ESTIMATION OF COVARIANCE MATRIX DISTANCES IN THE HIGH DIMENSION LOW SAMPLE SIZE REGIME

Malik Tiomoko

Université Paris Sud, Université ParisSaclay
Laboratoire des signaux et systèmes

Romain Couillet[*]

GIPSA-Lab, Université de Grenoble-Alpes
Centrale Supélec, Université ParisSaclay

ABSTRACT

A broad family of distances between two covariance matrices $C_1, C_2 \in \mathbb{R}^{p\times p}$, among which the Frobenhius, Fisher, Battacharrya distances as well as the Kullback-Leibler, Rényi and Wasserstein divergence for centered Gaussian distributions can be expressed as functionals $\frac{1}{p}\sum_{i=1}^{p} f\left(\lambda_i(C_1^{-1}C_2)\right)$ or $\frac{1}{p}\sum_{i=1}^{p} f\left(\lambda_i(C_1C_2)\right)$ of the eigenvalue distribution of $C_1^{-1}C_2$ or C_1C_2. Consistent estimates of such distances based on few (n_1, n_2) samples $x_i \in \mathbb{R}^p$ having covariance C_1, C_2 have been recently proposed using random matrix tools in the regime where $n_1, n_2 \sim p$. These estimates however demand that $n_1, n_2 > p$ for most functions f. The article proposes to alleviate this limitation using a polynomial approximation approach. The proposed method is supported by simulations in practical applications.

***Index Terms*—** Random Matrix Theory, Statistical Inference, Covariance Matrix

1. INTRODUCTION

Evaluating the distance between covariance matrices is at the core of many machine learning and signal processing applications. They are notably used for covariance features-based classification (for instance in brain signal or hyperspectral image classification), as well as for dimensionality reduction and representation of high dimensional points. Denote $C_1, C_2 \in \mathbb{R}^{p\times p}$ two large dimensional covariance matrices for which we would like to compute the distance $D(C_1, C_2)$ based on few (p-dimensional) sample vectors. We assume that D can be written as a linear functional $\frac{1}{p}\sum_{i=1}^{p} f\left(l_i\right)$ of the eigenvalue distribution of either $C_1^{-1}C_2$ ($l_i = \lambda_i(C_1^{-1}C_2)$) (as with the Fisher, Battacharrya, Kullback Leibler, Rényi divergences) or C_1C_2 ($l_i = \lambda_i(C_1C_2)$) (for the Wasserstein distance).

The natural estimation of $D(C_1, C_2)$ based on n_a samples x_i^a, $a = 1, 2$, is traditionally performed via $D(\hat{C}_1, \hat{C}_2)$, where $\hat{C}_a = \frac{1}{n_a}\sum_{i=1}^{n_a} x_i^{(a)} x_i^{(a)\mathsf{T}}$. This is justified by the law of large number since $D(\hat{C}_1, \hat{C}_2) \xrightarrow{\text{a.s.}} D(C_1, C_2)$ as $n_1, n_2 \gg p$. However, the estimate induces dramatic erroneous when $n_1, n_2 \sim p$ and even diverges for $n_a < p$.

To deal with the $n_1, n_2 \sim p$ scenario, [1] and [2] proposed a random matrix improved consistent estimate for $D(C_1, C_2)$ for all aforementioned metrics. The main idea behind these works consists in three steps: i) write the sought-for distance $D(C_1, C_2)$ as a complex integral involving the Stieltjes transform of the limiting eigenvalue distribution of $C_1^{-1}C_2$ (or C_1C_2), ii) exploit a functional identity relating the Stieltjes transforms of the limiting *population* and *sample* eigenvalue distributions (e.g., of $C_1^{-1}C_2$ and $\hat{C}_1^{-1}\hat{C}_2$) using the results from [3] and iii) find (if possible) a closed form solution for the resulting complex integral.

