ESTIMATING THE NUMBER OF CORRELATED COMPONENTS BASED ON RANDOM PROJECTIONS

Christian Lameiro, Tanuj Hasija, Tim Marrinan, and Peter J. Schreier

Signal and System Theory Group, Universität Paderborn, Germany, http://sst.upb.de

ABSTRACT

Estimating the number of correlated components between two data sets is a challenging task in the case of small sample support. Typically, a rank-reduction preprocessing step based on principal component analysis (PCA) is carried out on each data set individually to reduce the dimensionality before analyzing correlation between the data sets. However, PCA retains the components with the largest variance within a data set, and therefore fails when these components are not the ones that account for the correlation between the data sets. To overcome this, we propose an alternative technique that, instead of projecting the data into a single subspace, uses a large number of random projections.

Index Terms— Correlation analysis, small sample support, random projection, KL divergence.

1. INTRODUCTION

We consider the problem of estimating the number of correlated components between two data sets. This is a common problem in different areas such as biomedicine [1], climate science [2], and array processing [3]. It becomes very challenging when the number of available samples is small compared to the dimension of the data sets. This setting is called *small sample support* and is a common scenario in practice.

Correlation can be assessed by canonical correlation analysis (CCA) [4]. In CCA, each data set is linearly transformed into canonical variables, which represent the underlying correlation structure. Specifically, the ith pair of canonical variables exhibit the highest correlation under the constraint of being uncorrelated with the previous $i-1$ canonical variables. The correlations between the canonical variables are the canonical correlations. These correlations have to be estimated from samples, which is the main challenge in case of small sample support: When the number of samples is small, the sample canonical correlations highly overestimate the true ones [5]. This can be overcome by a dimensionality-reduction preprocessing step, but how this reduction is achieved has a strong influence on the performance of the scheme. A typical rank-reduction technique is principal component analysis (PCA) [6, 7]. In [6] white noise is assumed, and the number of components kept by PCA is chosen independently of the CCA step. Colored noise is allowed by the techniques in [7], where the PCA and CCA ranks are jointly obtained, providing very good results for a wide variety of scenarios. However, PCA is in general not the optimal rank-reduction approach for this problem. First, PCA retains the components with largest variance in the observed data, which do not necessarily correspond to the correlated components. Second, the signal subspace estimated by PCA may leak into the estimated noise subspace due to the small number of samples and colored noise [8].

Alternatives to PCA have also been studied in the past. In [9], diagonal loading together with results from random matrix theory are applied. However, the proposed approach only identifies whether or not the data sets are correlated, and it does not estimate the number of correlated signals. In [10, 11] techniques based on cross-validation (CV) and sparse CCA (SCCA), respectively, are proposed as alternatives to PCA-based approaches. Even though these techniques can overcome some of the issues associated with PCA, they still have some limitations as pointed out in [10, 11]. For example, the CV techniques [10] are more sensitive to the variance of the independent components, while SCCA [11] is computationally demanding. In this paper we propose a new technique that overcomes some of these limitations. The proposed approach is based on projecting the data into a large number of random low-dimensional subspaces. Canonical correlation analysis is then applied to each of these reduced-rank representations and the model order is selected by combining the information extracted from each subspace. There is, however, no free lunch, as the proposed approach is computationally more demanding and requires higher signal-to-noise ratio (SNR) than the techniques in [7].

2. PROBLEM STATEMENT

We consider the two channel model

$$\mathbf{x} = \mathbf{A}_x \mathbf{s}_x + \mathbf{n}_x \ , \ \mathbf{y} = \mathbf{A}_y \mathbf{s}_y + \mathbf{n}_y \tag{1}$$

In the above model, $\mathbf{A}_x \in \mathbb{R}^{n \times (d+f_x)}$ and $\mathbf{A}_y \in \mathbb{R}^{m \times (d+f_y)}$ are the deterministic but unknown mixing matrices, where d

This work was supported by the German Research Foundation (DFG) under grant SCHR 1384/3-2.

