Improved Estimation of Canonical Vectors in Canonical Correlation Analysis

Nicholas Asendorf
Department of Electrical and Computer Engineering
University of Michigan
Ann Arbor, Michigan 48105
Email: asendorf@umich.edu

Raj Rao Nadakuditi
Department of Electrical and Computer Engineering
University of Michigan
Ann Arbor, Michigan 48105
Email: rajnrao@umich.edu

***Abstract*—Canonical Correlation Analysis (CCA) is a multidimensional algorithm for two datasets that finds linear transformations, called canonical vectors, that maximize the correlation between the transformed datasets. However, in the low-sample high-dimension regime these canonical vector estimates are extremely inaccurate. We use insights from random matrix theory to propose a new algorithm that can reliably estimate canonical vectors in the sample deficient regime. Through numerical simulations we showcase that our new algorithm is robust to both limited training data and overestimating the dimension of the signal subspaces.**

I. Introduction

Canonical correlation analysis (CCA) is a joint multidimensional dimensionality reduction algorithm for two exactly two datasets [1]. CCA finds a linear transformation for each dataset such that the correlation between the two transformed features is maximized. The basis vectors that span the subspaces of these transformations are called canonical vectors. While CCA itself is not a data fusion algorithm, the correlated features that it returns may be used in data fusion algorithms. Such data fusion algorithms are becoming a necessity with the increased ability to capture high-dimensional multi-modal datasets, arising in fields such as computer vision [2]–[5] and medical signal processing [6]–[9].

However, in many applications the covariance matrices needed to solve CCA are unknown and must be estimated from training data. When using sample covariance estimates from fewer samples than the combined dimensions of the datasets, empirical CCA falsely reports a perfect correlation between the datasets and random linear transformations for each dataset [10]. In this low-sample, high-dimensionality regime, empirical CCA fails to reliably identify correlations between the datasets. However, using insights from random matrix theory, Nadakuditi [11] proposed informative CCA (ICCA), which overcomes this performance loss to reliably identify correlations in the sample deficient regime.

In this paper, we consider the accuracy of the canonical vectors returned by empirical CCA and ICCA and propose an improved estimate that we name ICCA+. Throughout, we assume that each dataset is modeled with a low-rank signal-plus-noise data model, which is ubiquitous in signal processing applications. We begin by deriving the CCA population canonical vectors if all parameters are known. From this analysis we see that the canonical vectors are a linear combination of signal vectors that form the linear signal subspace of each dataset. Empirical CCA, ICCA, and ICCA+ all use different weights of an estimated signal subspace. Importantly, ICCA+ relies on insights from random matrix theory that quantify the accuracy of the signal subspace estimate, resulting in the most accurate estimate. While the ICCA canonical vectors are suboptimal, they greatly improve upon the canonical vectors of empirical CCA, which are random in the low-sample low-SNR regime.

This paper is organized as follows. We provide the linear low-rank signal-plus-noise data model in Section II. We then derive the solution of CCA and empirical CCA in Section III. In Section IV, we provide the necessary insights from random matrix theory to derive ICCA and our new algorithm ICCA+ in sections V and VI, respectively. We validate our main results on synthetic data in Section VII and provide concluding remarks in Section VIII.

II. Data Model

Throughout this paper, we consider the follow ubiquitous low-rank signal-plus-noise data model. We assume that we are given n observations of each dataset, which we stack columnwise to form two data matrices

$$X = [x_1, \dots, x_n], \;\; Y = [y_1, \dots, y_n]. \tag{1}$$

For $i = 1, \dots, n$, let $x_i \in \mathbb{C}^{p\times 1}$ and $y_i \in \mathbb{C}^{q\times 1}$ be modeled as

$$x_i = U_x s_{x,i} + z_{x,i}, \;\; y_i = U_y s_{y,i} + z_{y,i}, \tag{2}$$

where $U_x^H U_x = I_{k_x}$, $U_y^H U_y = I_{k_y}$, $z_{x,i} \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0, I_p)$ and $z_{y,i} \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0, I_q)$. Furthermore, assume that

$$s_{x,i} \sim \mathcal{CN}(0, \Theta_x), \;\; s_{y,i} \sim \mathcal{CN}(0, \Theta_y),$$

where

$$\Theta_x = \mathbf{diag}\left(\left(\theta_1^{(x)}\right)^2, \dots, \left(\theta_{k_x}^{(x)}\right)^2\right) \tag{3a}$$

$$\Theta_y = \mathbf{diag}\left(\left(\theta_1^{(y)}\right)^2, \dots, \left(\theta_{k_y}^{(y)}\right)^2\right). \tag{3b}$$

