Statistica Sinica **33** (2023), 149-168
doi:https://doi.org/10.5705/ss.202020.0327

EIGENVALUE DISTRIBUTION OF A HIGH-DIMENSIONAL DISTANCE COVARIANCE MATRIX WITH APPLICATION

Weiming Li, Qinwen Wang and Jianfeng Yao

Shanghai University of Finance and Economics,
Fudan University and The Chinese University of Hong Kong (Shenzhen)

Abstract: We introduce a new random matrix model called the distance covariance matrix, the normalized trace of which is equivalent to the distance covariance. We first derive a deterministic limit for the eigenvalue distribution of the distance covariance matrix when the dimensions of the vectors and the sample size tend to infinity simultaneously. This limit is valid when the vectors are independent or weakly dependent through a finite-rank perturbation. It is also universal and independent of the distributions of the vectors. Furthermore, the top eigenvalues of the distance covariance matrix are shown to obey an exact phase transition when the dependence of the vectors is of finite rank. This finding enables the construction of a new detector for weak dependence, where classical methods based on large sample covariance matrices or sample canonical correlations may fail in the considered high-dimensional framework.

Key words and phrases: Distance covariance, distance covariance matrix, eigenvalue distribution, finite-rank perturbation, nonlinear correlation, spiked models.

1. Introduction

Székely, Rizzo and Bakirov (2007) introduced the concept of the *distance covariance* $\mathcal{V}(\mathbf{x},\mathbf{y})$ of two random vectors $(\mathbf{x},\mathbf{y}) \in \mathbb{R}^p \times \mathbb{R}^q$ as a measure of their dependence. It is defined through an appropriately weighted L_2-distance between the joint characteristic function $\phi_{\mathbf{x},\mathbf{y}}(s,t)$ of $(\mathbf{x},\mathbf{y})$ and the product of their marginal characteristic functions $\phi_{\mathbf{x}}(s)\phi_{\mathbf{y}}(t)$, namely

$$\mathcal{V}(\mathbf{x},\mathbf{y}) = \left\{ \frac{1}{c_p c_q} \iint_{\mathbb{R}^p \times \mathbb{R}^q} \frac{|\phi_{\mathbf{x},\mathbf{y}}(s,t) - \phi_{\mathbf{x}}(s)\phi_{\mathbf{y}}(t)|^2}{\|s\|^{1+p}\|t\|^{1+q}} dsdt \right\}^{1/2}, \tag{1.1}$$

where the normalization constants are $c_d = \pi^{(1+d)/2}/\Gamma((1+d)/2)$ $(d = p, q)$. Clearly, $\mathcal{V}(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x}$ and $\mathbf{y}$ are independent.

For a collection of independent and identically distributed (i.i.d.) observa-

Corresponding author: Qinwen Wang, School of Data Science, Fudan University, Shanghai 200433, China. E-mail: wqw@fudan.edu.cn.

tions $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ from the population $(\mathbf{x}, \mathbf{y})$, Székely, Rizzo and Bakirov (2007) proposed the *sample distance covariance* $\mathcal{V}_n(\mathbf{x}, \mathbf{y})$ as

$$\mathcal{V}_n(\mathbf{x}, \mathbf{y}) = \{S_{1,n} + S_{2,n} - 2S_{3,n}\}^{1/2}, \tag{1.2}$$

where

$$S_{1,n} = \frac{1}{n^2} \sum_{k,\ell=1}^{n} \|\mathbf{x}_k - \mathbf{x}_\ell\| \|\mathbf{y}_k - \mathbf{y}_\ell\|,$$

$$S_{2,n} = \frac{1}{n^2} \sum_{k,\ell=1}^{n} \|\mathbf{x}_k - \mathbf{x}_\ell\| \frac{1}{n^2} \sum_{k,\ell=1}^{n} \|\mathbf{y}_k - \mathbf{y}_\ell\|,$$

$$S_{3,n} = \frac{1}{n^3} \sum_{k,\ell,m=1}^{n} \|\mathbf{x}_k - \mathbf{x}_\ell\| \|\mathbf{y}_k - \mathbf{y}_m\|.$$

One remarkable result (Székely, Rizzo and Bakirov (2007, Thm. 2)) is that whenever $\mathbb{E}[\|\mathbf{x}\| + \|\mathbf{y}\|] < \infty$, $\mathcal{V}_n(\mathbf{x}, \mathbf{y})$ converges almost surely to $\mathcal{V}(\mathbf{x}, \mathbf{y})$ as $n \to \infty$. Based on this, a powerful statistic,

$$T_n = \frac{n\mathcal{V}_n^2(\mathbf{x}, \mathbf{y})}{S_{2,n}}, \tag{1.3}$$

was developed to test the independence hypothesis,

$$H_0 : \mathbf{x} \text{ is independent of } \mathbf{y}, \tag{1.4}$$

by establishing the following: (i) under H_0, $T_n \xrightarrow{\mathcal{D}} Q$, a countable mixture of independent chi-squared distributions; and (ii) if $\mathbf{x}$ and $\mathbf{y}$ are dependent, $T_n \to \infty$ in probability. This asymptotic theory for T_n was established for the large sample asymptotics, where the two dimensions (p, q) are fixed, and the sample size n tends to infinity.

When the dimensions (p, q) of the two vectors become large, Székely and Rizzo (2013) observed that the above test becomes invalid, owing to a non-negligible bias of the squared sample distance covariance $\mathcal{V}_n^2(\mathbf{x}, \mathbf{y})$, and then proposed a bias-corrected version $\tilde{\mathcal{V}}_n^2(\mathbf{x}, \mathbf{y})$ as a substitution. Using this correction, the *sample distance correlation* $\tilde{R}_n(\mathbf{x}, \mathbf{y}) = \tilde{\mathcal{V}}_n(\mathbf{x}, \mathbf{y}) / [\tilde{\mathcal{V}}_n(\mathbf{x}, \mathbf{x}) \tilde{\mathcal{V}}_n(\mathbf{y}, \mathbf{y})]^{1/2}$ is employed to test the independence hypothesis, the null distribution of which is established in a specific asymptotic scheme, where n is kept fixed and p and q both grow to infinity. We refer to this scheme as the fixed-n asymptotic regime. However, a recent paper Zhu et al. (2020) reported that even the test based

on $\tilde{R}_n(\mathbf{x}, \mathbf{y})$ may lose power when detecting nonlinear correlations if all the dimensions (p, q, n) grow to infinity. In particular, they demonstrated that for high-dimensional vectors, their squared sample distance covariance $\tilde{\mathcal{V}}_n^2(\mathbf{x}, \mathbf{y})$ is asymptotically equivalent to the summation of their squared component-wise (linear) cross sample covariances. This implies that the distance covariance can only capture linear correlations in high-dimensional regimes.

