Random Matrix-Improved Estimation of the Wasserstein Distance between two Centered Gaussian Distributions

Malik Tiomoko[1], Romain Couillet[1,2,*]
[1]CentraleSupélec, Université ParisSaclay, [2]GIPSA-lab, Université Grenoble-Alpes

***Abstract*— This article proposes a method to consistently estimate functionals $\frac{1}{p}\sum_{i=1}^{p} f(\lambda_i(C_1C_2))$ of the eigenvalues of the product of two covariance matrices $C_1, C_2 \in \mathbb{R}^{p\times p}$ based on the empirical estimates $\lambda_i(\hat{C}_1\hat{C}_2)$ ($\hat{C}_a = \frac{1}{n_a}\sum_{i=1}^{n_a} x_i^{(a)} x_i^{(a)\mathsf{T}}$), when the size p and number n_a of the (zero mean) samples $x_i^{(a)}$ are similar. As a corollary, a consistent estimate of the Wasserstein distance (related to the case $f(t) = \sqrt{t}$) between centered Gaussian distributions is derived.**

The new estimate is shown to largely outperform the classical sample covariance-based "plug-in" estimator. Based on this finding, a practical application to covariance estimation is then devised which demonstrates potentially significant performance gains with respect to state-of-the-art alternatives.

I. Introduction

Many machine learning and signal processing applications require an adequate framework to compare statistical objects, starting with probability distributions. The Wasserstein distance, initially inspired by Monge [1] and later by Kantorovich [2] in a transport theory analogy, provides a natural notion of dissimilarity for probability measures and finds a wide spectrum of applications in image analysis [3], shape matching [4], computer vision [5], etc.

However, computing the Wasserstein distance is expensive as it requires to minimize a cost function taking the form of an integral over the space of probability measures. Despite recent advances [6], where regularized approximations that reduce this numerical cost are proposed, the latter is still involved in general. Special cases exist for which the Wasserstein distance assumes a closed form, particularly when the underlying distributions are zero-mean Gaussian with covariance matrices C_1 and C_2. The closed-form formula however involves the eigenvalues of C_1C_2 and thus depends on the unknown population covariance matrices C_1 and C_2. Assuming the observation of $n_1, n_2 \gg p$ samples with covariances C_1, C_2, respectively, C_1C_2 is conventionally approximated by its empirical version $\hat{C}_1\hat{C}_2$. As we will show, this induces a dramatic estimation bias in practical applications where p is rather large or, equivalently, n_1, n_2 rather small, a standard assumption in big data applications.

Based on recent advances in random matrix theory, this article proposes a new consistent estimate for the Wasserstein distance between two centered Gaussian distributions when the dimension p of the samples is of the same order of magnitude as their numbers n_1, n_2. This work enters the scope of Mestre's seminal ideas [7] on the estimation of functionals $\frac{1}{p}\sum_{i=1}^{p} f(\lambda_i(C))$ of the eigenvalue distribution of population covariance matrices C, which can be related to the (limiting) eigenvalue distribution of the sample estimates $\hat{C}$ via a complex integration trick. We recently extended this work to the estimation of functionals of the eigenvalue distribution of F-matrices in [8], i.e., matrices of the form $C_1^{-1}C_2$, and applied to the estimation of the natural geodesic Fisher distance, Battacharrya distance, and Rényi/Kullbach-Leibler divergences between Gaussian distributions.

Our main contribution is the extension of [7], [8] to functionals f of the eigenvalues of *products* C_1C_2 of population covariance matrices. The Wasserstein distance falls within this scope for $f(t) = \sqrt{t}$. Unlike [8], where the functionals of interest ($f(t) = t, \log(t), \log^2(t)$) are amenable to explicit evaluations of the complex integrals, the present $f(t) = \sqrt{t}$ scenario is more technically involved and gives rise to real non-explicit, yet numerically computable, integrals.