However, although consistent for $n_1, n_2 \sim p$, these improved estimators still demand that $n_1, n_2 > p$ for all functions $f(z)$ having a singularity at $z = 0$ (e.g., $1/z$, $\log(z)$, $\log^2(z)$, $\sqrt{z}$). For functions f having no singularity at zero, this constraint may be relaxed (with some extra care though) as discussed in [1], but almost no usual covariance matrix distance falls into this scenario (except for the Frobenius distance, which can at any rate be treated easily without resorting to these elaborate methods).

Based on a polynomial approximation of the functions of interest, this article proposes to retrieve consistent estimates for the challenging $n_1, n_2 < p$ scenario. Besides, a closed-form and numerically convenient expression of the proposed estimator is derived for any (local) polynomial approximation of the functions.

The article is organized as follows. In section 2, we will explain briefly the general idea of the estimation and the problem induced by the challenging case $n_1 < p$ and $n_2 < p$. A solution based on polynomial approximation is then proposed in section 3 to cover the case $n_2 < p$. As an application, a dimensionality reduction framework shows the consistency of the proposed estimator in section 4.

Reproducibility. A set of Matlab codes for the various estimators introduced and studied in this article are available at https://github.com/maliktiomoko/RMTCovDistHDLSS.

2. PRELIMINARIES

2.1. Models and Assumption

For $a \in \{1, 2\}$, let $X_a = [x_1^{(a)}, \ldots, x_{n_a}^{(a)}]$ be n_a independent and identically distributed random vectors with $x_i^{(a)} = C_a^{\frac{1}{2}}\tilde{x}_i^{(a)}$, where $\tilde{x}_i^{(a)} \in \mathbb{R}^p$ has zero mean, unit variance and finite fourth order moment entries. This holds in particular for $x_i^{(a)} \sim \mathcal{N}(0, C_a)$. In order to control the growth rates of n_1, n_2, p, we make the following assumption:

Assumption 1 (Growth Rates). *As $n_a \to \infty$, $p/n_a \to c_a \in (0, \infty)$ and $\limsup_p \max\{\|C_a^{-1}\|, \|C_a\|\} < \infty$ for $\|\cdot\|$ the operator norm.*

We define the sample covariance estimate $\hat{C}_a$ of C_a as

$$\hat{C}_a \equiv \frac{1}{n_a} X_a X_a^{\mathsf{T}} = \frac{1}{n_a}\sum_{i=1}^{n_a} x_i^{(a)} x_i^{(a)\mathsf{T}}.$$

Our objective is to estimate, under the difficult regime where $p > n_a$, the distance $D(C_1, C_2)$ between the covariance matrices

[*]Couillet's work is supported by the ANR Project RMT4GRAPH (ANR-14-CE28-0006) and the IDEX GSTATS Chair at University Grenoble Alpes.

C_1 and C_2 of the form:

$$D(C_1, C_2) = \frac{1}{p} \sum_{i=1}^{p} f(l_i)$$

where $l_i = l_i^-$ are the eigenvalues of $C_1^{-1} C_2$ or $l_i = l_i^+$ are the eigenvalues of $C_1 C_2$. As discussed at length in [1], this form comprises, among others, the Fisher, Frobenius, Wasserstein and Battacharrya distances, along with the Kullbach-Liebler and Rényi divergences.

The proposed estimate relies on random matrix theory and particularly on the Stieltjes transform $m_\theta(z)$ of a probability distribution θ defined as the complex function:

$$m_\theta : \mathbb{C} \setminus \mathrm{supp}(\theta) \to \mathbb{C}, \quad z \mapsto \int (\lambda - z)^{-1} d\theta(\lambda).$$

The Stieltjes transform is here used to create a link between the population covariance eigenvalue distribution ν_p to the sample eigenvalue distribution μ_p defined by:

$$\nu_p = \frac{1}{p} \sum_{i=1}^{p} \delta_{l_i}, \quad \mu_p = \frac{1}{p} \sum_{i=1}^{p} \delta_{\hat{l}_i}$$

where $\hat{l}_i = \hat{l}_i^-$ are the eigenvalues of $\hat{C}_1^{-1} \hat{C}_2$ or $\hat{l}_i = \hat{l}_i^+$ are the eigenvalues of $\hat{C}_1 \hat{C}_2$.