In order to detect nonlinear correlations between $\mathbf{x}$ and $\mathbf{y}$ when all the dimensions (p, q, n) grow to infinity, we propose a new random matrix model, called the *distance covariance matrix* (DCM). Specifically, denoting two data matrices as $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, the DCM of $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathbf{S}_{xy} \triangleq \mathbf{P}_n \mathbf{D}_x \mathbf{P}_n \mathbf{D}_y \mathbf{P}_n, \tag{1.5}$$

where

$$\mathbf{D}_x \triangleq \frac{1}{p}\mathbf{X}'\mathbf{X} + \frac{1}{pn}\sum_{i=1}^{n} ||\mathbf{x}_i||^2 \mathbf{I}_n, \quad \mathbf{D}_y \triangleq \frac{1}{q}\mathbf{Y}'\mathbf{Y} + \frac{1}{qn}\sum_{i=1}^{n} ||\mathbf{y}_i||^2 \mathbf{I}_n, \tag{1.6}$$

and

$$\mathbf{P}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n' \tag{1.7}$$

is a projection matrix. The DCM $\mathbf{S}_{xy}$ is closely connected to the distance covariance $\mathcal{V}(\mathbf{x}, \mathbf{y})$. As discussed in Section 2, a normalized trace of $\mathbf{S}_{xy}$ is asymptotically equivalent to the empirical distance covariance $\mathcal{V}_n(\mathbf{x}, \mathbf{y})$. Therefore, we believe that the spectrum of $\mathbf{S}_{xy}$ might contain information on the nonlinear dependence between $\mathbf{x}$ and $\mathbf{y}$. To this end, we investigate the first-order asymptotic behavior of the whole spectrum of the DCM $\mathbf{S}_{xy}$ under the two-sample *Marčenko–Pastur asymptotic regime*,

$$(n, p, q) \to \infty, \quad (c_{n1}, c_{n2}) := \left(\frac{p}{n}, \frac{q}{n}\right) \to (c_1, c_2) \in (0, \infty)^2. \tag{1.8}$$

Interestingly, we find that instead of the normalized trace of $\mathbf{S}_{xy}$, its largest eigenvalues have the ability to detect certain nonlinear correlations between the two high-dimensional random vectors $\mathbf{x}$ and $\mathbf{y}$.

This study contributes to the literature in three ways. Our first result shows that the test statistic T_n developed in Székely, Rizzo and Bakirov (2007) for the independence hypothesis H_0 degenerates to the unit in the Marčenko–Pastur asymptotic regime. This extends a similar finding in Székely and Rizzo (2013) for their fixed-n asymptotic regime. Therefore, the statistic T_n cannot be applied to

test the independence hypothesis H_0 in the Marčenko–Pastur asymptotic regime (1.8).

As our second result, we derive a deterministic limiting distribution F for the eigenvalue distribution of $\mathbf{S}_{xy}$. This means, in particular, that an arbitrary eigenvalue statistic of the form $n^{-1}\sum_i g(\lambda_i)$, where (λ_i) denotes the eigenvalues of $\mathbf{S}_{xy}$, with some smooth function g, converges to $\int g(x)dF(x)$. The limiting distribution F is valid when the vectors are independent or weakly dependent, corresponding to a finite-rank perturbation of the independence. An important property is that this limit is *universal*, in the sense that it does not depend on the respective distributions of the vectors.

Third, to demonstrate the usefulness of our limiting eigenvalue distribution, we apply the theory to detect a deviation from the independence hypothesis by considering a family of finite-rank nonlinear dependence alternatives. We investigate both the global and local spectral behaviors of $\mathbf{S}_{xy}$. Globally, because the dependence is of finite rank, the limiting distribution of the eigenvalues remains the same as that in the independence case, that is, the universal limit. However at a local scale, the largest eigenvalues of $\mathbf{S}_{xy}$ converge to some limits outside the support of this universal limit, as long as the strength of the dependence is beyond some critical value. Moreover, the locations of these outlying limits can be completely determined through the model parameters. Actually, these results under finite-rank dependence parallel what is now known as Baik–Ben–Arous–Péché transition in random matrix theory; see Baik , Ben-Arous and Péché (2005), Baik and Silverstein (2006), and Paul (2007). Thus we conclude that the largest eigenvalues of $\mathbf{S}_{xy}$ can be used to detect such a dependence structure. In addition, we propose an estimator for the rank of the dependence. This estimator is based on the ratios of the largest adjacent eigenvalues of $\mathbf{S}_{xy}$. Its performance is assessed using simulation experiments.

Technically, our theoretical strategy for deriving the universal limit under independence is to first derive a system of equations for the corresponding Stieltjes transform in the Gaussian case. Indeed, when the vectors $\mathbf{x}$ and $\mathbf{y}$ are Gaussian, the DCM $\mathbf{S}_{xy}$ is orthogonally invariant; we can thus assume, without loss of generality, that the two population covariance matrices are diagonal, which greatly simplifies the analysis. In a second step, we use a generalization of Lindeberg's substitution method to obtain an accurate estimate for the difference between the Stieltjes transforms from Gaussian vectors and those of non-Gaussian ones. This difference is small enough that the limiting distribution for the global spectrum of $\mathbf{S}_{xy}$ is actually universal, regardless of the underlying distributions of the vectors.

The rest of the paper is organized as follows. Section 2 presents our model assumptions and discusses the relation between the DCM $\mathbf{S}_{xy}$ and the sample distance covariance $\mathcal{V}_n(\mathbf{x}, \mathbf{y})$. Section 3 establishes the limiting spectral distribution of $\mathbf{S}_{xy}$ under the Marčenko–Pastur asymptotic regime (1.8) when $\mathbf{x}$ and $\mathbf{y}$ are independent. Section 4 applies this theory to detect the finite-rank nonlinear dependence between two high-dimensional vectors. All proofs of our technical results are gathered in the online Supplementary Material.

2. DCM

Let $\mathbf{M}_p$ be a $p \times p$ symmetric or Hermitian matrix with eigenvalues $(\lambda_j)_{1 \le j \le p}$. Its spectral distribution is the probability measure

$$F^{\mathbf{M}_p} = \frac{1}{p} \sum_{j=1}^{p} \delta_{\lambda_j},$$

where δ_b denotes the Dirac mass at b. For a probability measure μ on the real line (equipped with its Borel σ-algebra), its Stieltjes transform s_μ is a map from $\mathbb{C}^+$ onto itself,

$$s_\mu(z) = \int_{\mathbb{R}} \frac{1}{x - z} d\mu(x), \qquad z \in \mathbb{C}^+,$$

where $\mathbb{C}^+ \triangleq \{z \in \mathbb{C} : \Im(z) > 0\}$.

Our asymptotic study of the spectrum of the DCM $\mathbf{S}_{xy}$ is developed under the following assumptions.

Assumption 1. *The dimensions (n, p, q) tend to infinity, as in* (1.8).

Assumption 2. *The data matrices $\mathbf{X} = (\mathbf{x}_i) \in \mathbb{R}^{p \times n}$ and $\mathbf{Y} = (\mathbf{y}_i) \in \mathbb{R}^{q \times n}$ admit the following independent components model:*

$$\mathbf{X} = \mathbf{\Sigma}_x^{1/2} \mathbf{W}_1 \quad \textit{and} \quad \mathbf{Y} = \mathbf{\Sigma}_y^{1/2} \mathbf{W}_2,$$

where $\mathbf{\Sigma}_x \in \mathbb{R}^{p \times p}$ and $\mathbf{\Sigma}_y \in \mathbb{R}^{q \times q}$ denote the population covariance matrices of $\mathbf{x}$ and $\mathbf{y}$, respectively, and $(\mathbf{W}_1', \mathbf{W}_2') = (w_{ij})$ is an array of i.i.d. random variables satisfying

$$\mathbb{E}(w_{11}) = 0, \quad \mathbb{E}(w_{11}^2) = 1, \quad \mathbb{E}|w_{11}|^\gamma < \infty,$$

for some $\gamma \ge 4$.

