High Dimension Low Sample Size Asymptotic Analysis of Canonical Correlation Analysis

Sungwon Lee*

Abstract
An asymptotic behavior of CCA is studied when dimension d grows and the sample size n is fixed (i.e., under the HDLSS situation). In particular, we are interested in the conditions for which CCA works or fails in the HDLSS situation. This paper presents a conjecture about those conditions, which is supported by extensitve simulation study.

Key Words: HDLSS Asymptotic, CCA

1. Introduction

Canonical correlation analysis (CCA) introduced in [4] is a standard statistical tool to explore the relationship between two sets of random variables. Consider d_X- and d_Y-dimensional random vectors $X^{(d_X)}$ and $Y^{(d_Y)}$,

$$\left(X^{(d_X)}\right)^T = \begin{bmatrix} X_1, & X_2, & \ldots, & X_{d_X} \end{bmatrix}, \; \left(Y^{(d_Y)}\right)^T = \begin{bmatrix} Y_1, & Y_2, & \ldots, & Y_{d_Y} \end{bmatrix}.$$

CCA first seeks a pair of d_X- and d_Y-dimensional weights vectors $\psi_{X1}^{(d_X)}$ and $\psi_{Y1}^{(d_Y)}$ such that two random variables, one being the linear combination of $X_1, X_2, \ldots, X_{d_X}$ weighted by the elements of $\psi_{X1}^{(d_X)}$ and the other being that of $Y_1, Y_2, \ldots, Y_{d_X}$ weighted by the elements of $\psi_{Y1}^{(d_Y)}$, have a maximal correlation,

$$(\psi_{X1}^{(d_X)}, \psi_{Y1}^{(d_Y)}) = \underset{\text{Var}(\langle \psi_{X1}^{(d_X)}, X^{(d_X)}\rangle) = \text{Var}(\langle \psi_{Y1}^{(d_Y)}, Y^{(d_Y)}\rangle) = 1}{\text{argmax}} \text{Cov}(\langle \psi_{X1}^{(d_X)}, X^{(d_X)}\rangle, \langle \psi_{Y1}^{(d_Y)}, Y^{(d_Y)}\rangle). \tag{1}$$

Requiring the norms of the weight vectors $\psi_{X1}^{(d_X)}$ and $\psi_{Y1}^{(d_Y)}$ to be one, the equation (1) can be written as an equivalent form of,

$$(\psi_{X1}^{(d_X)}, \psi_{Y1}^{(d_Y)}) = \underset{\|\psi_{X1}^{(d_X)}\|_2 = \|\psi_{Y1}^{(d_Y)}\|_2 = 1}{\text{argmax}} \frac{\text{Cov}(\langle \psi_{X1}^{(d_X)}, X^{(d_X)}\rangle, \langle \psi_{Y1}^{(d_Y)}, Y^{(d_Y)}\rangle)}{\sqrt{\text{Var}(\langle \psi_{X1}^{(d_X)}, X^{(d_X)}\rangle)}\sqrt{\text{Var}(\langle \psi_{Y1}^{(d_Y)}, Y^{(d_Y)}\rangle)}}. \tag{2}$$

For convenience, denote the objective function in the right hand side of (2) by $\rho_P(\psi^{(d_X)}, \psi^{(d_Y)})$,

$$\begin{aligned} &\rho : R^{d_X} \times R^{d_Y} \mapsto R \\ &\rho_P(\psi^{(d_X)}, \psi^{(d_Y)}) = \frac{\text{Cov}(\langle \psi^{(d_X)}, X^{(d_X)}\rangle, \langle \psi^{(d_Y)}, Y^{(d_Y)}\rangle)}{\sqrt{\text{Var}(\langle \psi^{(d_X)}, X^{(d_X)}\rangle)}\sqrt{\text{Var}(\langle \psi^{(d_Y)}, Y^{(d_Y)}\rangle)}}. \end{aligned}$$

Subsequent weights vectors $\psi_{Xi}^{(d_X)}$ and $\psi_{Yi}^{(d_Y)}$, for $i = 1, 2, \ldots, \min(d_X, d_Y)$, are found by maximizing the objective function $\rho_P(\psi^{(d_X)}, \psi^{(d_Y)})$,

$$(\psi_{Xi}^{(d_X)}, \psi_{Yi}^{(d_Y)}) = \underset{\|\psi_{Xi}^{(d_X)}\|_2 = \|\psi_{Yi}^{(d_Y)}\|_2 = 1}{\text{argmax}} \rho_P(\psi_{Xi}^{(d_X)}, \psi_{Yi}^{(d_Y)}), \; i = 1, 2, \ldots, \min(d_X, d_Y),$$

*Department of Statistics, University of Pittsburgh, PA 15260, U.S.A.

under the constraint that,

$$\begin{aligned}\text{Cov}(\langle\psi_{Xi}^{(d_X)}, X^{(d_X)}\rangle, \langle\psi_{Xj}^{(d_X)}, X^{(d_X)}\rangle) &= \text{Cov}(\langle\psi_{Yi}^{(d_Y)}, Y^{(d_Y)}\rangle, \langle\psi_{Yj}^{(d_Y)}, Y^{(d_Y)}\rangle)\\ &= \text{Cov}(\langle\psi_{Xi}^{(d_X)}, X^{(d_X)}\rangle, \langle\psi_{Yj}^{(d_Y)}, Y^{(d_Y)}\rangle)\\ &= \text{Cov}(\langle\psi_{Yi}^{(d_Y)}, Y^{(d_Y)}\rangle, \langle\psi_{Xj}^{(d_X)}, X^{(d_X)}\rangle)\\ &= 0,\ i = 1, 2, \ldots, \min(d_X, d_Y),\ j = 1, 2, .., i-1 \text{ for each } i.\end{aligned}$$

The ith pair of weight vectors $\psi_{Xi}^{(d_X)}$ and $\psi_{Yi}^{(d_Y)}$ are usually called the ith pair of canonical weight vectors (or canonical loadings). The correlation ρ evaluated at the ith pair $\psi_{Xi}^{(d_X)}$ and $\psi_{Yi}^{(d_Y)}$, denoted by $\rho_i^{(d_X,d_Y)}$, is called the ith canonical correlation coefficient, that is, $\rho_i^{(d_X,d_Y)} = \rho_P(\psi_{Xi}^{(d_X)}, \psi_{Yi}^{(d_Y)})$.

