Improved detection of correlated signals in low-rank-plus-noise type datasets using Informative Canonical Correlation Analysis (ICCA)

Nicholas Asendorf, *Member, IEEE,* Raj Rao Nadakuditi, *Member, IEEE*

***Abstract*—We consider two matrix-valued datasets that are modeled as low-rank-correlated-signal-plus-Gaussian-noise. When empirical canonical correlation analysis (CCA) is used to infer these latent correlations, there is a broad regime where this inference will fail, which was classified by Bao and collaborators in the limit of high dimensionality and sample size. This regime includes the setting, previously considered by Pezeshki and collaborators, where the sample size is less than the combined dimensionality of the datasets.**

We revisit this detection problem by first observing that the empirically estimated canonical correlation coefficients are the singular values of the inner products between the right singular vectors of the two datasets. Motivated by random matrix theory insights, we propose an algorithm, which we label Informative CCA (ICCA), that infers the presence of latent correlations by considering the singular values of only the "informative" right singular vectors of each dataset. We establish fundamental detection limits for ICCA and show that it dramatically outperforms empirical CCA in broad regimes where empirical CCA provably fails. We extend our theoretical analysis to the setting where the datasets have randomly missing data and for more general noise models. Finally, we validate our theoretical results with numerical simulations and a real-world experiment.

***Index Terms*—Canonical correlation analysis, Detection algorithms, Random matrix theory**

I. Introduction

Canonical correlation analysis (CCA) is a classical joint multidimensional dimensionality reduction algorithm for inferring or learning latent correlations present in two datasets [1]. CCA learns a linear transformation for each dataset such that the transformed features have maximal correlation. Often, canonical correlation analysis is the first step in algorithms that aim to fuse the information in the datasets to improve inference in the context of tasks involving the detection, estimation, classification, and prediction of correlated signals. CCA has been used in in machine learning [2], [3], [4], [5], [6], medical signal processing, [7], [8], [9], [10], [11], [12], [13], [14], [15], economics [16], climatology [17], [18], [19], and classical signal processing like Wiener filters [20] and array processing [21].

In practice, when the population covariance and cross-covariance matrices of the two datasets are unknown, they must be estimated from data. Empirical CCA relies on plug-in estimates for these quantities. When the number of samples is large relative to the combined dimensionality of the two datasets, empirical CCA performs well in the sense that the empirical canonical correlation coefficients can be used to reliably infer the presence of correlated signals buried in noise. However, when the sample size is less than the combined dimensionality of the datasets the nthe empirical canonical correlation coefficients will deterministically equal one, *irrespective of whether there is a correlated signal in the datasets* [22]. This observation led Pezeshki, Scharf et al to correctly conclude that in this regime

> ... the empirical canonical correlations are defective and may not be used as estimates of canonical correlations between random variables.

He et al. [21] used extensive simulations to make a similar observation about the deficiencies of empirical CCA in the sample size limited regime.

Recently, Bao et al [23] rigorously studied the limiting behavior of the empirical canonical correlation coefficients for the setting where the population covariance matrix is arbitrary but the cross-covariance matrix is low rank. Their work establishes the fundamental asymptotic limits of empirical CCA based detection of correlated signals in noise in the general setting where the dimensionality of the system is of the same order as the number of samples used to form the empirical covariance and cross-covariance matrices. A conclusion from this analysis is the existence of a phase transition threshold, which separates the regime where the low-rank signals are correlated and can be detected using empirical CCA from a regime where they remain correlated but cannot be detected using empirical CCA. More importantly, this phase transition threshold depends explicitly on the degree of correlation between the signals. Particularly, as the correlation decreases, the minimum (eigen) signal-to-noise ratio (SNR) above which reliable detection is possible increases. Moreover, when there are not enough samples relative to the combined dimensionality of the system, which includes the regime studied by Pezeshki et al [22], then the empirical correlation coefficients will tend to one regardless of whether there is a correlated signal or not, thereby crippling the inferential utility of empirical CCA based detection for the kinds of high-dimensional "large-p-relatively-small-n" type problems that arise in modern signal processing and machine learning [24], [25], [26], [27], [28].

N. Asendorf is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48019 USA e-mail: asendorf@umich.edu

R.R. Nadakuditi is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48019 USA e-mail: rajnrao@umich.edu

These results might convey to a practitioner that it is not theoretically possible to detect the presence of correlated signals in two datasets for modern and emerging high dimensional inferential problems. It is against this backdrop that we revisit this problem from first principles. We consider the setting where the individual matrix-valued datasets can be modeled as low-rank-correlated-signals-plus-noise type matrices; this model is motivated by the ubiquity and success of low-rank models in practice [29], [30], [31], [32], [33]. We utilize the results of Bao et al [23] and Pezeshki et al [22] to establish the fundamental limits of empirical CCA for this model. We then reconsider the connection between the canonical correlation coefficients and angles between subspaces, and propose a simple modification to empirical CCA, which builds on the work in [34], that we label Informative CCA (ICCA). We show that empirical CCA infers the presence of latent correlations by considering the singular values of a matrix formed using all of the right singular vectors of the individual signal-plus-noise matrices. In the regime where empirical CCA fails, a subset of the right singular vectors of the individual matrices are “informative”, i.e., positively correlated with the latent signal singular vectors. This insight motivates our development of ICCA which infers the presence of latent correlations by considering the singular values of a matrix formed using only these “informative” principal right singular vectors of the individual signal-plus-noise matrices. In the setting where the noise matrix is i.i.d. Gaussian, we also provide a principled approach, that leverages using results from random matrix theory [35], for selecting the number of informative components.

We then establish the fundamental limits of inference using ICCA and bring into sharp focus phase transitions that separate a regime where ICCA reliably infers (in an asymptotic sense that we make precise) the presence of a correlated signal from a regime where the correlation is present but ICCA fails. By comparing the derived fundamental limits of ICCA with the fundamental limits for empirical CCA derived by Bao et al [23], we are able to show that ICCA provably succeeds in reliably detecting correlations in the sample deficient regime where empirical CCA provably fails. Throughout this paper, when we say an algorithm “provably succeeds” we will mean that the difference between the test statistic for the pure noise setting versus correlated signal setting is almost surely positive. When we say that that the algorithm “provably fails” we mean that this difference is almost surely zero. The analysis also reveals that the detection performance of ICCA does not depend on the correlation coefficient, a very nice benefit over empirical CCA whose performance does depend on the correlation coefficient. We show that our algorithm extends readily to the widely considered missing data setting [36], [37], [38], [39], [40] and our analysis reveals that the benefits of ICCA over empircal CCA hold for this setting as well as for a class of generalized noise models such as those considered in [41].

The ICCA algorithm itself is relatively straightforward and we suspect that many practitioners have used it or are already using it because they have observed numerically that it “works” when empirical CCA does not. The main contribution of this work is the establishment of a principled, mathematically rigorous framework that justifies the use of ICCA based detection of correlated signals and the development of rigorous performance guarantees for when we expect ICCA to succeed and the sorts of performance improvements we can expect relative to empirical CCA. To the best of our knowledge, this is novel; in a subsequent paper we will provide performance guarantees for canonical vectors estimated using ICCA. The analysis of ICCA uses results established in [41], [42] and extends them to consider the new test statistic proposed herein. To that end, Theorem V.1, which is a crucial tool used to prove Theorem V.2, might be of independent interest to readers with theoretical leanings. In addition to our many main results, we present some theoretical conjectures that we believe to be true but which we were not able to prove. Proving these conjectures will require a precise characterization of the limiting behavior of of the empirical canonical correlation coefficients and their fluctuations and the fluctuation behavior of the ICCA test statistic. We provide empirical evidence to lend credence to our conjectures and hope that this will provide theoretically inclined readers with an impetus for bridging this gap.

This paper is organized as follows. We provide the linear low-rank-correlated-signal-plus-noise data model in Section II. We then derive the solution of CCA in Section III and show how to estimate the number of correlated components from its solution. In Section IV, we derive the empirical version of CCA using sample covariance matrices, highlight the connection to angles between subspaces, and exploit that connection to present our proposed ICCA algorithm. We then provide statistical tests to estimate the number of correlated components present in the datasets for both empirical CCA and ICCA. We summarize our main theoretical results in Section V, highlighting fundamental detection limits for empirical CCA (based on the work of Bao et al [23]) and ICCA (based on results we derive in this paper). We extend this analysis to the missing data and non-Gaussian noise settings. We verify these theoretical results both on simulated data and real-world datasets in Section VI. Finally, we provide concluding remarks in Section VII.

II. Setup

We assume that we are given n observations of each dataset, which we stack columnwise to form two data matrices

$$X = [x_1, \dots, x_n] \tag{1a}$$

$$Y = [y_1, \dots, y_n]. \tag{1b}$$

It is important to note that the number of observations of each dataset must be the same and that the observations come in pairs. 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

$$x_i = U_x s_{x,i} + z_{x,i} \tag{2a}$$

$$y_i = U_y s_{y,i} + z_{y,i}, \tag{2b}$$

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

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

and define the singular values of $\widetilde{K}_{xy}$ as $\kappa_1, \dots, \kappa_{\min(k_x,k_y)}$. Under this model, we define the following covariance matrices

$$\mathbb{E}\left[x_i x_i^H\right] = U_x \Theta_x U_x^H + I_p =: R_{xx} \tag{4a}$$

$$\mathbb{E}\left[y_i y_i^H\right] = U_y \Theta_y U_y^H + I_q =: R_{yy} \tag{4b}$$

$$\mathbb{E}\left[x_i y_i^H\right] = U_x K_{xy} U_y^H =: R_{xy}. \tag{4c}$$

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 $Z_x = [z_{x,1}, \dots, z_{x,n}]$, $Z_y = [z_{y,1}, \dots, z_{y,n}]$, $V_x = [v_{x,1}, \dots, v_{x,n}]$, and $V_y = [v_{y,1}, \dots, v_{y,n}]$, we may write our data matrices in (1) as the sum of a low-rank signal matrix and noise matrix

$$X = U_x \Theta_x^{1/2} V_x^H + Z_x \tag{5a}$$

$$Y = U_y \Theta_y^{1/2} V_y^H + Z_y. \tag{5b}$$

III. Canonical Correlation Analysis and its variants

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

$$\rho_{\text{cca}} = \max_{w_x, w_y} \frac{w_x^H R_{xy} w_y}{\sqrt{w_x^H R_{xx} w_x}\sqrt{w_y^H R_{yy} w_y}}, \tag{6}$$

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 the optimization problem

$$\begin{aligned} \max_{w_x, w_y} \quad & w_x^H R_{xy} w_y \\ \text{subject to} \quad & w_x^H R_{xx} w_x = 1 \\ & w_y^H R_{yy} w_y = 1. \end{aligned} \tag{7}$$

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 (7) results in the following optimization problem

$$\begin{aligned} \max_{\widetilde{w}_x, \widetilde{w}_y} \quad & \widetilde{w}_x^H R_{xx}^{-1/2} R_{xy} R_{yy}^{-1/2} \widetilde{w}_y \\ \text{subject to} \quad & \widetilde{w}_x^H \widetilde{w}_x = 1 \\ & \widetilde{w}_y^H \widetilde{w}_y = 1. \end{aligned} \tag{8}$$

Examining the optimization problem in (8), we can immediately see that the solution to CCA may be solved via the SVD of the matrix

$$C_{\text{cca}} = R_{xx}^{-1/2} R_{xy} R_{yy}^{-1/2}. \tag{9}$$

Define $C_{\text{cca}} = FKG^T$ as the SVD of C_{cca} where F is an unitary $p \times p$ matrix with columns $f_1, \dots, f_p$, G is a unitary $q \times q$ matrix with columns $g_1, \dots, g_q$, and $K = \mathbf{diag}(k_1, \dots, k_{\min(p,q)})$ is a $p \times q$ matrix whose diagonal elements are the singular values of C_{cca}. Therefore, the solution to (8) is

$$\begin{aligned} \widetilde{w}_x &= f_1 \\ \widetilde{w}_y &= g_1 \\ \rho_{\text{cca}} &= k_1. \end{aligned}$$

We can obtain higher order canonical correlations and vectors by taking successive singular value and vector pairs. From this solution, it is clear that the number of non-zero canonical correlation coefficients is exactly equal to the rank of C_{cca}. Recalling the definitions in (4), R_{xx} and R_{yy} are non-singular so we have that

$$\begin{aligned} \text{\# canonical correlation} \quad &= \text{rank}(C_{\text{cca}}) \\ \text{coefficients} \quad &= \text{rank}(R_{xx}^{-1/2} R_{xy} R_{yy}^{-1/2}) \\ &= \text{rank}(R_{xy}) \\ &= \text{rank}(K_{xy}) \\ &=: k. \end{aligned}$$

Therefore, when we know all parameters, k is exactly the number of non-zero singular values of K_{xy}. We note that $k \leq \min(k_x, k_y)$. Our objective is to infer the number of non-zero canonical correlation coefficients from data.