Assumption 3. *The spectral norms of $(\mathbf{\Sigma}_x, \mathbf{\Sigma}_y)$ are uniformly bounded, and their spectral distributions $(H_{xp}, H_{yq}) \triangleq (F^{\mathbf{\Sigma}_x}, F^{\mathbf{\Sigma}_y})$ converge weakly to two probability*

Table 1. Empirical mean and standard deviation of the test statistic T_n from 1,000 independent replications with $p/n = q/n = 1/2$ and $p \in \{50, 100, 200, 400\}$. Independent standard normal vectors are used for $\mathbf{x}$ and $\mathbf{y}$.

$p = 50$		$p = 100$		$p = 200$		$p = 400$	
mean	sd	mean	sd	mean	sd	mean	sd
1.0104	0.0075	1.0048	0.0036	1.0026	0.0018	1.0013	0.0009

distributions (H_x, H_y)*, which are referred as* population spectral distributions *(PSD).*

Our first result concerns the connection between our DCM $\mathbf{S}_{xy}$ defined in (1.5) and the sample distance covariance $\mathcal{V}_n(\mathbf{x}, \mathbf{y})$ defined in (1.2).

Theorem 1. *Suppose that Assumptions* 1–3 *hold, with some* $\gamma > 5$*. Then, we have*

$$\mathcal{V}_n^2(\mathbf{x}, \mathbf{y}) = \frac{1}{2n^2}\sqrt{\frac{pq}{\gamma_x \gamma_y}}\mathrm{tr}\mathbf{S}_{xy} + o_p(1). \tag{2.1}$$

Theorem 1 demonstrates that the squared sample distance covariance $\mathcal{V}_n^2(\mathbf{x}, \mathbf{y})$ is asymptotically equal to the normalized trace of the DCM $\mathbf{S}_{xy}$. As a first application of the DCM $\mathbf{S}_{xy}$, we use this approximation to establish the degeneracy of the test statistic T_n given in (1.3) for testing the independence hypothesis (1.4) under the Marčenko–Pastur asymptotic framework.

Theorem 2. *Suppose that Assumptions* 1–3 *hold, with some* $\gamma > 5$*. Then, under the null hypothesis* H_0*, we have* $T_n \to 1$ *in probability.*

A simple simulation experiment is conducted to exhibit the degeneracy of T_n for two independent standard normal vectors. The dimension-to-sample size ratios are fixed to be $p/n = q/n = 1/2$, the values of p $(= q)$ range from 50 to 400, and the number of independent replications is 1,000. As shown in Table 1, as p increases, the empirical mean and standard deviation of T_n converge to one and zero, respectively. Consequently, the test established in Székely, Rizzo and Bakirov (2007) using the chi-squared approximation has a much inflated size tending to one when the dimensions are large compared to the sample size.

3. Limiting Spectral Distribution of S_{xy} when x and y are Independent

This section presents the first-order convergence of the empirical spectral distribution $F^{\mathbf{S}_{xy}}$ of the DCM $\mathbf{S}_{xy}$ when $\mathbf{x}$ and $\mathbf{y}$ are independent.

Theorem 3. *Suppose that Assumptions* 1–3 *hold. Then, almost surely, the empirical spectral distribution* $F^{\mathbf{S}_{xy}}$ *converges weakly to a limiting spectral distri-*

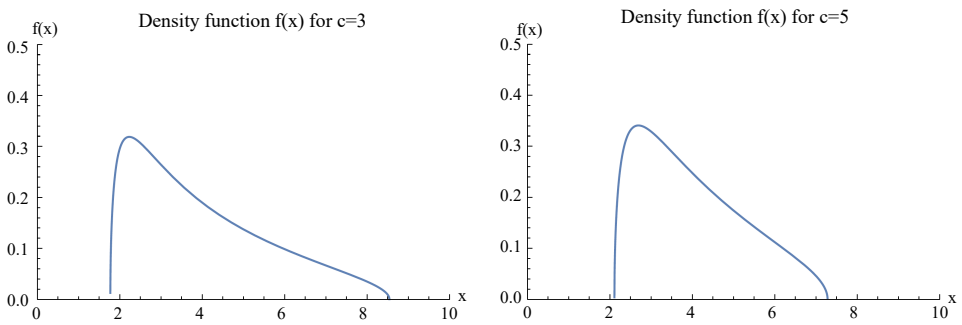

Figure 1. Density curves of LSDs for $c = 3$ (left) and $c = 5$ (right). The PSDs are $H_x = H_y = \delta_1$.

bution (LSD) F, the Stieltjes transform of which $s = s(z)$, is a solution to the following system of equations:

$$\begin{cases} s = \dfrac{wm - 1}{z}, \\ w = \displaystyle\int ts + \frac{ts}{1 + tc_1^{-1}sm} dH_x(t), \\ m = \displaystyle\int t + \frac{t}{1 + tc_2^{-1}w} dH_y(t), \end{cases} \tag{3.1}$$

where $w = w(z)$ and $m = m(z)$ are two auxiliary analytic functions. The solution is also unique on the set

$$\{s(z) : s(z) \in \mathbb{C}^+, w(z) \in \mathbb{C}^+, m(z) \in \mathbb{C}^-, z \in \mathbb{C}^+\}. \tag{3.2}$$

Remark 1. The two auxiliary functions $w(z)$ and $m(z)$ are the limits of $w_n(z)$ and $m_n(z)$, respectively, defined in (B.9) of the Supplementary Material. Their construction accounts for the signs of their imaginary parts, as in (3.2).

Next, we show how to calculate the LSD F using the system of equations (3.1). Considering the case where the two populations $\mathbf{x}$ and $\mathbf{y}$ are of the same dimension and both have identity covariance matrices, we thus have

$$c_1 = c_2 = c \quad \text{and} \quad H_x = H_y = \delta_1. \tag{3.3}$$

For this case, a closed-form solution to the system (3.1) does exist; that is, the Stieltjes transform $s = s(z)$ of the LSD F satisfies the following

$$c^2 - s + 2cs - 4c^2 s + s^2 + c^2 sz - 2s^2 z + 2cs^2 z + s^3 z - s^3 z^2 = 0. \tag{3.4}$$

Substituting $z = x + \mathrm{i}v$ and $s = s_u + \mathrm{i}s_v$ into (3.4), and then letting $v \downarrow 0$, we obtain the following system of equations by separating the real and imaginary parts on the left-hand side of (3.4):

$$\begin{cases} s_v^2 = \dfrac{c^2 - s_u + 2cs_u - 4c^2 s_u + c^2 x s_u + s_u^2 - 2xs_u^2 + 2cxs_u^2 + xs_u^3 - x^2 s_u^3}{1 - 2x + 2cx + 3xs_u - 3x^2 s_u}, \\ s_v^2 = \dfrac{1 - 2c + 4c^2 - c^2 x - 2s_u + 4xs_u - 4cxs_u - 3xs_u^2 + 3x^2 s_u^2}{-x + x^2}. \end{cases} \tag{3.5}$$

Cancelling the variable s_u from (3.5), we obtain three solutions for s_v^2 as a function of x. These three functions indeed have closed forms, but are lengthy, and we omit their explicit expressions here. Then, for each real value of x, only one solution of s_v^2 is real and nonnegative, which corresponds to the density function $f(x)$ of the LSD F, that is, $f(x) = \sqrt{s_v^2}/\pi$. Using this approach, in Figure 1, we plot two LSDs for the setting in (3.3) corresponding to $c = 3$ and $c = 5$. However, in general, when there is no closed-form solution for (3.1), we rely on numerical approximations for the limiting Stieltjes transform $s(z)$ and the underlying limiting density function. These methods are used in the illustration below, and also in the simulation experiments in Section 4.