In practice, we collect two sets of obervations of d_X- and d_Y-dimensional random vectors $X^{(d_X)}$ and $Y^{(d_Y)}$ on a common set of samples in a $d_X \times n$ matrix $\mathbf{X}^{(d_X)}$ and a $d_Y \times n$ matrix $\mathbf{Y}^{(d_Y)}$, respectively. We row-center $\mathbf{X}^{(d_X)}$ and $\mathbf{X}^{(d_X)}$ and let $\hat{\mathbf{\Sigma}}_X^{(d_X)}$, $\hat{\mathbf{\Sigma}}_Y^{(d_Y)}$ and $\hat{\mathbf{\Sigma}}_{XY}^{(d_X,d_Y)}$ be a covariance matrix of $X^{(d_X)}$, a covariance matrix of $Y^{(d_Y)}$ and a cross-covariance matrix of $X^{(d_X)}$ and $Y^{(d_Y)}$,

$$\hat{\mathbf{\Sigma}}_X^{(d_X)} = \frac{1}{n}\mathbf{X}^{(d_X)}\left(\mathbf{X}^{(d_X)}\right)^T,\ \hat{\mathbf{\Sigma}}_Y^{(d_Y)} = \frac{1}{n}\mathbf{Y}^{(d_Y)}\left(\mathbf{Y}^{(d_Y)}\right)^T,\ \hat{\mathbf{\Sigma}}_{XY}^{(d_X,d_Y)} = \frac{1}{n}\mathbf{X}^{(d_X)}\left(\mathbf{Y}^{(d_Y)}\right)^T.$$

For the case where the sample size n is greater than d_X and d_Y, the estimation of sample canonical weight vectors $(\hat{\psi}_{Xi}^{(d_X)}, \hat{\psi}_{Yi}^{(d_Y)})$ and sample canonical correlation coefficients $\hat{\rho}_i^{(d_X,d_Y)}$ are done through singular value decomposition of the matrix $\hat{\mathbf{R}}^{(d_X,d_Y)}$,

$$\begin{aligned}\hat{\mathbf{R}}^{(d_X,d_Y)} &= \left(\hat{\mathbf{\Sigma}}_X^{(d_X)}\right)^{-\frac{1}{2}}\hat{\mathbf{\Sigma}}_{XY}^{(d_X,d_Y)}\left(\hat{\mathbf{\Sigma}}_Y^{(d_Y)}\right)^{-\frac{1}{2}},\\ \text{SVD}(\hat{\mathbf{R}}^{(d_X,d_Y)}) &= \sum_{i=1}^{\min(d_X,d_Y)}\hat{\lambda}_{Ri}^{(d_X,d_Y)}\hat{\eta}_{RXi}^{(d_X)}\left(\hat{\eta}_{RYi}^{(d_Y)}\right)^T,\end{aligned} \tag{3}$$

where $\hat{\lambda}_{Ri}^{(d)}$ is a sample singular value with $\hat{\lambda}_{R1}^{(d)} \geq \hat{\lambda}_{R2}^{(d)} \geq \cdots \geq \hat{\lambda}_{R\min(d_X,d_Y)}^{(d)} \geq 0$, and $(\hat{\eta}_{RXi}^{(d_X)}, \hat{\eta}_{RYi}^{(d_Y)})$ is a pair of left and right sample singular vectors corresponding to $\hat{\lambda}_{Ri}^{(d)}$. Then, the ith sample canonical correlation coefficient $\hat{\rho}_i^{(d)}$ is found to be,

$$\hat{\rho}_i^{(d_X,d_Y)} = \hat{\lambda}_{Ri}^{(d_X,d_Y)}.$$

The ith pair of canonical weight vectors $\hat{\psi}_{Xi}^{(d_X)}$ and $\hat{\psi}_{Yi}^{(d_Y)}$ are obtained by unscaling and normalzing the ith pair of sample singular vectors $\hat{\eta}_{RXi}^{(d_X)}$ and $\hat{\eta}_{RYi}^{(d_Y)}$,

$$\hat{\psi}_{Xi}^{(d_X)} = \frac{\left(\hat{\mathbf{\Sigma}}_X^{(d_X)}\right)^{-\frac{1}{2}}\hat{\eta}_{RXi}^{(d_X)}}{\left\|\left(\hat{\mathbf{\Sigma}}_X^{(d_X)}\right)^{-\frac{1}{2}}\hat{\eta}_{RXi}^{(d_X)}\right\|_2},\ \hat{\psi}_{Yi}^{(d_Y)} = \frac{\left(\hat{\mathbf{\Sigma}}_Y^{(d_Y)}\right)^{-\frac{1}{2}}\hat{\eta}_{RYi}^{(d_Y)}}{\left\|\left(\hat{\mathbf{\Sigma}}_Y^{(d_Y)}\right)^{-\frac{1}{2}}\hat{\eta}_{RYi}^{(d_Y)}\right\|_2}. \tag{4}$$

The projection of the data matrix $\mathbf{X}^{(d_X)}$ onto the ith sample canonical weight vector $\hat{\psi}_{Xi}^{(d_X)}$ gives the canonical scores (or canonical variables) of $\mathbf{X}^{(d_X)}$ with respect to $\hat{\psi}_{Xi}^{(d_X)}$ and similarly for $\mathbf{X}^{(d_X)}$. Although powerful, CCA has several disadvantages. first, use of CCA is practically restricted to the case of two sets of data even if there is an attempt to generalize it to more than two sets of data [11]. Second, CCA components are estimable only if the sample size n is greater than d_X and d_Y. It is well know that, when $n < \max(d_X, d_Y)$, one can construct an infinite number of sample canonical weight vector pairs with their correlation of one. Moreover, overfitting is often a problem even when $n > d_X$ and d_Y. Hence, CCA is often considered not relible in high-dimensional data sets. We, however, will show that, even in the case where sample size n is less than d_X or d_Y, some sample canonical weight vectors is estimable and furthremore consistent under a certain condition.

As high-dimensional data are increasingly common these days, where a large number of variables are measured for each object, there is a strong need to investigate the behavior of estimates resulting from the application of standard statistical tools such as CCA to a high-dimensional case (that is, scalability of those tools). In studies in which dimension d is allowed to go to infinity, three scenarios are typically considered [10],

- Low Dimension High Sample Size (LDHSS): Both dimension d and sample size n go to infinity but n increases much faster than d, which can be summarized as $d/n \to 0$. These problems are similar to conventional asymptotics where $n \to \infty$ with n being fixed.
- High Dimension High Sample Size (HDHSS): In this case, sample size and dimension grow together in the sense that $d/n \to c$ for some constant c. The bahavior of eigenvalues of a sample covariance matrix under this high-dimensional situation were studied in [2, 5, 9] primarily using random matrix theories.
- High Dimension Low Sample Size (HDLSS): In this setting, the sample size is fixed and the dimension grows in the sense that $d/n \to \infty$. An important finding in this high-dimensional setting was studied in [1]. They showed that the first eigenvector of the sample covariance matrix converges consistently to its population counterpart in the spiked model, where the leading eigenvalue is considerably larger than the remaining eigenvalues. An intesting geometric structure of HDLSS data were revealed in [3].