IV. Detection of correlated signals using empirical and informative CCA

In many applications, we do not know the covariance matrices R_{xx}, R_{yy}, and R_{xy} *a priori* and hence the quantities in (4) are unknown. Consequently, we cannot determine the number of canonical correlation coefficients by examining the rank of R_{xy}. Instead, given the data matrices in (1), we form estimates of our unknown covariance matrices via

$$\begin{aligned} \widehat{R}_{xx} &= \frac{1}{n} X X^H \\ \widehat{R}_{yy} &= \frac{1}{n} Y Y^H \\ \widehat{R}_{xy} &= \frac{1}{n} X Y^H. \end{aligned}$$

Define the data SVDs of the matrices in (1) as

$$\begin{aligned} X &= \widehat{U}_x \widehat{\Sigma}_x \widehat{V}_y^H \\ Y &= \widehat{U}_y \widehat{\Sigma}_y \widehat{V}_y^H \end{aligned}$$

and trimmed matrices

$$\widetilde{U}_x = \widehat{U}_x\left(:, 1 : \min(p, n)\right) \tag{10a}$$

$$\widetilde{V}_x = \widehat{V}_x\left(:, 1 : \min(p, n)\right) \tag{10b}$$

$$\widetilde{U}_y = \widehat{U}_y\left(:, 1 : \min(q, n)\right) \tag{10c}$$

$$\widetilde{V}_y = \widehat{V}_y\left(:, 1 : \min(q, n)\right). \tag{10d}$$

Substituting the SVDs of sample analogs of the matrices in (9), reveals the insight ([34, Eq. (6)]) that the matrix C_{cca} can be estimated as

$$\widehat{C}_{\text{cca}} = \widetilde{U}_x \widetilde{V}_x^H \widetilde{V}_y \widetilde{U}_y^H. \tag{11}$$

We denote the singular values of this matrix by $\widehat{\rho}_{\text{cca}}^{(j)}$ for $j = 1, \ldots, \min(p, q)$; these are precisely the empirical CCA correlation coefficients. Empirical CCA can return up to $\min(p, q)$ canonical correlations; however, we know from the data model in (2) that X and Y have k_x and k_y underlying signals, respectively. As k_x and k_y are unknown, let $\widehat{k}_x$ and $\widehat{k}_y$ be estimates of the number of underlying signals in each dataset. As a consequence of (6) we define the plug-in estimates of ρ_{cca}, as the $\min(\widehat{k}_x, \widehat{k}_y)$ singular values of $\widehat{C}_{\text{cca}}$ as

$$\widehat{\rho}_{\text{cca}}^{(1)}, \ldots, \widehat{\rho}_{\text{cca}}^{(\min(\widehat{k}_x, \widehat{k}_y))}. \tag{12}$$

For now, we assume that we are given $\widehat{k}_x$ and $\widehat{k}_y$, but we will return to the problem of estimating these parameters from data. To estimate the canonical vectors, we use the corresponding left and right singular vectors of $\widehat{C}_{\text{cca}}$, f_i and g_i to form

$$w_x^{(i)} = \widehat{R}_{xx}^{-1/2} f_i \tag{13a}$$

$$w_y^{(i)} = \widehat{R}_{yy}^{-1/2} g_i. \tag{13b}$$

When the number of samples is less than the combined dimension of the datasets ($n < p+q$), the largest singular value of $\widehat{C}_{\text{cca}}$ is deterministically one [22], regardless of whether an underlying correlation actually exists between the datasets. This is a very unfortunate property of empirical CCA as many of the motivating applications operate in this low-sample, high-dimensionality regime. A key observation in [34] shows that the singular values of $\widehat{C}_{\text{cca}}$ are exactly the same as the singular values of $\widetilde{V}_x^H \widetilde{V}_y$. This is a $\min(p, n) \times \min(q, n)$ matrix that uses all right singular vectors of each dataset corresponding to a non-zero singular value. However, under the low-rank signal-plus-noise model, [34] shows that only a few of the right singular vectors actually contain *informative* signal. Therefore, by trimming $\widetilde{V}_x$ and $\widetilde{V}_y$ to have only $\widehat{k}_x$ and $\widehat{k}_y$ columns, we can avoid the performance loss of empirical CCA in the sample deficient regime. Define the trimmed data SVDs

$$\mathring{U}_x = \widehat{U}_x\left(:, 1 : \widehat{k}_x\right) \tag{14a}$$

$$\mathring{V}_x = \widehat{V}_x\left(:, 1 : \widehat{k}_x\right) \tag{14b}$$

$$\mathring{U}_y = \widehat{U}_y\left(:, 1 : \widehat{k}_y\right) \tag{14c}$$

$$\mathring{V}_y = \widehat{V}_y\left(:, 1 : \widehat{k}_y\right). \tag{14d}$$

Given these definitions, we define the informative CCA (ICCA) matrix

$$\widehat{C}_{\text{icca}} = \mathring{U}_x \mathring{V}_x^H \mathring{V}_y \mathring{U}_y^H. \tag{15}$$

Similar to empirical CCA, define the top $\min(\widehat{k}_x, \widehat{k}_y)$ singular values of $\widehat{C}_{\text{icca}}$ as

$$\widehat{\rho}_{\text{icca}}^{(1)}, \ldots, \widehat{\rho}_{\text{icca}}^{(\min(\widehat{k}_x, \widehat{k}_y))}. \tag{16}$$

To estimate the ICCA canonical vectors, we use the corresponding left and right singular vectors of $\widehat{C}_{\text{icca}}$, f_i and g_i, to form

$$w_x^{(i)} = \widehat{R}_{xx}^{-1/2} f_i \tag{17a}$$

$$w_y^{(i)} = \widehat{R}_{yy}^{-1/2} g_i. \tag{17b}$$

For completeness, we will address the problem of estimating the canonical vectors in a subsequent paper.

A. New Statistical Tests for Correlation Detection

Given the canonical correlation estimates from CCA and ICCA, we can estimate the number of canonical correlations using the test statistics

$$\widehat{k}_{\text{cca}} = \sum_{i=1}^{\min(p,q)} \mathbb{1}_{\left\{\left(\widehat{\rho}_{\text{cca}}^{(i)}\right)^2 > \tau_{\text{cca}}^{\alpha}\right\}} \tag{18a}$$

$$\widehat{k}_{\text{icca}} = \sum_{i=1}^{\min(\widehat{k}_x, \widehat{k}_y)} \mathbb{1}_{\left\{\left(\widehat{\rho}_{\text{icca}}^{(i)}\right)^2 > \tau_{\text{icca}}^{\alpha}\right\}}, \tag{18b}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function and

$$\tau_{\text{cca}}^{\alpha} = F_{\text{cca}}^{-1}(1 - \alpha) \tag{19a}$$

$$\tau_{\text{icca}}^{\alpha} = F_{\text{icca}}^{-1}(1 - \alpha). \tag{19b}$$

Here F_{cca} and F_{icca} are the distributions of the square of the largest singular value of $\widehat{C}_{\text{cca}}$ and $\widehat{C}_{\text{icca}}$ for the null setting where $\widetilde{V}_x$ and $\widetilde{V}_y$ are the $\min(n, p)$ and $\min(n, q)$ columns of two independent Haar (or isotropically random) distributed $n \times n$ matrices. The exact distribution of the squared singular values of $\widehat{C}_{\text{cca}}$ and $\widehat{C}_{\text{icca}}$ in the null model is given in [43]. The distributions of the square of the largest singular value of $\widehat{C}_{\text{cca}}$ and $\widehat{C}_{\text{icca}}$ in the null model may be approximated to second-order by the Tracy-Widom law [24] as

$$\tau_{\text{cca}}^{\alpha} \approx \sigma_{p,q,n} \mathsf{TW}_{\mathbb{R},\mathbb{C}}^{-1}(1 - \alpha) + \mu_{p,q,n}, \tag{20a}$$

$$\tau_{\text{icca}}^{\alpha} \approx \sigma_{kxhat,\widehat{k}_y,n} \mathsf{TW}_{\mathbb{R},\mathbb{C}}^{-1}(1 - \alpha) + \mu_{\widehat{k}_x,\widehat{k}_y,n}, \tag{20b}$$

where $\sigma_{n,p,q}$ is a scaling parameter and $\mu_{n,p,q}$ is a centering parameter and $\mathsf{TW}_{\mathbb{R},\mathbb{C}}$ is the appropriate Tracy-Widom law for either real or complex data. See Tables III and IV of [35] for values of these parameters as well as algorithms to determine $\widehat{k}_x$ and $\widehat{k}_y$. For a similar high dimensional analysis, see [25].

V. Main Results

In this section we summarize the main theoretical results for empirical CCA and our new ICCA algorithm. We derive parameter regimes where the estimates in (18) correctly infer the presence of correlated signals. We then extend these results to include the cases when the data matrices have missing entries and when the additive noise of the data model in (2) is not Gaussian.

A. Empirical CCA

Proposition V.1. *Let $n, p, q \to \infty$ such that $p/n \to c_x$ and $q/n \to c_y$. Let $p + q \leq n$. Then the largest singular value of $\widehat{C}_{cca}$ generated from data modeled in (2) behaves as*

$$\widehat{\rho}_{cca}^{(1)} = 1.$$

Proof. See [22], page 996. □

A consequence of Proposition V.1 is that the sample canonical correlation coefficients obtained using empirical CCA will equal one even when there is no correlation signal present in the datasets. We now establish the limiting behavior of the sample canonical correlation coefficients for the regime where they are not deterministically equal to one. The following proposition is a result presented in [23], adopted using our notation.

Proposition V.2. *Let $n, p, q \to \infty$ such that $p/n \to c_x$ and $q/n \to c_y$. Assume that $p+q < n$. For $i = 1, \ldots, \min(k_x, k_y)$, let $\widehat{\rho}_{cca}^{(i)}$ be the i-th largest singular value of $\widehat{C}_{cca}$ generated from data modeled in (2). Then these singular values behave as*

$$\widehat{\rho}_{cca}^{(i)} \xrightarrow{a.s.} \begin{cases} \sqrt{\kappa_i^2 \left(1 - c_x + \frac{c_x}{\kappa_i^2}\right)\left(1 - c_y + \frac{c_y}{\kappa_i^2}\right)} & \text{if } \kappa_i^2 \geq r_c \\ \sqrt{d_r} & \text{if } \kappa_i^2 < r_c \end{cases} \tag{21}$$

where κ_i are the singular values of $\widetilde{K}_{xy}$ and

$$r_c = \frac{c_x c_y + \sqrt{c_y c_y (1 - c_x)(1 - c_y)}}{(1 - c_x)(1 - c_y) + \sqrt{c_x c_y (1 - c_x)(1 - c_y)}} \tag{22}$$

and

$$d_r = c_x + c_y - 2c_x c_y + 2\sqrt{c_x c_y (1 - c_x)(1 - c_y)}. \tag{23}$$

Proof. See [23] for original proof our see our appendix for a proof providing the necessary mathematical manipulations to transform our data model to the one used in [23]. □

Proposition V.2 brings into sharp focus the existence of a phase transition that separates a regime where the sample canonical correlation coefficients can be used to infer the presence of a correlated signal from a regime where it cannot. The phase transition boundary depends on the correlation between the signals in the two datasets via the κ_i quantity. We will empirically verify this result by [23] and use it as a baseline for CCA to which we will compare our new estimate, ICCA.

The empirical CCA estimate of the number of correlated components is given by (18). Since $\tau_{\text{cca}}^{\alpha} \to \sqrt{d_r}$, we expect (18) to return the number of singular values greater than $\sqrt{d_r}$. Let

$$k_{\text{eff}}^{\text{cca}} = \sum_{i=1}^{k} \mathbb{1}_{\{\kappa_i^2 > r_c\}}, \tag{24}$$

denote the effective number of correlated components detectable by empirical CCA. Note that, by definition, $k_{\text{eff}}^{\text{cca}} \leq k$. Proposition V.2 shows that $k_{\text{eff}}^{\text{cca}} < k$ whenever some of the singular values are below the phase transition threshold r_c. These observations lead to the following conjecture.

Conjecture V.1. *In the same setting as Proposition V.2, we have that*

$$\mathbb{P}\left(\widehat{k}_{cca} = k_{\mathit{eff}}^{cca}\right) \to 1 \quad \text{if } \kappa_k^2 > r_c \text{ and } n > p + q$$

where κ_i are the singular value of $\widetilde{K}_{xy}$ and r_c is given in (22).

B. ICCA

We now provide analogous performance guarantees for ICCA and establish conditions that can be used to identify broad regimes where ICCA reliably detects correlations but empirical CCA, based on Proposition V.2, does not. Our analysis of ICCA relies on the characterization of the entries of the matrix $\mathring{V}_x^H \mathring{V}_y$. To that end, we provide the following intermediate theorem.

Theorem V.1. *Let $\mathring{V}_x$ and $\mathring{V}_y$ be the trimmed right singular vectors defined in (14) of the data matrices generated from the data model in (2). In the asymptotic setting of Theorem A.1 with $p/n \to c_x$ and $q/n \to c_y$*

$$\left| \left[\mathring{V}_x^H \mathring{V}_y \right]_{ij} \right| \xrightarrow{a.s.} \left| k_{ij}^{xy} \right| \alpha_{x,i} \alpha_{y,j},$$

where

$$\alpha_{x,i} = \sqrt{1 - \frac{c_x + \theta_i^{(x)}}{\theta_i^{(x)}(\theta_i^{(x)} + c_x)}} \qquad \text{if } \theta_i^{(x)} > c_x^{1/4}, \tag{25a}$$

and

$$\alpha_{y,j} = \sqrt{1 - \frac{c_y + \theta_j^{(y)}}{\theta_j^{(y)}(\theta_j^{(y)} + c_x)}} \qquad \text{if } \theta_j^{(y)} > c_y^{1/4}, \tag{25b}$$

and k_{ij}^{xy} are the entries of K_{xy}.

Proof. See Appendix for proof and Theorem A.1. □

Theorem V.1 allows us to establish the following result.

Theorem V.2. *Let $p, q, n \to \infty$ with $p/n \to c_x$ and $q/n \to c_y$. For $i = 1, \ldots, \min(k_x, k_y)$, let $\widehat{\rho}_{icca}^{(i)}$ be the i-th largest singular value of $\widehat{C}_{icca}$ defined as in (15). Then, for data modeled as in (2), we have that if*

$$\min_{i=1,\ldots,k_x} \theta_i^{(x)} > c_x^{1/4} \text{ and } \min_{i=1,\ldots,k_y} \theta_i^{(y)} > c_y^{1/4},$$

then

$$\widehat{\rho}_{icca}^{(k)} > 0 \text{ almost surely,}$$

where $k = \mathit{rank}(K_{xy})$.