Numerical illustrations of Theorem 3 are conducted under two models:

Model 1 : $H_x = H_y = \delta_1$, $c_1 = c_2 = 1$, $z_{11} \sim N(0, 1)$;

Model 2 : $H_x = 0.5\delta_{0.5} + 0.5\delta_1$, $H_y = 0.5\delta_{0.25} + 0.5\delta_{0.75}$, $c_1 = 2, c_2 = 1$, and $z_{11} \sim (\chi_v^2 - v)/\sqrt{2v}$, a standardized chi-squared distribution with degrees of freedom $v = 2$.

The PSDs in the first model are simple point masses, and the system (3.1) defining the LSD simplifies to a single equation $(z^2 - z)s^3 - s^2 + (3 - z)s - 1 = 0$ (letting $c = 1$ in (3.4)). The second model is a bit more elaborate. The PSDs are mixtures of two point masses, and the innovations z_{ij} follow a chi-squared distribution with heavy tails.

To exhibit the LSDs defined by Models 1 and 2, we simply approximate their density functions by $\hat{f}(x) = \Im s(x + \mathrm{i}/10^4)/\pi$, for $x \in \mathbb{R}$. This approximation is justified by the inversion formula of the Stieltjes transforms, that is, $f(x) = \lim_{\varepsilon \to 0^+} \Im s(x + \mathrm{i}\varepsilon)/\pi$, provided the limit exists; see Theorem B.10 in Bai and Silverstein (2010). Obviously, our approximation takes $\varepsilon = 10^{-4}$, which is small enough for the illustration here. Next, for any given $z = x + \mathrm{i}/10^4$, we numerically solve the system of equations in (3.1), and select the unique solution

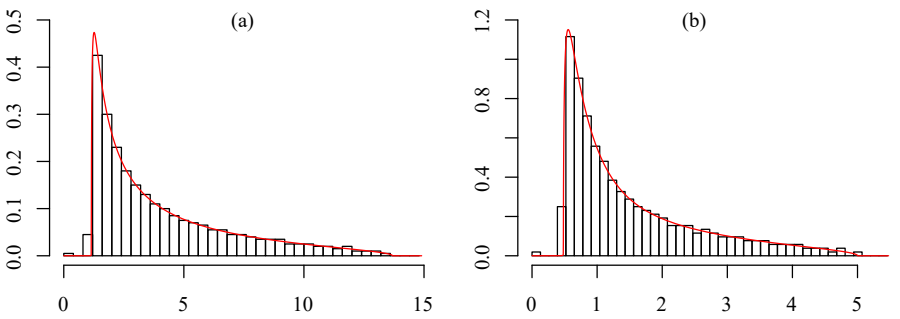

Figure 2. Histogram of eigenvalues of the matrix $\mathbf{S}_{xy}$ under Model 1 (left panel), with dimensions $p = q = n = 500$, and Model 2 (right panel), with dimensions $p = 2q = 2n = 800$. The solid line curves are the corresponding densities of the LSDs.

$(s(z), w(z), m(z))$ satisfying (3.2), which is done automatically in the software Mathematica. Finally, taking the imaginary part of $s(z)/\pi$ gives $\hat{f}(x)$.

In this simulation experiment, the empirical PSDs are chosen as their limiting PSDs and the dimensions are $(p, q, n) = (500, 500, 500)$ for Model 1, and $(p, q, n) = (800, 400, 400)$ for Model 2. All eigenvalues are collected from 100 independent replications. The averaged histograms of the eigenvalues of $\mathbf{S}_{xy}$ from these replications are depicted in Figure 2, which shows that these empirical distributions match well the limiting density curves predicted in Theorem 3.

4. Application to the Detection of Dependence between Two High-Dimensional Vectors

Theorem 3 determines a universal limit for the bulk spectrum of the DCM when the two sets of samples are independent. Here, a natural question arises: how will this bulk limit evolve when they become dependent? Apparently, if their inherent dependence is very strong, the spectral limit of the DCM will differ from the universal limit in Theorem 3. Here, we study a special type of weak dependence, namely, finite-rank dependence. This concept parallels the idea of finite-rank perturbation or spiked population models in high-dimensional statistics, which are widely studied in connection with high-dimensional PCA, factor modeling, and the signal detection problem (Johnstone and Paul (2018)). A striking finding from our work is that such finite-rank nonlinear dependence can be detected using the largest eigenvalues of the DCM, which existing methods based on the sample covariances, sample correlations, or sample canonical correlations are not able to do.

4.1. Extreme eigenvalues of DCM under finite-rank dependence

Specifically, we consider two dependent populations $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{z} \in \mathbb{R}^q$, defined as follows:

(i) For a fixed $m \in \mathbb{N}$, let $(\mathbf{u}_k)_{1\le k\le m}$ and $(\mathbf{v}_k)_{1\le k\le m}$ be two independent sequences of i.i.d. vectors distributed uniformly on the unit spheres in $\mathbb{R}^q$ and $\mathbb{R}^p$, respectively.

(ii) Given the sequences $(\mathbf{u}_k)$ and $(\mathbf{v}_k)$, the population $\mathbf{z}$ is defined as

$$\mathbf{z} = \varepsilon \left(\sum_{k=1}^{m} \theta_k \mathbf{u}_k \mathbf{v}_k' \right) \mathbf{x} + \mathbf{y}, \tag{4.1}$$

where

(1) $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$ satisfy Assumptions (b) and (c);

(2) ε is a standardized random variable with a finite fourth moment.

(3) $0 < \theta_m < \cdots < \theta_1 < \infty$ are m constants representing the strengths of the dependence between $\mathbf{x}$ and $\mathbf{z}$.

Remark 2. The pair of random vectors $(\mathbf{x}, \mathbf{z})$ in (4.1) are nonlinearly dependent; that is, they are uncorrelated, but dependent. To see this, consider a particular case such that ε is a random sign taking values 1 or -1 with equal probability. Then, it is easy to see that the random sign ε put on the vector $\mathbf{x}$ implies the lack of correlation between the vectors. To establish their dependence, simple algebra shows that

$$\mathbb{E}\big(\|\mathbf{x}\|^2\big)\mathbb{E}\big(\|\mathbf{z}\|^2\big) = \frac{1}{p}\|\theta\|^2\mathbb{E}^2\big(\|\mathbf{x}\|^2\big) + \mathbb{E}\|\mathbf{x}\|^2\mathbb{E}\|\mathbf{y}\|^2,$$
$$\mathbb{E}\big(\|\mathbf{x}\|^2\|\mathbf{z}\|^2\big) = \frac{1}{p}\|\theta\|^2\mathbb{E}\|\mathbf{x}\|^4 + \mathbb{E}\|\mathbf{x}\|^2\mathbb{E}\|\mathbf{y}\|^2.$$

Here, $\|\theta\|^2 = \theta_1^2 + \cdots + \theta_m^2$. Unless $\mathbf{x}$ is a constant vector, $\mathbb{E}\|\mathbf{x}\|^4 > \mathbb{E}^2\big(\|\mathbf{x}\|^2\big)$, and thus the vectors $\mathbf{x}$ and $\mathbf{z}$ are dependent.

