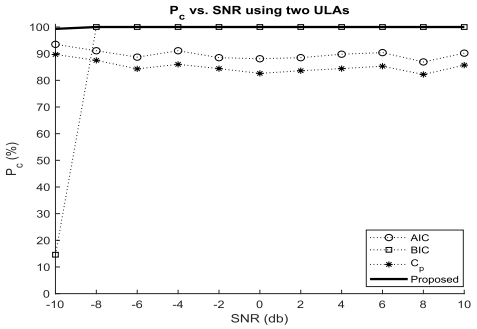

FIGURE 5. Investigating the effect of noise power on CCA order estimation using 300 snapshots collected from two ULAs spaced 10 meters apart. The number of common impinging sources is 8. We see that our proposed method (bold black) outperforms all order estimation methods.

under-estimated scenarios, the variance of the MSE decreases as truncation approaches the number of sources. Whereas for over-estimated scenarios, although the variance of the MSE shrinks as the truncation increases, the actual MSE value approaches a fixed non-zero bias. We can see from the average MSE curve (red curve in Fig. 4), the best truncation rank coincides with the number of sources $r = 10$. This observation illustrates the bias/variance trade-off associated with model order selection [47].

Now that the significance of accurate order estimation for the CCA-DOA experimental setup has been established, we focus on comparing the performance of our proposed CCA order estimation criterion with existing criteria, namely AIC, BIC, and C_p as they were described in Section II. We will compare robustness with respect to: signal-to-noise ratio (SNR), number of common sources, and number of nuisance sources. Performance of order estimation accuracy is measured as the probability of correct detection (P_c) calculated over $T = 1000$ repetitions. P_c is computed as $100 \times \frac{\#\text{ correct estimations}}{T}$.

First, we investigated the effect of noise power on P_c by setting the SNR to values between $(-10, 10)$dB. Through experiments we find that the least acceptable number of snapshots for reasonable performance from all methods is $n = 300$. We consider counting 8 sources using two uniform lineary arrays (ULAs) placed 10 meters apart, each array is comprised of 16 sensors with between-sensor distance of $d = \lambda/2$. Figure 5 shows P_c versus SNR for all four methods. We see that our proposed criterion shows more resistance to noise compared to BIC and has higher detection accuracy at high SNR compared to BIC, AIC, and C_p. We also performed the same experiment using two arrays with randomly placed sensors and average sensor distance of $\lambda/2$. Figure 6 shows that for $-6dB$ SNR and higher, our proposed method outperforms other methods.

Next, we estimate CCA orders as a function of the number of common sources (i.e., the true r). We will use SNR=5dB to ensure all methods work reasonably well, based on the observations from Figs. 5 and 6. We had to double the number of snapshots to keep C_p competitive. Therefore, in the following experiments $n = 600$, which had little impact on the performance of AIC, BIC, and our proposed method. All other experiment conditions remained unchanged. Tables 1 and 2 show the performance of all 4 methods, respectively for ULAs and randomly placed sensors. The number of common sources range between $r = 1, \ldots, 10$. We see that our proposed method always outperforms the other methods.

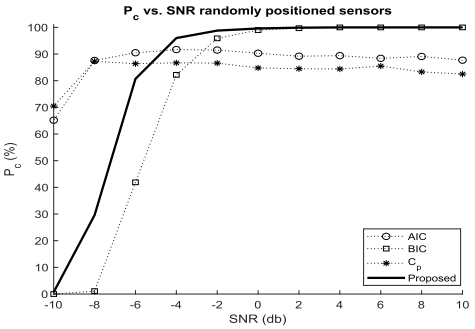

FIGURE 6. Similar to experiments of Fig. 5, except using randomly placed sensors with average between-sensor distance of $\lambda/2$. The performances drop compared to ULA, however, our proposed method outperforms all methods in nominal SNR values.

TABLE 3. Order selection accuracy (%) vs. number of interfering signals ($r_x - r = r_y - r$).

$r_x - r$	AIC	BIC	C_p	Prop.
1	90.2	**100**	83.7	**100**
2	93.6	**100**	86.2	**100**
3	90.9	**100**	84.5	**100**
4	92.3	**100**	86.2	**100**
5	92.5	**100**	87.7	**100**
6	92.4	**100**	87.2	**100**
7	92.3	**100**	86	**100**
8	91.6	56.1	85.5	**100**

Finally, we would like to investigate the effect of $r_x - r$ signals at x and $r_y - r$ signals at y that are local to each of these arrays and are not amongst the r common signals. The ability to distinguish common sources from local interferences is unique to CCA and is therefore one of the motivations of CCA-DOA [29]. We assume a constant signal-to-interference (SIR) ratio of $1.5dB$, which is calculated as the ratio between the sum of the powers of common sources to the sum of the powers of uncommon sources for each array. Table 3 shows the accuracy calculated over 1000 repetitions, when the number of common source $r = 4$, and the number of uncommon sources ($r_x - r$ and $r_y - r$) varies between $1, \ldots, 8$. We can see that our proposed method, which relies on the sum-of-squares of the lowest canonical correlations, is much more robust to variations in the number of interfering (uncommon) signals.