Proof. See Appendix. □

Analogous to the empirical CCA setting, let

$$k_{\text{eff}}^{\text{icca}} = \sum_{i=1}^{k} \mathbb{1}_{\{\theta_i^{(x)} > c_x^{1/4} \text{and } \theta_i^{(y)} > c_y^{1/4}\}}, \tag{26}$$

denote the effective number of correlated components detectable by ICCA. Note that, by definition, $k_{\text{eff}}^{\text{icca}} \leq k$. Comparing Theorem V.2 and Proposition V.2, it is easy to see that $k_{\text{eff}}^{\text{icca}} \geq k_{\text{eff}}^{\text{cca}}$. The ICCA estimate of the number of correlated components is given by (18). Since $\tau_{\text{icca}}^{\alpha} \to 0$, whenever $\widehat{k}_x$ and $\widehat{k}_y$ are consistent estimates of k_x and k_y, we are led to the following conjecture.

Conjecture V.2. *In the setting of Theorem V.2, let $\widehat{k}_{icca}$ denote the estimate of the number of correlated components as given by (18). Then we have that*

$$\mathbb{P}\left(\widehat{k}_{icca} = k_{eff}^{icca}\right) \to 1,$$

C. Extension to Missing Data

We now consider the setting where our data matrices X and Y have missing entries. In such as setting, our matrices are modeled similar to (5) but with additional masking matrices

$$X = \left(U_x \Theta_x^{1/2} V_x^H + Z_x\right) \odot M_x \tag{27a}$$

$$Y = \left(U_y \Theta_y^{1/2} V_y^H + Z_y\right) \odot M_y \tag{27b}$$

where

$$M_{ij}^x = \begin{cases} 1 & \text{with probability } \gamma_x \\ 0 & \text{with probability } 1-\gamma_x \end{cases},$$

$$M_{ij}^y = \begin{cases} 1 & \text{with probability } \gamma_y \\ 0 & \text{with probability } 1-\gamma_y \end{cases}$$

and $\odot$ denotes the Hadamard or element-wise product. Throughout this section we make the following assumption on the entries of U_x, U_y, V_x, V_y. Recall the definitions for Θ_x and Θ_y in (3). This assumption ensures that the columns of these matrices are not "spiked".

Assumption V.1. *In the missing data setting, assume that the columns of U_x, U_y, V_x, and V_y satisfy a 'low-coherence' condition in the following sense: we suppose that there exist non-negative constants $\eta_{u,x}$, $C_{u,x}$, $\eta_{u,y}$, $C_{u,y}$, $\eta_{v,x}$, $C_{v,x}$,$\eta_{v,y}$, $C_{v,y}$ independent of p, q and n, such that for $i = 1, \ldots, k_x$ and $\jmath = 1, \ldots, k_y$,*

$$\max_i \|u_i^{(x)}\|_\infty \leq \eta_{u,x} \frac{\log^{C_{u,x}} p}{\sqrt{p}}, \quad \max_i \|u_j^{(y)}\|_\infty \leq \eta_{u,y} \frac{\log^{C_{u,y}} q}{\sqrt{q}}$$

$$\max_i \|v_i^{(x)}\|_\infty \leq \eta_{v,x} \frac{\log^{C_{v,x}} n}{\sqrt{n}}, \quad \max_i \|v_j^{(y)}\|_\infty \leq \eta_{v,y} \frac{\log^{C_{v,y}} n}{\sqrt{n}}.$$

In the missing data setting, our optimization problem remains unchanged from (6), whose solution is given via the eigenvalue decomposition of $\widehat{C}_{\text{cca}}$ in (11). The only difference comes from the fact that the sample covariance matrices are formed via the data matrices in (27), which contain missing elements. In the same manner of Section V-B, we wish to provide performance guarantees in the presence of missing data. The theorem below characterizes this behavior. We proceed as in [44] and extensively use the results derived there. The two theorems are very similar except that in the case of missing data, we simply replace Θ_x with $\gamma_x\Theta_x$ and Θ_y with $\gamma_y\Theta_y$. Therefore, missing data has the effect of decreasing the SNR of our problem.

Theorem V.3. *Let $p, q, n \to \infty$ with $p/n \to c_x > 0$ and $q/n \to c_y > 0$ and assume the coherence conditions given in Assumption V.1. Given data modeled in (27), we have that if*

$$\min_{i=1,\ldots,\widehat{k}_x} \theta_i^{(x)} > \frac{c_x^{1/4}}{\sqrt{\gamma_x}} \textit{ and } \min_{i=1,\ldots,\widehat{k}_y} \theta_i^{(y)} > \frac{c_y^{1/4}}{\sqrt{\gamma_y}}$$

then

$$\widehat{\rho}_{icca}^{(k)} > 0 \textit{ almost surely},$$

where $k = rank(K_{xy})$.

Proof. See Appendix. □

Conjecture V.3. *For the same setting as in Theorem V.3, we have that*

$$\mathbb{P}\left(\widehat{\rho}_{cca}^{(k)} > \sqrt{d_r}\right) \to 1 \quad \textit{if} \min_{i=1,\ldots,k} \mathring{\kappa}_i^2 > r_c \textit{ and } n > p + q$$

where r_c and d_r are given in (22) and (23), respectively, and $\mathring{\kappa}_i$ are the singular values of

$$\left(\gamma_x\Theta_x + I_{k_x}\right)^{-1/2} \left(\gamma_x\Theta_x\right)^{1/2} P_{xy} \left(\gamma_y\Theta_y\right)^{1/2} \left(\gamma_y\Theta_y + I_{k_y}\right)^{-1/2}.$$

Proving Conjectures V.1, V.2, and V.3 requires a finer understanding of fluctuations of the test statistic than we presently have.

D. Generalized Noise Model

Finally, we provide a performance guarantee for non-Gaussian noise matrix models in (2).

Assumption V.2. *Let $Z_x^{(n)} = [z_{x,1}, \ldots, z_{x,n}]$ be the $p \times n$ matrix formed by stacking n observations of our noise. Let $Z_x^{(n)}$ have singular values $\sigma_1\left(Z_x^{(n)}\right) \geq \cdots \geq \sigma_p\left(Z_x^{(n)}\right)$. Let $\mu_{Z_x^{(n)}}$ be the empirical singular value distribution defined by the probability measure*

$$\mu_{Z_x^{(n)}} = \frac{1}{p}\sum_{i=1}^{p} \delta_{\sigma_i\left(Z_x^{(i)}\right)}.$$

We assume that the probability measure $\mu_{Z_x^{(n)}}$ converges almost surely weakly as $p, n \to \infty$ with $p/n \to c_x$ to a non-random compactly supported probability measure μ_{Z_x} that is supported on $[a_x, b_x]$. We assume that $\sigma_1 \xrightarrow{a.s.} b_x$.

Similarly, we assume that the empirical singular value distribution for the noise matrix of Y converges almost surely to the non-random compactly supported probability measures μ_{Z_y} that is supported on $[a_y, b_y]$ and that $\sigma_1 \xrightarrow{a.s.} b_y$.

Corollary V.1. *As in Assumption V.2, let μ_{Z_x} and μ_{Z_y} be the non-random compactly supported probability measures modeling the singular values of the noise matrices X and Y. Let b_x and b_y be the supremums of the supports, respectively. Let $p, q, n \to \infty$ with $p/n \to c_x$ and $q/n \to c_y$. Relax the constraint in (2) that the noise is Gaussian but instead drawn from the probability measures above. The ICCA canonical correlation estimates behave as*

$$\mathbb{P}\left(\widehat{\rho}_{icca}^{(k)} > 0\right) \to 1$$

if

$$\min_{i=1,\ldots,k_x} \theta_i^{(x)} > \frac{1}{D_{\mu_X}(b_x^+)} \textit{ and } \min_{i=1,\ldots,k_y} \theta_i^{(y)} > \frac{1}{D_{\mu_Y}(b_y^+)}$$

where D_{μ_X} and D_{μ_Y}, the D-transforms of μ_X and μ_Y, are the functions, depending on c_x and c_y, defined by

$$\begin{aligned} D_{\mu_X}(z) =: \quad & \left[\int \frac{z}{z^2 - t^2} d\mu_X(t)\right] \times \\ & \left[c_x \int \frac{z}{z^2 - t^2} d\mu_X(t) + \frac{1 - c_x}{z}\right] \text{ for } z > b_x \\ D_{\mu_Y}(z) =: \quad & \left[\int \frac{z}{z^2 - t^2} d\mu_Y(t)\right] \times \\ & \left[c_y \int \frac{z}{z^2 - t^2} d\mu_Y(t) + \frac{1 - c_y}{z}\right] \text{ for } z > b_y. \end{aligned}$$

Define the notation

$$D_\mu(b^+) =: \lim_{z \downarrow b} D_\mu(z).$$

Proof. This result follows from the proof of Theorem V.2 using the analysis in [41]. □

This is the more general result to Theorem V.2 as it is applicable to non-Gaussian noise. See [44] for a discussion on computing D-transforms in practice.

VI. Empirical Results

A. Simulated Data

We first showcase the accuracy of the detection boundary for both empirical CCA and ICCA described in Proposition V.2 and Theorem V.2. We consider a rank-1 setting ($k_x = k_y = 1$) and generate data from (2) for fixed $p = q = 150$ over various number of samples n, signal-to-noise ratio (SNR) $\theta = \theta_1^{(x)} = \theta_1^{(y)}$, and various $\rho = P_{xy}$. In this setting, there is only one correlated signal so $k = 1$ and the detection boundary becomes a phase transition. We then form matrices X and Y and compute $\widehat{\rho}_{\text{cca}}^{(1)}$ and $\widehat{\rho}_{\text{icca}}^{(1)}$ from the SVD of of $\widehat{C}_{\text{cca}}$ and $\widehat{C}_{\text{icca}}$, respectively. Using these correlation estimates, we compute the estimated number of correlated components via (18) for a significance level of $\alpha = 0.01$. For a fixed set of parameters (n, θ, ρ) we repeat the above process for 10000 trials and determine the percentage of trials where we detect $\widehat{k}_{\text{cca}} = 1$ and $\widehat{k}_{\text{icca}} = 1$. In all simulations, we use Algorithm 2 of [35] to estimate $\widehat{k}_x$ and $\widehat{k}_y$ using a significance level of $\alpha = 0.01$. Figure 1 plots the $\log_{10}$ of this percentage for empirical CCA and ICCA for two values of ρ. On each plot, we overlay the empirical CCA detection boundary given by Proposition V.2 using a solid white line and the ICCA detection boundary given by Theorem V.2 using a dashed white line.

From this figure, we see that for smaller ρ, it is more difficult for empirical CCA to detect the presence of the correlated signal. However, ICCA is very robust to the underlying correlation; the ICCA detection boundary in Theorem V.2 does not depend on the value of ρ. We also verify Proposition V.1 showing that when $n < 300$, it is impossible to detect the presence of correlated signals using empirical CCA because $\widehat{\rho}_{\text{cca}} = 1$ deterministically. With ICCA, we avoid this undesirable property and can still detect the presence of a correlated signal for very small n and θ. This figure also provides evidence for Conjectures V.1 and V.2.

Next, we explore the minimum $1/c$ for $c = c_x = c_y$ needed to reliably detect $k = 1$ correlated signal in the experiment setting described for Figure 1. As $c = p/n = q/n$, the minimum $1/c$ is equivalent to the minimum number of samples needed for fixed dimensions. Using the theoretical phase transitions in Proposition V.2 and Theorem V.2, we have that this critical value of c is $c_{\text{crit}} = \theta^4$ for ICCA and $c_{\text{crit}} = \min\left(\frac{r_c^{\text{crit}}}{1+r_c^{\text{crit}}}, 0.5\right)$ for empirical CCA, where

$$r_c^{\text{crit}} = \left(\frac{-\rho + \sqrt{\rho^2 + 4\theta^2\rho^2(1 + \theta^2\rho^2)}}{2(1 + \theta^2\rho^2)}\right)^2.$$

Figure 2 plots level sets of c_{crit} for empirical CCA and ICCA for various values of $\theta = \theta_1^{(x)} = \theta_1^{(y)}$ and $\rho = P_{xy}$. Recall that if $c > 0.5$, empirical CCA fails entirely, so for comparison we only show contour lines for $1/c = 10$ and $1/c = 3$.

From this figure, we once again observe that the performance of ICCA is independent of the value of $\rho = P_{xy}$, while the performance of empirical CCA is highly dependent on the correlation. This figure allows us to showcase that ICCA is theoretically better than empirical CCA in all parameter regimes as ICCA can achieve the same performance of empirical CCA given fewer samples at a lower SNR.