Suppose we have an i.i.d. sample $(\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n)$ from the population $(\mathbf{x}, \mathbf{z}) \in \mathbb{R}^p \times \mathbb{R}^q$ defined in (4.1). Denote by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ the two data matrices with sizes $p \times n$ and $q \times n$, respectively. Similarly to the matrices in (1.6), we define two matrices $\mathbf{D}_x$ and $\mathbf{D}_z$ as

$$\mathbf{D}_x = \frac{1}{p}\mathbf{X}'\mathbf{X} + \kappa_x \mathbf{I}_n \quad \text{and} \quad \mathbf{D}_z = \frac{1}{q}\mathbf{Z}'\mathbf{Z} + \kappa_z \mathbf{I}_n,$$

where

$$\kappa_x \triangleq (pn)^{-1}\sum_{i=1}^{n}||\mathbf{x}_i||^2 \quad \text{and} \quad \kappa_z \triangleq (qn)^{-1}\sum_{i=1}^{n}||\mathbf{z}_i||^2.$$

The corresponding DCM is written as

$$\mathbf{S}_{xz} \triangleq \mathbf{P}_n\mathbf{D}_x\mathbf{P}_n\mathbf{D}_z\mathbf{P}_n.$$

We examine the spectral properties of $\mathbf{S}_{xz}$ for the dependent pair $(\mathbf{x}, \mathbf{z})$ defined in (4.1). First, because the rank of the perturbation is finite, we show that the limiting spectral distribution of $\mathbf{S}_{xz}$ remains as if the two populations are independent.

Theorem 4. *Suppose that Assumptions* 1–3 *hold for model* (4.1). *The limiting spectral distribution of* $\mathbf{S}_{xz}$ *is given by the same* F *defined in Theorem* 3.

According to Theorem 4, the global behavior of the eigenvalues of the DCM $\mathbf{S}_{xz}$ are not useful for distinguishing such weak dependence from the independence scenario. In the following, we examine the top eigenvalues of $\mathbf{S}_{xz}$, and show that the weak dependence structure is encoded in these top eigenvalues. Thus, detecting this weak dependence becomes possible using these top eigenvalues. First, we introduce some notation. We denote

$$\lambda_+ = \limsup_{n\to\infty}\|\mathbf{S}_{xy}\|,$$

which is finite. On (λ_+, ∞), define the function

$$g(\lambda) = -\int t dH_x(t)\int \frac{w(\lambda)}{c_2 + tw(\lambda)}dH_y(t), \quad \lambda > \lambda_+, \tag{4.2}$$

where $w(z)$ is given in (3). It is easy to verify that $g(\lambda) > 0$, $g'(\lambda) < 0$, and $\lim_{\lambda\to+\infty} g(\lambda) = 0$. Next, define

$$\theta_0 := \lim_{\lambda\downarrow\lambda_+}[g(\lambda)]^{-1/2}. \tag{4.3}$$

Therefore, g is a one-to-one, strictly decreasing, and nonnegative function from (λ_+, ∞) to $(1/\theta_0^2, 0)$.

Theorem 5. *Suppose that Assumptions* 1–3 *hold for model* (4.1) *and, for some* $k \in \{1, \ldots, m\}$, $\theta_k > \theta_0$. *Then, the* kth *largest eigenvalue* $\lambda_{n,k}$ *of the DCM* $\mathbf{S}_{xz}$ *converges almost surely to a limit*

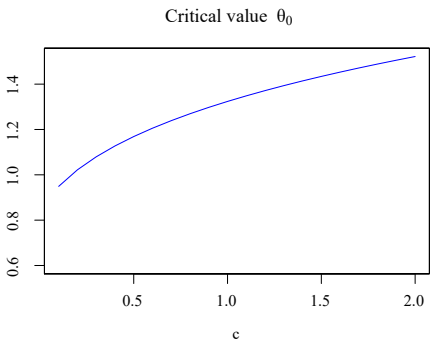

Figure 3. Critical value θ_0 for $c_1 = c_2 = c$ and $H_x = H_y = \delta_1$.

$$\lambda_k = g^{-1}\left(\frac{1}{\theta_k^2}\right) > \lambda_+, \tag{4.4}$$

where g^{-1} denotes the functional inverse of g.

Remark 3. In general, the function g and the critical value θ_0 have no analytic formulae; however both can be found numerically for any given model setting. In some cases, for example, the setting considered in (3.3), the function $g(\lambda)$ given in (4.2) is a solution to

$$cg^3(\lambda) + (1+4c)g^2(\lambda) + g(\lambda)(3+4c-c\lambda) + 2 = 0. \tag{4.5}$$

In fact, there are three solutions to (4.5), all of which have explicit, but lengthy expressions. Here, we choose the one that monotonically decreases to zero as λ tends to infinity, which is our target function $g(\lambda)$. Then, the critical value $\theta_0 = [g(\lambda_+)]^{-1/2}$ can be obtained accordingly. Note that the right edge λ_+ of the LSD F can be derived theoretically by setting the density function $f(x)$ to zero. As an illustration, we show the relation between the value θ_0 and the ratio c for the case (3.3) in Figure 3.

The limit λ_k in (4.4) is outside the support of the LSD F. A technical point here is that Theorem 5 does not tell us what happens to $\lambda_{n,k}$ if $\theta_k \le \theta_0$. By assuming the convergence of the largest eigenvalue of the base component $\mathbf{S}_{xy}$ to the right edge point of the LSD, we can establish the following *exact phase transition* for the top eigenvalues $\lambda_{n,k}$ $(1 \le k \le m)$.

Corollary 1. *In addition to Assumptions* 1–3 *for model* (4.1)*, suppose that the largest eigenvalue of the DCM $\mathbf{S}_{xy}$ converges to λ_+, which is the right edge point*

of the LSD F. Then, for $k = 1, \ldots, m$,

$$\lambda_{n,k} \xrightarrow{a.s.} \begin{cases} \lambda_k & \text{if } \theta_k > \theta_0, \\ \lambda_+ & \text{if } \theta_k \le \theta_0, \end{cases}$$

where θ_0 and λ_k are given in (4.3) *and* (4.4)*, respectively.*

Corollary 1 follows directly from the proof of Theorem 5 and the classic interlacing theorem. It implies that the value θ_0 is the exact critical value for the phase transition of the top eigenvalues of the DCM $\mathbf{S}_{xz}$. Note that the convergence of the largest eigenvalue of the (null) DCM $\mathbf{S}_{xy}$ to λ_+ is needed and assumed here to ensure the convergence of those sub-critical spike eigenvalues, that is, $\theta_k \le \theta_0$, to the same right edge point λ_+. On the other hand, it is very likely that this largest eigenvalue does converge. However, the proof for such convergence is lengthy and technical, and thus left for future investigation.