In the remainder of the article, Section II introduces the main model and assumptions, Section III provides our key technical result and its corollary to the Wasserstein distance estimation, and a practical application to covariance matrix estimation is finally proposed in Section IV.

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

II. Model and Main Objective

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 $n_a \to \infty$, $p/n_a \to c_a \in (0, 1)$ 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

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

The Wasserstein distance $D_W(C_1, C_2)$ between two zero-mean Gaussian distributions with covariances C_1 and C_2, respectively, assumes the form [9, Remark 2.31]:

$$D_W(C_1, C_2) = \mathrm{tr}(C_1) + \mathrm{tr}(C_2) - 2\mathrm{tr}\left[(C_1^{\frac{1}{2}} C_2 C_1^{\frac{1}{2}})^{\frac{1}{2}}\right]. \quad (1)$$

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

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

It is easily shown that, under Assumption 1,

$$\frac{1}{p}\mathrm{tr}\hat{C}_a - \frac{1}{p}\mathrm{tr}C_a \to 0$$

almost surely. But estimating $\mathrm{tr}\left[(C_1^{\frac{1}{2}}C_2C_1^{\frac{1}{2}})^{\frac{1}{2}}\right]$ is more involved: this is the focus of the article. Up to a normalization by p, this term can be written under the functional form:

$$\frac{1}{p}\mathrm{tr}\left[(C_1^{\frac{1}{2}}C_2C_1^{\frac{1}{2}})^{\frac{1}{2}}\right] = \frac{1}{p}\sum_{i=1}^{n}\sqrt{\lambda_i(C_1C_2)} \equiv D(C_1, C_2; \sqrt{\cdot}) \tag{2}$$

with $\lambda_i(X)$ the i-th smallest eigenvalue of X.

Our objective is to estimate the more generic form

$$D(C_1, C_2; f) \equiv \frac{1}{p}\sum_{i=1}^{n} f(\lambda_i(C_1C_2)) \tag{3}$$

for $f : \mathbb{R} \to \mathbb{R}$ a real function admitting a complex-analytic extension. To this end, we shall relate the eigenvalues $\lambda_i(C_1C_2)$ to $\lambda_i(\hat{C}_1\hat{C}_2)$ through the *Stieltjes transform* ($m_\theta(z) \equiv \int \frac{d\theta(\lambda)}{\lambda - z}$ for measure θ and $z \in \mathbb{C}$) of their associated normalized counting measures

$$\mu_p = \frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i(\hat{C}_1\hat{C}_2)}, \quad \nu_p = \frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i(C_1C_2)}.$$

In particular, $m_{\mu_p}(z) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i - z}$ for $\lambda_i = \lambda_i(\hat{C}_1\hat{C}_2)$.

With these notations, we are in position to introduce our main results.

III. Main results

The following theorem provides a consistent estimate for the metric $D(C_1, C_2; f)$ defined in (3).

Theorem 1. *Let $\Gamma \subset \{z \in \mathbb{C}, \mathrm{real}[z] > 0\}$ be a contour surrounding $\cup_{p=1}^{\infty}\mathrm{supp}(\mu_p)$. Then, under Assumption 1,*

$$D(C_1, C_2; f) - \hat{D}(X_1, X_2; f) \xrightarrow{\text{a.s.}} 0$$

where

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

and, recalling $m_{\mu_p}(z) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i - z}$ for $\lambda_i = \lambda_i(\hat{C}_1\hat{C}_2)$, $\varphi_p(z) = \frac{z}{1 - \frac{p}{n_1} - \frac{p}{n_1}zm_{\mu_p}(z)}$, $\psi_p(z) = 1 - \frac{p}{n_2} - \frac{p}{n_2}zm_{\mu_p}(z)$.

The result of Theorem 1 is very similar to [8, Theorem 1] established for functionals of the eigenvalues of $C_1^{-1}C_2$. The main difference lies in the expression of the function $\varphi_p(z)$.