2.2. Previous Results

Theorem 1 from [1] and [4] provide an estimate of these distances in the "easy" regime where $\lim p/n_a < 1$.

Theorem 1. *Let $f : \mathbb{C} \to \mathbb{C}$ be analytic on a contour $\Gamma \subset \{z \in \mathbb{C}, \mathcal{R}[z] > 0\}$ surrounding μ_p. Then,*

$$\int f d\nu_p - \frac{1}{2\pi i} \oint_\Gamma f\left(\frac{\varphi_p(z)}{\psi_p(z)}\right) \left[\frac{\psi_p'(z)}{\psi_p(z)} - \frac{\varphi_p'(z)}{\varphi_p(z)}\right] \frac{\psi_p(z) dz}{c_2} \xrightarrow{\text{a.s.}} 0$$

where

$$\varphi_p(z) = \begin{cases} z(1 + c_1 z m_{\mu_p}(z)), & \mu_p = \mu_p^- \\ \frac{z}{1 - c_1 - c_1 z m_{\mu_p}(z)}, & \mu_p = \mu_p^+ \end{cases}$$

$$\psi_p(z) = 1 - c_2 - c_2 z m_{\mu_p}(z).$$

To understand the generalization of Theorem 1 proposed in the present article, we need to recall the main steps of its proof (here for $\mu_p = \mu_p^-$).

Sketch of Proof. Using the Cauchy integral formula, we have

$$D(C_1, C_2) = \frac{1}{p} \sum_{i=1}^{p} f(l_i) = \int f(t) \nu_p(dt) \tag{1}$$

$$= \frac{1}{2\pi \imath} \int \left[\oint_{\Gamma_\nu} \frac{f(z)}{z - t} \right] \nu_p(dt) = \frac{-1}{2\pi \imath} \oint_{\Gamma_\nu} f(z) m_{\nu_p}(z) dz.$$

Thus, estimating $D(C_1, C_2)$ is equivalent to relating m_{ν_p} to m_{μ_p}. Since X_1 and X_2 are independent, we can condition first on X_2. By [3], the limiting eigenvalue distribution of $C_2 \hat{C}_1^{-1}$, denoted ζ, can be written as a function $C_2 C_1^{-1}$, and similarly for the limiting eigenvalue distributions of $\hat{C}_2 \hat{C}_1^{-1}$ and $C_2 \hat{C}_1^{-1}$. This entails the two equations:

$$z m_{\mu_p}(z) = \varphi_p(z) m_{\zeta_p}(\varphi_p(z)) + o_p(1) \tag{2}$$

$$m_{\nu_p}(z/\Psi_p(z)) = m_{\zeta_p}(z) \Psi_p(z) + o_p(1). \tag{3}$$

Through the changes of variable $z \to \varphi_p(z)$ and $\omega \to \Psi_p(\omega)$ applied in $\oint_{\Gamma_\nu} f(z) m_{\nu_p}(z) dz$, the result follows. □

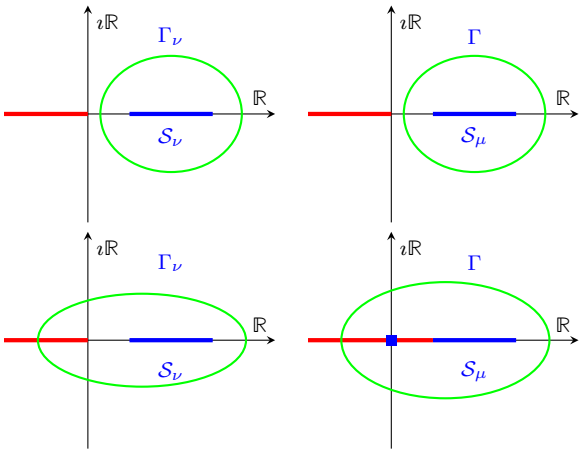

Fig. 1. Illustration of the contours maps $\Gamma \mapsto \Gamma_\nu$ (from right to left) by the variable changes leading up to Theorem 1. (Top) $n_2 > p$. (Bottom) $n_2 < p$. For $n_2 < p$, the left real crossing of Γ_ν is necessarily negative (even if the mass at $\{0\}$ of $\mathcal{S}_\mu$ were not included in Γ). In case of singularities or branch-cuts (shown in red for the $\log(z)$ and $\sqrt{z}$ functions), the contours are invalid.