V. CONCLUSION

A model selection criterion for determining the number of canonical coordinates (i.e., dimensionality estimation for canonical correlation analysis) was proposed in this paper. It has the particularity of using the sum of squares of the smallest canonical correlation coefficients between two sets of multivariate vectors as a goodness-of-fit term instead of deriving it using the likelihood approach. The penalty term was derived based on the approximate asymptotic distribution of the proposed goodness-of-fit term, which when combined together can be used as a simple to use and reliable model selection criterion for estimating the number of significant component in CCA. Numerical simulations were conducted to show the effectiveness of our proposed CCA model order selection criterion compared to convectional criteria. Our proposed method was shown to outperform conventional criteria in the context of counting the number of common sources impinging on a two channel array setup. We attribute this performance to the simplicity of the proposed goodness-of-fit fit term, which is robust against variations in the canonical correlation coefficients.

APPENDIX A
PROOF OF PROPOSITION 1 CONSISTENCY OF $\hat{C}$

Proof: To prove Proposition 1 we will show that

$$\sqrt{n}\text{vec}\left(\hat{C} - C\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Omega})$$

where $\mathbf{\Omega}$ is an $pq \times pq$ covariance matrix and $\xrightarrow{d}$ means convergence in distribution. Using the definitions of C and $\hat{C}$ from (2) and (7):

$$\text{vec}\left(\hat{C} - C\right) = \text{vec}\left(S_{xx}^{-1/2} S_{xy} S_{yy}^{-\top/2} - \mathbf{\Sigma}_{xx}^{-1/2} \mathbf{\Sigma}_{xy} \mathbf{\Sigma}_{yy}^{-\top/2}\right)$$

add and subtract $S_{xx}^{-1/2} \mathbf{\Sigma}_{xy} S_{yy}^{-\top/2}$ and $\mathbf{\Sigma}_{xx}^{-1/2} \mathbf{\Sigma}_{xy} S_{yy}^{-\top/2}$:

$$\begin{aligned}\text{vec}\left(\hat{C} - C\right) = &\text{vec}\left(S_{xx}^{-1/2}\left(S_{xy} - \mathbf{\Sigma}_{xy}\right) S_{yy}^{-\top/2}\right) \\ &+ \text{vec}\left(\left(S_{xx}^{-1/2} - \mathbf{\Sigma}_{xx}^{-1/2}\right) \mathbf{\Sigma}_{xy} S_{yy}^{-\top/2}\right) \\ &+ \text{vec}\left(\mathbf{\Sigma}_{xx}^{-1/2} \mathbf{\Sigma}_{xy}\left(S_{yy}^{-\top/2} - \mathbf{\Sigma}_{yy}^{-\top/2}\right)\right) \quad (24)\end{aligned}$$

Using the Kronecker product $\otimes$, (24) can be restated as:

$$\begin{aligned}\text{vec}\left(\hat{C} - C\right) = &\left(S_{yy}^{-1/2} \otimes S_{xx}^{-1/2}\right) \text{vec}\left(S_{xy} - \mathbf{\Sigma}_{xy}\right) \\ &+ \left[\left(S_{yy}^{-1/2} \mathbf{\Sigma}_{xy}^{\top}\right) \otimes I_p\right] \text{vec}\left(S_{xx}^{-1/2} - \mathbf{\Sigma}_{xx}^{-1/2}\right) \\ &+ \left[I_q \otimes \left(\mathbf{\Sigma}_{xx}^{-1/2} \mathbf{\Sigma}_{xy}\right)\right] \text{vec}\left(S_{yy}^{-\top/2} - \mathbf{\Sigma}_{yy}^{-\top/2}\right) \quad (25)\end{aligned}$$

From the law of large numbers, as $n \to \infty$

$$\begin{aligned} S_{yy}^{-1/2} \otimes S_{xx}^{-1/2} &\xrightarrow{p} \mathbf{\Sigma}_{yy}^{-1/2} \otimes \mathbf{\Sigma}_{xx}^{-1/2} \text{ [bounded]} \\ \left(S_{yy}^{-1/2} \mathbf{\Sigma}_{xy}^{\top}\right) \otimes I_p &\xrightarrow{p} \left(\mathbf{\Sigma}_{yy}^{-1/2} \mathbf{\Sigma}_{xy}^{\top}\right) \otimes I_p \text{ [bounded]} \\ \text{vec}\left(S_{xy} - \mathbf{\Sigma}_{xy}\right) &\xrightarrow{p} 0 \\ \text{vec}\left(S_{xx}^{-1/2} - \mathbf{\Sigma}_{xx}^{-1/2}\right) &\xrightarrow{p} 0 \\ \text{vec}\left(S_{yy}^{-1/2} - \mathbf{\Sigma}_{yy}^{-1/2}\right) &\xrightarrow{p} 0 \end{aligned}$$