Assume that $z_{x,i}$ and $z_{y,i}$ are mutually independent and independent from both $s_{x,i}$ and $s_{y,i}$. Finally, assume that

$$\mathbb{E}\left[s_{x,i} s_{y,i}^H\right] =: K_{xy} = \Theta_x^{1/2} P_{xy} \Theta_y^{1/2},$$

where the entries of P_{xy} are $-1 \leq |\rho_{kj}| \leq 1$ and represent the correlation between $s_{x,i}^{(k)}$ and $s_{y,i}^{(j)}$. For reasons to be made clear later, define

$$\widetilde{K}_{xy} = \left(\Theta_x + I_{k_x}\right)^{-1/2} K_{xy} \left(\Theta_y + I_{k_y}\right)^{-1/2}. \tag{4}$$

Under this model, we define the following covariance matrices

$$\mathbb{E}\left[x_i x_i^H\right] = U_x \Theta_x U_x^H + I_p =: R_{xx} \tag{5a}$$

$$\mathbb{E}\left[y_i y_i^H\right] = U_y \Theta_y U_y^H + I_q =: R_{yy} \tag{5b}$$

$$\mathbb{E}\left[x_i y_i^H\right] = U_x K_{xy} U_y^H =: R_{xy}. \tag{5c}$$

Note that we may write $s_{x,i} = \Theta_x^{1/2} v_{x,i}$ and $s_{y,i} = \Theta_y^{1/2} v_{y,i}$ where $v_{x,i} \sim \mathcal{CN}(0, I_{k_x})$ and $v_{y,i} \sim \mathcal{CN}(0, I_{k_y})$ are independent random vectors. Defining $Z_n^x = [z_{x,1}, \ldots, z_{x,n}]$, $Z_n^y = [z_{y,1}, \ldots, z_{y,n}]$, $V_x = [v_{x,1}, \ldots, v_{x,n}]$, and $V_y = [v_{y,1}, \ldots, v_{y,n}]$, we may write our data matrices in (1) as the sum of a low-rank signal matrix and noise matrix

$$X = U_x \Theta_x^{1/2} V_x^H + Z_n^x, \quad Y = U_y \Theta_y^{1/2} V_y^H + Z_n^y. \tag{6}$$

Finally, denote the singular values of Z_n^x and Z_n^y as

$$\sigma_1(Z_n^x) \geq \cdots \geq \sigma_p(Z_n^x), \quad \sigma_1(Z_n^y) \geq \cdots \geq \sigma_q(Z_n^y)$$

where without loss of generality we let $p < n$ and $q < n$ to simplify the definition of the empirical singular value distribution. Let $\mu_{Z_n^x}$ and $\mu_{Z_n^y}$ be the empirical singular value distribution defined as

$$\mu_{Z_n^x} = \frac{1}{p} \sum_{i=1}^{p} \delta_{\sigma_i(Z_n^x)}, \quad \mu_{Z_n^y} = \frac{1}{q} \sum_{i=1}^{q} \delta_{\sigma_i(Z_n^y)}.$$

Assume that the probability measures $\mu_{Z_n^x}$ and $\mu_{Z_n^y}$ converge almost surely as $p, q, n \to \infty$ with $p/n \to c_x$ and $q/n \to c_y$ to non-random compactly supported probability measures μ_{Z_x} and μ_{Z_y} respectively. Finally, we assume that $\sigma_1(Z_n^x) \xrightarrow{\text{a.s.}} b_x$ and $\sigma_1(Z_n^y) \xrightarrow{\text{a.s.}} b_y$.

III. Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a dimensionality reduction algorithm that finds linear projections for x_i and y_i such that in the projected spaces, the variables are maximally correlated. Specifically, CCA solves the following optimization problem

$$\rho_{\text{cca}} = \max_{w_x, w_y} \frac{w_x^H R_{xy} w_y}{\sqrt{w_x^H R_{xx} w_x} \sqrt{w_y^H R_{yy} w_y}}, \tag{7}$$

where w_x and w_y are called canonical vectors and ρ_{cca} is called the canonical correlation coefficient. Notice that we can scale w_x and w_y and still achieve the same objective function. Therefore, we may constrain the canonical variates to have unit norm, resulting in

$$\begin{aligned} \max_{w_x, w_y} \quad & w_x^H R_{xy} w_y \\ \text{subject to} \quad & w_x^H R_{xx} w_x = 1, w_y^H R_{yy} w_y = 1. \end{aligned} \tag{8}$$

Substituting the change of variables $\widetilde{w}_x = R_{xx}^{1/2} w_x$ and $\widetilde{w}_y = R_{yy}^{1/2} w_y$ in (8) results in the following optimization problem

$$\begin{aligned} \max_{\widetilde{w}_x, \widetilde{w}_y} \quad & \widetilde{w}_x^H R_{xx}^{-1/2} R_{xy} R_{yy}^{-1/2} \widetilde{w}_y \\ \text{subject to} \quad & \widetilde{w}_x^H \widetilde{w}_x = 1, \widetilde{w}_y^H \widetilde{w}_y = 1. \end{aligned} \tag{9}$$

Examining the optimization problem in (9), we can immediately see that the solution to CCA may be solved via the SVD of the matrix

$$C_{\text{cca}} = R_{xx}^{-1/2} R_{xy} R_{yy}^{-1/2} \tag{10}$$

Define $C_{\text{cca}} = FKG^T$ as the SVD of C_{cca} where F is an unitary $p \times p$ matrix with columns $f_1, \ldots, f_p$, G is a unitary $q \times q$ matrix with columns $g_1, \ldots, g_q$, and $K = \mathbf{diag}(k_1, \ldots, k_{\min(p,q)})$ is a $p \times q$ matrix whose diagonal elements are the singular values of C_{cca}. The solution to (9) is

$$\widetilde{w}_x = f_1, \quad \widetilde{w}_y = g_1, \quad \rho = k_1.$$

and the canonical vectors are

$$w_x = R_{xx}^{-1/2} \widetilde{w}_x, \quad w_y = R_{yy}^{-1/2} \widetilde{w}_y. \tag{11}$$

We can obtain higher order canonical correlations and vectors by taking successive singular value and vector pairs.

A. Population canonical vectors

We first determine the population canonical vectors of our data model in (2). To do so, we need the singular vectors of

$$\begin{aligned} C_{\text{cca}} &= R_{xx}^{-1/2} R_{xy} R_{yy}^{-1/2} \\ &= U_x \left(\Theta_x + I_{k_x}\right)^{-1/2} K_{xy} \left(\Theta_y + I_{k_y}\right)^{-1/2} U_y^H \\ &= U_x \widetilde{K}_{xy} U_y^H. \end{aligned}$$

Define $U_{\widetilde{K}} K_{\widetilde{K}} V_{\widetilde{K}}^H$ as the SVD of $\widetilde{K}_{xy}$. First observe that the rank of C_{cca} is $k =: \min(k_x, k_y)$. Defining the matrices of the canonical vectors $W_x = [w_x^{(1)}, \ldots, w_x^{(k)}]$ and $W_y = [w_y^{(1)}, \ldots, w_y^{(k)}]$, we have that

$$W_x = U_x \left(\Theta_x + I_{k_x}\right)^{-1/2} U_{\widetilde{K}} \tag{12a}$$

$$W_y = U_y \left(\Theta_y + I_{k_y}\right)^{-1/2} V_{\widetilde{K}}. \tag{12b}$$

Therefore, we see that the individual canonical vectors $w_x^{(i)}$ and $w_y^{(i)}$ are linear combinations of the signal subspaces U_x and U_y, respectively. The weights of these linear combinations are dependent on Θ_x, Θ_y, $U_{\widetilde{K}}$, and $V_{\widetilde{K}}$.