Proof. The proof of Theorem 1 is based on the same approach as for [10, Theorem 1]. One first creates a link between the Stieltjes transform m_{ν_p} and $D(C_1, C_2; f)$ using Cauchy's integral formula:

$$\begin{aligned}\frac{1}{p}\sum_{i=1}^{p} f(\lambda_i(C_1C_2)) &= \int f(t)d\nu_p(t) \\ &= \frac{1}{2\pi i}\int\left[\oint_{\Gamma_\nu}\frac{f(z)}{z-t}dz\right]d\nu_p(t) \\ &= \frac{-1}{2\pi i}\oint_{\Gamma_\nu} f(z)m_{\nu_p}(z)dz \end{aligned} \tag{4}$$

with Γ_ν a contour surrounding the support $\mathrm{supp}(\nu_p)$ of ν_p. To relate the unknown m_{ν_p} to the observable m_{μ_p}, we proceed as follows. By first conditioning on $\hat{C}_1$, $\hat{C}_1^{\frac{1}{2}}\hat{C}_2\hat{C}_1^{\frac{1}{2}}$ is seen as a sample covariance matrix for the samples $\hat{C}_1^{\frac{1}{2}}C_2^{\frac{1}{2}}\tilde{x}_i^{(2)}$, for which [11] allows one to relate m_{μ_p} to the Stieltjes transform of the eigenvalue distribution ζ_p of $C_2^{\frac{1}{2}}\hat{C}_1C_2^{\frac{1}{2}}$. The latter is yet another sample covariance matrix for the samples $C_2^{\frac{1}{2}}C_1^{\frac{1}{2}}\tilde{x}_i^{(1)}$; exploiting [11] again creates the connection from m_{ζ_p} to m_{ν_p}. This entails the two equations:

$$zm_{\mu_p}(z) = \varphi_p(z)m_{\zeta_p}(\varphi_p(z)) + o_p(1) \tag{5}$$

$$m_{\nu_p}\left(\frac{z}{\Psi_p(z)}\right) = m_{\zeta_p}(z)\Psi_p(z) + o_p(1). \tag{6}$$

where $\Psi_p(z) \equiv 1 - \frac{p}{n_2} - \frac{p}{n_2}zm_{\zeta_p}(z)$. Successively plugging (5)–(6) into (4) by means of two successive appropriate changes of variables, we obtain Theorem 1. □

Theorem 1 takes the form of a complex integral which, for generic choices of f, needs be numerically evaluated. In the specific case of present interest where $f(z) = \sqrt{z}$, this complex integral can be evaluated as follows.

Theorem 2. *Let $\lambda_1 \le \ldots \le \lambda_p$, with $\lambda_i \equiv \lambda_i(\hat{C}_1\hat{C}_2)$, and define $\{\xi_i\}_{i=1}^p$ and $\{\eta_i\}_{i=1}^p$ the (increasing) 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 $\lambda = (\lambda_1, \ldots, \lambda_p)^\mathsf{T}$, $\Lambda = \mathrm{diag}(\lambda)$ and $\sqrt{\cdot}$ is understood entry wise. Then, under Assumption 1,*

$$D(C_1, C_2; \sqrt{\cdot}) - \hat{D}(X_1, X_2; \sqrt{\cdot}) \xrightarrow{\text{a.s.}} 0$$

where, if $n_1 \neq n_2$,

$$\begin{aligned}\hat{D}(X_1, X_2; \sqrt{\cdot}) = {} & 2\sqrt{n_1n_2}\frac{1}{p}\sum_{j=1}^{p}\sqrt{\lambda_j} \\ & + \frac{2n_2}{\pi p}\sum_{j=1}^{p}\int_{\xi_j}^{\eta_j}\sqrt{-\frac{\varphi_p(x)}{\psi_p(x)}}\psi_p'(x)dx\end{aligned}$$

with φ_p, ψ_p defined in Theorem 1 and, if $n_1 = n_2$,

$$\hat{D}(X_1, X_2; \sqrt{\cdot}) = \frac{2n_1}{p}\sum_{j=1}^{p}\left(\sqrt{\lambda_j} - \sqrt{\xi_j}\right).$$