2.3. On the need for $n_2 > p$ and $n_1 > p$ in Theorem 1

Since distances involving the eigenvalues of $C_1^{-1} C_2$ are estimated from the empirical matrix $\hat{C}_1^{-1} \hat{C}_2$, the constraint $n_1 > p$ is inevitable to ensure the existence of $\hat{C}_1^{-1}$. The requirement for $n_2 > p$ is less immediate. The two variable changes discussed in the proof of Theorem 1 are only licit if they realize a mapping from a contour Γ enclosing the limiting support $\mathcal{S}_\mu$ of μ_p and a valid contour Γ_ν enclosing the limiting support $\mathcal{S}_\nu$ of ν_p *while* enclosing no additional singular points of the function $f(z)$ (otherwise the Cauchy formula used in (1) is incorrect). But for $n_2 < p$, it can be proved (see details in [1]) that the pre-image of Γ_μ by the variable changes wraps around $\mathcal{S}_\nu$ *and* around zero (with leftmost real crossing depending on the ratio n_2/p). This is a problem for all functions $f(z)$ singular at $z = 0$. In particular, $1/z$, $\log(z)$, $\log^2(z)$ and $\sqrt{z}$ are examples of such invalid functions which, for some, additionally have a branch-cut terminating at zero (that no valid contour may cross). This discussion is most conveniently illustrated in Figure 1.

Unfortunately, there seems to be no simple workaround to this situation. We propose in this article to (partially) solve the problem by introducing *entire* functions (thus analytical over $\mathbb{C}$) as substitutes for the locally non-analytic functions $\log(z)$, $\log^2(z)$ and $\sqrt{z}$ intervening in the distance $D(C_1, C_2)$ formulation.

3. MAIN RESULTS

3.1. Proposed estimation

Our approach consists in approximating (arbitrarily closely) the analytic functions f under study that present singularities around zero by entire functions, and particularly degree-N polynomials $\tilde{f}_N(z)$ defined by $\tilde{f}_N(z) = \sum_{n=0}^{N} a_n z^n$.

Our central argument relies in the fact that, since $\|C_a\|$ and $\|C_a^{-1}\|$ are bounded (as per Assumption 1), the limiting support $\mathcal{S}_\nu$ of ν_p is a compact set *strictly* away from zero. As such, one needs not approximate f on the whole $\mathbb{R}^+$ half-line (which would still pose problems in the vicinity of zero) but only on a subset $[a, b] \subset (0, \infty)$

over which polynomials are universal approximators.

This gives rise to the following immediate extension of Theorem 1.

Theorem 2. *Let Γ be a contour surrounding the limiting support $\mathcal{S}_\mu$ of μ_p. Then, for all $\epsilon > 0$, there exists N and $\tilde{f}_N(z) = \sum_{n=0}^{N} a_n z^n$ a polynomial of order N, such that, under the notations of Theorem 1,*

$$\limsup_{n,p} |D(C_1, C_2; f) - \hat{D}(X_1, X_2; \tilde{f}_N)| < \epsilon$$

almost surely, where

$$\hat{D}(X_1, X_2; \tilde{f}_N) = \frac{n_2}{2\pi i p} \oint_\Gamma \tilde{f}_N \left(\frac{\varphi_p(z)}{\psi_p(z)}\right) \left[\frac{\varphi_p'(z)}{\varphi_p(z)} - \frac{\psi_p'(z)}{\psi_p(z)}\right] \psi_p(z) dz - a_0 \frac{1-c_2}{c_2}.$$

The result immediately follows from our discussion above and Theorem 1. To avoid the numerical integration and proper choice of a contour in the statement of Theorem 2, we next provide closed-form expressions for the estimate $\hat{D}(X_1, X_2; \tilde{f}_N)$. For simplicity of exposition, we consider here only the case $n_2 < n_1$.