B. Empirical CCA canonical vector estimates

In many applications, we do not know the covariance matrices R_{xx}, R_{yy}, and R_{xy} *a priori*. Therefore, we cannot find the population canonical vectors in by forming C_{cca}, but instead must estimate this matrix from training data. Given the data matrices in (1), we form the sample covariance matrices $\widehat{R}_{xx} = \frac{1}{n} X X^H$, $\widehat{R}_{yy} = \frac{1}{n} Y Y^H$, and $\widehat{R}_{xy} = \frac{1}{n} X Y^H$. Define the data SVDs of the matrices in (1) as

$$X = \widehat{U}_x \widehat{\Sigma}_x \widehat{V}_y^H, \quad Y = \widehat{U}_y \widehat{\Sigma}_y \widehat{V}_y^H$$

and trimmed matrices

$$\widetilde{U}_x = \widehat{U}_x\left(:, 1:\min(p,n)\right),\ \widetilde{V}_x = \widehat{V}_x\left(:, 1:\min(p,n)\right)$$
$$\widetilde{U}_y = \widehat{U}_y\left(:, 1:\min(q,n)\right),\ \widetilde{V}_y = \widehat{V}_y\left(:, 1:\min(q,n)\right).$$

Substituting the SVDs of sample analogs of the matrices in (10), reveals the insight ([11, Eq. (6)]) that the matrix C_{cca} can be estimated as

$$\widehat{C}_{\text{cca}} = \widetilde{U}_x \widetilde{V}_x^H \widetilde{V}_y \widetilde{U}_y^H. \tag{14}$$

To estimate the population canonical vectors, we use the corresponding left and right singular vectors of $\widehat{C}_{\text{cca}}$, denoted by f_i and g_i respectively, and substitute the sample covariance matrices into (11) to achieve the estimates

$$\widehat{w}_{x,i}^{\text{cca}} = \widehat{R}_{xx}^{-1/2} f_i, \quad \widehat{w}_{y,i}^{\text{cca}} = \widehat{R}_{yy}^{-1/2} g_i. \tag{15}$$

Notice that the inner matrix product of $\widehat{C}_{\text{cca}}$ in (14) is $\widetilde{V}_x^H \widetilde{V}_y$. Define the SVD of this $\min(p,n) \times \min(q,n)$ matrix as $\widetilde{U}_{\widetilde{K}} \widetilde{K}_{\widetilde{K}} \widetilde{V}_{\widetilde{K}}^H$. Using this definition, the sample covariance matrices, and the identity $f_i = \widetilde{U}_x \widetilde{U}_{\widetilde{K}}(:,i)$, we have that under the data model in (2) the empirical CCA population canonical vector estimate is

$$\widehat{w}_{x,i}^{\text{cca}} = \widetilde{U}_x \widetilde{\Sigma}_x^{-1} \widetilde{U}_{\widetilde{K}}(:,i).$$

A similar expression may be found for $\widehat{w}_{y,i}^{\text{cca}}$. Stacking these empirical CCA canonical vectors estimates in a matrix yields

$$\widehat{W}_x^{\text{cca}} = \widetilde{U}_x \left(\widetilde{\Sigma}_x\right)^{-1} \widetilde{U}_{\widetilde{K}}, \quad \widehat{W}_y^{\text{cca}} = \widetilde{U}_y \left(\widetilde{\Sigma}_y\right)^{-1} \widetilde{V}_{\widetilde{K}}. \tag{16}$$

We note here that the singular values of the individual data matrices may be used to estimate the SNRs via $\widehat{\Theta}_x = \widetilde{\Sigma}_x^2 - I$ and $\widehat{\Theta}_y = \widetilde{\Sigma}_y^2 - I$.

IV. Pertinent Results from Random Matrix Theory

In empirical CCA, the canonical vector estimates in (16) use the full data SVDs, which assumes that the rank of underlying signals are of rank $\min(p,n)$ and $\min(q,n)$, respectively. This is obviously quite incorrect given our data model in (2). Important advances in random matrix theory allow us to quantify the accuracy of the singular values and singular vectors of the individual data matrices, X and Y, used in the empirical CCA canonical vector estimates. These insights from random matrix theory motivate a new algorithm to estimate the canonical vectors. In this section, we provide the pertinent results needed for our new estimation algorithm.

Let X be a data matrix as in (1) modeled as in (2). For this section, we drop the subscript dependent on x to ease notation and assume, without loss of generality, that $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_k$. Let $\widehat{U}\widehat{\Sigma}\widehat{V}^H$ be the SVD of X. We define the D-transform of a probability measure μ, depending on c, for $z > b$ as

$$D_\mu(z) =: \left[\int \frac{z}{z^2 - t^2} d_\mu(t)\right] \times \left[c_x \int \frac{z}{z^2 - t^2} d_{\mu(t)} + \frac{1 - c_x}{z}\right] \tag{17}$$

In the proposition below, $D_\mu^{-1}(\cdot)$ will denote its function inverse on $[b, +\infty)$.