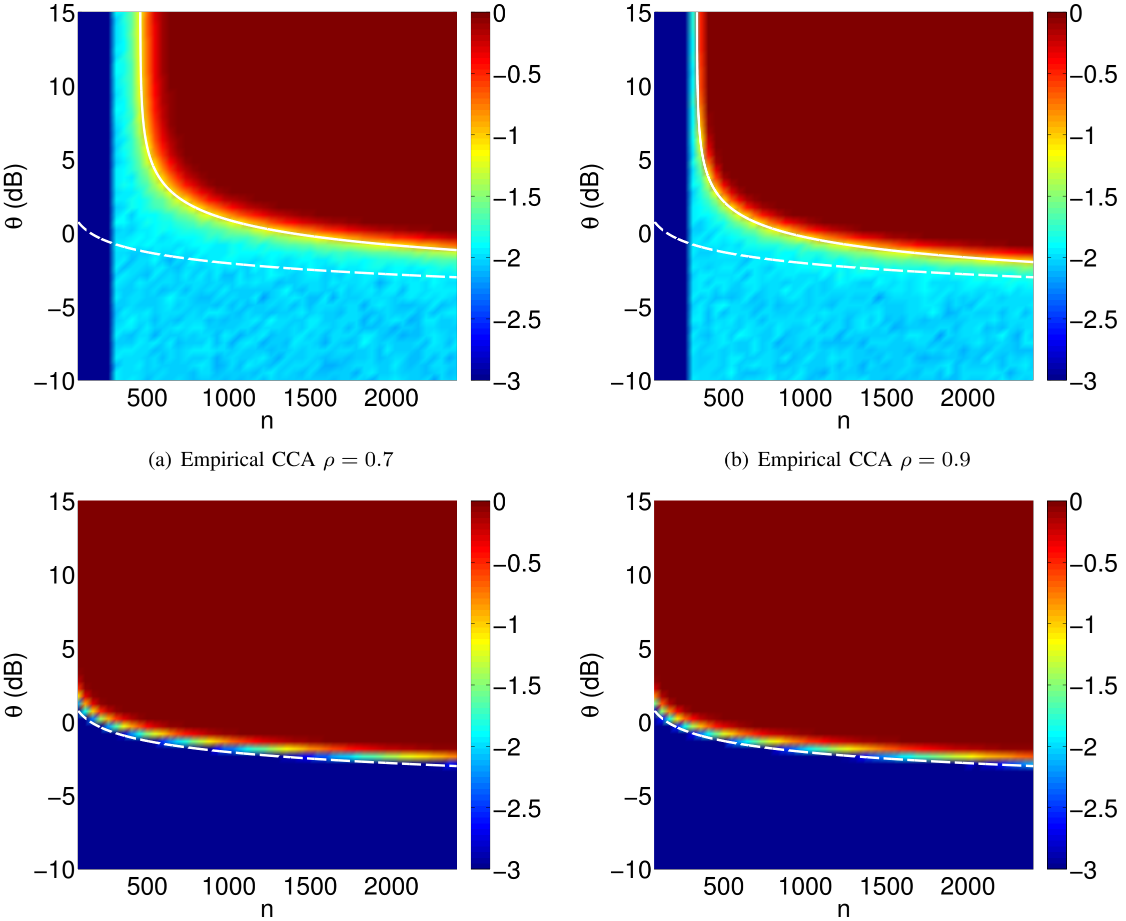
B. Simulated Missing Data

Next, we demonstrate the accuracy of the performance limits for both empirical CCA and ICCA in the setting of missing data described in Theorem V.3 and Conjecture V.3. Again, we consider a rank-1 setting ($k_x = k_y = 1$) but generate data from (27) for fixed $p = q = 150$ over various number of samples n, signal-to-noise ratio (SNR) $\theta = \theta_1^{(x)} = \theta_1^{(y)}$, various $\rho = P_{xy}$ (so that $k = 1$), and also various percentages of missing data $\gamma = \gamma_x = \gamma_y$. In all simulations, we use Algorithm 2 of [35] to estimate $\widehat{k}_x$ and $\widehat{k}_y$ using a significance level of $\alpha = 0.01$. We stack the data into matrices X and Y and compute $\widehat{\rho}_{\text{cca}}^{(1)}$ and $\widehat{\rho}_{\text{icca}}^{(1)}$ from the SVD of of $\widehat{C}_{\text{cca}}$ and $\widehat{C}_{\text{icca}}$, respectively. Using these correlation estimates, we compute the estimated number of correlated components via (18) for a significance level of $\alpha = 0.01$. For a fixed set of parameters (n, θ, ρ, γ) we repeat the above process for 10000 trials and determine the percentage of trials where we detect $\widehat{k}_{\text{cca}} = 1$ and $\widehat{k}_{\text{icca}} = 1$. Figure 3 plots the $\log_{10}$ of this percentage for empirical CCA and ICCA, respectively. We overlay the ICCA performance boundary given by Theorem V.3 in a dashed line and the empirical CCA performance boundary given by Conjecture V.3 in a solid line.

From these figures, we observe that Theorem V.3 and Conjecture V.3 accurately predict the phase transition for both empirical CCA and ICCA in the presence of missing data for a wide array of parameters. Even in the presence of missing data, ICCA can detect the presence of the correlated signal in the low-sample regime ($n < p + q$) where empirical CCA deterministically fails. In this missing data setting, we once again observe that the value of ρ affects the phase transition for empirical CCA but not for ICCA; it is harder for CCA to detect signals with small correlations.

C. Controlled Flashing Lights Experiment

To verify the effectiveness of ICCA for real world applications, we conducted a controlled experiment consisting of 5



(a) Empirical CCA $\rho = 0.7$

(b) Empirical CCA $\rho = 0.9$

(c) ICCA $\rho = 0.7$

(d) ICCA $\rho = 0.9$

Fig. 1. We generate data from (2) for $p = q = 150$, $k_x = k_y = 1$, $k = 1$, and various $\rho = P_{xy}$ and sweep over $\theta = \theta_1^{(x)} = \theta_1^{(y)}$ and n. We compute $\widehat{k}_x$ and $\widehat{k}_y$ using Algorithm 2 of [35] for a significance value of $\alpha = 0.01$. Using these estimates, we compute $\widehat{\rho}_{\text{cca}}^{(1)}$ as the largest singular value of $\widehat{C}_{\text{cca}}$ as in (12) and $\widehat{\rho}_{\text{icca}}^{(1)}$ as the largest singular value of $\widehat{C}_{\text{icca}}$ as in (16). We then estimate the number of correlated signals $\widehat{k}_{\text{cca}}$ and $\widehat{k}_{\text{icca}}$ via (18) for a significance level of $\alpha = 0.01$. We repeat this for 10000 trials and compute the percentage of trials where $\widehat{k}_{\text{cca}} = 1$ and $\widehat{k}_{\text{icca}} = 1$. We plot $\log_{10}$ of these percentages for multiples values of θ and n. We plot the theoretical phase transition of empirical CCA (given in Proposition V.2 that relies on [23]) in a solid white line and the theoretical phase transition of ICCA (given in Theorem V.2) in a dashed white line.

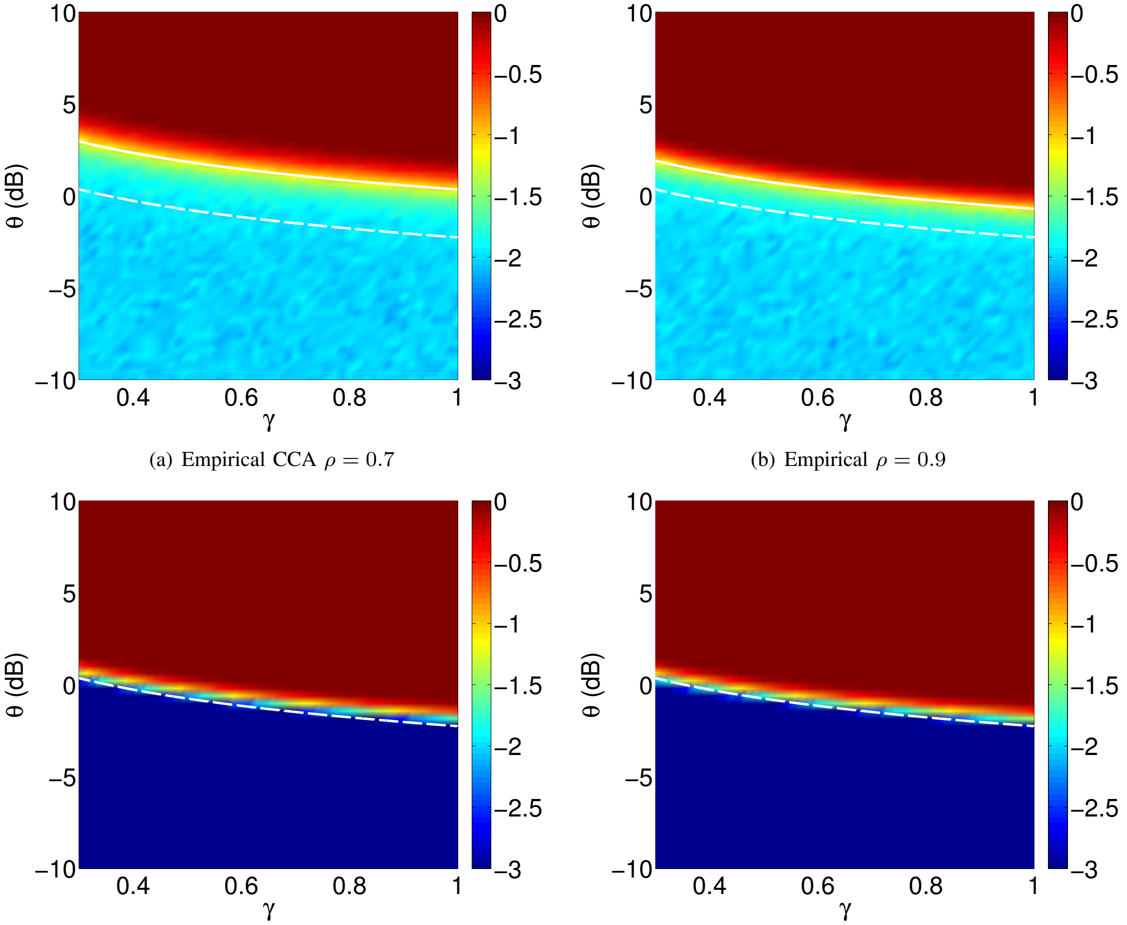
stationary flashing lights and two stationary iPhone cameras[1]. Figure 4 shows the left and right camera views at one time point of our experiment and manually identifies each source. The 5 sources are a blue flashing police light (BPL) outlined in the green rectangle, one phone with a flashing strobe light (PH1) outlined in the dark blue rectangle, another phone with a flashing strobe light (PH2) outlined in a red rectangle, a tablet with a flashing screen (T1) outlined in the magenta rectangle, and a red flashing police light (RPL) outlined in the cyan rectangle. From left to right, the left camera can see BPL, PH1, and PH2. From left to right, the right camera can see PH2, T1, and RPL. Therefore, both cameras share the common signal of PH2.

To synchronize the cameras we used the RecoLive MultiCam iPhone app [2]. After turning on all light sources, we recorded 30 seconds of video at 30 frames per second. The resolutions of the iPhone's cameras were both 1920×1080 pixels. To post-process the video data, we first converted the video streams to grayscale and then downsampled each spatial dimension by a factor of 8, resulting in a resolution of 240×135. We then vectorized each image and stacked the 900 frames into data matrices, both of dimension 32400×900. Finally, we subtract the mean from each dataset so that we may run empirical CCA and ICCA on the zero-mean datasets, X_{left} and Y_{right}.

First, we run principal component analysis (PCA) on X_{left} and Y_{right} to identify the number of signals in each dataset. We know from our setup that each camera has 3 independent sources. Figure 5 plots the singular values of X_{left} and Y_{right}. However, PCA does not provided any information about whether the identified signals are correlated across cameras.

[1] For a video demonstration of this experiment, visit `https://www.youtube.com/watch?v=WYlC2XgBDXs`.
For similar experiments on an audio-audio dataset, visit `https://www.youtube.com/watch?v=lQzO1OS7PEs` and an audio-video dataset visit
`https://www.youtube.com/watch?v=8E83P-_oVgg`.

[2] http://recolive.com/en/



(a) Empirical CCA $\rho = 0.7$

(b) Empirical $\rho = 0.9$

(c) ICCA $\rho = 0.7$

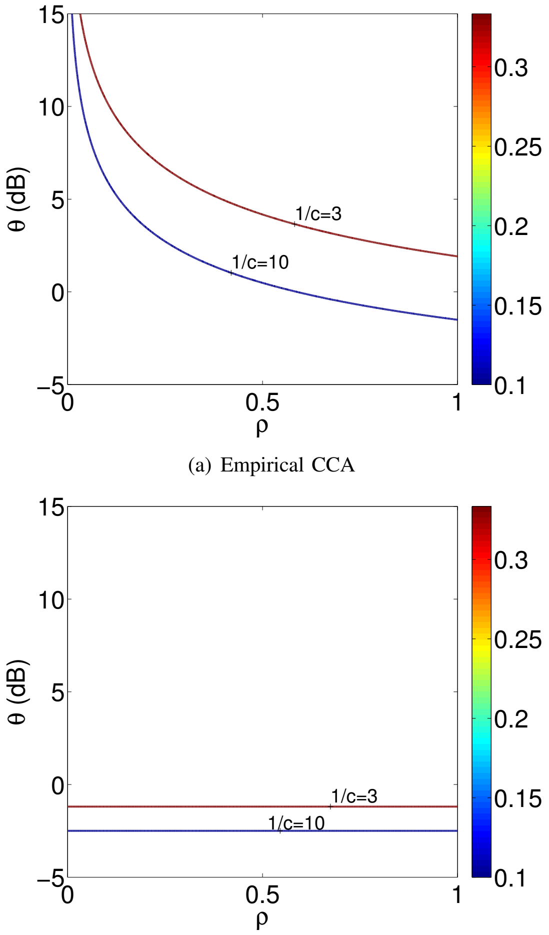
(d) ICCA $\rho = 0.9$

Fig. 3. We generate data from (27) for $p = q = 150$, $k_x = k_y = 1$, $k = 1$, $n = 1200$, and various $\rho = P_{xy}$ and sweep over $\theta = \theta_1^{(x)} = \theta_1^{(y)}$ and $\gamma = \gamma_x = \gamma_y$. We compute $\widehat{k}_x$ and $\widehat{k}_y$ as using Algorithm 2 of [35] for a significance value of $\alpha = 0.01$. Using these estimates, we compute $\widehat{\rho}_{\text{cca}}^{(1)}$ as the largest singular value of $\widehat{C}_{\text{cca}}$ as in (12) and $\widehat{\rho}_{\text{icca}}^{(1)}$ as the largest singular value of $\widehat{C}_{\text{icca}}$ as in (16). We then estimate the number of correlated signals $\widehat{k}_{\text{cca}}$ and $\widehat{k}_{\text{icca}}$ via (18) for a significance level of $\alpha = 0.01$. We repeat this for 10000 trials and compute the percentage of trials where $\widehat{k}_{\text{cca}} = 1$ and $\widehat{k}_{\text{icca}} = 1$. We plot $\log_{10}$ of these percentages for multiples values of θ and n. We plot the theoretical performance limit of empirical CCA (given in Conjecture V.3) in a solid white line and the theoretical performance boundary of ICCA (given in Theorem V.3) in a dashed white line.