4.2. Monte Carlo experiments

This section examines the finite-sample properties of the outlier eigenvalues of $\mathbf{S}_{xz}$. To simplify the exposition, we consider only the rank-one situation ($m = 1$) in this section. Higher dependence ranks with $m > 1$ are discussed in Section 4.3. Three models are considered under normal populations:

Model 4 : $H_x = H_y = \delta_1,\ c_1 = c_2 = 2;$

Model 5 : $H_x = H_y = \delta_1,\ c_1 = 0.1,\ c_2 = 0.2;$

Model 6 : $H_x = 0.5\delta_{0.5} + 0.5\delta_1,\ H_y = 0.5\delta_1 + 0.5\delta_{1.5},\ c_1 = 1, c_2 = 2.$

Models 4 and 5 are both standard normal populations, with different dimension-to-sample size ratios. Model 6 is more general by employing two discrete PSDs. All statistics are calculated using 1,000 independent replications.

We begin with the convergence of the largest eigenvalue of $\mathbf{S}_{xz}$ under Model 4. Theoretically, the largest eigenvalue becomes an outlier when $\theta > \theta_0 = 1.52$ (see Figure 3 for the critical value). The parameter θ is thus set to $\theta = 0, 1, 2, 3$. The sample size n ranges from 100 to 1,600. The empirical mean and standard deviation of the largest eigenvalue are shown in Table 2. It shows that, for $\theta = 0$ and 1 (second to fifth columns), the largest eigenvalue increases with a decreasing standard error as n grows and is close to $\lambda_+ = 9.95$, the right edge point of F. When $\theta = 2$ and 3 (last four columns), the largest eigenvalue converges to its theoretical limit $\lambda = 10.6875$ for $\theta = 2$, and $\lambda = 15.0123$ for $\theta = 3$. These results fully coincide with the conclusions of Theorem 5.

Table 2. Empirical mean and standard deviation of the largest eigenvalue under Model 4. The setting is $c_{n1} = c_{n2} = 2$ with varying n and 1,000 independent replications. The right edge point of the LSD is $\lambda_+ = 9.95$.

	$\theta = 0$		$\theta = 1$		$(\theta, \lambda) = (2, 10.6875)$		$(\theta, \lambda) = (3, 15.0123)$	
n	mean	sd	mean	sd	mean	sd	mean	sd
100	9.5732	0.3126	9.6443	0.3419	10.7285	0.7055	15.1056	1.5770
200	9.7247	0.1972	9.7486	0.2013	10.7219	0.5048	15.0821	1.1099
400	9.8094	0.1302	9.8209	0.1239	10.7114	0.3500	15.0446	0.7458
800	9.8587	0.0769	9.8729	0.0796	10.7079	0.2531	14.9985	0.5505
1,600	9.8950	0.0479	9.8966	0.0502	10.6985	0.1745	15.0249	0.3794

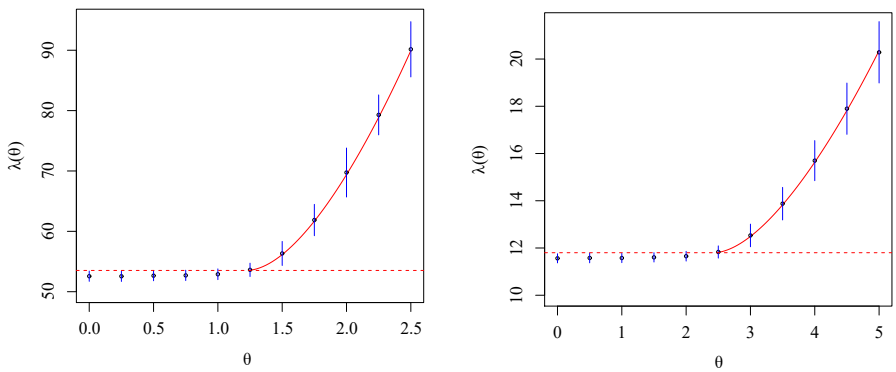

Figure 4. The average of the largest eigenvalue under Model 5 (left panel) and Model 6 (right panel) from 1,000 independent replications, with ±1 standard deviations (blue bars). The solid red line is the limiting curve of the function $\lambda(\theta)$, and the dashed red line represents the right boundary of the LSD's support.

Next, we study the evolution of the outlier limit $\lambda(\theta)$ in functions of the dependence strength θ. Models 5 and 6 are considered, with the dimensions fixed at $(p, q, n) = (200, 400, 2000)$ for Model 5, and at $(p, q, n) = (800, 800, 400)$ for Model 6. The parameter θ ranges from 0 to 2.5 for Model 5, and from 0 to 5 for Model 6. Figure 4 displays the average of the largest eigenvalue with ±1 standard deviations (vertical bars). The dashed red lines mark the right boundary of F, and the solid red lines are the theoretical curves of $\lambda = \lambda(\theta)$. Both graphs in Figure 4 exhibit a common trend that the largest eigenvalue departs from the bulk when θ crosses a critical value and goes up with an increasing standard deviation.

Lastly, we compare the performance of using the largest eigenvalues of our DCM model with that of a high-dimensional canonical correlation analysis (CCA) (Yang and Pan (2015); Bao et al. (2019)) for detecting dependence between two

groups of random samples. As is well known, a direct application of CCA often fails to detect dependence when the two sample sets are dependent, but uncorrelated. Thus, Yang and Pan (2015) suggest transforming the data in a suitable way before applying a CCA, if one has some prior knowledge of the dependence structure. We refer to this variant of CCA as TCCA in the following.

Model 5 is employed in this experiment. The parameter settings are $\theta = 2, 4, 10$ and $(p, q, n) = (100, 200, 1000)$. For the TCCA method, we use the exponential function $f(x) = e^x$ to transform each coordinate of the sample vectors, and then conduct the CCA procedure. In this way, the two sets of transformed data are linearly correlated.

Histograms of the bulk eigenvalues and the largest eigenvalue are plotted in Figure 5. In the left panel, the eigenvalues are from the DCM $\mathbf{S}_{xz}$. Clearly, the empirical SD of the bulk eigenvalues (black strips) is perfectly predicted by its LSD density curve (red lines). Moreover, the largest eigenvalues (blue strips) are centered at $\lambda = 69.83, 187.5$, and 1041.5 (blue lines) for $\theta = 2, 4$ and 10, respectively, which are clearly separated from the bulks. Similar statistics from the CCA are shown in the middle panel. This demonstrates that the largest eigenvalues for $\theta = 2, 4, 10$ are all centered at $\lambda = 0.49$, which is smaller than the right edge point $\lambda_+ = 0.5$ of the LSD. The results from the TCCA are plotted in the right panel, where the largest eigenvalues are centered at $\lambda = 0.49, 0.50, 0.52$ for $\theta = 2, 4, 10$, respectively. On the other hand, Figure 6 reports the sequences of sample ratios $\{\lambda_{n,i+1}/\lambda_{n,i}\}$ with ± 2 standard deviations. For $\theta = 2, 4, 10$, the first ratio $\{\lambda_{n,2}/\lambda_{n,1}\}$ from the DCM model is well separated from the rest, while those from the CCA and TCCA models have no clear separation. Therefore, the nonlinear correlation between $\mathbf{x}$ and $\mathbf{z}$ is entirely captured by the DCM model, whereas the CCA and TCCA both fail to identify it efficiently. Note that the TCCA method has some potential for the detection because, on average, the largest eigenvalue from the TCCA surpasses the right edge limit 0.5 of the LSD as the parameter θ increases. However, its power is weak compared with that of our proposed method for the studied cases. Some other transforms are also tested under the same settings, such as polynomial functions, Box–Cox transforms, and trigonometric functions. Their performance is either comparable with, or less superior to that of the exponential function.