Thus, $\text{vec}(\hat{C} - C) \xrightarrow{p} 0$ as $n \to \infty$. Moreover, by the central limit theorem, $\sqrt{n}\text{vec}\left(S_{xy} - \mathbf{\Sigma}_{xy}\right)$, $\sqrt{n}\text{vec}\left(S_{xx}^{-1/2} - \mathbf{\Sigma}_{xx}^{-1/2}\right)$, and $\sqrt{n}\text{vec}\left(S_{yy}^{-1/2} - \mathbf{\Sigma}_{yy}^{-1/2}\right)$ each converges to a fixed multivariate Normal distribution. This implies that $\sqrt{n}vec(\hat{C} - C)$ is a mixture of Gaussians and $vec\left(\hat{C} - C\right) = O\left(n^{-1/2}\right)$. Thus, $E\left[\hat{C} - C\right] = 0$. Therefore, we have

$$\sqrt{n}\text{vec}\left(\hat{C} - C\right) \xrightarrow{p} 0, \quad (26)$$

In other words: Now, using (26) and by replacing in the first term of (25) $\left(S_{yy}^{-1/2} \otimes S_{xx}^{-1/2}\right)$ with $\left(\mathbf{\Sigma}_{yy}^{-1/2} \otimes \mathbf{\Sigma}_{xx}^{-1/2}\right)$, in the second term $\left[\left(S_{yy}^{-1/2}\mathbf{\Sigma}_{xy}^{\top}\right) \otimes I_p\right]$ with $\left[\left(\mathbf{\Sigma}_{yy}^{-1/2}\mathbf{\Sigma}_{xy}^{\top}\right) \otimes I_p\right]$, and in the third term $\left[I_q \otimes \left(S_{xx}^{-1/2}\mathbf{\Sigma}_{xy}\right)\right]$ with $\left[I_q \otimes \left(\mathbf{\Sigma}_{xx}^{-1/2}\mathbf{\Sigma}_{xy}\right)\right]$, we have:

$$\begin{aligned}\sqrt{n}\mathrm{vec}\left(\hat{C}-C\right) &= \sqrt{n}\left(I_q \otimes \mathbf{\Sigma}_{xx}^{-1/2}\right)\\ &\times \mathrm{vec}\left[\mathbf{\Sigma}_{xy}\left(S_{yy}^{-T/2}-\mathbf{\Sigma}_{yy}^{-T/2}\right)\right]\\ &+\sqrt{n}\left(\mathbf{\Sigma}_{yy}^{-1/2} \otimes I_p\right)\\ &\times \mathrm{vec}\left[\left(S_{xx}^{-1/2}-\mathbf{\Sigma}_{xx}^{-1/2}\right)\mathbf{\Sigma}_{xy}\right]\\ &+\sqrt{n}\left(\mathbf{\Sigma}_{xx}^{-1/2} \otimes \mathbf{\Sigma}_{yy}^{-1/2}\right)\\ &\times \mathrm{vec}\left(S_{xy}-\mathbf{\Sigma}_{xy}\right)\\ &+O_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned} \tag{27}$$

Defining

$$\mathbf{\Psi} = \left[\left(I_q \otimes \mathbf{\Sigma}_{xx}^{-1/2}\right), \left(\mathbf{\Sigma}_{yy}^{-1/2} \otimes I_p\right), \left(\mathbf{\Sigma}_{xx}^{-1/2} \otimes \mathbf{\Sigma}_{yy}^{-1/2}\right)\right],$$

it follows that

$$\begin{aligned}\sqrt{n}\mathrm{vec}\left(\hat{C}-C\right) = \sqrt{n}\mathbf{\Psi}&\begin{bmatrix}\mathrm{vec}\left[\mathbf{\Sigma}_{xy}\left(S_{yy}^{-T/2}-\mathbf{\Sigma}_{yy}^{-T/2}\right)\right]\\ \mathrm{vec}\left[\left(S_{xx}^{-1/2}-\mathbf{\Sigma}_{xx}^{-1/2}\right)\mathbf{\Sigma}_{xy}\right]\\ \mathrm{vec}\left(S_{xy}-\mathbf{\Sigma}_{xy}\right)\end{bmatrix}\\ &+O_p\left(\frac{1}{\sqrt{n}}\right)\end{aligned} \tag{28}$$