To identify correlated pixels between the cameras, we run empirical CCA and ICCA after each new video frame. For frame ℓ, we construct the $32400 \times \ell$ submatrices X_{left}^{ℓ} and Y_{right}^{ℓ} by taking the matrix of the first ℓ original vectorized frames and zero meaning it. We then use these matrices as the input to empirical CCA and ICCA. Using our knowledge of 3 sources present in each camera, we set $\widehat{k}_x = \widehat{k}_y = 3$. Figure 6 plots the top 3 correlation coefficients returned by empirical CCA and ICCA over the first 800 frames. Intuitively, empirical CCA returns perfect correlation as we have only a few frames but a large dimension (pixels).

Using these singular values returned by empirical CCA and ICCA, we can set a threshold via (20) to determine which ones indicate the presence of a correlated signal between the datasets. Examining Figure 6, we can easily accomplish this for ICCA as the top two singular values separate from the third. However, as we operate in the sample deficient regime, we cannot set such a threshold for empirical CCA to detect the presence of correlated signals. We overlay the thresholded unit-norm canonical vectors (defined in (13) and (17)) onto the original images in Figure 7 for both empirical CCA and ICCA. From this figure, we observe that the empirical CCA canonical vectors appear to be very random and noisy. The ICCA canonical vectors correctly identify both sources of correlation in our dataset.

Given that our experiment setup has only one shared flashing light, it is initially surprising that ICCA returns a two large singular values. Examining the ICCA canonical vector overlay in Figure 7, we observe that this correlation corresponds to RPL and BPL. Figure 8 examines the right singular vectors returned by PCA corresponding to RPL and BPL. We observe that these light sources have approximately the same period and even though they were started at random times, they are in approximate antiphase, making them correlated. This is especially interesting because neither camera can see both sources, but ICCA is still able to reveal a latent correlation inherent in the period and phase of these lights.



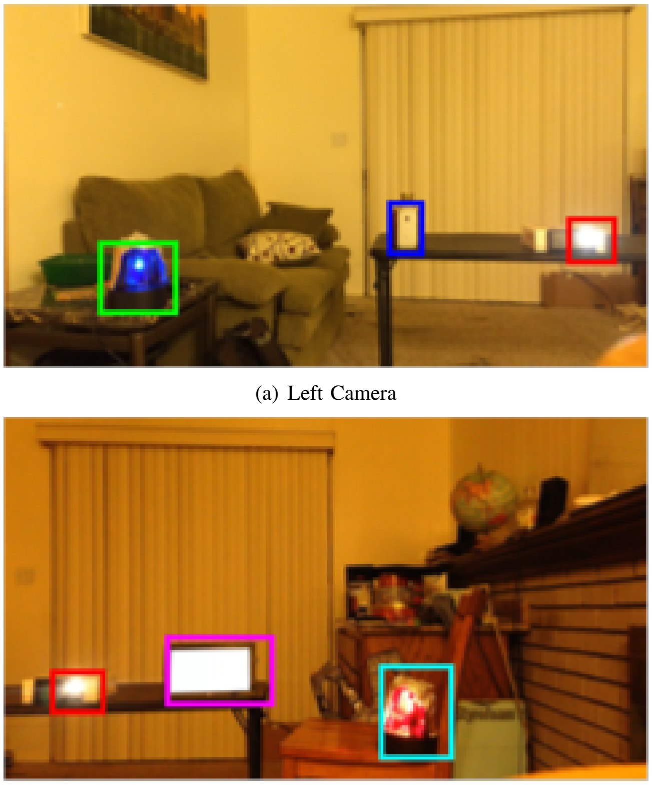
(a) Empirical CCA

(b) ICCA

Fig. 2. Contour lines for minimum $1/c$ necessary for reliable detection of $k = 1$ correlated component. The quantity $1/c = n/p$ is equivalent to the number of samples per dimension of data. For $c = c_x = c_y$, Figure 2(a) plots the contours for empirical CCA using the limits given in Proposition V.2 and Figure 2(b) plots the ICCA contours using the limits given in Theorem V.2. We plot the contours for $1/c = 10$ and $1/c = 3$. These plots clearly demonstrate that the ICCA limits are independent of $\rho = P_{xy}$ while those for empirical CCA are highly dependent on $\rho = P_{xy}$. For a fixed number of samples (fixed c), ICCA can reliably detect the presence of a correlated signal at lower SNR values than empirical CCA.

D. Controlled Flashing Lights with Missing Data

Using the same dataset in the previous section, we add missing data to each frame independently[3]. We set $\gamma = \gamma_x = \gamma_y = 0.75$ so that about 25% of the pixels are set to 0. We generate the missing pixels independently for each camera and for each frame. We then process the data exactly as above without missing data. We note that in this setup, our light sources do not obey the low-coherence condition, but we still run ICCA to demonstrate its robustness. Particularly, source PH1 occupies only a small number of pixels so that it has a very spiked signal and violates the low-coherence assumption the most. In this missing data framework, PCA cannot detect this source. However, this source is independent of all other signals as so we will still be able to detect all correlated signal in the setting of Theorem V.3.



(a) Left Camera

(b) Right Camera

Fig. 4. Left and right camera views of our experiment with boxes manually identifying each source. Both cameras share a common flashing phone, outlined in a red rectangle. Each camera has two independent sources besides the shared flashing phone.

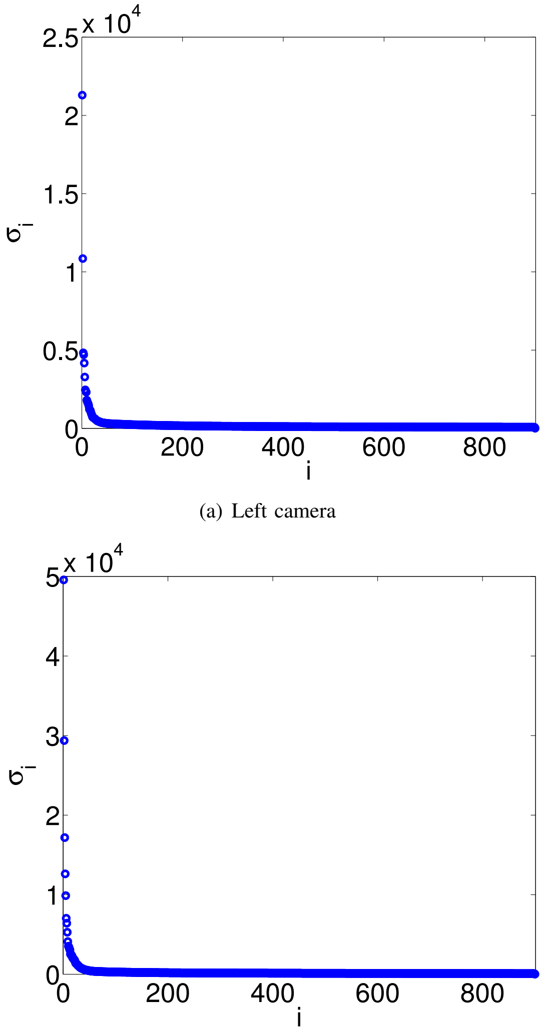
Figure 9 overlays the thresholded canonical vectors (defined in (13) and (17)) corresponding to the top 2 singular values for both empirical CCA and ICCA after 800 frames. Unsurprisingly, empirical CCA is still unable to detect the two correlated signals because in this regime the top singular values are deterministically one and the corresponding canonical vectors are uninformative. However, ICCA is able to detect our correlated signals even in the presence of missing data. The colored pixels clearly identify our two sources of correlation.

Figure 10 plots the top 3 singular values returned by empirical CCA and ICCA. Unsurprisingly, the singular values reported by CCA are 1 and uninformative. However, once we collect enough frames, there are two large singular values reported by ICCA that identify the two sources of correlation in our dataset. Similar to the above discussion, we can set a threshold via (20) to determine which ones indicate the presence of a correlated signal between the datasets.

VII. Conclusion

In this paper we explored the problem of detecting correlations present in exactly two datasets when the covariance and cross-covariance matrices are unknown and estimated from training data. We showcased that the standard algorithm, empirical CCA, fails to detect such correlations when the number

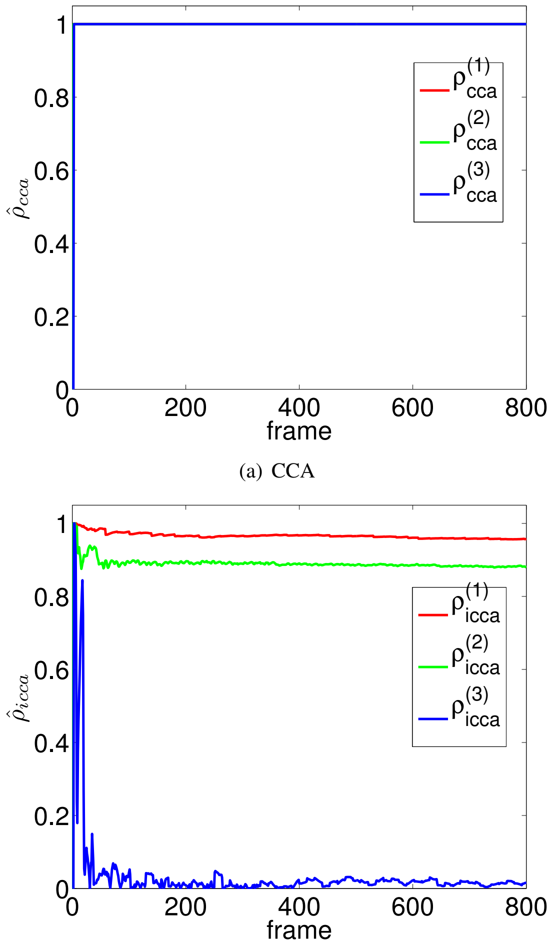
[3] For a video demonstration of this experiment, please visit `https://www.youtube.com/watch?v=vhi3T4S8riE`



(a) Left camera

(b) Right camera

Fig. 5. Singular value spectra of X_{left} and Y_{right} for the flashing light experiment.



(a) CCA

(b) ICCA

Fig. 6. (a) Top three singular values returned by empirical CCA as defined in (12) for the flashing light demonstration. As we are in the sample deficient regime, these singular values are deterministically 1. (b) Top three singular values returned by ICCA as defined in (16) for the flashing light demonstration. ICCA correctly identifies two sources of correlation.

of training samples is limited. Motivated by insights from random matrix theory, we presented informative CCA (ICCA), which can reliably detect correlations present in low-rank-correlated-signal-plus-noise type datasets. We then extended this analysis to the case of missing data and showcased the improved detection performance of ICCA on both synthetic and real-world examples.

This paper assumed a low-rank-correlated-signal-plus-noise data model, which is ubiquitous in signal processing applications. We note that depending on the application, the linear, low-rank-correlated-signal-plus-noise data model may be inappropriate. In such a setting, kernel CCA (KCCA) [45], [46] uses the kernel trick to first map the data into a higher dimensional space before running CCA. The performance analysis of such kernel methods for non-linear data models is important future work. Proving Conjectures V.1, V.2, and V.3 remains an open problem and important area of future work. Finally, in a future paper we will characterize the accuracy of the empirical canonical vector estimates and provide a new estimate that uses insights from random matrix theory.