While still assuming an integral form (when $n_1 \neq n_2$), this formulation no longer requires the arbitrary choice of a contour Γ and significantly reduces the computational time to estimate $D(C_1, C_2, \sqrt{\cdot})$. For $n_1 = n_2$, a case of utmost practical interest, the expression is completely explicit and computationally only requires to evaluate the eigenvalues ξ_j of $\Lambda - \frac{1}{n_1}\sqrt{\lambda}\sqrt{\lambda}^\mathsf{T}$. The latter being a (negative definite) rank-1 perturbation of Λ, by Weyl's interlacing lemma [12], the ξ_j's are interlaced with the λ_j's as

$$\xi_1 \le \lambda_1 \le \xi_2 \le \ldots \le \xi_p \le \lambda_p.$$

As the λ_j's are of order $O(1)$ with respect to p, $|\lambda_j - \xi_j| \le |\lambda_j - \lambda_{j-1}| = O(p^{-1})$, therefore explaining why the expression of $\hat{D}(X_1, X_2; \sqrt{\cdot})$ is of order $O(1)$.

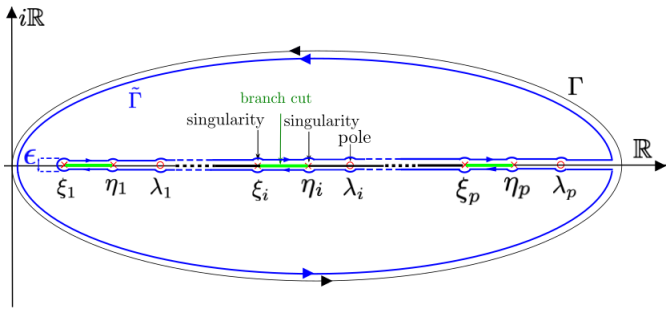

Fig. 1. Deformation of the initial contour Γ (in black) into the new contour $\tilde{\Gamma}$ (in blue). The branch cuts are represented in green (i.e., real z's for which the argument of $\varphi(z)\psi(z)$ is negative).

Proof. The ξ_i and η_i, as defined in the theorem statement, are the respective zeros of the rational functions $1-\frac{p}{n_1}-\frac{p}{n_1}zm_{\mu_p}(z)$ and $1-\frac{p}{n_2}-\frac{p}{n_2}zm_{\mu_p}(z)$ (see [10, Appendix B]). Thus, φ_p and ψ_p can be expressed under the rational form:

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

Evaluating the estimate from Theorem 1 for $f(z)=\sqrt{z}$ then requires to evaluate a complex integral involving rational functions and square roots of rational functions. Since the complex square root is multivalued, a careful control of "branch-cuts" is required. To perform this calculus, we deform the integration contour Γ of Theorem 1 into $\tilde{\Gamma}$ as per Figure 1. In the case $n_1\neq n_2$, the closed null-integral contour $\tilde{\Gamma}$ (blue in Figure 1) is the sum of the sought-for integral over Γ and of four extra components:

1) Integrals over ϵ-radius circles around ξ_i: those are null in the limit $\epsilon\to 0$, as confirmed by a change of variable $z=\xi_i+\epsilon e^{\imath\theta}$ which allows one to bound the integrand;
2) Integrals over the real axis (in the $\epsilon\to 0$ limit):