Proposition IV.1. *Let $p, n \to \infty$ such that $p/n \to c$. Let $\widehat{u}_i$ and $\widehat{v}_i$ be the left and right singular vectors, respectively, associated with the singular value $\widehat{\sigma}_i$. Then for ach $1 \leq i \leq k$*

$$\alpha_i =: |\langle \widehat{u}_i, u_i \rangle|^2 \xrightarrow{a.s.} \begin{cases} \frac{-2\varphi_{\mu_Z}(\rho)}{\theta_i^2 D'_{\mu_Z}(\rho)} & \theta_i^2 > 1/D_{\mu_Z}(b^+) \\ 0 & \textit{otherwise} \end{cases} \tag{18}$$

where $\rho = D_{\mu_Z}^{-1}\left(1/\theta_i^2\right)$, $\widetilde{\mu}_Z = c\mu_Z + (1-c)\delta_0$, and for any probability measure μ,

$$\varphi_\mu(z) =: \int \frac{z}{z^2 - t^2} d\mu(t). \tag{19}$$

Proposition IV.1 gives asymptotic results for the accuracy of the singular vectors of a data matrix. These results rely on the D-transform of the noise matrix Z. This transform is invariant to orthogonal transformations, and so only depends on the singular values of Z. In a future journal paper, we will explain how to use the singular values of Z to estimate φ_μ, D_μ and D'_μ as similarly done in [12]. These estimates will allow us to estimate α_i given only data.

V. Informative CCA (ICCA)

As an alternative to empirical CCA, we consider informative CCA (ICCA) [11], an algorithm that first trims the singular vectors of the individual datasets to only include *informative* singular vectors. Let $\widehat{k}_x$ and $\widehat{k}_y$ be estimates of the number of informative components in X and Y respectively. Then define the trimmed data matrices

$$\mathring{U}_x = \widehat{U}_x\left(:, 1:\widehat{k}_x\right), \mathring{U}_y = \widehat{U}_y\left(:, 1:\widehat{k}_y\right) \tag{20a}$$

$$\mathring{V}_x = \widehat{V}_x\left(:, 1:\widehat{k}_x\right), \mathring{V}_y = \widehat{V}_y\left(:, 1:\widehat{k}_y\right) \tag{20b}$$

$$\mathring{\Sigma}_x = \widehat{\Sigma}_x\left(1:\widehat{k}_x, 1:\widehat{k}_x\right), \mathring{\Sigma}_y = \widehat{\Sigma}_y\left(1:\widehat{k}_y, 1:\widehat{k}_y\right). \tag{20c}$$

With these definitions, we define the ICCA matrix

$$\widehat{C}_{\text{icca}} = \mathring{U}_x \mathring{V}_x^H \mathring{V}_y \mathring{U}_y^H \tag{21}$$

with SVD $\widehat{U}_{\widetilde{K}} \widehat{\Sigma}_{\widetilde{K}} \widehat{V}_{\widetilde{K}}$. Substituting these definitions in (12) yields the ICCA population canonical vector estimates

$$\widehat{W}_x^{\text{icca}} = \mathring{U}_x \left(\widehat{\Theta}_x + I_{\widehat{k}_x}\right)^{-1/2} \widehat{U}_{\widetilde{K}} \tag{22a}$$

$$\widehat{W}_y^{\text{icca}} = \mathring{U}_y \left(\widehat{\Theta}_y + I_{\widehat{k}_y}\right)^{-1/2} \widehat{V}_{\widetilde{K}}. \tag{22b}$$

Similar to the population canonical vectors, the ICCA canonical vector estimates correctly take only a linear combination of a few signal vectors. The weighting is dependent on the estimates $\widehat{\Theta}_x$, $\widehat{\Theta}_y$, $\widehat{U}_{\widetilde{K}}$, and $\widehat{V}_{\widetilde{K}}$.