Appendix

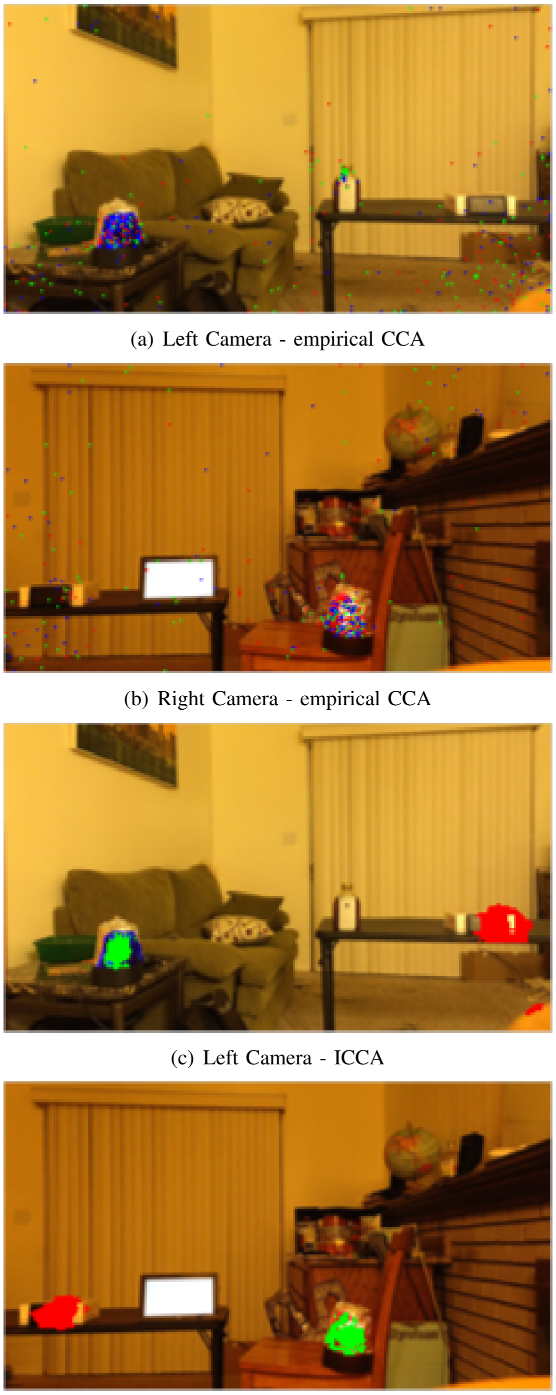
Proof of Proposition V.2

Bao et al. [23] proved this result for a slightly simplified model. Here we provide the linear transformations to recover their model. We may write our data matrices X and Y jointly via,

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} R_{xx} & R_{xy} \\ R_{xy}^H & R_{yy} \end{bmatrix}^{1/2} \begin{bmatrix} W_1 \\ W_2 \end{bmatrix}$$

where W_1 is a $p \times n$ matrix with independent $\mathcal{N}(0,1)$ entries and W_2 is an independent $q \times n$ matrix with independent $\mathcal{N}(0,1)$. As $p+q < n$, R_{xx} and R_{yy} are non-singular. Define

$$\begin{aligned} \begin{bmatrix} \widetilde{X} \\ \widetilde{Y} \end{bmatrix} &= \begin{bmatrix} R_{xx}^{-1/2} & 0 \\ 0 & R_{yy}^{-1/2} \end{bmatrix}^{1/2} \begin{bmatrix} X \\ Y \end{bmatrix} \\ &= \begin{bmatrix} I_p & R_{xx}^{-1/2} R_{xy} R_{yy}^{-1/2}. \\ R_{yy}^{-1/2} R_{xy}^H R_{xx}^{-1/2} & I_q \end{bmatrix} \begin{bmatrix} W_1 \\ W_2 \end{bmatrix}. \end{aligned}$$



(a) Left Camera - empirical CCA

(b) Right Camera - empirical CCA

(c) Left Camera - ICCA

(d) Right Camera - ICCA

Fig. 7. (a)-(b) Top 3 threholded empirical CCA canonical vectors overlayed on the original scene after 800 frames as computed in (13). The red pixels correspond to the vector with the highest correlation, the green pixels correspond to the vector with the second highest correlation, and the blue pixels correspond to the vector with the third highest correlation in Figure 6(a). We use a threshold of $\log(p)/\sqrt{p}$ where $p = 32400$ pixels. (c)-(d) Top 2 threholded ICCA canonical vectors overlayed on video after 800 frames as computed in (17). The red pixels correspond to the vector with the highest correlation and the green pixels correspond to the vector with the second highest correlation in Figure 6(b). We again use a threshold of $\log(p)/\sqrt{p}$ where $p = 32400$ pixels.

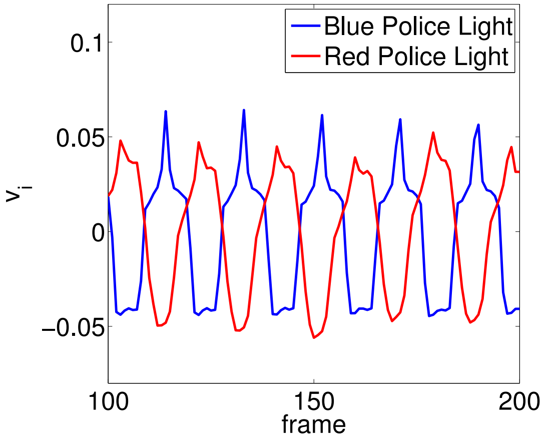


Fig. 8. A portion of the right singular vectors of X_{left} (blue) and Y_{right} (red) corresponding the flashing police lights in each camera view. Both sources have very similar periods and are approximately in antiphase.

With the definitions of the covariance matrices in (4), we have that $R_{xx}^{-1/2}R_{xy}R_{yy}^{-1/2}$

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

From this expression, it is clear why we defined $\widetilde{K}_{xy}$ as we originally did. Let $U_{\widetilde{K}_{xy}}KV_{\widetilde{K}_{xy}}$ be the SVD of $\widetilde{K}_{xy}$, where K is the $k_x \times k_y$ matrix with κ_j along the diagonal. Define $F = \left[\left(U_xU_{\widetilde{K}_{xy}}\right)\ \left(U_xU_{\widetilde{K}_{xy}}\right)^{\perp}\right]$ and $G = \left[\left(U_yV_{\widetilde{K}_{xy}}\right)\ \left(U_yV_{\widetilde{K}_{xy}}\right)^{\perp}\right]$. Then

$$\begin{aligned}\begin{bmatrix}\widetilde{\widetilde{X}}\\ \widetilde{\widetilde{Y}}\end{bmatrix} &= \begin{bmatrix}F^H & 0\\ 0 & G^H\end{bmatrix}^{1/2}\begin{bmatrix}\widetilde{X}\\ \widetilde{Y}\end{bmatrix}\\ &= \begin{bmatrix}F^HR_{yy}^{-1/2} & 0\\ 0 & G^HR_{yy}^{-1/2}\end{bmatrix}^{1/2}\begin{bmatrix}X\\ Y\end{bmatrix}\\ &= \begin{bmatrix}I_p & K\\ K^H & I_q\end{bmatrix}^{1/2}\begin{bmatrix}W_1\\ W_2\end{bmatrix}.\end{aligned}$$

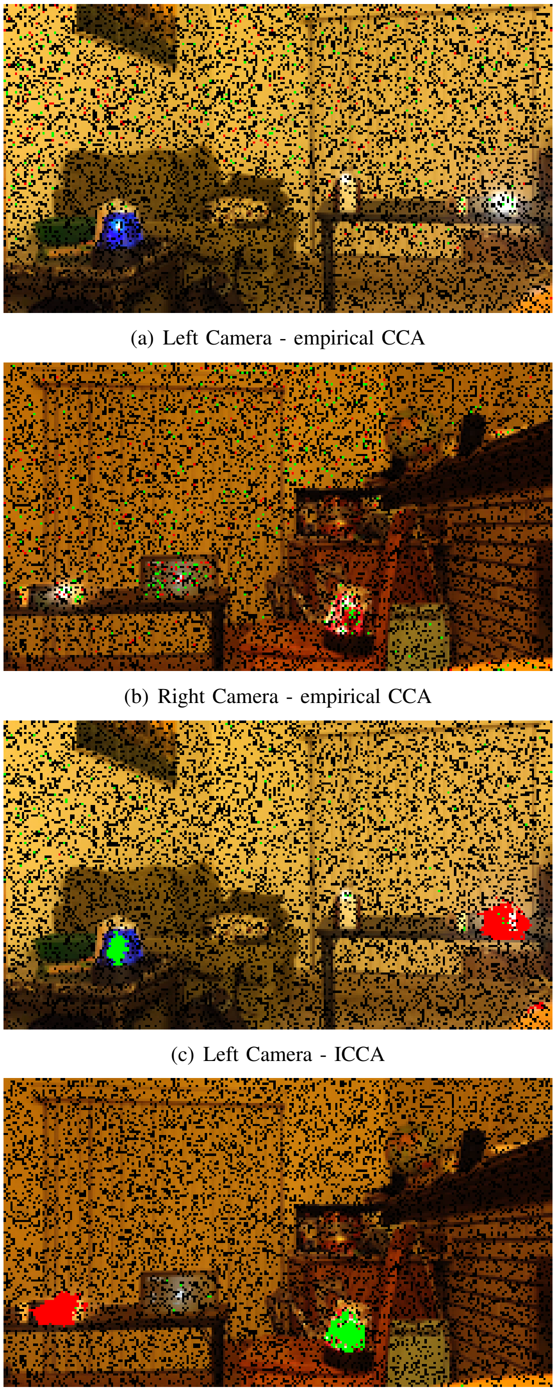
Transforming X and Y to $\widetilde{\widetilde{X}}$ and $\widetilde{\widetilde{Y}}$ preserves the canonical correlation estimates because our transformation matrix is nonsingular. After this transformation, we follow the proof from Bao et al. [23] with $\sqrt{r_i} = \kappa_i$.

Theorem needed to prove Theorem V.1

Theorem A.1. *Let $\widetilde{u}_i$ and $\widetilde{v}_i$ be the left and right singular vectors associated with the i-th singular value, $\widetilde{\theta}_i$, of the $p \times n$ matrix*

$$\widetilde{X} = \sum_{i=1}^{k}\underbrace{\theta_i u_i v_i^H}_{P} + X.$$

Assume that X satisfies the hypotheses in Assumption V.2 and suppose that $\theta_i > c^{1/4}$ for $c = p/n$. Let w be an arbitrary unit



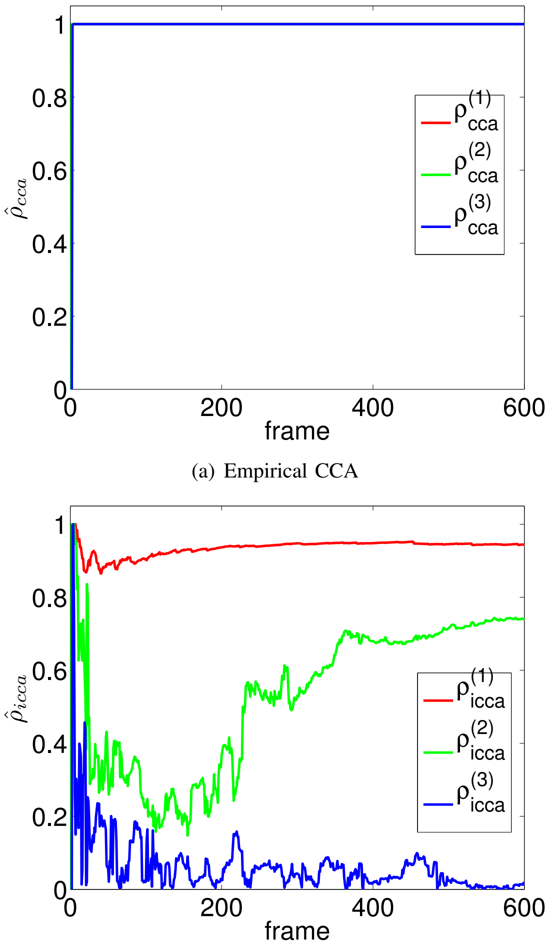
(a) Left Camera - empirical CCA

(b) Right Camera - empirical CCA

(c) Left Camera - ICCA

(d) Right Camera - ICCA

Fig. 9. (a)-(b) Top 2 threholded empirical CCA canonical vectors overlayed on missing data video as computed in (13). Again we use the threshold $\log(p)/\sqrt{p}$ for $p = 32400$. The red pixels correspond to the vector with the highest correlation and the green pixels correspond to the vector with the second highest correlation in Figure 10(a). (c)-(d) Top 2 threholded ICCA canonical vectors overlayed on missing data video as computed in (17). Again we use the threshold $\log(p)/\sqrt{p}$ for $p = 32400$. The red pixels correspond to the vector with the highest correlation and the green pixels correspond to the vector with the second highest correlation in Figure 10(a). For all figures, $\gamma_x = \gamma_y = 0.75$ so that 25% of our pixels are missing. We note that the middle source of the left camera violates the low-coherence assumption in Assumption V.1 and so Theorem V.3 and Conjecture V.3 provide no guarantees for detecting correlations based on this source.



(a) Empirical CCA

(b) ICCA

Fig. 10. (a) Top three singular values returned by empirical CCA as defined in (12). As we are in the sample deficient regime, these singular values are deterministically 1. (b) Top three singular values returned by ICCA as defined in (16). ICCA correctly identifies two sources of correlation. As our data matrices now have missing data, it takes more frames for ICCA to identify the two sources of correlations. For both figures, $\gamma_x = \gamma_y = 0.75$ so that 25% of our pixels are missing. This figure is analogous to Figure 6, which observes all data so that $\gamma_x = \gamma_y = 1$.

norm vector that is orthogonal to u_i for some $i \in \{1, \ldots, k\}$. Then as $n, p \to \infty$ such that $p/n \to c$, we have that

$$\langle w, \widetilde{u}_i \rangle \xrightarrow{a.s.} 0.$$

Proof. The result of this theorem may be of interest outside of this paper for analysis of similar low-rank signal-plus-noise matrix models. We will use this theorem to prove Theorem V.1. We begin with a technical lemma needed to prove the theorem.