In this chapter, we are going to study the asymptotic behavior of the sample canonical weight vectors and canonical correlation coefficients of CCA under the HDLSS setting, where dimension d is allowed to grow with sample size n being fixed.

Literature in the HDLSS asymptotic study of CCA is very limited, while the behavior of PCA components under the similar high-dimensional condition is well-studied in [6, 7]. This might be in part because CCA is not as widely used as PCA, which is almost an indispensible tool for dimension reduction of high-dimensional data prevalent these days, and in part due to the complicated estimation steps involving an inverse operator as in (3), which makes the analysis not straightforward. A relevant work is first addressed in [8], where the asymptotic behavior of sample singular vectors and singular values are analysed under a HDLSS setting. In [10], the similar study of CCA is elaborated on, but their proof should have considered the fact that an infinite sum of quantities converging to zero does not neccessarily approach to zero. The HDLSS asymptotic behavior of CCA components in this chapter will be studied in relatively a simple population structure and serves as a groundwork for further analysis.

2. Assumptions and Definitions

Without loss of generality for the case where the dimensions of two random vectors $X^{(d_X)}$ and $Y^{(d_Y)}$ grow in a sense that $d_X/d_Y \to 1$, we set $d_X = d_Y$ and consider two random vectors $X^{(d)}$ and $Y^{(d)}$ of a same dimension with mean zero. We assume that covariance structure of $X^{(d)}$ and $Y^{(d)}$ follows a simple spiked model as in [1], where the leading eigenvalues of their covariance matrix is considerably larger than the rest. In specific, let $\mathbf{\Sigma}_X^{(d)}$ and $\mathbf{\Sigma}_Y^{(d)}$ be the covariance matrices of $X^{(d)}$ and $Y^{(d)}$. Then, a spiked model can be easily understood via eigendecomposition of $\mathbf{\Sigma}_X^{(d)}$ and $\mathbf{\Sigma}_Y^{(d)}$,

$$\mathbf{\Sigma}_X^{(d)} = \sum_{i=1}^{d} \lambda_{Xi}^{(d)} \xi_{Xi}^{(d)} \left(\xi_{Xi}^{(d)}\right)^T, \ \mathbf{\Sigma}_Y^{(d)} = \sum_{j=1}^{d} \lambda_{Yj}^{(d)} \xi_{Yj}^{(d)} \left(\xi_{Yj}^{(d)}\right)^T, \tag{5}$$

where $\lambda_{Xi}^{(d)}$ is an polpulation eigenvalue (or population PC variance) with $\lambda_{X1}^{(d)} \geq \lambda_{X2}^{(d)} \geq \cdots \geq \lambda_{Xd}^{(d)} \geq 0$, $\xi_{Xi}^{(d)}$ is an population eigenvector (or population PC direction) with $\|\xi_{Xi}^{(d)}\|_2 = 1$ and $\langle \xi_{Xi}^{(d)}, \xi_{Xj}^{(d)} \rangle = 0$ for $i \neq j$ and similarly for $\lambda_{Yj}^{(d)}$ and $\xi_{Yj}^{(d)}$. Here, we set,

$$\begin{aligned} \lambda_{X1}^{(d)} = \sigma_X^2 d^\alpha \text{ and } \lambda_{Xi}^{(d)} = \tau_X^2 \text{ for } i = 2, 3, \ldots, d, \\ \lambda_{Y1}^{(d)} = \sigma_Y^2 d^\alpha \text{ and } \lambda_{Yj}^{(d)} = \tau_Y^2 \text{ for } j = 2, 3, \ldots, d, \end{aligned} \tag{6}$$

where one sees that the leading eigenvalues $\lambda_{X1}^{(d)}$ and $\lambda_{Y1}^{(d)}$ become dominating the rest as $d \to \infty$. We now set up the population canonical components. We assume that the two random vector is related by a pair of canonical weight vectors with its canonical correlation coefficient of ρ. The population canonical weight vector $\psi_X^{(d)}$ in the $X^{(d)}$ part is a linear combination of two eigenvectors $\xi_{X1}^{(d)}$ and $\xi_{X2}^{(d)}$ without loss of generality ($\xi_{X2}^{(d)}$ can be replaced with $\xi_{Xi}^{(d)}$ for any i) and similarly for the other population canonical weight vector $\psi_Y^{(d)}$ in the $Y^{(d)}$ part,

$$\psi_X^{(d)} = \cos\theta_X \xi_{X1}^{(d)} + \sin\theta_X \xi_{X2}^{(d)}, \ \psi_Y^{(d)} = \cos\theta_Y \xi_{Y1}^{(d)} + \sin\theta_Y \xi_{Y2}^{(d)}. \tag{7}$$

Note that the angle between $\psi_X^{(d)}$ and $\xi_{X1}^{(d)}$ is θ_X and that the angle between $\psi_Y^{(d)}$ and $\xi_{Y1}^{(d)}$ is θ_Y as $\langle \psi_X^{(d)}, \xi_{X1}^{(d)} \rangle = \cos\theta_X$ and $\langle \psi_Y^{(d)}, \xi_{Y1}^{(d)} \rangle = \cos\theta_Y$. At this point, we apply the change of basis to the spaces of $X^{(d)}$ and $Y^{(d)}$ so that the eigenvectors $\{\xi_{Xi}^{(d)}\}_{i=1}^d$ and $\{\xi_{Yi}^{(d)}\}_{j=1}^d$ are represented by the standard basis $\{e_i^{(d)}\}_{k=1}^d$. Then, the canonical weight vectors $(\psi_X^{(d)}, \psi_Y^{(d)})$ given in (7) is rewritten as,

$$\psi_X^{(d)} = \cos\theta_X e_1^{(d)} + \sin\theta_X e_2^{(d)}, \ \psi_Y^{(d)} = \cos\theta_Y e_1^{(d)} + \sin\theta_Y e_2^{(d)},$$

and the covariance structures given in (5) and (6) are described as,

$$\mathbf{\Sigma}_X^{(d)} = \underset{d\times d}{\operatorname{diag}}(\sigma_X^2 d^\alpha, \ \tau_X^2, \ \tau_X^2, \ \ldots, \ \tau_X^2), \ \mathbf{\Sigma}_Y^{(d)} = \underset{d\times d}{\operatorname{diag}}(\sigma_Y^2 d^\alpha, \ \tau_Y^2, \ \tau_Y^2, \ \ldots, \ \tau_Y^2), \tag{8}$$

where diag($\bullet$) is a square matrix with entries of $\bullet$ in the main diagonal and 0 off of it. With these population covariance structures and canonical components, the multivariate version of the corallory **??** gives the cross-covariance structure of $X^{(d)}$ and $Y^{(d)}$ as follows,