Denoting by $\mathbf{\Phi}$ the covariance

$$\mathbf{\Phi} = \lim_{n\to\infty}\mathrm{cov}\begin{bmatrix}\sqrt{n}\mathrm{vec}\left[\mathbf{\Sigma}_{xy}\left(S_{yy}^{-T/2}-\mathbf{\Sigma}_{yy}^{-T/2}\right)\right]\\ \sqrt{n}\mathrm{vec}\left[\left(S_{xx}^{-1/2}-\mathbf{\Sigma}_{xx}^{-1/2}\right)\mathbf{\Sigma}_{xy}\right]\\ \sqrt{n}\mathrm{vec}\left(S_{xy}-\mathbf{\Sigma}_{xy}\right)\end{bmatrix} \tag{29}$$

The asymptotic distribution of $\hat{C}$ follows directly as

$$\sqrt{n}\mathrm{vec}\left(\hat{C}-C\right) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Psi}\mathbf{\Phi}\mathbf{\Psi}^{\top}).$$

■

APPENDIX B
PROOF OF PROPOSITION 2 CONSISTENCY OF $\hat{F}_r$ and $\hat{G}_r$

Proof: From proposition 1, $\hat{C}$ is a $\sqrt{n}$-consistent estimator of C and it can be expressed as

$$\hat{C} = C + \epsilon\frac{\hat{C}-C}{\epsilon} = C + \epsilon D \tag{30}$$

where the perturbation of C is of order $1/\sqrt{n}$ [34], [38], [48]; that is $\epsilon D = O_p\left(\frac{1}{\sqrt{n}}\right)$. Therefore

$$\begin{aligned}\hat{C}G_rK^{-1} &= (C+\epsilon D)\,G_rK^{-1}\\ &= F_r + \epsilon D G_r K^{-1}\end{aligned} \tag{31}$$

and

$$\begin{aligned}K^{-1}F_r^{\top}\hat{C} &= K^{-1}F_r^{\top}(C+\epsilon D)\\ &= G_r^{\top} + \epsilon K^{-1}F_r^{\top}D\end{aligned} \tag{32}$$

Since both matrices F_r and G_r are orthogonal and non-random and as K^{-1} is bounded because $k_r > \epsilon$, it implies that both second terms of (31) and (32) are still bounded by $O_p(\frac{1}{\sqrt{n}})$. Then $\hat{F}_r = F_r + O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\hat{G}_r = G_r + O_p\left(\frac{1}{\sqrt{n}}\right)$, so that both $\hat{F}_r$ and $\hat{G}_r$ are $\sqrt{n}$-consistent estimators for F_r and G_r, respectively. ■

APPENDIX C
PROOF OF COROLLARY 1 CONSISTENCY OF THE ORTHOGONAL PROJECTION MATRICES

Proof: From proposition 2, $\hat{F}_r = F_r + O_p\left(\frac{1}{\sqrt{n}}\right)$, therefore

$$\begin{aligned}\hat{F}_{p-r}\hat{F}_{p-r}^{\top} &= I_p - \hat{F}_r\hat{F}_r^{\top}\\ &= I_p - \left[F_r + O_p\left(\frac{1}{\sqrt{n}}\right)\right]\left[F_r + O_p\left(\frac{1}{\sqrt{n}}\right)\right]^{\top}\\ &= I_p - F_rF_r^{\top} + O_p\left(\frac{1}{\sqrt{n}}\right)\\ &= F_{p-r}F_{p-r}^{\top} + O_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned} \tag{33}$$

The proof of (16) is similar. ■

APPENDIX D
PROOF OF PROPOSITION 3 ASYMPTOTIC DISTRIBUTION OF VEC $\left(\hat{Q}_F\hat{C}\hat{Q}_G\right)$

Proof: Using Corollary 1, we have

$$\begin{aligned}\mathrm{vec}\left(\hat{Q}_F\hat{C}\hat{Q}_G\right) &= \mathrm{vec}\left(\hat{Q}_F\hat{C}(\hat{Q}_G-Q_G)\right)\\ &\quad+\mathrm{vec}\left(\hat{Q}_F\hat{C}Q_G\right)\\ &= \mathrm{vec}\left(\hat{Q}_F\hat{C}Q_G\right) + O_p\left(\frac{1}{\sqrt{n}}\right)\\ &= \mathrm{vec}\left((\hat{Q}_F-Q_F)\hat{C}Q_G\right)\\ &\quad+\mathrm{vec}\left(Q_F\hat{C}Q_G\right) + O_p\left(\frac{1}{\sqrt{n}}\right)\\ &= \mathrm{vec}\left(Q_F\hat{C}Q_G\right) + O_p\left(\frac{1}{\sqrt{n}}\right)\\ &= \mathrm{vec}\left(Q_F(\hat{C}-C)Q_G\right) + O_p\left(\frac{1}{\sqrt{n}}\right)\end{aligned} \tag{34}$$

where the last line is obtained using that fact that $Q_FCQ_G = 0$. Therefore vec $\left(\hat{Q}_F\hat{C}\hat{Q}_G\right)$ has the same asymptotic distribution as vec $\left(Q_F\hat{C}Q_G\right)$.