Theorem 3 (Case $C_1 C_2$)**.** *Let $0 < \lambda_1 \leq \ldots \leq \lambda_{n_2}$ be the n_2 non zero eigenvalues of $\hat{C}_1 \hat{C}_2$, and define $\{\xi_i\}_{i=1}^{n_2-1}$ and $\{\eta_i\}_{i=1}^{n_2}$ the (increasing non zero) eigenvalues of $\Lambda - \frac{1}{n_1}\sqrt{\lambda}\sqrt{\lambda}^\mathsf{T}$ and $\Lambda - \frac{1}{n_2}\sqrt{\lambda}\sqrt{\lambda}^\mathsf{T}$, respectively, where $\lambda = (\lambda_1, \ldots, \lambda_{n_2})^\mathsf{T}$ and $\Lambda = \mathrm{diag}(\lambda)$. Then,*

$$\begin{aligned}
&\hat{D}(X_1, X_2; \tilde{f}_N) = \\
&a_0 + \frac{a_n}{c_2} \sum_{j=1}^{n_2-1} \xi_j \left(1 + \frac{c_1}{c_2} - c_1\right) \\
&+ \sum_{n=2}^{N} \left[\frac{a_n}{c_2} \sum_{j=1}^{n_2-1} \frac{1}{(n-2)!} \frac{\partial^{n-2}}{\partial z^{n-2}} \left(A_n^{\xi_j}(z) C_n^{\xi}(z)\right)|_{z=\zeta_j}\right] \\
&+ \sum_{n=1}^{N} \left[\frac{a_n}{c_2} \sum_{j=1}^{n_2} \frac{1}{(n-1)!} \frac{\partial^{n-1}}{\partial z^{n-1}} \left(A_n^{\eta_j}(z) C_n^{\eta}(z)\right)|_{z=\eta_j}\right]
\end{aligned}$$

where

$$\begin{aligned}
A_n^{\xi_j}(z) &= \left(\frac{\varphi(z)}{\psi(z)}(z - \xi_j)\right)^n \frac{\psi(z)}{z - \xi_j} \\
A_n^{\eta_j}(z) &= \left(\frac{\varphi(z)}{\psi(z)}(z - \eta_j)\right)^n \psi(z) \\
C_n^{\xi}(z) &= \frac{-1}{n-1} \frac{\varphi'(z)}{\varphi(z)}, \quad C_n^{\eta}(z) = \frac{-1}{n} \frac{\psi'(z)}{\psi(z)}.
\end{aligned}$$

Theorem 4 (Case $C_1^{-1} C_2$)**.** *Let $0 < \lambda_1 \leq \ldots \leq \lambda_{n_2}$ the n_2 non zero eigenvalues of $\hat{C}_1^{-1} \hat{C}_2$, and define $\{\xi_i\}_{i=1}^{n_2-1}$ and $\{\eta_i\}_{i=1}^{n_2}$ the (increasing non zero) eigenvalues of $\Lambda - \frac{1}{p-n_1}\sqrt{\lambda}\sqrt{\lambda}^\mathsf{T}$ and $\Lambda - \frac{1}{n_2}\sqrt{\lambda}\sqrt{\lambda}^\mathsf{T}$, respectively, where $\lambda = (\lambda_1, \ldots, \lambda_{n_2})^\mathsf{T}$ and $\Lambda = \mathrm{diag}(\lambda)$. Then,*

$$\begin{aligned}
&\hat{D}(X_1, X_2; \tilde{f}_N) = \\
&a_0 + \frac{a_n}{c_2} \sum_{j=1}^{n_2-1} \xi_j \left(1 + \frac{c_1}{c_2} - c_1\right) \\
&+ \sum_{n=2}^{N} \left[\frac{a_n}{c_2} \sum_{j=1}^{n_2-1} \frac{1}{(n-2)!} \frac{\partial^{n-2}}{\partial z^{n-2}} \left(A_n^{\xi_j}(z) C_n(z)\right)|_{z=\xi_j} \right. \\
&\left. + \sum_{j=1}^{n_2} \frac{a_n}{p} (1 - n_1 + n_2 - p) \left(-\frac{c_1}{c_2}\lambda_j\right)^n\right]
\end{aligned}$$