VI. ICCA+

We expect the ICCA canonical vector estimates in (22) to greatly outperform the empirical CCA estimates in (16). However, we still expect the estimates in (22) to be suboptimal because they substitute the signal subspace estimates, $\widehat{U}_x$ and $\widehat{U}_y$, without considering their accuracy, which we quantified in Proposition IV.1. The population, empirical CCA, and ICCA canonical vector estimates all take a linear combination of the known or estimated signal subspace. With this observation, we consider the following canonical vector estimates

$$\widehat{w}_{x,i}^{\text{icca+}} = \mathring{U}_x \, \mathbf{diag}\left(\lambda_{x,i}^{\text{opt}}\right) \widehat{U}_{\widetilde{K}}\left(:,i\right) \tag{23a}$$

$$\widehat{w}_{y,i}^{\text{icca+}} = \mathring{U}_y \, \mathbf{diag}\left(\lambda_{y,i}^{\text{opt}}\right) \widehat{V}_{\widetilde{K}}\left(:,i\right), \tag{23b}$$

where $\lambda_{x,i}^{\text{opt}} = \left[\lambda_{x,i}^{(1)}, \ldots, \lambda_{x,i}^{(k_x)}\right]$ and $\lambda_{y,i}^{\text{opt}} = \left[\lambda_{y,i}^{(1)}, \ldots, \lambda_{y,i}^{(k_y)}\right]$ and are the solutions to the following optimization problems

$$\lambda_{x,i}^{\text{opt}} = \underset{\lambda_x}{\text{argmin}} \left\| w_x^{(i)} - \widehat{U}_x \, \mathbf{diag}(\lambda_x) \widehat{U}_{\widetilde{K}}\left(:,i\right) \right\|_F \tag{24a}$$

$$\lambda_{y,i}^{\text{opt}} = \underset{\lambda_y}{\text{argmin}} \left\| w_y^{(i)} - \widehat{U}_y \, \mathbf{diag}(\lambda_x) \widehat{V}_{\widetilde{K}}\left(:,i\right) \right\|_F. \tag{24b}$$

This matrix approximation is similar to [12], which examines the optimal approximation to a signal matrix from noisy observations. Nadakuditi shows that the classical Eckart-Young-Mirsky (EYM) low-rank matrix approximation is suboptimal when trying to estimate a low-rank signal matrix from a low-rank signal-plus-noise matrix. The EYM approximation is the optimal low-rank approximation of the low-rank signal-plus-noise matrix but *not* the low-rank signal matrix. Similarly here, the ICCA estimates find the best representation of noisy canonical vectors and not the true underlying canonical vectors. Instead we want the optimal estimates of the population canonical vectors. The following proposition provides the closed form, deterministic answer to the optimization problem for the ICCA+ weights.

Proposition VI.1. *The solutions to (24) are given by*

$$\lambda_{x,i}^{opt} = \mathbf{diag}\left(\mathring{U}_x^H U_x \left(\Theta_x + I_{k_x}\right)^{-1/2}\right) \tag{25a}$$

$$\lambda_{y,i}^{opt} = \mathbf{diag}\left(\mathring{U}_y^H U_y \left(\Theta_y + I_{k_y}\right)^{-1/2}\right). \tag{25b}$$

The first key observation of this proposition is that the optimal weights are the same for each subsequent canonical vectors (i.e. independent of i). The second key observation is that the optimal weights are dependent on the matrix products $\mathring{U}_x^H U_x$ and $\mathring{U}_y^H U_y$. The diagonal elements of these matrices are the accuracies of the estimated components of our signal subspaces that we asymptotically quantified in Proposition IV.1. It makes sense then that the optimal weights tell us to place less weight on inaccurately estimated signal subspaces.

In a future journal paper, we will provide the asymptotic limit of the optimal weights in Proposition VI.1. The analysis relies on the asymptotic limit of the subspace accuracy provided in Proposition IV.1.

Proposition VI.1 highlights some key differences between the ICCA and ICCA+. While ICCA places positive weight on all subspace components, ICCA+will place place zero weight on any subspace component that is *uninformative*. We expect ICCA+ to perform better than ICCA in this uninformative regime since it places no weight on estimated subspaces that are simply noise.