$$\mathbf{\Sigma}_{XY}^{(d)} = \begin{bmatrix} \frac{\rho\sigma_X^2\sigma_Y^2 d^{2\alpha}\cos\theta_X\cos\theta_Y}{AB} & \frac{\rho\sigma_X^2 d^\alpha \tau_Y^2 \cos\theta_X \sin\theta_Y}{AB} & \underset{1\times(d-2)}{\mathbf{0}} \\ \frac{\rho\tau_X^2\sigma_Y^2 d^\alpha \sin\theta_X\cos\theta_Y}{AB} & \frac{\rho\tau_X^2\tau_Y^2\sin\theta_X\sin\theta_Y}{AB} & \underset{1\times(d-2)}{\mathbf{0}} \\ \underset{(d-2)\times 1}{\mathbf{0}} & \underset{(d-2)\times 1}{\mathbf{0}} & \underset{(d-2)\times(d-2)}{\mathbf{0}} \end{bmatrix}, \tag{9}$$

where

$$A = \sqrt{\sigma_X^2 d^\alpha \cos^2\theta_X + \tau_X^2 \sin^2\theta_X}, \ B = \sqrt{\sigma_Y^2 d^\alpha \cos^2\theta_Y + \tau_Y^2 \sin^2\theta_Y}.$$

Then the covariance and cross-covariance structure of $X^{(d)}$ and $Y^{(d)}$ is succintly described by the covariance structure of the concatenated random vector $T^{(2d)}$,

$$T^{(2d)} = \begin{bmatrix} X^{(d)} \\ Y^{(d)} \end{bmatrix}, \ \mathbf{\Sigma}_T^{(2d)} = \begin{bmatrix} \mathbf{\Sigma}_X^{(d)} & \mathbf{\Sigma}_{XY}^{(d)} \\ \left(\mathbf{\Sigma}_{XY}^{(d)}\right)^T & \mathbf{\Sigma}_Y^{(d)} \end{bmatrix}. \tag{10}$$

To make the analysis a bit easy, we are going to work with a different representation of $X^{(d)}$ and $Y^{(d)}$. Let $Z^{(d)}$ be the $2d$-dimensional standard normal random vector. Then, $T^{(2d)}$ can be expressed as,

$$T^{(2d)} = \begin{bmatrix} X^{(d)} \\ Y^{(d)} \end{bmatrix} = \left(\mathbf{\Sigma}_T^{(2d)}\right)^{\frac{1}{2}} Z^{(2d)}, \ Z^{(2d)} \sim N\left(\underset{2d\times 1}{0}, \underset{2d\times 2d}{\mathbf{I}}\right). \tag{11}$$

We state some definitions used in the estimation. Since the dimensionality d is much larger than the sample size n in the HDLSS setting, the estimation step (3) of canonical components is problematic as the sample covariance matrices $\hat{\mathbf{\Sigma}}_X^{(d)}$ and $\hat{\mathbf{\Sigma}}_Y^{(d)}$ are singular. There are two ways to handle this singularity situation. The first one is to add a minute perturbation of $\epsilon \mathbf{I}$ for a small $\epsilon > 0$ to $\hat{\mathbf{\Sigma}}_X^{(d)}$ and $\hat{\mathbf{\Sigma}}_Y^{(d)}$ and the second is to use a pseudoinverse such as Moore-Penrose pseudoinverse. We use the pseudoinverse obtained from the eigendecomposition of the sample covariance matrices,

$$\hat{\mathbf{\Sigma}}_X^{(d)} = \sum_{i=1}^{n} \hat{\lambda}_{Xi}^{(d)} \hat{\xi}_{Xi}^{(d)} \left(\hat{\xi}_{Xi}^{(d)}\right)^T, \ \hat{\mathbf{\Sigma}}_Y^{(d)} = \sum_{j=1}^{n} \hat{\lambda}_{Yj}^{(d)} \hat{\xi}_{Yj}^{(d)} \left(\hat{\xi}_{Yj}^{(d)}\right)^T, \tag{12}$$

where $\hat{\lambda}_{Xi}^{(d)}$ is an sample eigenvalue (or sample PC variance) with $\hat{\lambda}_{X1}^{(d)} \geq \hat{\lambda}_{X2}^{(d)} \geq \cdots \geq \hat{\lambda}_{Xd}^{(n)} \geq 0$, $\hat{\xi}_{Xi}^{(d)}$ is an sample eigenvector (or sample PC direction) with $\|\hat{\xi}_{Xi}^{(d)}\|_2 = 1$ and $\langle \hat{\xi}_{Xi}^{(d)}, \hat{\xi}_{Xj}^{(d)} \rangle = 0$ for $i \neq j$ and similarly for $\hat{\lambda}_{Yj}^{(d)}$ and $\hat{\xi}_{Yj}^{(d)}$. The pseudoinverse we employ is defined as,

$$\left(\hat{\mathbf{\Sigma}}_X^{(d)}\right)^{-1} = \sum_{i=1}^{n} \left(\hat{\lambda}_{Xi}^{(d)}\right)^{-1} \hat{\xi}_{Xi}^{(d)} \left(\hat{\xi}_{Xi}^{(d)}\right)^T, \ \left(\hat{\mathbf{\Sigma}}_Y^{(d)}\right)^{-1} = \sum_{j=1}^{d} \left(\hat{\lambda}_{Yj}^{(d)}\right)^{-1} \hat{\xi}_{Yj}^{(d)} \left(\hat{\xi}_{Yj}^{(d)}\right)^T. \tag{13}$$

Then, the sample canonical correlation coefficient $\hat{\rho}_i^{(d)}$ is found as an ith sample singular value from the SVD of the matrix $\hat{R}^{(d)}$ defined in (4). The sample canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$ and $\hat{\psi}_{Yi}^{(d)}$ corresponding to $\hat{\rho}_i^{(d)}$ are obtained from (4) using the pseudoinverses (13).

The success and failure of CCA can be described by the consistency of the sample canonical weight vectors $\hat{\psi}_X^{(d)}$ and $\hat{\psi}_Y^{(d)}$ with their population counterpart $\psi_X^{(d)}$ and $\psi_Y^{(d)}$ under the limiting operation of $d \to \infty$ and n fixed. Using the angle as a measure of consistency, we say that $\hat{\psi}_X^{(d)}$ (similarly $\hat{\psi}_Y^{(d)}$) is,

- Consistent with $\psi_X^{(d)}$ if angle($\hat{\psi}_X^{(d)}$, $\psi_X^{(d)}$) $\to 0$ as $d \to \infty$,
- Inonsistent with $\psi_X^{(d)}$ if angle($\hat{\psi}_X^{(d)}$, $\psi_X^{(d)}$) $\to a$, for $0 < a < \pi/2$, as $d \to \infty$,
- Strongly inonsistent with $\psi_X^{(d)}$ if angle($\hat{\psi}_X^{(d)}$, $\psi_X^{(d)}$) $\to \pi/2$ as $d \to \infty$.