where

$$\begin{aligned}
A_n^{\xi_j}(z) &= \left(\tfrac{\varphi(z)}{\psi(z)}(z - \xi_j)\right)^n \tfrac{\psi(z)}{z - \xi_j} \\
C_n(z) &= \tfrac{-1}{n-1} \left[\tfrac{\varphi'(z)}{\varphi(z)} - \tfrac{1}{z}\right]
\end{aligned}$$

Proof. The proofs of Theorems 3 and 4 are similar. For conciseness we only sketch the proof of the Theorem 4. The ξ_i and η_i are the respective zeros of the rational functions $1 - \frac{p}{n_2} - \frac{p}{n_2} z m(z)$ and $1 + \frac{p}{n_1} z m(z)$. Thus, φ_p and ψ_p can be expressed as the following rational functions:

$$\varphi_p(z) = (1 - c_1)\, z \frac{\prod_{i=1}^{n_2} z - \eta_i}{\prod_{i=1}^{n_2} z - \lambda_i}, \quad \psi_p(z) = \frac{z \prod_{i=1}^{n_2-1} z - \xi_i}{\prod_{i=1}^{n_2} z - \lambda_i}$$

and their derivatives as

$$\frac{\varphi_p'(x)}{\varphi_p(x)} - \frac{\psi_p'(x)}{\psi_p(x)} = \sum_{i=1}^{n_2} \frac{1}{x - \eta_i} - \sum_{i=1}^{n_2-1} \frac{1}{x - \xi_i}.$$

The complex integral can thus be written as $\sum_{n=1}^{N} \frac{a_n}{2\pi\imath c_2} \oint I_n(z) dz$ with the integrand $I_n(z)$ given by the rational function:

$$\begin{aligned}
I_n(z) = (1 - c_1)^n &\frac{\prod_{i=1}^{n_2} (z - \eta_i)^n}{\prod_{i=1}^{n_2-1} (z - \xi_i)^{n-1} \prod_{i=1}^{n_2} (z - \lambda_i)^n} \\
&\times \left[\sum_{i=1}^{n_2} \frac{1}{z - \eta_i} - \sum_{i=1}^{n_2-1} \frac{1}{z - \xi_i}\right]
\end{aligned}$$

for which we need to evaluate the residue at poles:

- the $(n-1)$-th and n-th order pole at ξ_j with residue (for $n \geq 2$) by:

$$\begin{aligned}
R_1 = &\sum_{j=1}^{n_2-1} \lim_{z \to \xi_j} \frac{1}{(n-2)!} \frac{\partial^{n-2}}{\partial x^{n-2}} \left[I_n(z)(z - \xi_j)^{n-1}\right] \\
&+ \sum_{j=1}^{n_2-1} \lim_{z \to \xi_j} \frac{1}{(n-1)!} \frac{\partial^{n-1}}{\partial x^{n-1}} \left[I_n(z)(z - \xi_j)^n\right]
\end{aligned}$$

- the 1st order pole in λ_j with residue:

$$R_3 = \sum_{j=1}^{n_2} \lim_{z \to \lambda_j} I_n(z)(z - \lambda_j)$$

- the 1st order pole in η_j with residue (for $n = 0$):

$$R_4 = \sum_{j=1}^{n_2} \lim_{z \to \eta_j} \psi(z).$$

Putting all residues together and exploiting further relations involving the functions φ_p and ψ_p and their derivatives (see [1] for details) entails the result. □