4.3. A consistent estimator for the order of finite-rank dependence

Assume that among the m dependence strengths $(\theta_k)_{1\le k\le m} := \boldsymbol{\theta}$, there are m_0 strengths above the critical value θ_0 given in (4.3). According to Corollary 1, the m_0 largest eigenvalues $\lambda_{n,k}$ of the DCM $\mathbf{S}_{xz}$ will converge almost surely to

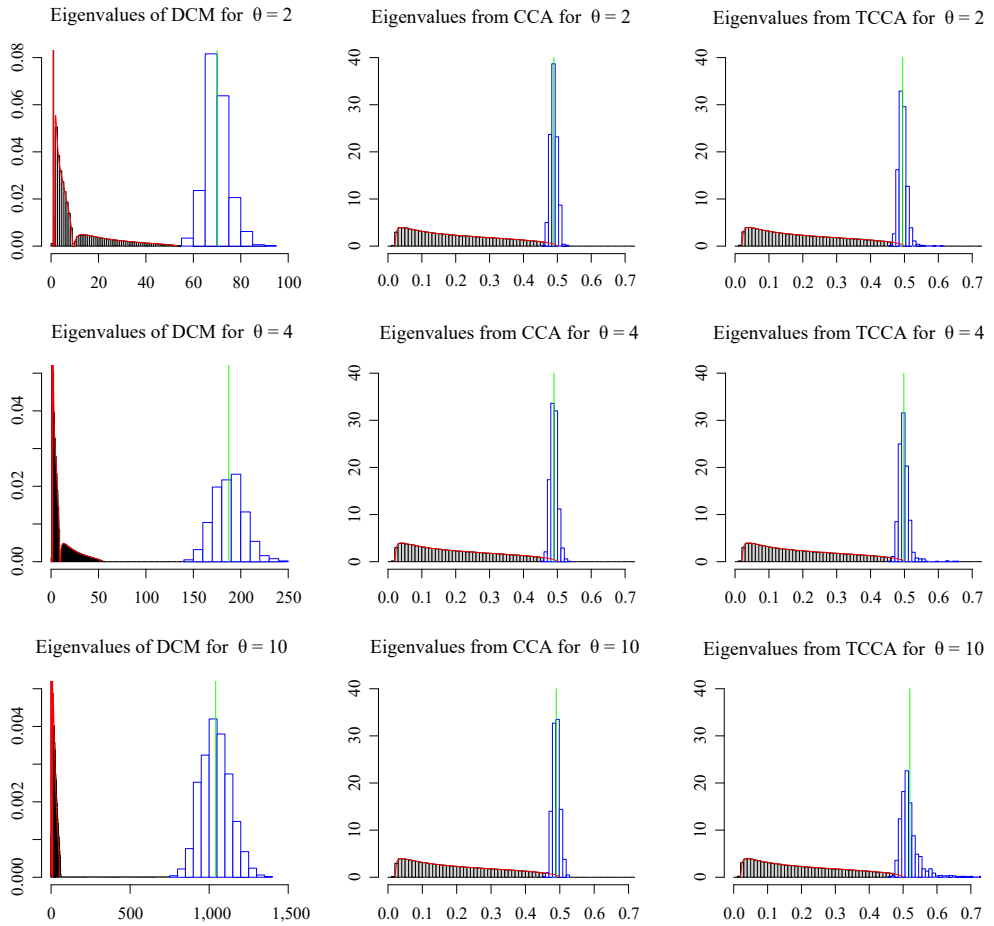

Figure 5. Histograms of bulk eigenvalues (black strips) and the largest eigenvalue (blue strips) from 1,000 independent replications under Model 5, with $\theta = 2, 4$, and 10. The red solid line curves are LSD densities, and the green vertical lines show the averages of the largest eigenvalues. The plots in the left panel are based on the DCM $\mathbf{S}_{xz}$, and those in the middle and left panels are based on the CCA and TCCA, respectively. The dimensions are $(p, q, n) = (100, 200, 1000)$.

m_0 limits λ_k, for $1 \le k \le m_0$, which are outside the support of F and given in (4.4). At the same time, the following eigenvalues of any given number, say s, $\lambda_{n,m_0+1}, \ldots, \lambda_{n,m_0+s}$ will all converge to the right edge λ_+ of the LSD F. The rank m_0 corresponds to the detectable rank of the weak dependence considered here. In a sense, the remaining $m - m_0$ dependence strengths $\{\theta_{m_0+1}, \ldots, \theta_m\}$ below the critical value θ_0 are too weak for detection. Following the popular ratio estimator for the number of factors or spikes developed in Onatski (2010) and

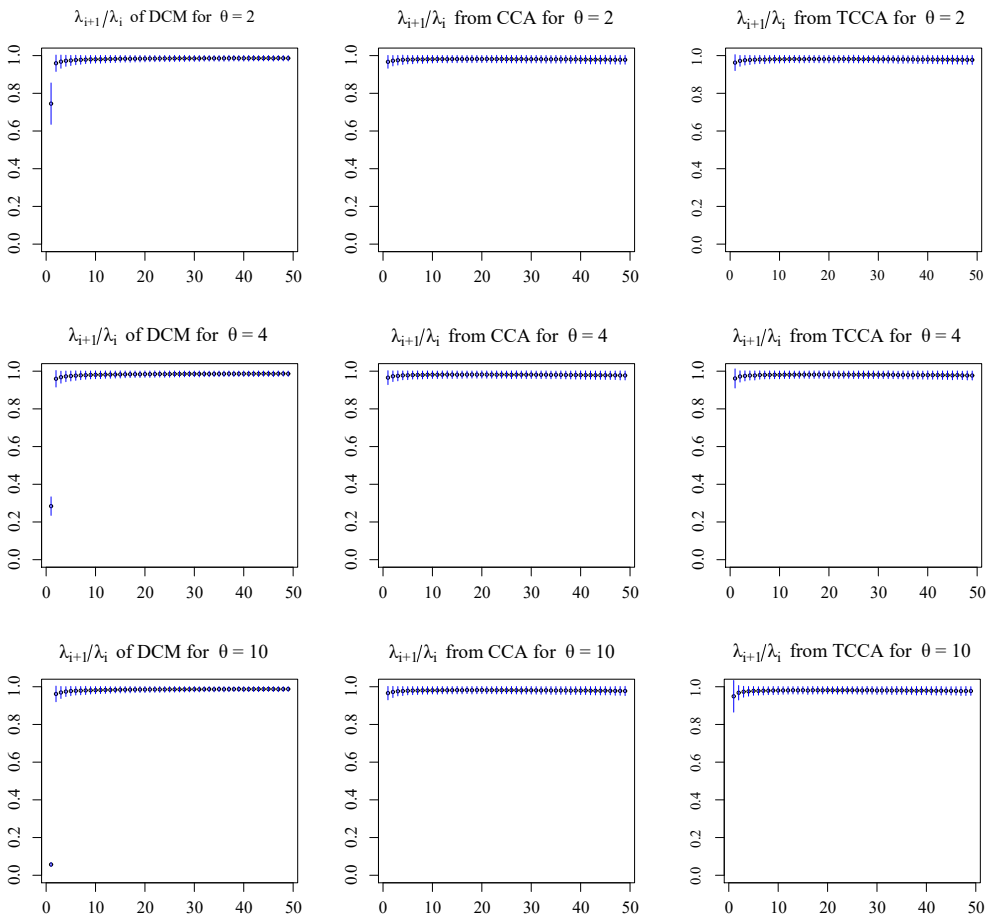

Figure 6. The average of the sequences of sample ratios $\{\lambda_{n,i+1}/\lambda_{n,i}\}$ from 1,000 independent replications under Model 5, with $\theta = 2, 4$. The plots in the left panel are based on the DCM $\mathbf{S}_{xz}$, and those in the middle and right panels are based on the CCA and TCCA, respectively. The dimensions are $(p, q, n) = (100, 200, 1000)$.