VII. Empirical Results

In this section, we explore the accuracy of empirical CCA, ICCA, and ICCA+ canonical vectors. We consider a rank-2 setting where $k_x = k_y = 2$, $p = 200$, $q = 250$, $\Theta_x = \Theta_y = \mathbf{diag}(16, 4)$, $P_{xy} = \mathbf{diag}(0.9, 0.5)$, $V_K = I_2$, and

$$U_K = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}. \tag{26}$$

In this setup,

$$U_{\widetilde{K}} = \begin{bmatrix} -0.8559 & -0.5172 \\ -0.5172 & 0.8559 \end{bmatrix}.$$

In all experiments, we consider the accuracy of a canonical vector estimate, $\widehat{w}^{(i)}$ as

$$\text{accuracy}_i = \frac{|\langle \widehat{w}^{(i)}, w_x^{(i)} \rangle|^2}{\|w_x^{(i)}\|_2^2 \|\widehat{w}^{(i)}\|_2^2}. \tag{27}$$

All simulations average over 500 trials where each trial generates new noise matrices and signal matrices. We show results for only canonical vectors corresponding to X; similar results may be obtained for canonical vectors corresponding to Y. For ICCA+, we use an oracle value for the necessary parameters in Proposition VI.1. The future journal version of this paper will find data driven estimates for all parameters.

Figure 1 explores the effect of the number of samples on the canonical vectors. Here we see when $n < p+q$, the empirical CCA canonical vectors behave as if they are totally random. As n increases, the accuracy of the empirical CCA canonical vectors improves slowly. As the figure shows, both ICCA an ICCA+ are able to avoid the performance loss of empirical CCA, especially in the low sample regime. In this setup, we observe a similar performance for ICCA and ICCA+ across all values of n.

Figure 2 explores the effect of the value of Θ_x on the accuracy of the canonical vectors. As $n \approx p+q$, the canonical vectors of empirical CCA are essentially random and do not improve with an increase in SNR, which is very undesirable. However, ICCA and ICCA+ are able to avoid this performance loss and the accuracy of their canonical vectors increase with θ, as desired. Again, the performance of ICCA and ICCA+ are essentially identical.

Figure 3 explores the effect of the value of $\widehat{k}_x$ on the accuracy of the canonical vectors. Again, empirical CCA returns essentially random canonical vectors. The canonical vectors of both ICCA and ICCA+ improve upon empirical CCA, however, ICCA+ is more robust to overestimation of $\widehat{k}_x$. In this setup $k_x = 2$ and we see that ICCA+ is able to place less weight on these inaccurate signal subspaces that are included because $\widehat{k}_x$ overestimates k_x.

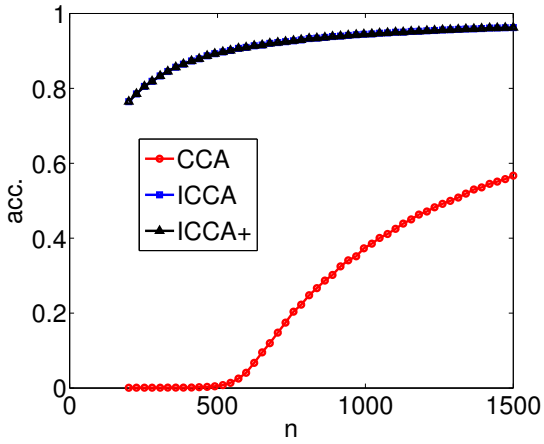

Fig. 1. Accuracy plots as a function of n for a rank-2 setting where $k_x = k_y = 2$, $p = 200$, $q = 250$, $\Theta_x = \Theta_y = \mathbf{diag}(16, 4)$, $P_{xy} = \mathbf{diag}(0.9, 0.5)$, $V_K = I_2$, and non-identity U_K defined in (26). Accuracy is defined in (27). We plot the accuracy of the first canonical vector for CCA, ICCA, and ICCA+.