Although Theorems 3 and 4 are difficult to interpret, they are numerically easy to implement: the terms involving the k-th derivative can indeed be evaluated iteratively by remarking the following identities for any rational functional function $A(z) = \frac{\prod_i z-a_i}{\prod_i z-b_i}$ and $C(z) = \sum_i \frac{1}{z-c_i} - \frac{1}{z-d_i}$ ($a_i, b_i, c_i, d_i \in \mathbb{R}$):

$$\frac{\partial^n}{\partial z^n}(A(z)C(z)) = \sum_{k=1}^{n} \binom{k}{n} \frac{\partial^k}{\partial z^k} A(z) \frac{\partial^{n-k}}{\partial z^{n-k}} C(z)$$

$$\frac{\partial^n}{\partial z^n} A(z) = \frac{\partial^{n-1}}{\partial z^{n-1}} A(z)B(z)$$

$$= \sum_{k=1}^{n-1} \binom{k}{n-1} \frac{\partial^k}{\partial x^k} A(z) \frac{\partial^{n-1-k}}{\partial x^{n-1-k}} B(z)$$

where $B(z) = \sum_i \frac{1}{z-a_i} - \frac{1}{z-b_i}$, from which we obtain a recursive expression for the derivatives of $A(z)$.

4. SIMULATIONS AND APPLICATIONS

4.1. Applications to synthetic Gaussian data

We confirm the consistency of the proposed estimates through simulations on synthetic Gaussian data, here for the Wasserstein distance (that is, for $f(z) = \sqrt{z}$ and $\mu_p = \mu_p^+$) and for the $p = 2n_2 = 1.5n_1$. Table 1 compares the proposed estimator to the traditional sample covariance matrix plugin estimate $D_{\mathrm{SCM}} = D(\hat{C}_1, \hat{C}_2, f)$. It is clear in this setting that, while D_{SCM} to be largely erroneous, the proposed estimate performs well down to rather small values of p. Somewhat surprisingly, polynomials of high orders, which threaten the stability of the method (due to the not-so-low probability of occurrence of large eigenvalues for $\hat{C}_1\hat{C}_2$), display a stable behavior even for small values of p. This suggests the possibility, in practice, to run the proposed estimator for rather large values of N.

p	$D(f)$	$D(\tilde{f}_{10})$	$\hat{D}(\tilde{f}_{10})$	$D(\tilde{f}_{20})$	$\hat{D}(\tilde{f}_{20})$	D_{SCM}
64	0.1325	0.1379	0.1478	**0.1328**	0.1307	0.8477
128	0.1364	0.1437	0.1388	0.1382	0.1432	0.8367
256	0.1352	0.1425	0.1464	**0.1364**	0.1377	0.8446
512	0.1342	0.1424	0.1453	**0.1348**	0.1375	0.8412
1024	0.1345	0.1427	0.1431	**0.1352**	**0.1356**	0.8403
2048	0.1347	0.1430	0.1434	**0.1353**	**0.1361**	0.8405

Table 1. Estimates of Wasserstein distance $D(\cdot) = D_{\mathrm{W}}(C_1, C_2, \cdot)$ versus proposed estimate $\hat{D}(\cdot) = \hat{D}_{\mathrm{W}}(X_1, X_2, \cdot)$ and sample covariance (SCM) plugin estimate $D_{\mathrm{SCM}} = D_{\mathrm{W}}(\hat{C}_1, \hat{C}_2, f)$; with $C_1 = I_p$, C_2 a random (p-dimensional) standard Wishart matrix with $2p$ degrees of freedom, $x_i^{(a)} \sim \mathcal{N}(0, C_a)$; $c_1 = 1.5$ and $c_2 = 2$. Averaged over 10 trials. The function $\tilde{f}_N$ is the polynomial that best fits (in the least-squares sense) $\sqrt{D}$ for D sampled logarithmically on the support of the eigenvalues of $\hat{C}_1\hat{C}_2$. In **boldface**, distance estimates within 1% of relative error.