$$\begin{aligned}A_2&=\frac{n_2}{\pi p}\sum_{j=1}^p\int_{\xi_j+\epsilon}^{\eta_j-\epsilon}\sqrt{-(\varphi_p\psi_p)(x)}\left[2\frac{\psi_p'(z)}{\psi_p(z)}\right.\\&\left.-\left(\frac{\varphi_p'(z)}{\varphi_p(z)}+\frac{\psi_p'(z)}{\psi_p(z)}\right)\right]dx\\&=\frac{2n_2}{\pi p}\sum_{j=1}^p\int_{\xi_j}^{\eta_j}\sqrt{-\varphi_p\psi_p(x)}\left[\frac{\psi_p'(z)}{\psi_p(z)}\right]dx\\&-\frac{n_2}{\pi p}\sum_{j=1}^p\int_{\xi_j+\epsilon}^{\eta_j-\epsilon}\frac{\sqrt{-\varphi_p\psi_p(x)}}{\varphi_p\psi_p(x)}\left[\frac{d}{dx}\left(\varphi_p(x)\psi_p(x)\right)\right]dx\\&=\frac{2n_2}{\pi p}\sum_{j=1}^p\int_{\xi_j}^{\eta_j}\sqrt{-\frac{\varphi_p(x)}{\psi_p(x)}}\psi_p'(x)dx\\&-2\frac{n_2}{\pi p}\sum_{j=1}^p\frac{1}{\sqrt{\epsilon\frac{d}{dx}\left(\frac{1}{(\varphi_p\psi_p(x))}\right)(\eta_j)}}+o(\epsilon)\end{aligned}$$

3) Integrals over the ϵ-radius circles around η_j, with $\epsilon\to 0$

$$A_3=2\frac{n_2}{\pi p}\sum_{j=1}^p\frac{1}{\sqrt{\epsilon\frac{d}{dx}\left(\frac{1}{(\varphi_p\psi_p(x))}\right)(\eta_j)}}+o(\epsilon)$$

which thus compensates the last (ϵ-diverging) term in A_2.

4) Residues in the λ_j poles

$$A_4=2\frac{n_2}{p}\lim_{z\to\lambda_j}\sum_{j=1}^p\sqrt{(\varphi_p\psi_p)(z)}=2\frac{n_2}{p}\sqrt{\frac{n_1}{n_2}}\sum_{j=1}^p\sqrt{\lambda_j}.$$

Putting these terms together entails the result of the theorem for the case where $n_1\neq n_2$. For $n_1=n_2$, it suffices to take the limit of the expression as $\xi_j\to\eta_j$. This yields:

$$\begin{aligned}\hat{D}(X_1,X_2;\sqrt{\cdot})&=\frac{2n_1}{p}\sum_{j=1}^p\sqrt{\lambda_j}\\&+\frac{2n_1}{p}\sum_{j=1}^p\frac{1}{\pi}\lim_{t\to\xi_j}\int_{\xi_j}^t\sqrt{-\frac{\varphi_p(x)}{\psi_p(x)}}\psi_p'(x)dx\\&=\frac{2n_1}{p}\sum_{j=1}^p\sqrt{\lambda_j}\\&-\frac{2n_1}{p}\sum_{j=1}^p\frac{1}{2\pi\imath}\lim_{\epsilon\to 0}\oint_{\Gamma_{\xi_j}^\epsilon}\sqrt{-\varphi_p\psi_p(x)}\frac{\psi_p'(x)}{\psi_p(x)}dx\end{aligned}$$

where $\Gamma_{\xi_j}^\epsilon$ is an ϵ-radius circular contour around ξ_j. The second equality is obtained by deforming the real integral in the complex plane (see [13] for complex analysis details). The result unfolds by letting $x=\xi_i+\epsilon e^{\imath\theta}$.

□

Consequently, we obtain the following n,p-consistent estimate for the Wasserstein distance $D_W(C_1,C_2)$ of (1).

Corollary 1 (Consistent Estimate of $D_W(C_1,C_2)$)**.** *Under Assumption 1,*

$$\frac{1}{p}D_W(C_1,C_2)-\left[\frac{1}{p}\mathrm{tr}(\hat{C}_1+\hat{C}_2)-2\hat{D}(X_1,X_2;\sqrt{\cdot})\right]\xrightarrow{\text{a.s.}}0 \tag{7}$$

for $\hat{D}(X_1,X_2;\sqrt{\cdot})$ given by Theorem 2.