Strong inconsistency implies that the estimate $\hat{\psi}_X^{(d)}$ and $\hat{\psi}_Y^{(d)}$ become completely oblivious of its population structure and reduce to arbitrary quantities, as indicated in the fact that $pi/2$ is indeed a largest angle possible between two vectors.

3. Conjecture and Interpretation

3.1 Conjecture

Let $X^{(d)}$ and $Y^{(d)}$ be the d-dimensional random vectors from the multivariate Gaussian distributions with mean 0 and the simple spiked covariance matrices $\mathbf{\Sigma}_X^{(d)}$ and $\mathbf{\Sigma}_Y^{(d)}$ described in (5) and (6). With the population canonical correlation coefficient ρ for $0 \leq \rho \leq 1$, define the population canonical weight vectors $\psi_X^{(d)}$ and $\psi_Y^{(d)}$ as,

$$\psi_X^{(d)} = \cos\theta_X \xi_{X1}^{(d)} + \sin\theta_X \xi_{X2}^{(d)},\ \psi_Y^{(d)} = \cos\theta_Y \xi_{Y1}^{(d)} + \sin\theta_Y \xi_{Y2}^{(d)}$$

so that the angle between $\psi_X^{(d)}$ and $\xi_{X1}^{(d)}$ is θ_X, and the angle between $\psi_Y^{(d)}$ and $\xi_{Y1}^{(d)}$ is θ_Y. Then, the cross-covariance matrix $\mathbf{\Sigma}_{XY}^{(d)}$ of $X^{(d)}$ and $Y^{(d)}$ is found as in 9. The two random variables $X^{(d)}$ and $Y^{(d)}$ can be written in a equivalent form,

$$\begin{bmatrix} X^{(d)} \\ Y^{(d)} \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma}_X^{(d)} & \mathbf{\Sigma}_{XY}^{(d)} \\ \left(\mathbf{\Sigma}_{XY}^{(d)}\right)^T & \mathbf{\Sigma}_Y^{(d)} \end{bmatrix} Z^{(2d)}, \tag{14}$$

where $Z^{(2d)}$ is a $2d$-dimensional standard normal random vector. The data matrix whose columns consist of n i.i.d. samples from the distribution 14 is written as,

$$\begin{bmatrix} \mathbf{X}^{(d)} \\ \mathbf{Y}^{(d)} \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma}_X^{(d)} & \mathbf{\Sigma}_{XY}^{(d)} \\ \left(\mathbf{\Sigma}_{XY}^{(d)}\right)^T & \mathbf{\Sigma}_Y^{(d)} \end{bmatrix} \mathbf{Z}^{(2d)}, \tag{15}$$

where the columns of $\mathbf{Z}^{(2d)}$ consist of n i.i.d. samples from $2d$-dimensional standard normal ditribution. Denote by z_1 and z_2 the first and $(d+1)$th rows of $\mathbf{Z}^{(2d)}$ corresponding to the first rows of $\mathbf{X}^{(d)}$ and $\mathbf{Y}^{(d)}$ respectively. Then, as $d \to \infty$ with the sample size n being fixed, the limiting behaviors of the sample canonical correlation coefficient $\hat{\rho}_i^{(d)}$ and its corresponding sample canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$ and $\hat{\psi}_{Yi}^{(d)}$ obtained from the data 15 are as follows,

Conjecture 1. *(i)* $\alpha > 1$

$$\begin{gathered}
\text{angle}\left(\hat{\psi}_{X1}^{(d)}, \psi_X^{(d)}\right) \xrightarrow[d\to\infty]{P} \theta_X,\ \text{angle}\left(\hat{\psi}_{Y1}^{(d_Y)}, \psi_Y^{(d)}\right) \xrightarrow[d\to\infty]{P} \theta_Y,\ \hat{\rho}_1^{(d)} \xrightarrow[d\to\infty]{D} \frac{\langle m_1, m_2 \rangle}{\|m_1\|_2 \|m_2\|_2}, \\
\text{angle}\left(\hat{\psi}_{Xi}^{(d)}, \psi_X^{(d)}\right) \xrightarrow[d\to\infty]{P} 0,\ \text{angle}\left(\hat{\psi}_{Yi}^{(d)}, \psi_Y^{(d)}\right) \xrightarrow[d\to\infty]{P} 0,\ \hat{\rho}_i^{(d)} \xrightarrow[d\to\infty]{P} 0,\ i = 2, 3, \ldots, n,
\end{gathered}$$

where

$$\begin{aligned}
m_1 &= (\sqrt{C_1} A_1^2 + \sqrt{C_2} B_1^2) z_1 + (\sqrt{C_1} A_1 A_2 + \sqrt{C_2} B_1 B_2) z_2, \\
m_2 &= (\sqrt{C_1} A_1 A_2 + \sqrt{C_2} B_1 B_2) z_1 + (\sqrt{C_1} A_1^2 + \sqrt{C_2} B_1^2) z_2,
\end{aligned}$$

where

$$\begin{aligned}
& z_1, z_2 \overset{i.i.d.}{\sim} N\left(\underset{n\times 1}{0}, \underset{n\times n}{\mathbf{I}}\right), \\
& C_1 = \frac{\sigma_X^2 + \sigma_Y^2 + \sqrt{\left(\sigma_X^2\right)^2 - 2\sigma_X^2\sigma_Y^2 + 4\sigma_X^2\sigma_Y^2\rho^2 + \left(\sigma_Y^2\right)^2}}{2},
\end{aligned}$$

$$C_2 = \frac{\sigma_X^2 + \sigma_Y^2 - \sqrt{\left(\sigma_X^2\right)^2 - 2\sigma_X^2\sigma_Y^2 + 4\sigma_X^2\sigma_Y^2\rho^2 + \left(\sigma_Y^2\right)^2}}{2},$$

$$A_1 = \frac{C_1 - \sigma_Y^2}{\rho\sigma_X\sigma_Y} / \sqrt{\left(\frac{C_1 - \sigma_Y^2}{\rho\sigma_X\sigma_Y}\right)^2 + 1},\ A_2 = 1/\sqrt{\left(\frac{C_1 - \sigma_Y^2}{\rho\sigma_X\sigma_Y}\right)^2 + 1},$$

$$B_1 = \frac{C_2 - \sigma_Y^2}{\rho\sigma_X\sigma_Y} / \sqrt{\left(\frac{C_2 - \sigma_Y^2}{\rho\sigma_X\sigma_Y}\right)^2 + 1},\ B_2 = 1/\sqrt{\left(\frac{C_2 - \sigma_Y^2}{\rho\sigma_X\sigma_Y}\right)^2 + 1}.$$