is the number of components correlated between the two data sets, and f_x and f_y are the unknown number of independent components in channel $\mathbf{x}$ and $\mathbf{y}$, respectively. The sources $\mathbf{s}_x$ and $\mathbf{s}_y$ are Gaussian with cross-covariance matrix $\mathbf{R}_{s_x s_y} = \mathrm{diag}\,(\rho_1\sigma_{x,1}\sigma_{y,1}, \ldots, \rho_d\sigma_{x,d}\sigma_{y,d}, 0, \ldots, 0)$, where $|\rho_i| \leq 1$ is the correlation coefficient between the ith signal pair, and $\sigma^2_{x,i}$ and $\sigma^2_{y,i}$ are the corresponding variances. The additive noise terms $\mathbf{n}_x$ and $\mathbf{n}_y$ are independent and Gaussian with arbitrary and unknown covariance matrices. In this setting, we are interested in the following problem.

Problem: *Given M independent and identically distributed (i.i.d.) samples from model* (1)*, with M possibly of the order of the data dimensions n, m, determine the number d of correlated components.*

3. REVIEW OF CCA

The correlation between $\mathbf{x}$ and $\mathbf{y}$ can be assessed by the canonical correlations. These can be obtained by the singular value decomposition (SVD) of the coherence matrix, $\mathbf{C} = \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}\mathbf{R}_{yy}^{-1/2}$, where $\mathbf{R}_{xx}$, $\mathbf{R}_{yy}$, and $\mathbf{R}_{xy}$ are the covariance and cross-covariance matrices, respectively. Let $\mathbf{X} \in \mathbb{R}^{n\times M}$ and $\mathbf{Y} \in \mathbb{R}^{m\times M}$ be the matrices containing the observations from each channel. The sample canonical correlations are given by the SVD of the sample coherence matrix, which is obtained by replacing $\mathbf{R}_{xx}$, $\mathbf{R}_{xy}$, and $\mathbf{R}_{yy}$ with their sample counterparts. These sample canonical correlations can then be used to estimate the number d of correlated components, provided M is large compared to m and n, which is typically carried out by hypothesis testing (HT) [12] or information-theoretic criteria (ITC) [13, 14].

In the case of small sample support, the sample canonical correlations significantly overestimate the true correlations and, hence, cannot be directly used to infer the number of correlated components (see [5, 7] for further details). This is typically overcome by representing the data in some low-dimensional subspace, but a new problem arises: What is the optimal subspace? The typical approach is PCA, whereby the data is projected onto the dominant eigenvectors of the sample covariance matrices. CCA can then be performed on these low-rank descriptions. However, as explained earlier, PCA is not the optimal dimensionality reduction, since the components retained this way are not necessarily the ones that account for the correlation between the data sets.

4. PROPOSED APPROACH

4.1. Main idea

For a given rank r, suppose that we choose random projections. The reduced-rank description of $\mathbf{X}$ and $\mathbf{Y}$ in these random subspaces is

$$\tilde{\mathbf{X}} = \mathbf{P}^T\mathbf{X}\ ,\ \tilde{\mathbf{Y}} = \mathbf{Q}^T\mathbf{Y} \tag{2}$$

where $\mathbf{P} \in \mathbb{R}^{n\times r}$ and $\mathbf{Q} \in \mathbb{R}^{m\times r}$ are each an orthonormal basis of some r-dimensional subspaces. To solve our problem, we will consider as test statistic t the largest sample canonical correlation between $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ that is due to noise. Let us also assume that we are able to determine its probability density function, denoted as $f_d(t)$ (notice its dependence on the number d of correlated components). We could then obtain a trivial estimator of d by placing a threshold on the sample canonical correlations below which components are deemed to be due to noise. However, the performance of such an estimator would strongly depend on the random projectors $\mathbf{P}$ and $\mathbf{Q}$: If $\mathbf{P}$ and $\mathbf{Q}$ do not sufficiently overlap with the correlated-signal subspaces, it will not be possible to separate these from spurious correlations.