Lemma A.1. *Let $U = [u_1, \ldots, u_k] \in \mathbb{C}^{p \times k}$ and $V = [v_1, \ldots, v_k] \in \mathbb{C}^{n \times k}$ be independent matrices with orthonormal columns. Let $X \in \mathbb{C}^{p \times n}$ satisfy the hypotheses in Assumption V.2. Then as $n, p \to \infty$ with $p/n \to c$, for $i \neq j$*

$$u_i^H \left(z^2 I_n - XX^H\right)^{-1} u_j \xrightarrow{a.s.} 0.$$

Similarly, for all i, j,

$$u_i^H \left(z^2 I_n - XX^H\right)^{-1} X v_j \xrightarrow{a.s.} 0.$$

Proof. The proof of Lemma 4.1 in [41] proves both of these statements. □

We now are in a position to prove Theorem A.1. If $w \in \textbf{span}(u_1, \dots, u_k)$, then Theorem 2.10 c) of [41] proves our result. If $w \notin \textbf{span}(u_1, \dots, u_k)$, then we may write

$$w = w_u + w_u^{\perp},$$

where $w_u \in \textbf{span}(u_1, \dots, u_k)$ and $w_u^{\perp}$ is in the orthocomplement of $\textbf{span}(u_1, \dots, u_k)$. Therefore applying Theorem 2.10 c) of [41]

$$\langle w, \widetilde{u}_i \rangle = \langle w_u, \widetilde{u}_i \rangle + \langle w_u^{\perp}, \widetilde{u}_i \rangle \xrightarrow{\text{a.s.}} \langle w_u^{\perp}, \widetilde{u}_i \rangle ,$$

so we only must focus on $w_u^{\perp}$.

Based on their definitions, $\widetilde{X}\widetilde{X}^H\widetilde{u}_i = \widetilde{\theta}_i^2\widetilde{u}_i$ and $\widetilde{X}^H\widetilde{u}_i = \widetilde{\theta}_i\widetilde{v}$. Using the fact that $\widetilde{X} = P + X$, we have

$$\left(PP^H + PX^H + XP^H + XX^H\right)\widetilde{u}_i = \widetilde{\theta}_i^2\widetilde{u}_i \tag{28}$$

and $\left(X^H + P^H\right)\widetilde{u}_i = \widetilde{\theta}_i\widetilde{v}$. Multiplying both sides of this second expression by P, we have

$$PX^H\widetilde{u}_i + PP^H\widetilde{u}_i = \widetilde{\theta}_i P\widetilde{v}_i.$$

Substituting this expression in (28) gives

$$\widetilde{\theta}_i P\widetilde{v}_i + XP^H\widetilde{u}_i + XX^H\widetilde{u}_i = \widetilde{\theta}_i^2\widetilde{u}_i.$$

Rearranging terms gives

$$\widetilde{u}_i = \left(\widetilde{\theta}_i^2 I_p - XX^H\right)^{-1}\left(\widetilde{\theta}_i P\widetilde{v}_i + XP^H\widetilde{u}_i\right).$$

Therefore, we have the equivalences for $\langle w_u^{\perp}, \widetilde{u}_i \rangle$

$$\begin{aligned} &= w_u^{\perp H}\left(\widetilde{\theta}_i^2 I_p - XX^H\right)^{-1}\left(\widetilde{\theta}_i P\widetilde{v}_i + XP^H\widetilde{u}_i\right) \\ &= \widetilde{\theta}_i w_u^{\perp H}\left(\widetilde{\theta}_i^2 I_p - XX^H\right)^{-1}\sum_{j=1}^{k}\theta_j\langle v_j, \widetilde{v}_i\rangle u_j \\ &\quad + \widetilde{\theta}_i w_u^{\perp H}\left(\widetilde{\theta}_i^2 I_p - XX^H\right)^{-1} X\sum_{j=1}^{k}\theta_j\langle u_j, \widetilde{u}_i\rangle v_j. \end{aligned}$$

By Theorem 2.7 c) in [41], we have that for $i \neq j$, $|\langle u_j, \widetilde{u}_i\rangle| \xrightarrow{\text{a.s.}} 0$ and $|\langle v_j, \widetilde{v}_i\rangle| \xrightarrow{\text{a.s.}} 0$. Therefore

$$\begin{aligned} \langle w_u^{\perp}, \widetilde{u}_i\rangle = &\left(\widetilde{\theta}_i\theta_i\langle v_i, \widetilde{v}_i\rangle\right) w_u^{\perp H}\left(\widetilde{\theta}_i^2 I_p - XX^H\right)^{-1} u_i \\ &+ \left(\widetilde{\theta}_i\theta_i\langle u_i, \widetilde{u}_i\rangle\right) w_u^{\perp H}\left(\widetilde{\theta}_i^2 I_p - XX^H\right)^{-1} Xv_i. \end{aligned} \tag{29}$$

By Lemma A.1,

$$w_u^{\perp H}\left(\widetilde{\theta}_i^2 I_p - XX^H\right)^{-1} u_i \xrightarrow{\text{a.s.}} 0$$

and

$$w_u^{\perp H}\left(\widetilde{\theta}_i^2 I_p - XX^H\right)^{-1} Xv_i \xrightarrow{\text{a.s.}} 0.$$

Therefore, $\langle w_u^{\perp}, \widetilde{u}_i\rangle \xrightarrow{\text{a.s.}} 0$. □

PROOF OF THEOREM V.1

The entries of the matrix $\mathring{V}_x^H\mathring{V}_y$ are the inner products between the columns of $\mathring{V}_x$ and $\mathring{V}_y$

$$\left|\left(\mathring{V}_x^H\mathring{V}_y\right)\right|_{ij} = \left|\mathring{V}_x^H(:,i)\mathring{V}_y(:,j)\right|.$$

Notice that we may write

$$\begin{aligned} \mathring{V}_x(:,i) &= aV_y(:,j) + bw_y \\ V_y(:,j) &= k_{ij}^{xy}V_x(:,i) + cw_x \end{aligned} \tag{30}$$

for some arbitrary unit-norm vector w_x that is orthogonal to $V_x(:,i)$, some arbitrary unit-norm vector w_y that is orthogonal to $V_y(:,j)$, and constants a, b, and c. With these observations, we have

$$\begin{aligned} \mathring{V}_x^H(:,i)\mathring{V}_y(:,j) \quad &= \left(aV_y(:,j) + bw_y\right)^H\mathring{V}_y(:,j) \\ &= aV_y^H(:,j)\mathring{V}_y(:,j) + bw_y^H\mathring{V}_y(:,j). \end{aligned}$$

By Theorem A.1, $w_y^H\mathring{V}_y(:,j) \xrightarrow{\text{a.s.}} 0$. As derived in theorem 5.2 in [34],

$$\left|V_y^H(:,j)\mathring{V}_y(:,j)\right| \xrightarrow{\text{a.s.}} \alpha_{y,j} =: \begin{cases} \sqrt{1 - \frac{c_y + \theta_j^{(y)}}{\theta_j^{(y)}(\theta_j^{(y)} + c_x)}} & \theta_j^{(y)} > c^{1/4} \\ 0 & \text{otherwise} \end{cases}.$$

Therefore, $\left|\mathring{V}_x^H(:,i)\mathring{V}_y(:,j)\right| \xrightarrow{\text{a.s.}} a\alpha_{y,j}$. Using the expression for $\mathring{V}_x(:,i)$ in (30), we observe that

$$V_y(:,j)^H\mathring{V}_x(:,i) = a.$$

Using the expression for $V_y(:,j)$ in (30), we have

$$\begin{aligned} a \quad &= V_y(:,j)^H\mathring{V}_x(:,i) \\ &= \left(k_{ij}^{xy}V_x(:,i) + cw_x\right)^H\mathring{V}_x(:,i) \\ &= k_{ij}^{xy}V_x^H(:,i)\mathring{V}_x(:,i) + cw_x^H\mathring{V}_x(:,i). \end{aligned}$$

By Theorem A.1, $w_x^H\mathring{V}_x(:,i) \xrightarrow{\text{a.s.}} 0$. As derived in theorem 5.1 of [34],

$$\left|V_y(:,j)\mathring{V}_y(:,j)\right| \xrightarrow{\text{a.s.}} =: \alpha_{x,i}\begin{cases} \sqrt{1 - \frac{c_x + \theta_i^{(x)}}{\theta_i^{(x)}(\theta_i^{(x)} + c_x)}} & \theta_i^{(x)} > c^{1/4} \\ 0 & \text{otherwise} \end{cases}.$$

Therefore, $|a| \xrightarrow{\text{a.s.}} \left|k_{ij}^{xy}\right|\alpha_{x,i}$. Therefore,

$$\left|\mathring{V}_x^H(:,i)\mathring{V}_y(:,j)\right| \xrightarrow{\text{a.s.}} \left|k_{ij}^{xy}\right|\alpha_{x,i}\alpha_{y,i}.$$

PROOF OF THEOREM V.2

For ICCA, recall that

$$\widehat{k}_{\text{icca}} = \sum_{i=1}^{\min(\widehat{k}_x,\widehat{k}_y)} \mathbb{1}_{\left\{\left(\widehat{\rho}_{\text{icca}}^{(i)}\right)^2 > \tau_{\text{icca}}^{\alpha}\right\}}.$$

When $\widehat{k}_x = k_x$ and $\widehat{k}_y = k_y$, the estimate of the number of correlated signals becomes

$$\widehat{k}_{\text{icca}} \xrightarrow{\text{a.s.}} \sum_{i=1}^{\min(k_x,k_y)} \mathbb{1}_{\left\{\left(\widehat{\rho}_{\text{icca}}^{(i)}\right)^2 > \tau_{\text{icca}}^{\alpha}\right\}}.$$

To prove the theorem, we want to show that

$$\widehat{\rho}_{\text{icca}}^{(k)} > 0 \text{ almost surely,}$$

under the conditions on Θ_x and Θ_y in the theorem statement. Momentarily, we assume that $k = \min(k_x, k_y)$. The singular values of $\overset{\circ}{V}_x^H \overset{\circ}{V}_y$ are ordered and so we must show that

$$\left(\widehat{\rho}_{\text{icca}}^{(\min(k_x,k_y))}\right)^2 > \tau_{\text{icca}}^{\alpha} \text{ almost surely.}$$

From Theorem V.1, we also know that

$$\left|\left[\overset{\circ}{V}_x^H \overset{\circ}{V}_y\right]_{ij}\right| \xrightarrow{\text{a.s.}} \left|k_{ij}^{xy}\right| \alpha_{x,i} \alpha_{y,i}.$$

Using this fact we define

$$\begin{aligned} A_x &= \mathbf{diag}(\alpha_{x,1}, \ldots, \alpha_{x,k_x}) \\ A_y &= \mathbf{diag}(\alpha_{y,1}, \ldots, \alpha_{y,k_y}) \end{aligned}$$

so that we may write

$$\overset{\circ}{V}_x^H \overset{\circ}{V}_y = A_x K_{xy} A_y + \Delta,$$

where $\Delta = [\delta_{ij}]$ such that $\delta_{ij} \xrightarrow{\text{a.s.}} 0$. Examining (25), we see that under the above conditions on Θ_x and Θ_y, A_x and A_y are both full rank. Define

$$\begin{aligned} \alpha_{x,\min} &= \min_{i=1\ldots,k_x} \alpha_{x,i} \\ \alpha_{y,\min} &= \min_{i=j\ldots,k_y} \alpha_{y,j}. \end{aligned}$$

By properties of singular values

$$\begin{aligned} \sigma_{\min}(A_x K_{xy} A_y) - \sigma_{\max}(\Delta) &\leq \sigma_{\min}(A_x K_{xy} A_y + \Delta) \\ &\leq \sigma_{\min}(A_x K_{xy} A_y) \\ &\quad + \sigma_{\max}(\Delta). \end{aligned}$$

Examining $\sigma_{\max}(\Delta)$, we observe that

$$\sigma_{\max}(\Delta) \leq \|\Delta\|_F = \sqrt{\sum_{i=1}^{k_x} \sum_{j=1}^{k_y} |\delta_{ij}|^2}.$$

Using the fact that $\delta_{ij} \xrightarrow{\text{a.s.}} 0$, we have that $\sigma_{\max}(\Delta) \xrightarrow{\text{a.s.}} 0$. Therefore, almost surely

$$\sigma_{\min}(A_x K_{xy} A_y) \leq \sigma_{\min}(A_x K_{xy} A_y + \Delta) \leq \sigma_{\min}(A_x K_{xy} A_y),$$

which implies that

$$\widehat{\rho}_{\text{icca}}^{(\min(k_x,k_y))} \xrightarrow{\text{a.s.}} \sigma_{\min}(A_x K_{xy} A_y)$$

By properties of singular values

$$\begin{aligned} \widehat{\rho}_{\text{icca}}^{(\min(k_x,k_y))} &= \sigma_{\min(k_x,k_y)}\left(\overset{\circ}{V}_x^H \overset{\circ}{V}_y\right) \\ &\xrightarrow{\text{a.s.}} \sigma_{\min(k_x,k_y)}\left(A_x K_{xy} A_y\right) \\ &\geq \sigma_{k_x}\left(A_x\right) \sigma_{\min(k_x,k_y)}\left(K_{xy}\right) \sigma_{k_y}\left(A_y\right) \\ &= \alpha_{x,\min} \kappa_{\min(k_x,k_y)} \alpha_{y,\min} > 0, \end{aligned}$$

and we have proved the desired result.

Proof of Theorem V.3

Defining $P_x = U_x V_x^H$ We may write (27) as

$$\begin{aligned} X &= \underbrace{P_x \odot M_x}_{\widehat{P}_x} + \underbrace{Z_x \odot M_x}_{\widehat{Z}_x} \\ &= \mathbb{E}\left[\widehat{P}_x\right] + \widehat{Z}_x + \Delta_{\widehat{P}_x} \\ &= \underbrace{\gamma_x P_x + \widehat{Z}_x}_{\widetilde{X}} + \Delta_{\widehat{P}_x}. \end{aligned}$$

Similarly, we may write $Y = \widetilde{Y} + \Delta_{\widehat{P}_y}$ where $\widetilde{Y} = \gamma_y P_y + \widehat{Z}_y$.