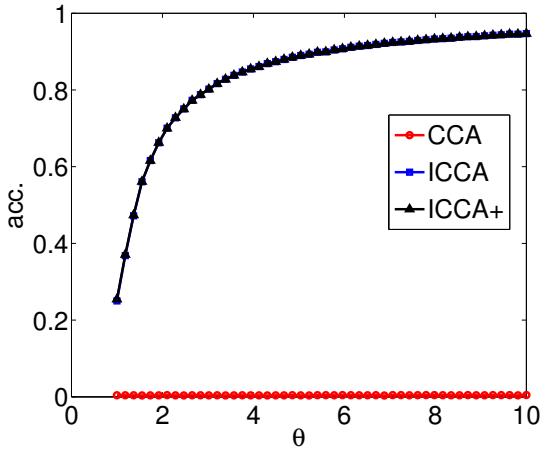

Fig. 2. This is the same setting as Figure 1 except we keep a fixed $n = 500$ and sweep θ such that $\Theta_x = \Theta_y = \mathbf{diag}(\theta, 0.8 \times \theta)$.

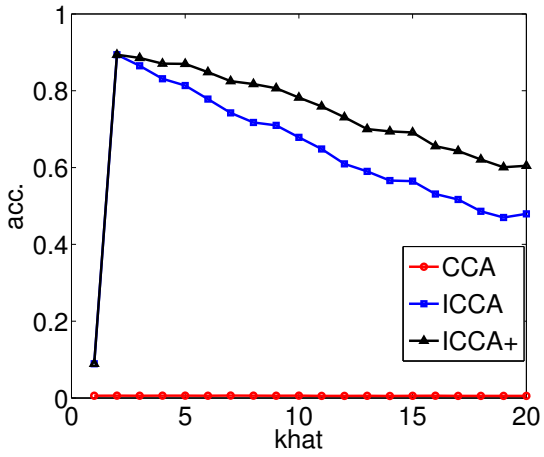

Fig. 3. This is the same setup as Figure 1 except we keep a fixed $n = 500$ and sweep of $\widehat{k}_x$.

VIII. Conclusion

In this paper we considered the problem of estimating canonical vectors in Canonical Correlation Analysis. We saw that in the low-sample, high dimensionality regime, the canonical vectors returned by empirical CCA are very inaccurate. We saw that Informative Canonical Correlation Analysis can overcome much of this performance loss by trimming data matrices to include only *informative* components. The canonical vectors of both empirical CCA and ICCA take a linear combination of the signal subspace. By optimizing this linear combination, our new algorithm, ICCA+, can systematically improve the accuracy of the canonical vectors and is more robust to overestimation of the rank of the signal subspace. In future work we will consider the theoretical performance of ICCA+, extend the algorithm to missing data, and provide rigorous proofs of all results.

Acknowledgment

This work was supported by ONR Young Investigator Award N000141110660, ONR Award N00014-15-1-2141, AFOSR Young Investigator Award FA9550-12-1-0266, NSF award CCF-1116115, ARfO MURI grant W911NF-11-1-0391, and ARO MURI grant W911NF-15-1-0479.

References

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

[2] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[3] P. Dhillon, D. P. Foster, and L. H. Ungar, "Multi-view learning of word embeddings via cca," in *Advances in Neural Information Processing Systems*, 2011, pp. 199–207.

[4] D. R. Hardoon, C. Saunders, S. Szedmak, and J. Shawe-Taylor, "A correlation approach for automatic image annotation," in *Advanced Data Mining and Applications*. Springer, 2006, pp. 681–692.

[5] K. Chaudhuri, S. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 129–136.

[6] M. U. Khalid and A.-K. Seghouane, "Improving functional connectivity detection in fmri by combining sparse dictionary learning and canonical correlation analysis," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*. IEEE, 2013, pp. 286–289.

[7] N. Correa, T. Adali, Y. Li, and V. Calhoun, "Canonical correlation analysis for data fusion and group inferences," *Signal Processing Magazine, IEEE*, vol. 27, no. 4, pp. 39–50, 2010.

[8] J. A. Seoane, C. Campbell, I. N. Day, J. P. Casas, and T. R. Gaunt, "Canonical correlation analysis for gene-based pleiotropy discovery," *PLoS computational biology*, vol. 10, no. 10, p. e1003876, 2014.

[9] M. Spuler, A. Walter, W. Rosenstiel, and M. Bogdan, "Spatial filtering based on canonical correlation analysis for classification of evoked or event-related potentials in eeg data," 2013.

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

[11] R. Nadakuditi, "Fundamental finite-sample limit of canonical correlation analysis based detection of correlated high-dimensional signals in white noise," in *Statistical Signal Processing Workshop (SSP), 2011 IEEE*. IEEE, 2011, pp. 397–400.

[12] ——, "Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage," 2014.