Remark 1 (Estimation of $\|C_1-C_2\|_F^2$)**.** *The Frobenius distance between two covariance matrices also falls under the scope of the present article for the function $f(z)=z$. Indeed,*

$$D_F(C_1,C_2)=\|C_1-C_2\|_F^2=\mathrm{tr}\left(C_1^2+C_2^2\right)-2\mathrm{tr}\left(C_1C_2\right).$$

Then under Assumption 1 and along with the fact that $\frac{1}{p}\mathrm{tr}C_1^2$ can be estimated consistently from $\frac{1}{p}\mathrm{tr}\hat{C}_1^2-\frac{1}{n_1p}(\mathrm{tr}\hat{C}_1)^2$,

$$\begin{aligned}\frac{1}{p}D_F(C_1,C_2)-\left[\frac{1}{p}\mathrm{tr}(\hat{C}_1^2+\hat{C}_2^2)-\frac{p}{n_1}\left(\frac{1}{p}\mathrm{tr}\hat{C}_1\right)^2\right.\\\left.-\frac{p}{n_2}\left(\frac{1}{p}\mathrm{tr}\hat{C}_2\right)^2-2\hat{D}(X_1,X_2;\cdot)\right]\xrightarrow{\text{a.s.}}0.\end{aligned}$$

In this case, $\hat{D}(X_1,X_2;\cdot)$ assumes the simple expression

$$\hat{D}(X_1,X_2;\cdot)=\frac{1}{p}\sum_{j=1}^p\lambda_j=\frac{1}{p}\mathrm{tr}\hat{C}_1\hat{C}_2$$

which follows from $\frac{1}{p}\mathrm{tr}\hat{C}_1\hat{C}_2 - \frac{1}{p}\mathrm{tr}C_1C_2 \xrightarrow{\text{a.s.}} 0$ *(by elementary probability arguments) or equivalently from a residue calculus based on Theorem 1 for* $f(z) = z$.

IV. Simulations and Applications

In this section, we first corroborate our theoretical findings by comparing the classical plug-in estimator to our proposed estimator on synthetic Gaussian data. We then provide an application of our results to improved covariance matrix estimation based on few samples.

A. Confirmation of our results on synthetic data

We here compare the classical plug-in estimate of the Wasserstein distance (that is (1) with C_a replaced by $\hat{C}_a$, $a = 1, 2$) with our proposed estimate in Corollary 1. Table I lists the results obtained for Toeplitz matrices C_1, C_2 estimated based on various values of p, n_1, n_2. While our proposed estimator is designed under a large p, n_1, n_2 assumption (as per Assumption 1), it achieves competitive performances even for small values of p, corroborating here our findings in [8] for other classes of covariance matrix distances.

p	$D_{\mathrm{W}}(C_1, C_2)$	Classical	Proposed
2	0.0110	0.0127	0.0120
4	0.0175	0.0198	0.0183
8	0.0208	0.0232	0.0206
16	0.0225	0.0280	0.0227
32	0.0233	0.0339	0.0234
64	0.0237	**0.0451**	0.0240
128	0.0239	**0.0667**	0.0244
256	0.0240	**0.1092**	0.0244
512	0.0241	**0.1953**	0.0245

(error $< 5\%$) **(error $> 50\%$)** **(error $> 100\%$)** **(error $> 300\%$)**

TABLE I

Estimators of the Wasserstein distance between C_1 and C_2 with $[C_1]_{ij} = .2^{|i-j|}$, $[C_2]_{ij} = .4^{|i-j|}$, $x_i^{(a)} \sim \mathcal{N}(0, C_a)$; $n_1 = 1024$ and $n_2 = 2048$ for different p. Averaged over 100 trials.

B. Application to covariance matrix estimation

As a concrete application, Theorem 1 may be used to improve the actual estimation of covariance matrices under a small number $n \sim p$ of sample data, as similarly performed in [14] for other covariance matrix distances.