Thus, we have:

$$\begin{aligned}\sqrt{n}\mathrm{vec}\left(\hat{Q}_F\hat{C}\hat{Q}_G\right) &= \mathrm{vec}\left(Q_F\hat{C}Q_G\right) + O_p(1)\\ &= \mathrm{vec}\left(Q_F(\hat{C}-C)Q_G\right) + O_p(1)\end{aligned}$$

$$= (Q_G \otimes Q_F)\sqrt{n}\text{vec}(\hat{C} - C) + O_p(1). \quad (35)$$

Using (35) we obtain

$$E\left\{\text{vec}\left(\hat{Q}_F\hat{C}\hat{Q}_G\right)\right\} = (Q_G \otimes Q_F)\sqrt{n}E\left\{\text{vec}(\hat{C} - C)\right\} = 0$$

and from Appendix A we have $E\left\{\text{cov}(\hat{C} - C)\right\} = \mathbf{\Omega}$, therefore:

$$\text{cov}\left\{\text{vec}\left(\hat{Q}_F\hat{C}\hat{Q}_G\right)\right\} = (Q_G \otimes Q_F)\mathbf{\Omega}(Q_G \otimes Q_F)$$

■

APPENDIX E
PROOF OF COROLLARY 2 DISTRIBUTION OF PROPOSED GOODNESS-OF-FIT

Proof: Note that

$$\begin{aligned} L(\hat{C}, r) &= n\text{vec}(K_{s-r})^\top \text{vec}(K_{s-r}) \\ &= n\text{tr}\left(\hat{K}_{s-r}^2\right) \\ &= n\text{tr}\left(\hat{F}_{p-r}^\top \hat{C}\hat{G}_{q-r}\hat{G}_{q-r}^\top \hat{C}\hat{F}_{p-r}\right) \\ &= n\text{tr}\left(\hat{Q}_F\hat{C}\hat{Q}_G\hat{C}\right) \\ &= \text{tr}\left(\| \sqrt{n}\hat{Q}_F\hat{C}\hat{Q}_G \|_F^2\right) \\ &= \| \sqrt{n}\text{vec}\left(\hat{Q}_F\hat{C}\hat{Q}_G\right) \|_F^2 \end{aligned}$$

Proposition 3 shows that $\sqrt{n}\text{vec}\left(\hat{Q}_F\hat{C}\hat{Q}_G\right)$ is normal with with covariance $\mathbf{\Omega}$. Using $\mathbf{\Omega}$ as defined in Proposition 3, it is easy to show that the alternative measure $\tilde{L}(\hat{C}, r)$

$$\begin{aligned} \tilde{L}(\hat{C}, r) &= n\text{vec}\left(\hat{Q}_F\hat{C}\hat{Q}_G\right)^\top \mathbf{\Omega}^{-1}\text{vec}\left(\hat{Q}_F\hat{C}\hat{Q}_G\right) \\ &= \|\sqrt{n}\mathbf{\Omega}^{-1/2}\text{vec}\left(\hat{Q}_F\hat{C}\hat{Q}_G\right)\|_F^2, \end{aligned} \quad (36)$$

has all equal non-zero unit eigenvalues, and therefore $\tilde{L}(\hat{C}, r)$ is $\chi^2_{(p-r)(q-r)}$ distributed. However, since $\mathbf{\Omega}$ is difficult to compute, we are motivated to use an alternative approach. We know that as $n \to \infty$ the goodness-of-fit term $\| \sqrt{n}\text{vec}\left(\hat{Q}_F\hat{C}\hat{Q}_G\right) \|_F^2$ follows a wighted mixture of χ_1^2 distributions $\sum_{i=1}^{\gamma} \omega_i\chi_1^2$, where ω_i are the eigenvalues of $\mathbf{\Omega}$ and $\gamma = (p - r)(q - r)$ corresponds to the degrees-of-freedom determined by the size of the null spaces of F and G. To find an approximation for the mean of the statistics, we use the Gamma function approximation [49]:

$$\begin{aligned} \sum_{i=1}^{\gamma} \omega_i\chi_1^2 &= \sum_{i=1}^{\gamma} \omega_i\Gamma\left(\frac{1}{2}, 2\right) = \sum_{i=1}^{\gamma} \Gamma\left(\frac{1}{2}, 2\omega_i\right) \\ &= \sum_{i=1}^{\gamma} \Gamma(K_i, \theta_i) \approx \Gamma(K, \theta) \end{aligned} \quad (37)$$

where [49]

$$K = \frac{(\sum \theta_i K_i)^2}{\sum \theta_i^2 K_i} \quad \text{and} \quad \theta = \frac{\sum \theta_i K_i}{K}.$$

for all $i = 1, \ldots, \gamma$, the values $K_i = \frac{1}{2}$ and $\theta_i = 2\omega_i$, which we assume are equal to further simplify the derivations. The mean of the asymptotic approximation of the distribution of $L(\hat{C}, r)$ is therefore given by $K\theta$. This is the quantity that has to be used to penalize the sum of the smallest canonical correlations. It is easy to show if the smallest canonical correlations are relatively close, the mean of the distribution is proportional to $\gamma = (p - r)(q - r)$. In Section III it is empirically shown that this approximation is reasonable. ■

APPENDIX F
PROOF OF PROPOSITION 4 CONSISTENCY OF PROPOSED CRITERION

Proof: In order to prove the strong consistency of

$$\hat{r} = \arg\min_r IC(r)$$

we first consider the case where $\hat{r} > r_0$, where r_0 is the true dimension:

$$\begin{aligned} IC(\hat{r}) - IC(r_0) &= L(\hat{C}, \hat{r}) - L(\hat{C}, r_0) \\ &\quad + c_n(p - \hat{r})(q - \hat{r}) \\ &\quad - c_n(p - r_0)(q - r_0) \\ &= L(\hat{C}, \hat{r}) - L(\hat{C}, r_0) \\ &\quad + c_n(\hat{r} - r_0)(\hat{r} + r_0 - p - q) \end{aligned} \quad (38)$$

and assuming that the law of iterative logarithms holds [50], then

$$\begin{aligned} IC(\hat{r}) - IC(r_0) &= O(\log\log(n)) \\ &\quad + c_n(\hat{r} - r_0)(\hat{r} + r_0 - p - q) \\ &= \log\log(n)\left[O(1)\right. \\ &\quad \left. + \frac{c_n(\hat{r} + r_0 - p - q)(\hat{r} - r_0)}{\log\log(n)}\right] \\ &\to \infty \end{aligned} \quad (39)$$

which implies that $IC(\hat{r}) > IC(r_0)$ a.s. Then for large n, $\hat{r}$ will not be bigger than r_0.
In case $\hat{r} < r_0$, we have

$$\begin{aligned} IC(\hat{r}) - IC(r_0) &= \tau n - c_n(\hat{r} + r_0 - p - q)(r_0 - \hat{r}) \\ &= n\left[\tau - \frac{c_n((\hat{r} + r_0 - p - q)(r_0 - \hat{r}))}{n}\right] \end{aligned} \quad (40)$$

gives $IC(\hat{r}) > IC(r_0)$, since $\tau = L(\hat{C}, \hat{r}) - L(\hat{C}, r_0) > 0$ and the second term on the right hand side in brackets converges to zero for large n. Combining both cases concludes the proof [41]. ■

REFERENCES

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.
[2] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.