Alternatively, by regarding the projectors $\mathbf{P}$ and $\mathbf{Q}$ as random matrices, the sample canonical correlations between $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$, given the observations $\mathbf{X}$ and $\mathbf{Y}$, will follow a certain distribution. Let $f_s(t|\mathbf{X}, \mathbf{Y})$ be the conditional distribution of t given the observations $\mathbf{X}$ and $\mathbf{Y}$, and assuming that the number of correlated components is s. That is, $f_s(t|\mathbf{X}, \mathbf{Y})$ is the conditional distribution of the $(s+1)$th largest sample canonical correlation, which coincides with the largest noise sample canonical correlation when $s = d$. Notice that $\mathrm{E}[f_s(t|\mathbf{X}, \mathbf{Y})] = f_d(t)$ for $s = d$, where the expectation is taken over $\mathbf{X}$ and $\mathbf{Y}$. Let $f_s(t)$ be the distribution of the statistic assuming the number of correlated components d is equal to s. The main idea is that for $r > d$, $f_s(t|\mathbf{X}, \mathbf{Y})$ will generally be closer to $f_s(t)$ when $s = d$ than when $s \neq d$. The value of d can then be determined by evaluating which of the observed sample canonical correlations follows its corresponding target distribution $f_s(t)$ most closely.

To this end, the conditional distribution $f_s(t|\mathbf{X}, \mathbf{Y})$ is required, which seems impossible to obtain analytically. Nevertheless, we can obtain this distribution empirically by using many random projections. That is, for each random projection we obtain rank-reduced descriptions as in (2) along with a set of sample canonical correlations. Repeating this procedure for many random projections, an empirical distribution for each of the sample canonical correlations can be obtained.

The proposed framework can be summarized as follows.

1. For a given rank r, we perform L random projections and obtain L r-dimensional descriptions of $\mathbf{X}$ and $\mathbf{Y}$ following (2). The random projectors are uniformly drawn from the r-dimensional Grassmann manifold.

2. For $s = 0, \ldots, r-1$, we obtain our statistic t as the $(s+1)$th largest sample canonical correlation on each rank-reduced description. This way, we obtain L observations of t, i.e., $t_1, \ldots, t_L$, from which we empirically obtain $f_s(t|\mathbf{X}, \mathbf{Y})$.

3. The number d of correlated components is then estimated as the value of s for which the empirically-determined distribution of the statistic most closely follows its corresponding target distribution $f_s(t)$.

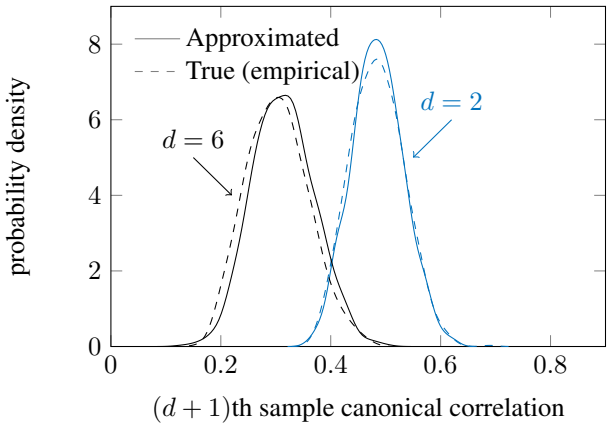

Fig. 1: True and approximated distributions of the statistic under the null hypothesis $M = 100$.