4.2. Application to dimensionality reduction

More interestingly in a statistical learning context, we now demonstrate the performance of the proposed estimator in a dimensionality reduction scenario. Consider m data points $X_1, \ldots, X_m$, each $X_i \in \mathbb{R}^{p \times n_i}$ being obtained from n_i independent p-dimensional centered Gaussian samples. For $i \leq m/2$, $\mathrm{E}[[X_i]_{\cdot j}[X_i]_{\cdot j}^\top] = C^{(i)} = C_1 = I_p$ and for $i > m/2$, $C^{(i)} = C_2$ a random (p-dimensional) standard Wishart matrix with $100p$ degrees of freedom (to avoid trivial classification).

The objective is to depict the data in a two-dimensional projection space in order to identify the two classes based on their covariance matrices. To this end, we use kernel PCA [5] for a Wasserstein distance metric, which consists in a (2D) principal component analysis of the kernel matrix $K \in \mathbb{R}^{m \times m}$ with $K_{ij} = \exp(-\frac{1}{2} D_W(C^{(i)}, C^{(j)}))$. Being unknown, the distances are approximated using either the classical estimate $D_W(\hat{C}^{(i)}, \hat{C}^{(j)})$ or the proposed estimator of the Wasserstein distance.

Figure 2 shows that the proposed estimator preserves the local neighborhood structure of the data where the classical estimate dramatically fails.

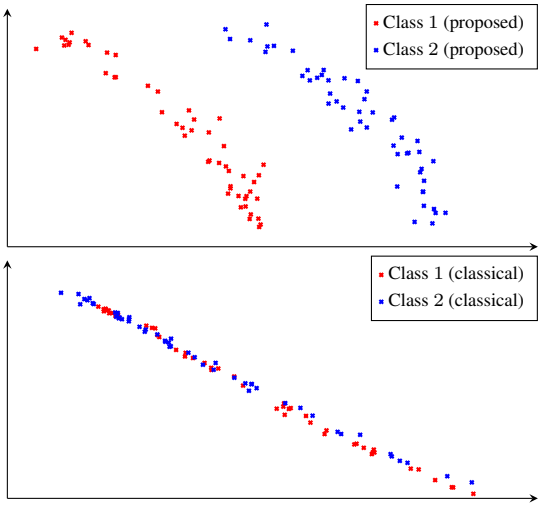

Fig. 2. First versus second eigenvectors of K for distinct uniformly random $n_i \in \{p/2, \ldots, p\}$, $p = 512$ and polynomial $\tilde{f}_N$ chosen similarly as in Table 1 with $N = 6$.

5. CONCLUDING REMARKS

The polynomial approximation approach proposed in this article breaks the stringent limitation that demands $p < n$ for a host of estimators, here for covariance distance estimation. Yet, it presents some technical difficulties when dealing with functions with high variations near zero (such as $\log^2(x)$). Combining polynomials $P(x)$ and other family of better behaving functions (e.g., $exp(-P(x))$) may leverage the problem.

Moreover, there is still a need to cover the case $p > n$ in estimators involving covariance matrix inverses for which $\hat{C}_1^{-1}\hat{C}_2$ is not even defined. Random projections and regularization methods can be devised to tackle this scenario, however possibly to the detriment of the estimator consistency.

6. REFERENCES

[1] Romain Couillet, Malik Tiomoko, Steeve Zozor, and Eric Moisan, "Random matrix-improved estimation of covariance matrix distances," *arXiv preprint arXiv:1810.04534*, 2018.

[2] Malik Tiomoko, Florent Bouchard, Guillaume Ginholac, and Romain Couillet, "Random matrix improved covariance estimation for a large class of metrics," *arXiv preprint arXiv:1902.02554*, 2019.

[3] J. W. Silverstein and Z. D. Bai, "On the empirical distribution of eigenvalues of a class of large dimensional random matrices," *Journal of Multivariate Analysis*, vol. 54, no. 2, pp. 175–192, 1995.

[4] Malik Tiomoko and Romain Couillet, "Random matrix-improved estimation of the wasserstein distance between two centered gaussian distributions," *arXiv preprint arXiv:1903.03447*, 2019.

[5] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller, "Kernel principal component analysis," in *International conference on artificial neural networks*. Springer, 1997, pp. 583–588.