[3] M. U. Khalid and A.-K. Seghouane, "Improving functional connectivity detection in fMRI by combining sparse dictionary learning and canonical correlation analysis," in *Proc. IEEE 10th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2013, pp. 286–289.
[4] M. A. A. Qadar and A.-K. Seghouane, "A projection CCA method for effective fMRI data analysis," *IEEE Trans. Biomed. Eng.*, to be published.
[5] M. A. Qadar, A. Aïssa-El-Bey, and A. K. Seghouane, "Two dimensional CCA via penalized matrix decomposition for structure preserved fMRI data analysis," *Digit. Signal Process.*, vol. 92, pp. 36–46, Sep. 2019.
[6] A. Pezeshki, M. R. Azimi-Sadjadi, and L. L. Scharf, "Undersea target classification using canonical correlation analysis," *IEEE J. Ocean. Eng.*, vol. 32, no. 4, pp. 948–955, Oct. 2007.
[7] J. D. Tucker and M. R. Azimi-Sadjadi, "Coherence-based underwater target detection from multiple disparate sonar platforms," *IEEE J. Ocean. Eng.*, vol. 36, no. 1, pp. 37–51, Jan. 2011.
[8] L. L. Scharf and J. K. Thomas, "Wiener filters in canonical coordinates for transform coding, filtering, and quantizing," *IEEE Trans. Signal Process.*, vol. 46, no. 3, pp. 647–654, Mar. 1998.
[9] J. Via, I. Santamaria, and J. Perez, "Deterministic CCA-based algorithms for blind equalization of FIR-MIMO channels," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3867–3878, Jul. 2007.
[10] A. K. Falcone, M. R. Azimi-Sadjadi, and J. A. Kankiewicz, "Dual-satellite cloud product generation using temporally updated canonical coordinate features," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 1046–1060, Apr. 2007.
[11] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 129–136.
[12] F. Zhou and F. De la Torre, "Generalized canonical time warping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 279–294, Feb. 2016.
[13] Y. Fujikoshi and L. G. Veitch, "Estimation of dimensionality in canonical correlation analysis," *Biometrika*, vol. 66, pp. 345–351, Aug. 1979.
[14] B. K. Gunderson and R. J. Muirhead, "On estimating the dimensionality in canonical correlation analysis," *J. Multivariate Anal.*, vol. 62, pp. 121–136, Jul. 1997.
[15] C. Lameiro and P. J. Schreier, "Cross-validation techniques for determining the number of correlated components between two data sets when the number of samples is very small," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2016, pp. 601–605.
[16] M. S. Bartlett, "Multivariate analysis," *Suppl. J. Roy. Stat. Soc.*, vol. 9, no. 2, pp. 176–197, 1947.
[17] D. Lawley, "Tests of significance in canonical analysis," *Biometrika*, vol. 46, nos. 1–2, pp. 59–66, 1959.
[18] S. S. Wilks, "Certain generalizations in the analysis of variance," *Biometrika*, vol. 24, nos. 3–4, pp. 471–494, 1932.
[19] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.
[20] A.-K. Seghouane and S.-I. Amari, "The AIC criterion and symmetrizing the Kullback–Leibler divergence," *IEEE Trans. Neural Netw.*, vol. 18, no. 1, pp. 97–106, Jan. 2007.
[21] A.-K. Seghouane, "Asymptotic bootstrap corrections of AIC for linear regression models," *Signal Process.*, vol. 90, no. 1, pp. 217–224, Jan. 2010.
[22] A.-K. Seghouane, "New AIC corrected variants for multivariate linear regression model selection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 47, no. 2, pp. 3012–3021, Apr. 2011.
[23] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[24] C. L. Mallows, "Some comments on CP," *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.
[25] Q. T. Zhang and K. M. Wong, "Information theoretic criteria for the determination of the number of signals in spatially correlated noise," *IEEE Trans. Signal Process.*, vol. 41, no. 4, pp. 1652–1663, Apr. 1993.
[26] A. Pezeshki, L. L. Scharf, J. K. Thomas, and B. D. Van Veen, "Canonical coordinates are the right coordinates for low-rank Gauss–Gauss detection and estimation," *IEEE Trans. Signal Process.*, vol. 54, no. 12, pp. 4817–4820, Dec. 2006.
[27] Y. Wang, H. Wang, and L. L. Scharf, "Scaled canonical coordinates for compression and transmission of noisy sensor measurements," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, 2013, pp. 409–413.
[28] I. Santamaria and J. Via, "Estimation of the magnitude squared coherence spectrum based on reduced-rank canonical coordinates," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 3, 2007, p. III-985.
[29] X. Wang, H. Ge, and I. P. Kirsteins, "Direction-of-arrival estimation using distributed arrays: A canonical coordinates perspective with limited array size and sample support," in *Proc. IEEE Int. Conf. Acoustic, Speech Signal Process. (ICASSP)*, Mar. 2010, pp. 2622–2625.
[30] J. Benesty, J. Chen, and Y. A. Huang, "A generalized MVDR spectrum," *IEEE Signal Process. Lett.*, vol. 12, no. 12, pp. 827–830, Dec. 2005.
[31] A. Bertrand and M. Moonen, "Distributed canonical correlation analysis in wireless sensor networks with application to distributed blind source separation," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4800–4813, Sep. 2015.
[32] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.
[33] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. Hoboken, NJ, USA: Wiley, 1982.
[34] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York, NY, USA: Wiley, 1958.
[35] P. Stoica, K. M. Wong, and Q. Wu, "On a nonparametric detection method for array signal processing in correlated noise fields," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 1030–1032, Apr. 1996.
[36] Y. Song, P. J. Schreier, D. Ramírez, and T. Hasija, "Canonical correlation analysis of high-dimensional data with very small sample support," *Signal Process.*, vol. 128, pp. 449–458, Nov. 2016.
[37] T. W. Anderson, "The asymptotic distribution of certain characteristic roots and vectors," in *Proc. 2nd Berkeley Symp. Math. Statist. Probab.*, J. Neyman, Ed. 1951, pp. 103–130.
[38] Z. Ratsimalahelo, "Rank test based on matrix perturbation theory," UFR Sci. Economique, Univ. Franche-Comté, Besançon, France, Tech. Rep. 04/2001, 2001.
[39] E. Bura and J. Yang, "Dimension estimation in sufficient dimension reduction: A unifying approach," *J. Multivariate Anal.*, vol. 102, no. 1, pp. 130–142, 2011.
[40] A.-K. Seghouane and N. Shokouhi, "Consistent estimation of dimensionality for data-driven methods in fMRI analysis," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 493–503, Feb. 2018.
[41] R. Nishii, "Maximum likelihood principle and model selection when the true model is unspecified," *J. Multivariate Anal.*, vol. 27, no. 2, pp. 392–403, 1988.
[42] J. G. Cragg and S. G. Donald, "Inferring the rank of a matrix," *J. Econometrics*, vol. 76, nos. 1–2, pp. 223–250, 1997.
[43] H. Ge, I. P. Kirsteins, and X. Wang, "Does canonical correlation analysis provide reliable information on data correlation in array processing?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2009, pp. 2113–2116.
[44] D. R. Brillinger, *Time Series: Data Analysis and Theory*, vol. 36. Philadelphia, PA, USA: SIAM, 1981.
[45] J. Choi, I. Song, and H. M. Kim, "On estimating the direction of arrival when the number of signal sources is unknown," *Signal Process.*, vol. 34, pp. 193–205, Nov. 1993.
[46] A. Khodayari-Rostamabad and S. Valaee, "Information theoretic enumeration and tracking of multiple sources," *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2765–2773, Jun. 2007.
[47] L. L. Scharf and C. Demeure, *Statistical Signal processing: Detection, Estimation, and Time Series Analysis*, vol. 63. Reading, MA, USA: Addison-Wesley, 1991.
[48] T. Kato, *Perturbation Theory for Linear Operators*. New York, NY, USA: Springer-Verlag, 1966.
[49] D. R. Jensen and H. Solomon, "A Gaussian approximation to the distribution of a definite quadratic form," *J. Amer. Stat. Assoc.*, vol. 67, no. 340, pp. 898–902, 1972.
[50] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On detection of the number of signals in presence of white noise," *J. Multivariate Anal.*, vol. 20, no. 1, pp. 1–25, 1986.