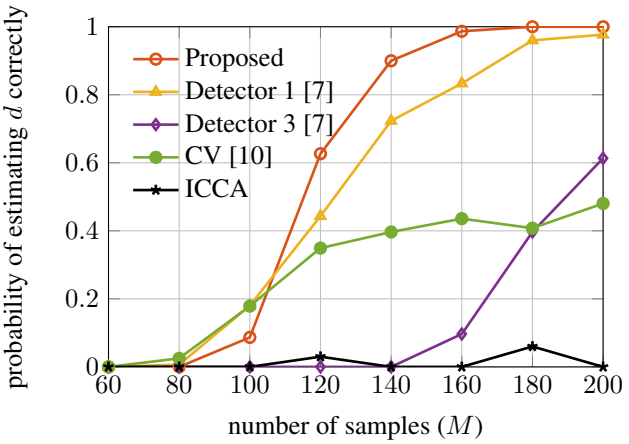

Fig. 2: Performance results for the first scenario.

In the above procedure there are still some questions that need to be answered: How can we evaluate how close the samples of the statistic follow the target distribution? How do we choose the rank r? How can we obtain the target distribution, i.e., the distribution of the largest noise sample canonical correlation? We address these questions in the next subsections.

4.2. Measuring the closeness to the target distribution

We first address the question how to assess the similarity between the distribution of the observed statistic and the target distribution. The closeness of two distributions can be measured by the J-divergence [15, 16], which is a symmetrized version of the Kullback-Leibler (KL) divergence. Specifically, the J-divergence between the distributions $f_1(x)$ and $f_2(x)$ is defined as

$$J = D\left(f_1(x)||f_2(x)\right) + D\left(f_2(x)||f_1(x)\right) \ , \qquad (3)$$

where $D\left(f_i(x)||f_j(x)\right)$ is the KL divergence from $f_i(x)$ to $f_j(x)$. For a given rank r we then proceed as follows. For $s = 0, \ldots, r-1$, we obtain the J-divergence between the conditional and target distributions, $f_s(t|\mathbf{X}, \mathbf{Y})$ and $f_s(t)$, respectively. We then select the number of correlated components as the value of s for which the J-divergence is minimum. In order to compute the J-divergence, we estimate the KL divergence using the algorithm proposed in [17].

Let us now focus on the question how to choose the rank r. This is a difficult task since the optimal rank depends in general on the true number of correlated components, which is precisely what we want to estimate. If $r < d$, part of the signal subspace is completely lost after the projections. Even if r is greater than d but not sufficiently greater than d, it may happen that the random projections do not keep enough signal variance for some components. This makes the observed distribution of their corresponding sample canonical correlations similar to that of the noise, and hence they cannot be detected. Finally, if r is too large compared to the number of samples M, the distribution of the observed sample canonical correlations corresponding to the correlated components will also be similar to those of the noise. To overcome these issues, we propose to estimate d for different ranks, that is, $r = 1, \ldots, r_{\max}$, and then choose our estimate by majority voting among the different ranks. We assume $d < 0.3M$ and set the maximum rank as $r_{\max} = \min(\lfloor 0.3M \rfloor, \min(n, m))$ [18], so that $r_{\max}$ is small compared to M.

4.3. Distribution of the target statistic

In order to apply the proposed approach, we need to determine the distribution $f_s(t)$ of the statistic, i.e., the distribution of the $(s+1)$th largest sample canonical correlation when the number d of correlated components is equal to s. This distribution is, however, unknown. Nevertheless, the noise sample canonical correlations are asymptotically independent of any model parameter other than the dimension and number of samples [19]. Therefore, their distribution can be estimated offline by running a large number of Monte-Carlo simulations. Following these lines, we obtained the distribution of the statistic by generating 1000 data sets following the model (1), with dimension $n = m$, d correlated components with unit variance and unit correlation coefficients, $f_x = f_y = 0$, and white noise with unit variance. This is performed offline for $d = 0, \ldots, n-1$ and different values of n and M.