(ii) $\alpha < 1$

$$\text{angle}\left(\hat{\psi}_{Xi}^{(d)}, \psi_X^{(d)}\right) \xrightarrow[d\to\infty]{P} 0,\ \text{angle}\left(\hat{\psi}_{Yi}^{(d)}, \psi_Y^{(d)}\right) \xrightarrow[d\to\infty]{P} 0,\ \hat{\rho}_i^{(d)} \xrightarrow[d\to\infty]{P} 1,\ i = 1, 2, \ldots, n.$$

3.2 Interpretation

The conjecture 1 implies that where $\hat{\psi}_{X1}^{(d)}$ and $\hat{\psi}_{Y1}^{(d)}$ converge to depend heavily on the size of the variance d^α of the population eigenvector $\xi_{X1}^{(d)}$ and $\xi_{Y1}^{(d)}$. That is, the estimates $\hat{\psi}_{X1}^{(d)}$ and $\hat{\psi}_{Y1}^{(d)}$ tend to converge to the eigenvectors $\xi_{X1}^{(d)}$ and $\xi_{Y1}^{(d)}$ when their eigenvalues $\sigma_X^2 d^\alpha$ and $\sigma_Y^2 d^\alpha$ become strong enough ($\alpha > 1$) as $d \to \infty$. Briefly, we summarize results. The sample canonical weight vector $\hat{\psi}_{X1}^{(d)}$ (similarly $\hat{\psi}_{Y1}^{(d)}$) is,

- Consistent with $\psi_X^{(d)}$ if $\alpha > 1$ and angle($\psi_X^{(d)}, \xi_{X1}^{(d)}) = 0$ as $d \to \infty$,
- Inonsistent with $\psi_X^{(d)}$ if $\alpha > 1$ and angle($\psi_X^{(d)}, \xi_{X1}^{(d)}) = \theta_X$, for $0 < \theta_X < \pi/2$, as $d \to \infty$,
- Strongly inonsistent with $\psi_X^{(d)}$ if $\alpha < 1$ or if $\alpha > 1$ and angle($\psi_X^{(d)}, \xi_{X1}^{(d)}) = \pi/2$ as $d \to \infty$.

The asymptotic behavior of the sample canonical correlation coefficient $\hat{\rho}_1^{(d)}$ is not straightforward to imagine. Let's take a simple example where $\sigma_X^2 = 1, \sigma_X^2 = 1, \tau_X^2 = 1$ and $\tau_Y^2 = 1$ in the spiked covariance structure in (5) and (6). In this case, referring to the conjecture 1, the sample canonical correlation coefficient $\hat{\rho}_1^{(d)}$ converges in probability to the following random quantity,

$$\hat{\rho}_1^{(d)} \xrightarrow[d\to\infty]{P} \frac{\langle m_1, m_2 \rangle}{\|m_1\|_2 \|m_2\|_2},$$

where

$$m_1 = \left(\frac{\sqrt{1+\rho} + \sqrt{1-\rho}}{2}\right) z_1 + \left(\frac{\sqrt{1+\rho} - \sqrt{1-\rho}}{2}\right) z_2,$$
$$m_2 = \left(\frac{\sqrt{1+\rho} - \sqrt{1-\rho}}{2}\right) z_1 + \left(\frac{\sqrt{1+\rho} + \sqrt{1-\rho}}{2}\right) z_2.$$

Note that z_1 and z_2 are samples from n-dimensional multivariate standard normal distribution. It can be easily verified that each element m_{1i} of m_1 (similarly for m_{2i} of m_2) follows a standard normal distribution,

$$m_{1,i} = \left(\frac{\sqrt{1+\rho} + \sqrt{1-\rho}}{2}\right) z_{1i} + \left(\frac{\sqrt{1+\rho} - \sqrt{1-\rho}}{2}\right) z_{2i} \sim N(0,1),$$
$$m_{2,i} = \left(\frac{\sqrt{1+\rho} + \sqrt{1-\rho}}{2}\right) z_{1i} + \left(\frac{\sqrt{1+\rho} - \sqrt{1-\rho}}{2}\right) z_{2i} \sim N(0,1),$$

which leads to,

$$\|m_1\|_2 \sim \sqrt{\chi_n^2},\ \|m_2\|_2 \sim \sqrt{\chi_n^2},$$

where χ_n^2 denotes the chi-square distribution with degree of freedom of n. Since the numerator part $\langle m_1, m_2 \rangle$ is not a degenerate random quantity, one sees that $\hat{\rho}_1^{(d)}$ does not converge to a trivial random variable such as 1.

Now increase the sample size n to see which value the sample canonical correlation coefficient $\hat{\rho}_1^{(d)}$ converges to. By the law of large numbers and noting that the elements $m_{1,i}$ and $m_{2,i}$ are from i.i.d. standard normal distribution,

$$\frac{\|m_1\|_2^2}{n} = \sum_{i=1}^{n} \frac{m_{1i}^2}{n} \xrightarrow[n\to\infty]{P} 1, \;\; \frac{\|m_2\|_2^2}{n} = \sum_{j=1}^{n} \frac{m_{2j}^2}{n} \xrightarrow[n\to\infty]{P} 1.$$

Furthurmore, noting that m_1 and m_2 are i.i.d. samples,

$$\begin{aligned}\frac{\langle m_1, m_2 \rangle}{n} &= \left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}\right)\left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right)\sum_{i=1}^{n}\frac{z_{1i}^2}{n} \\ &+ \left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}\right)\left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right)\sum_{i=1}^{n}\frac{z_{2i}^2}{n} \\ &+ \left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}\right)^2\sum_{i=1}^{n}\frac{z_{1i}z_{2i}}{n} \\ &+ \left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right)^2\sum_{i=1}^{n}\frac{z_{1i}z_{2i}}{n} \\ &\xrightarrow[n\to\infty]{P} 2\left(\frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}\right)\left(\frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}\right) = \rho,\end{aligned}$$

which confirms the conventional large sample asymptotic property of the statistic $\hat{\rho}_1^{(d)}$,

$$\hat{\rho}_1^{(d)} \xrightarrow[d,n\to\infty]{D} \rho.$$

4. Simulation

Simulation study in this section aims at verifing the asymptotic behavior of sample canonical correlation coefficients and their corresponding weight vectors given in the main theorom 1 as dimension d grows with sample size n fixed. We first state the parameter settings to be used. For the spiked covariance structures of the random variables $X^{(d)}$ and $Y^{(d)}$ described in (5) and (6), we set $\sigma_X^2 = \tau_X^{(d)} = \sigma_Y^2 = \tau_Y^{(d)} = 1$. The population caconical weight vectors described in (7) and population caconical correlation coefficient are set to be,

$$\psi_X^{(d)} = (\cos 0.75\pi)e_1^{(d)} + (\sin 0.75\pi)e_2^{(d)}, \;\; \psi_Y^{(d)} = (\cos 0.75\pi)e_1^{(d)} + (\sin 0.75\pi)e_2^{(d)}, \;\; \rho = 0.7.$$