ABD-KRIM SEGHOUANE received the Ph.D. degree in signal processing and control from Universite Paris Sud (Paris XI), France, in 2003.

From 2005 to 2012, he was a Researcher and subsequently a Senior Researcher with National ICT Australia (NICTA) Canberra Research Laboratory and an Adjunct Faculty with the College of Engineering and Computer Science, Australian National University (ANU). Prior to that, he was a Postdoctoral Researcher with INRIA Rocquencourt, France. He has been on the EEE Faculty with The University of Melbourne, since 2013. His research has been recognized via the Australian Research Council Future Fellowship. His research interests include statistical signal and image processing, including the application areas of medical imaging and physiological signal analysis.

Dr. Seghouane is currently an Elected Member of the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee (MLSPTC) elected first, in 2014 and re-elected, in 2017 for another term of three years. He was the Chair of the MLSP TC Data Competition Sub-committee, from 2014 to 2017. He was the General Co-Chair of the 2014 IEEE Workshop on Statistical Signal Processing, Gold Coast, Australia, the Data Competition Co-Chair of the 2017 and 2018 IEEE Workshop on Machine Learning for Signal Processing, Tokyo, Japan and Aalborg, Denmark, and the Organization Co-Chair of the 2017 IEEE International Symposium on Biomedical Imaging, Melbourne, Australia. He has served as an Associate Editor on the editorial board for the IEEE TRANSACTIONS ON IMAGE PROCESSING, from 2014 to 2018; and since 2018, serves as a Senior Editor of the area. He also serves on the editorial board of the IEEE TRANSACTIONS ON SIGNAL PROCESSING as an Associate Editor, since 2017.

NAVID SHOKOUHI received the bachelor's degree in electrical engineering from the Amirkabir University of Technology (Tehran Polytechnique), Iran, in 2011, and the Ph.D. degree in signal processing from The University of Texas at Dallas, USA, in 2016. Prior to that, he was a Postdoctoral Researcher with the Department of Electrical and Electronic Engineering, The University of Melbourne, from 2017 to 2018. He is currently a Machine Learning Researcher with Global Kinetics, Melbourne, Australia. His Ph.D. research interests include multispeaker recognition and diarization, which entails developing robust speaker recognition algorithms for co-channel signals. His Postdoctoral and current research interests include applying machine learning and signal processing techniques in image processing and spoken language technology.