In order to illustrate the closeness of the approximation, we depict in Fig. 1 the approximated distribution obtained by the aforementioned procedure, and the true distribution obtained by Monte-Carlo simulations using the exact parameters of the scenario. These are the following: $n = m = 10$, $f_x = f_y = 4$ uncorrelated components with variance 3, and two different number of correlated components, namely, $d = 2$ and $d = 6$, in both cases with variance 5 and correlation coefficients equally spaced in the interval $[0.7, 0.9]$. The noise is spatially colored, obtained by filtering white noise with vari-

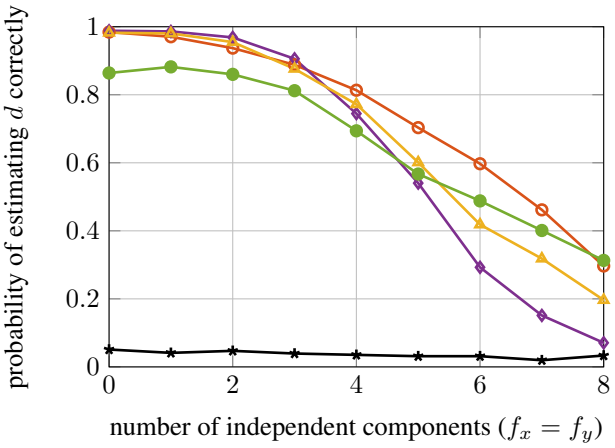

Fig. 3: Performance results for the second scenario. Refer to the legend of Fig. 2 for the meaning of the markers.

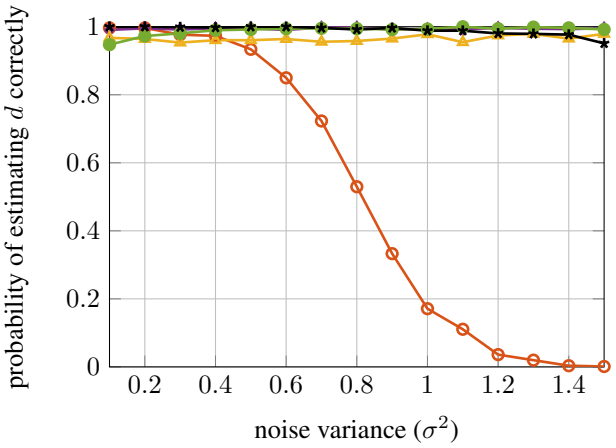

Fig. 4: Performance results for the third scenario. Refer to the legend of Fig. 2 for the meaning of the markers.

ance 1 with a moving-average (MA) filter with coefficients $\frac{1}{\sqrt{3}}[1,1,1]$. The estimated probability density functions are depicted in Fig. 1 for $d = 2$ and $d = 6$, where we observe that the approximate distributions closely follow the true distributions regardless of the model parameters.

5. NUMERICAL EXAMPLES

We generate the observations according to the two-channel model in (1), with each entry of the mixing matrices being independently drawn from a zero-mean Gaussian distribution with unit variance. We compare our techniques with the following competitors: Detector 1 and 3 from [7], which are based on PCA-CCA; CV technique from [10], and informative CCA (ICCA) from [6], also based on PCA-CCA.

We first consider the following scenario: $n = m = 150$, $f_x = f_y = 15$ uncorrelated components with variance 5, and $d = 10$ correlated signals with variance 5 and correlation coefficients equally spaced in the interval $[0.75, 0.95]$. The noise is spatially colored, obtained by filtering white noise with variance 1 with an MA filter with coefficients $\frac{1}{\sqrt{3}}[1,1,1]$. The probability of correctly detecting the exact number of correlated components is shown in Fig. 2. Except for very small number of samples ($M \leq 100$), where all the techniques perform poorly, the proposed approach outperforms the competing ones in this scenario.

The effect of a large number of uncorrelated components is further illustrated in Fig. 3, which shows the detection performance in a scenario with varying number of independent components. We consider: $n = m = 80$, $M = 60$, uncorrelated components with variance 5, and $d = 5$ correlated components with variance 5 and correlation coefficients equally spaced in the interval $[0.7, 0.95]$. The noise is colored following the same model as in the previous scenario. As the number of uncorrelated signals increases, the performance of all techniques decreases, but we observe that the proposed approach exhibits a lower decrease in detection probability. This corroborates the suitability of the proposed technique for scenarios where there is a strong presence of uncorrelated signals.