First we show that the maximum singular value of $\Delta_{\widehat{P}_x}$ and $\Delta_{\widehat{P}_y}$ converge almost surely to 0. Under the low-coherence assumption, we have that

$$\begin{aligned} \max_{ij} |P_x|_{ij} &\leq \max_i \theta_i^{(x)} \max_k \|u_k^{(x)}\|_\infty \max_\ell \|v_\ell^{(x)}\|_\infty \\ &= \max_i \theta_i^{(x)} \mathcal{O}\left(\frac{\log\, n, p \text{ factors}}{\sqrt{np}}\right). \end{aligned} \tag{31}$$

By assumption that $c_x > 0$, we have that $n = \mathcal{O}(p)$. This fact, coupled with the fact that $\theta_i^{(x)}$ is not dependent on n gives

$$\max_{ij} |P_x|_{ij} \leq \mathcal{O}\left(\frac{\log\, n \text{ factors}}{n}\right). \tag{32}$$

To characterize the largest singular value of $\Delta_{\widehat{P}_x}$, we want to use Latala's theorem [47], which states that for a matrix A with independent mean zero random entries with bounded fourth moment

$$\begin{aligned} \mathbb{E}\left[\sigma_1(A)\right] \leq \; C & \left[\max_i \left(\sum_j \mathbb{E}\left[A_{ij}^2\right]\right)^{1/2}\right. \\ & + \max_j \left(\sum_i \mathbb{E}\left[A_{ij}^2\right]\right)^{1/2} \\ & \left. + \left(\sum_{i,j} \mathbb{E}\left[A_{ij}^4\right]\right)^{1/4}\right] \end{aligned}$$

for some universal constant C that does not depend on n or p. Through basic calculation, one can show that

$$\begin{aligned} \mathbb{E}\left[\left(\Delta_{\widehat{P}_x}\right)_{ij}^2\right] &= \gamma_x\left(1 - \gamma_x\right)\left(P_x\right)_{ij}^2 \\ \mathbb{E}\left[\left(\Delta_{\widehat{P}_x}\right)_{ij}^4\right] &= \left(-3\gamma_x^4 + 6\gamma_x^3 - 4\gamma_x^2 + \gamma\right)\left(P_x\right)_{ij}^4. \end{aligned}$$

These expressions satisfy the conditions on Latala's theorem. Therefore, by substituting these expressions into Latala's theorem with the bound in (32), we have

$$\mathbb{E}\left[\sigma_1\left(\Delta_{\widehat{P}_x}\right)\right] \leq \mathcal{O}\left(\frac{\log\, n \text{ factors}}{\sqrt{n}}\right).$$

By concentration and convexity of the largest singular value (see [44] page 3015), we have that in our asymptotic regime

$$\sigma_1(\Delta_{\widehat{P}_x}) \xrightarrow{\text{a.s.}} 0.$$

Using a similar argument

$$\sigma_1(\Delta_{\widehat{P}_y}) \xrightarrow{\text{a.s.}} 0.$$

Therefore we have that

$$\begin{aligned} X &\to \gamma_x P_x + \widehat{Z}_x \\ Y &\to \gamma_y P_y + \widehat{Z}_y. \end{aligned}$$

Examining $\widehat{Z}_x$, we have

$$\begin{aligned} \mathbb{E}\left[\widehat{Z}_{ij}^{(x)}\right] = \quad & \mathbb{E}\left[\widehat{Z}_{ij}^{(x)} | M_{ij}^{(x)} = 0\right] \mathbb{P}\left(M_{ij}^{(x)} = 0\right) \\ & + \mathbb{E}\left[\widehat{Z}_{ij}^{(x)} | M_{ij}^{(x)} = 1\right] \mathbb{P}\left(M_{ij}^{(x)} = 1\right) = 0 \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}\left[\left(\widehat{Z}_{ij}^{(x)}\right)^2\right] = \quad & \mathbb{E}\left[\left(\widehat{Z}_{ij}^{(x)}\right)^2 | M_{ij}^{(x)} = 0\right] \mathbb{P}\left(M_{ij}^{(x)} = 0\right) + \\ & \mathbb{E}\left[\left(\widehat{Z}_{ij}^{(x)}\right)^2 | M_{ij}^{(x)} = 1\right] \mathbb{P}\left(M_{ij}^{(x)} = 1\right) \\ = \quad & 0 + \gamma_x. \end{aligned}$$

Therefore, $\widehat{Z}_{ij}^{(x)}$ are i.i.d. zero mean with variance γ_x and $\widehat{Z}_x \to \sqrt{\gamma_x} Z_x$. Using this observation,

$$\begin{aligned} X \quad & \xrightarrow{\text{a.s.}} \gamma_x P_x + \widehat{Z}_x \\ & \to \gamma_x P_x + \sqrt{\gamma_x} Z_x \\ & = \sqrt{\gamma_x}\left(\sqrt{\gamma_x} P_x + Z_x\right). \end{aligned}$$

Similarly, $Y \to \sqrt{\gamma_y}\left(\sqrt{\gamma_y} P_y + Z_y\right)$.

From this we can conclude that eigenvector expressions of the form $\langle u, \widehat{u} \rangle$ behave as if we replace Θ_x with $\gamma_x \Theta_x$ and Θ_y with $\gamma_y \Theta_y$. Consider

$$\begin{aligned} \Big| u_i^H \Big(zI - (Z_x + \Delta_{\widehat{P}_x}) \quad & (Z_x + \Delta_{\widehat{P}_x})^H\Big)^{-1} \cdot \\ & u_j - u_i^H \left(zI - Z_x Z_x^H\right)^{-1} u_j \Big|, \end{aligned} \tag{33}$$

which as a consequence of the variational characterization of the largest singular value is upper bounded by

$$\sigma_1 \left(\left(zI - (Z_x + \Delta_{\widehat{P}_x})(Z_x + \Delta_{\widehat{P}_x})^H\right)^{-1} - \left(zI - Z_x Z_x^H\right)^{-1} \right). \tag{34}$$

Following a similar argument in [44] (equation 34), we have that (34) is upper bounded by

$$\frac{3\sigma_z(Z_x)}{\Im w} \sigma_1(\Delta_{\widehat{P}_x}),$$

where $\Im w > 0$. Using the facts that $\sigma_1(Z_x) \xrightarrow{\text{a.s.}} \sqrt{\gamma_x} b_x$ by the above relationship and $\sigma_1(\Delta_{\widehat{P}_x}) \xrightarrow{\text{a.s.}} 0$ combined with Assumption V.2, we have that (33) converges to 0. Using a similar argument, one can prove the same result for quadratic forms with Z_y and $\Delta_{\widehat{P}_y}$.

Therefore, an analogous version of Theorem A.1 holds for the missing data section, as the quadratic forms used by Lemma A.1 still hold. Therefore, we prove the theorem following the same rank argument as Theorem V.2, except that we replace Θ_x with $\gamma_x \Theta_x$ and replace Θ_y with $\gamma_y \Theta_y$.

Acknowledgment

This work supported by an AFOSR Young Investigator Award FA9550-12-1-0266, ONR Young Investigator Award N00014-11-1-0660, US Army Research Office (ARO) under grant W911NF-11-1-0391 and NSF grant CCF-1116115.

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. Zhai, Y. Zhang, D.-Y. Yeung, H. Chang, X. Chen, and W. Gao, "Instance-specific canonical correlation analysis," *Neurocomputing*, 2015.

[5] 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.

[6] 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.

[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] M. R. Arbabshirani, M. Nakhkash, and H. Soltanian-Zadeh, "Comparison of canonical correlation analysis and ica techniques for fMRI," in *Communications, Control and Signal Processing (ISCCSP), 2010 4th International Symposium on*. IEEE, 2010, pp. 1–5.

[9] 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.

[10] D. Lin, J. Zhang, J. Li, V. Calhoun, and Y.-P. Wang, "Identifying genetic connections with brain functions in schizophrenia using group sparse canonical correlation analysis," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*. IEEE, 2013, pp. 278–281.

[11] 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.

[12] Y. Zhang, G. Zhou, J. Jin, M. Wang, X. Wang, and A. Cichocki, "L1-regularized multiway canonical correlation analysis for SSVEP-based BCI," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 21, no. 6, pp. 887–896, 2013.

[13] M. Nakanishi, Y. Wang, Y.-T. Wang, Y. Mitsukura, and T.-P. Jung, "Enhancing unsupervised canonical correlation analysis-based frequency detection of ssveps by incorporating background EEG," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 3053–3056.

[14] 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," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 22, no. 6, pp. 1097–1103, 2014.

[15] J. Kuzilek, V. Kremen, and L. Lhotska, "Comparison of jade and canonical correlation analysis for ECG de-noising," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 3857–3860.

[16] K. Todros and A. Hero, "Measure transformed canonical correlation analysis with application to financial data," in *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2012 IEEE 7th*. IEEE, 2012, pp. 361–364.

[17] D. S. Wilks, "Probabilistic canonical correlation analysis forecasts, with application to tropical pacific sea-surface temperatures," *International Journal of Climatology*, vol. 34, no. 5, pp. 1405–1413, 2014.

[18] A. J. Prera, K. M. Grimsrud, J. A. Thacher, D. W. McCollum, and R. P. Berrens, "Using canonical correlation analysis to identify environmental attitude groups: Considerations for national forest planning in the southwestern us," *Environmental management*, vol. 54, no. 4, pp. 756–767, 2014.

[19] J. L. Steward, Z. Haddad, S. Hristova-Veleva, and T. Vukicevic, "Assimilating scatterometer observations of tropical cyclones into an ensemble kalman filter system with a robust observation operator based on canonical-correlation analysis," in *SPIE Asia Pacific Remote Sensing*. International Society for Optics and Photonics, 2014, pp. 926 507–926 507.

[20] L. L. Scharf and J. K. Thomas, "Wiener filters in canonical coordinates for transform coding, filtering, and quantizing," *Signal Processing, IEEE Transactions on*, vol. 46, no. 3, pp. 647–654, 1998.

[21] H. Ge, I. Kirsteins, and X. Wang, “Does canonical correlation analysis provide reliable information on data correlation in array processing?” in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 2113–2116.

[22] 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.

[23] Z. Bao, J. Hu, G. Pan, and W. Zhou, “Canonical correlation coefficients of high-dimensional normal vectors: finite rank case,” *arXiv preprint arXiv:1407.7194*, 2014.

[24] I. M. Johnstone, “Multivariate analysis and jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence,” *Annals of statistics*, vol. 36, no. 6, p. 2638, 2008.

[25] Y. Fujikoshi and T. Sakurai, “High-dimensional asymptotic expansions for the distributions of canonical correlations,” *Journal of Multivariate Analysis*, vol. 100, no. 1, pp. 231–242, 2009.

[26] M. J. Wainwright, “Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso),” *IEEE transactions on information theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[27] Y. Yang, M. J. Wainwright, and M. I. Jordan, “On the computational complexity of high-dimensional bayesian variable selection,” *arXiv preprint arXiv:1505.07925*, 2015.

[28] L. Janson, R. F. Barber, and E. Candes, “Eigenprism: Inference for high-dimensional signal-to-noise ratios,” *arXiv preprint arXiv:1505.02097*, 2015.

[29] C. Eckart and G. Young, “The approximation of one matrix by another of lower rank,” *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[30] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[31] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[32] F. Bach and M. Jordan, “Kernel independent component analysis,” *The Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2003.

[33] E. J. Candes and Y. Plan, “Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements,” *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[34] 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.

[35] R. Nadakuditi and J. Silverstein, “Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 3, pp. 468–480, 2010.

[36] E. J. Candes, J. K. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Communications on pure and applied mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[37] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[38] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[39] E. J. Candes and Y. Plan, “Matrix completion with noise,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[40] E. J. Candès and T. Tao, “The power of convex relaxation: Near-optimal matrix completion,” *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[41] F. Benaych-Georges and R. R. Nadakuditi, “The singular values and vectors of low rank perturbations of large rectangular random matrices,” *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.

[42] ——, “The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices,” *Advances in Mathematics*, vol. 227, no. 1, pp. 494–521, 2011.

[43] A. Constantine and R. J. Muirhead, “Asymptotic expansions for distributions of latent roots in multivariate analysis,” *Journal of Multivariate Analysis*, vol. 6, no. 3, pp. 369–391, 1976.

[44] R. R. Nadakuditi, “Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage,” *Information Theory, IEEE Transactions on*, vol. 60, no. 5, pp. 3002–3018, 2014.

[45] M. Welling, “A first encounter with machine learning,” *Irvine, CA.: University of California*, pp. 1–93, 2011.

[46] S. Yu, B. De Moor, and Y. Moreau, “Learning with heterogenous data sets by weighted multiple kernel canonical correlation analysis,” in *Machine Learning for Signal Processing, 2007 IEEE Workshop on*. IEEE, 2007, pp. 81–86.

[47] R. Latała, “Some estimates of norms of random matrices,” *Proceedings of the American Mathematical Society*, vol. 133, no. 5, pp. 1273–1282, 2005.

Nicholas Asendorf is a data scientist in the Corporate Research Labs at 3M in St. Paul, Minnesota, USA. He received his Ph.D. in Electrical Engineering:Systems from the University of Michigan in 2015. His thesis research focused on data fusion and the need for data driven algorithms in statistical signal processing and machine learning, particularly in sample starved regimes. At 3M he applies recent advances in signal processing, machine learning, and data science to enhance existing products and develop new-to-the-world technologies.

Raj Rao Nadakuditi is an Associate Professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. He received his PhD in 2007 from the Massachusetts Institute of Technology and the Woods Hole Oceanographic Institution. He was awarded an Ofce of Naval Research Young Investigator Award in 2011, an Air Force Ofce of Scientic Research Young Investigator Award in 2012, the Signal Processing Society Best Young Author Paper Award in 2012 and the DARPA Young Faculty Award in 2014. His research focuses on developing theory for random matrices for applications in signal processing, machine learning, queuing theory and scattering theory.