Note that $\langle \psi_X^{(d)}, e_1^{(d)} \rangle = \langle \psi_Y^{(d)}, e_1^{(d)} \rangle = \cos 0.75\pi = 0.7071$, which implies that the angle between $\psi_X^{(d)}$ and $e_1^{(d)}$ is 135°. The population cross-covariance structure of $X^{(d)}$ and $Y^{(d)}$ can be accordingly defined as in (9). We perform 100 runs of simulations for each combination of different values of the following three sets,

- Sample size $n \in \{20, 80\}$,
- Dimension $d \in \{200, 500\}$,

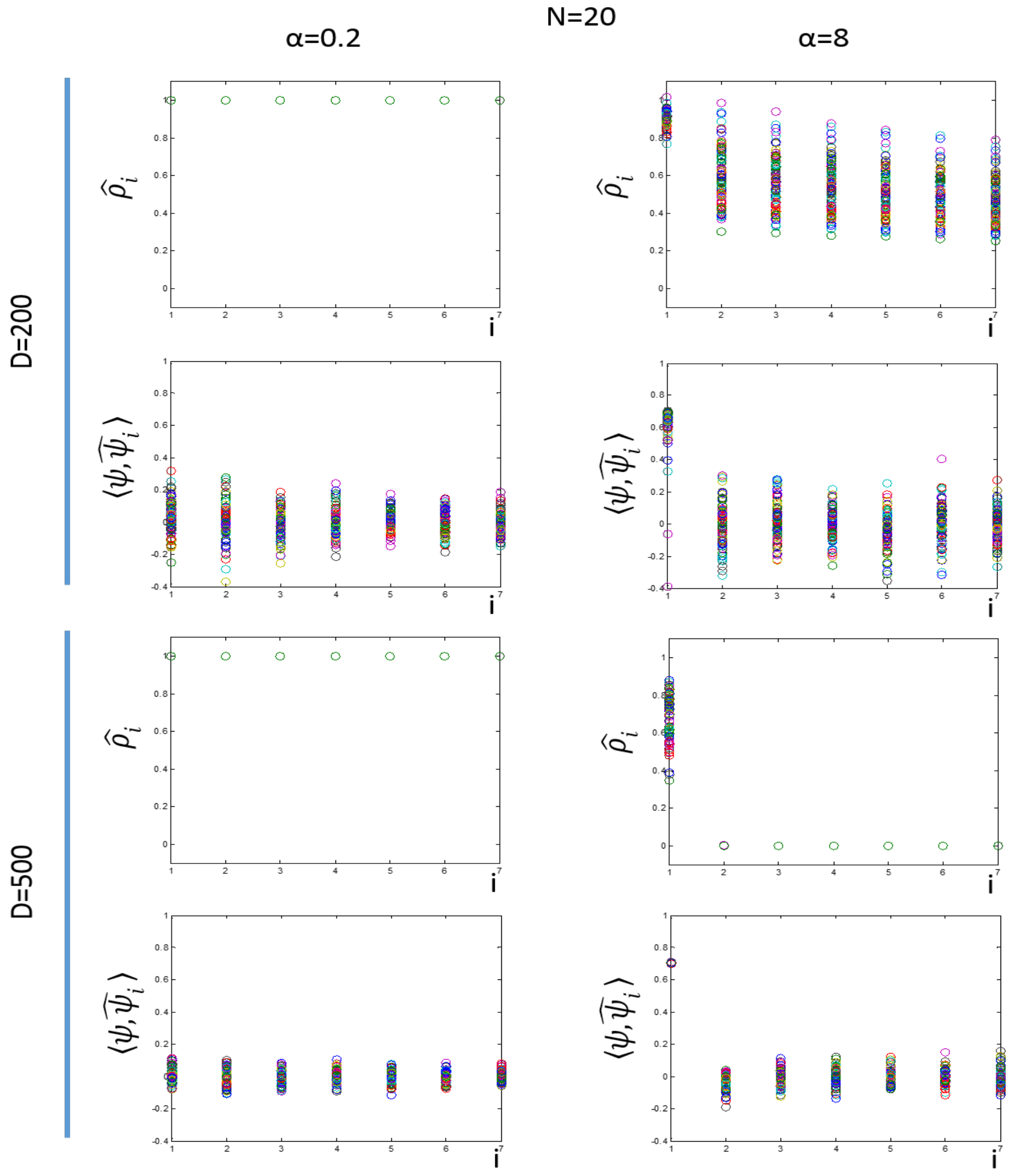

Figure 1. Estimated sample canonical correlation coefficients $\hat{\rho}_i^{(d)}$ and inner products of the sample left canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$ and the population canonical weight vector $\psi_{Xi}^{(d)}$, for $i = 1, 2, \ldots, 5$, obtained from 100 repetitions of simulations for different settings of dimension d and exponent α with a sample size of $n = 20$.

- Exponent $\alpha \in \{0.2, 8\}$.

Each case, estimates of the first 5 canonical correlation coefficients $\hat{\rho}_i^{(d)}$ and their corresponding canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$ and $\hat{\psi}_{Yi}^{(d)}$ are obtained. The estimated vectors $\hat{\psi}_{Xi}^{(d)}$ and $\hat{\psi}_{Yi}^{(d)}$, for $i = 1, 2, \ldots, 5$, are compared to the population canonical weight vector $\psi_X^{(d)}$ using their inner product. Here, we do not include results of $\hat{\psi}_{Yi}^{(d)}$ as they are similar as those of $\hat{\psi}_{Xi}^{(d)}$.