Finally, we would like to point out that the proposed techniques require larger SNR than, e.g., the techniques proposed in [7]. This is because part of the signal variance is lost when the data is projected into a random subspace. This effect is illustrated in Fig. 4. For this scenario, we consider unitary mixing matrices, so as to have better control over the SNR. The parameters are the following: $n = m = 50$, $M = 100$ samples, no uncorrelated components, $d = 3$ correlated signals with variance 10 and correlation coefficients $[0.7, 0.825, 0.95]$. The noise is white and we vary its variance from 0.1 to 1.5. While the performance of the competing techniques remains approximately constant, the proposed approach experiences a degradation in the detection performance. Nevertheless, in light of the simulation results, we can conclude that this is a promising approach. When the SNR is sufficiently large, the proposed technique can achieve better performance than existing ones when there is a high number of uncorrelated components with large variance. Overcoming the issues observed at low SNR and reducing the computational cost are interesting lines of further work.

6. CONCLUSIONS

We have proposed a new technique to estimate the number of correlated components between two high-dimensional data sets. As opposed to existing techniques, which find a single projection, the proposed technique projects the data onto a large number of random low-dimensional subspaces. Our simulations show that the proposed approach can outperform existing techniques in the presence of a large number of uncorrelated components with high variance.

7. REFERENCES

[1] Y. Levin-Schwartz, Y. Song, P. J. Schreier, V. D. Calhoun, and T. Adalı, "Sample-poor estimation of order and common signal subspace with application to fusion of medical imaging data," *NeuroImage*, vol. 134, pp. 486 – 493, 2016.

[2] J. M. Wallace, C. Smith, and C. S. Bretherton, "Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies," *Journal of Climate*, vol. 5, no. 6, pp. 561–576, June 1992.

[3] P. Stoica, M. Viberg, K. M. Wong, and Q. Wu, "Maximum-likelihood bearing estimation with partly calibrated arrays in spatially correlated noise fields," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 888–899, Apr. 1996.

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

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

[6] N. Asendorf and R. R. Nadakuditi, "Improved detection of correlated signals in low-rank-plus-noise type data sets using informative canonical correlation analysis (ICCA)," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3451–3467, June 2017.

[7] Y. Song, P. J. Schreier, D. Ramírez, and T. Hasija, "Canonical correlation analysis of high-dimensional data with very small sample support," *Signal Processing*, vol. 128, pp. 449–458, Nov. 2016.

[8] J. R. Guerci and J. S. Bergin, "Principal components, covariance matrix tapers, and the subspace leakage problem," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, no. 1, pp. 152–162, Jan. 2002.

[9] Y. Yang and G. Pan, "Independence test for high dimensional data based on regularized canonical correlation coefficients," *Ann. Statist.*, vol. 43, no. 2, pp. 467–500, Apr. 2015.

[10] C. Lameiro and P. J. Schreier, "Cross-validation techniques for determining the number of correlated components between two data sets when the number of samples is very small," in *Proc. of the Asilomar Conference on Signals, Systems and Computers*, Nov. 2016, pp. 601–605.

[11] C. Lameiro and P. J. Schreier, "A sparse CCA algorithm with application to model-order selection for small sample support," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 4721–4725.

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

[13] Q. T. Zhang and Kon Max Wong, "Information theoretic criteria for the determination of the number of signals in spatially correlated noise," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1652–1663, 1993.

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

[15] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.

[16] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, Feb. 1967.

[17] Q. Wang, S. R. Kulkarni, and S. Verdu, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3064–3074, Sept. 2005.

[18] Y. Song, P. J. Schreier, and N. J. Roseveare, "Determining the number of correlated signals between two data sets using PCA-CCA when sample support is extremely small," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 3452–3456.

[19] P. L. Hsu, "On the limiting distribution of the canonical correlations," *Biometrika*, vol. 32, no. 1, pp. 38–45, 1941.