Li, Wang and Yao (2017), we introduce a consistent estimator for the detectable dependence rank m_0 in model (4.1), as follows. Note that for $j = 1, \ldots, m_0$, the ratios $\lambda_{n,j+1}/\lambda_{n,j}$ converge almost surely to a number in $(0, 1)$, whereas for $j \geq m_0 + 1$, these ratios converge to one. Let $0 < d_n < 1$ be a sequence of positive and vanishing constants, and consider the following estimator for the dependence rank m_0:

$$\hat{m}_0 = \left\{ \text{first } j \geq 1 \ \text{ such that } \ \frac{\lambda_{n,j+1}}{\lambda_{n,j}} > 1 - d_n \right\} - 1.$$

Table 3. Frequencies of $\hat{m}_0$ under Model 5 with $\boldsymbol{\theta} = (4, 3, 2)$ and $m_0 = 3$ from 1,000 independent replications. The dimensional settings are $c_{n1} = 0.1, c_{n2} = 0.2$ and n ranging from 100 to 1,600.

	$\hat{m}_0 = 0$	$\hat{m}_0 = 1$	$\hat{m}_0 = 2$	$\hat{m}_0 = 3$	$\hat{m}_0 = 4$
$n =$ 100	0.045	0.649	0.293	0.013	0
$n =$ 200	0	0.144	0.676	0.176	0.004
$n =$ 400	0	0.020	0.406	0.561	0.013
$n =$ 800	0	0	0.057	0.942	0.001
$n =$ 1,600	0	0	0	0.995	0.005

Table 4. Frequencies of $\hat{m}_0$ under Model 6 with $\boldsymbol{\theta} = (4, 3, 2, 1)$ and $m_0 = 2$ from 1,000 independent replications. The dimensional settings are $c_{n1} = 1, c_{n2} = 2$ and n ranging from 100 to 1,600.

	$\hat{m}_0 = 0$	$\hat{m}_0 = 1$	$\hat{m}_0 = 2$	$\hat{m}_0 = 3$
$n =$ 100	0.122	0.743	0.135	0
$n =$ 200	0.016	0.625	0.357	0.002
$n =$ 400	0	0.409	0.584	0.007
$n =$ 800	0	0.179	0.813	0.008
$n =$ 1,600	0	0.039	0.953	0.008

Under conditions similar to those of Theorem 3.1 in Li, Wang and Yao (2017), one can show that $\hat{m}_0$ converges to m_0 almost surely.

It remains to set up an appropriate value for the tuning parameter d_n. Theoretically, any vanishing sequence $d_n \to 0$ is sufficient for the consistency of $\hat{m}_0$. Here, we follow the calibration proposed in Li, Wang and Yao (2017). Specifically, we empirically find $q_{n,p,q,0.5\%}$, the lower 0.5% quantile of $n^{2/3}\big(\nu_2/\nu_1 - 1\big)$, where ν_1 and ν_2 are the top two sample eigenvalues of the DCM $\mathbf{S}_{xy}$ under the null model with $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{y} \sim N(\mathbf{0}, \mathbf{I}_q)$. Then, we set $d_n = n^{-2/3}|q_{n,p,q,0.5\%}|$. Note that d_n vanishes at the rate $n^{-2/3}$. This tuned value of d_n is used for all the simulation experiments in this section.

We now examine the performance of $\hat{m}_0$ in finite-sample situations. Models 5 and 6 are adopted again when generating samples of $\mathbf{x}$ and $\mathbf{y}$. Under Model 5, we take $m = 3$ and $\boldsymbol{\theta} = (4, 3, 2)$. The critical value θ_0 is 1.2, and thus the detectable dependence rank is $m_0 = 3$. Under Model 6, we take $m = 4$ and $\boldsymbol{\theta} = (4, 3, 2, 1)$. In this case $\theta_0 = 2.5$ and $m_0 = 2$. The frequencies of $\hat{m}_0$ are calculated from 1,000 independent replications under the two models, with the sample size n ranging from 100 to 1,600. The results are shown in Tables 3 and 4, which verify the convergence of the proposed estimator.

Supplementary Material

An online Supplementary Material contains additional technical tools used in this paper and proofs of Theorems 1, 2, 3, 4 and 5.

Acknowledgments

The authors are grateful to Prof. Xiaofeng Shao for important discussions from which this research originated. Weiming Li's research was partially supported by the NSFC (No. 11971293 and No. 12141107). Qinwen Wang acknowledges support from the NSFC Grants (No. 11801085 and No. 12171099).

References

Bai, Z. D. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. 2nd Edition. BSbook., Springer, New York.

Baik, J., Ben-Arous, G. and Péché, S. (2005). Phase transition of the largest eigenvalue for non-null complex sample covariance matrices. *Ann. Probab.* **33**, 1643–1697.

Baik, J. and Silverstein, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate. Anal.* **97**, 1382–1408.

Bao, Z. G., Hu, J., Pan, G. M. and Zhou, W. (2019). Canonical correlation coefficients of high-dimensional Gaussian vectors: Finite rank case. *Ann. Statist.* **47**, 612–640.

Johnstone, I. and Paul, D.(2018). PCA in high dimensions: An orientation. *P. IEEE* **106**, 1277–1292.

Li, Z., Wang, Q. W. and Yao, J. F. (2017). Identifying the number of factors from singular values of a large sample auto-covariance matrix. *Ann. Statist.* **45**, 257–288.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econ. Stat.* **92**, 1004–1016.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17**, 1617–1642.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769–2794.

Székely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *J. Multivariate. Anal.* **117**, 193–213.

Yang, Y. and Pan, G. (2015). Independence test for high dimensional data based on regularized canonical correlation coefficients. *Ann. Statist.* **43**, 467–500.

Zhu, C., Zhang, X., Yao, S. and Shao, X. (2020). Distance-based and RKHS-based dependence metrics in high-dimension. *Ann. Statist.* **48**, 3366–3394.

Weiming Li

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China.

E-mail: li.weiming@shufe.edu.cn

Qinwen Wang
School of Data Science, Fudan University, Shanghai 200433, China.
E-mail: wqw@fudan.edu.cn
Jianfeng Yao
School of Data Science, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China.
E-mail: jeffyao@cuhk.edu.cn

(Received August 2020; accepted May 2021)