Figure 1 presents the simulation results for a small sample size of $n = 20$. For $\alpha = 0.2$, sample coefficients and vectors are almost of no use as the estimated vectors tend to be as far away as possible from the popultion direction (implied in the inner products of 0) with always perfect correlation. When

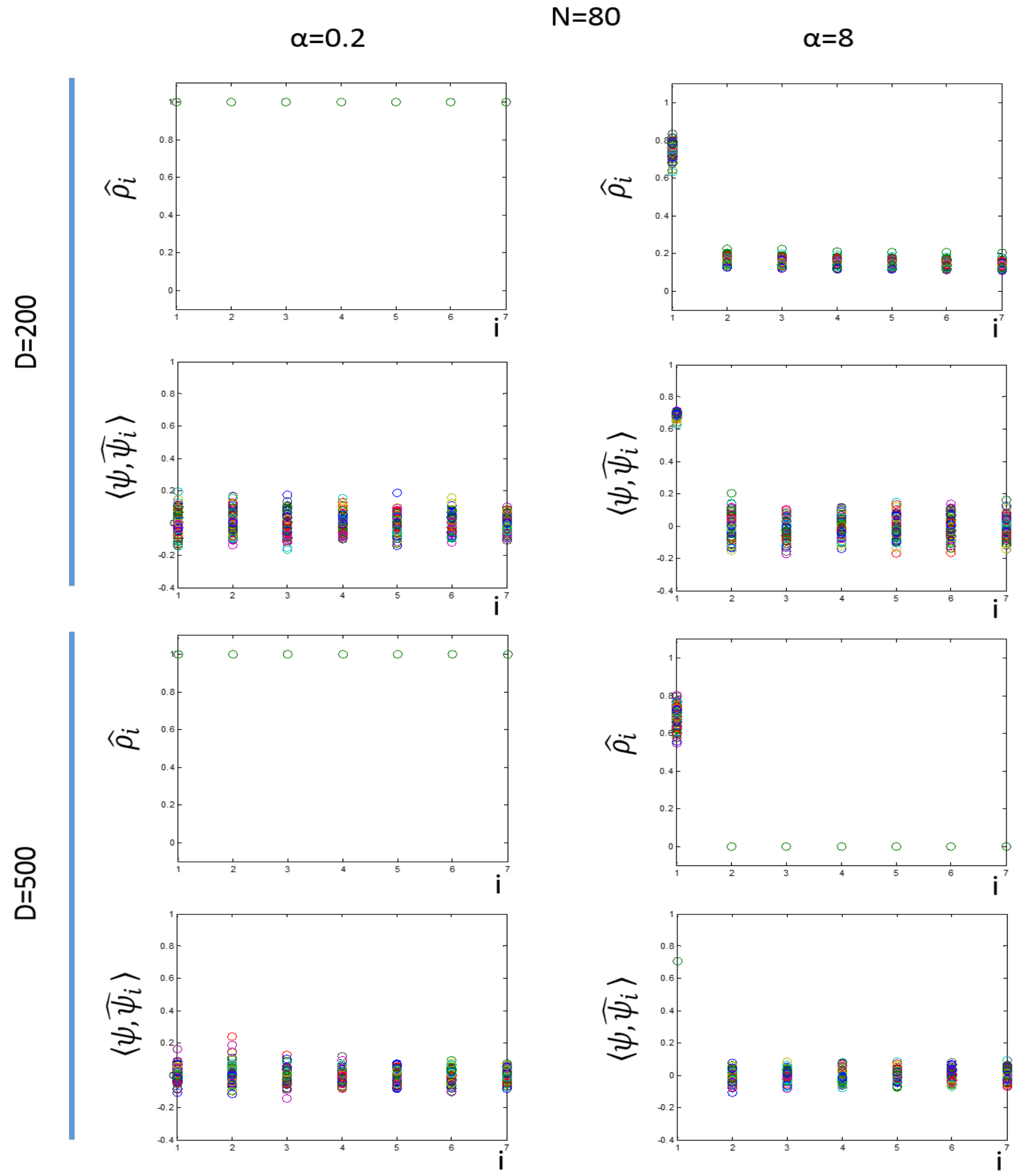

Figure 2. 100 estimated sample canonical correlation coefficients $\hat{\rho}_i^{(d)}$ and inner products of the sample left canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$ and the population canonical weight vector $\psi_{Xi}^{(d)}$, for $i = 1, 2, \ldots, 5$, obtained from 100 repetitions of simulations for different settings of dimension d and exponent α with a sample size of $n = 80$.

α increases to a high strength of 8, the first sample coefficient $\hat{\rho}_1^{(d)}$ approachs to the population direction whereas the rest degenerate to 0 as $d \to \infty$. The first left sample canonical weight vector $\hat{\psi}_{X1}^{(d)}$ converges to the direction $e_1^{(d)}$ (implied in the inner products of $\cos 0.75\pi$) containing dominant variability as $d \to \infty$ and the rest carry no information on the population direction with tending to deviate from it by a highest degree of 90°. Figure 2 illustrates the results for a larger sample size of $n = 80$. For the case of $\alpha = 0.2$, the behavior of $\hat{\rho}_i^{(d)}$ and $\hat{\psi}_{Xi}^{(d)}$ is similar as that in a small sample size case. However, for $\alpha = 8$, we see a noticeable decrease in variability of the first sample canonical correlation coefficient $\hat{\rho}_1^{(d)}$ around a true value of 0.7 and of the rest of $\hat{\rho}_i^{(d)}$ around 0. This implies that the usual large sample theory works for

$\hat{\rho}_1^{(d)}$. Diminishing variability is also observed for the sample canonical weight vectors $\hat{\psi}_{Xi}^{(d)}$, where the first sample vector $\hat{\psi}_{Xi}^{(d)}$ becomes almost identical to the largest variance direction $e_1^{(d)}$ and the rest diverge from the population canonical direction $\psi_X^{(d)}$.

References

[1] J. Ahn, J.S. Marron, K. Muller, and Y. Chi. The high dimension, low sample size geometric represetnation holds under mild conditions. *Biometrika*, 94(3):1–7, 2007.

[2] Z. Bai and Y. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability*, 21(3):1275–1294, 1993.

[3] P. Hall, J.S. Marron, and A. Neeman. Geometric representation of high dimension low sample size data. *Journal of the Royal Statistical Society*, 67(3):427–444, 2005.

[4] H. Hotelling. Relations between two sets of bariants. *Biometrika*, 28:321–377, 1936.

[5] I. Johnstone and A. Lu. Sparse principal components analysis. *Technical report, Stanford University*, 2009.

[6] S. Jung and J.S. Marron. Pca consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130, 2009.

[7] S. Jung, A. Sen, and J.S. Marron. Boundary behavior in high dimension, low sample size asymptotics of pca. *Journal of Multivariate Analysis*, 109:190–203, 2012.

[8] M. Lee. Continuum direction vectors in high dimensional low sample size data. *Dissertation, University of North Carolina at Chapel Hill*, pages 54–87, 2007.

[9] D. Paul. Asymptotics of the leading sample eigenvalues for a spiked covariance model. *Technical report, Stanford University*, 2005.

[10] D. Samarov. The analysis and advanced extensions of canonical correlation analysis. *Dissertation, University of North Carolina at Chapel Hill*, pages 121–176, 2009.

[11] D. Witten and R. Tibishirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.