The idea is as follows: we first particularize Theorem 1 and Theorem 2 to the case where one of the covariance matrices, say C_1, is known by taking $c_1 = 0$ (i.e., $n_1 \to \infty$ for all fixed p). This gives access to estimates for $D_W(M, C_2; \sqrt{\cdot})$ for all deterministic positive definite matrix M. We then minimize this estimated distance over M in order to estimate C_2 by means of a gradient descent approach.

For C_1 known, we redefine $\mu_p = \frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i(C_1\hat{C}_2)}$ and obtain, as a corollary of Theorem 1:

Theorem 3. *Let* $\Gamma \subset \{z \in \mathbb{C}, \mathrm{real}[z] > 0\}$ *a contour surrounding* $\cup_{p=1}^{\infty}\mathrm{supp}(\mu_p)$. *Then,*

$$D(C_1, C_2; f) - \frac{1}{2\pi i c_2}\oint_\Gamma F\left(-m_{\tilde{\mu}_p}(z)\right) dz \xrightarrow{\text{a.s.}} 0$$

with $m_{\tilde{\mu}_p}(z) = \frac{p}{n_2}m_{\mu_p}(z) + \frac{p-n_2}{n_2 z}$ *and* $F'(z) = f(\frac{1}{z})$.

Proof. For C_1 known ($c_1 \to 0$), $\varphi_p(z) = z$, and the estimator of Theorem 1 yields:

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

Using the relation $m_{\tilde{\mu}_p}(z) = -\frac{\psi_p(z)}{z}$, we then get

$$\hat{D}(X_1, X_2; f) = -\frac{1}{2\pi i c_2}\oint_\Gamma f\left(-\frac{1}{m_{\tilde{\mu}_p}(z)}\right) m'_{\tilde{\mu}_p}(z) z dz$$

and the result is immediate after an integration by parts. □

For $f(z) = \sqrt{z}$, one has $F(z) = 2\sqrt{z}$ and we obtain, with a similar proof as for Theorem 2,

$$D(C_1, C_2; \sqrt{\cdot}) - \hat{D}(C_1, X_2; \sqrt{\cdot}) \xrightarrow{\text{a.s.}} 0,$$
$$\hat{D}(C_1, X_2; \sqrt{\cdot}) = \frac{2}{\pi c_2}\sum_{j=1}^p \int_{\xi_j}^{\lambda_j} \sqrt{m_{\tilde{\mu}_p(x)}}dx.$$

Our objective is now to exploit the fact that

$$C_2 = \mathrm{argmin}_{M\succ 0} D_W(M, C_2) \tag{8}$$

where the minimization is over the open cone of positive definite matrices. Using the approximation $D(M, C_2; \sqrt{\cdot}) \simeq \hat{D}(M, X_2; \sqrt{\cdot})$, we are then tempted to minimize $\frac{1}{p}\mathrm{tr}(M + \hat{C}_2) - 2\hat{D}(M, X_2; \sqrt{\cdot})$ in place of $D_W(M, C_2)$. The former quantity however has a non zero probability to be negative, and we thus instead propose to estimate C_2 as:

$$\check{C}_2 = \mathrm{argmin}_M\, h(M)$$
$$h(M) = \left[\frac{1}{p}\mathrm{tr}(M + \hat{C}_2) - 2\hat{D}(M, X_2; \sqrt{\cdot})\right]^2.$$

To compute the gradient $\nabla h(M)$ of h at position M, one needs to evaluate the differential $\mathrm{D}h(M)[\xi]$, at M and in the direction ξ, in the Riemmanian manifold of $p \times p$ symmetric positive definite matrices (see [15], [14] to further technical details). We then use the relation $\mathrm{D}h(M)[\xi] = \langle \nabla h(M), \xi\rangle_M$ where $\langle\cdot,\cdot\rangle_.$ is the Riemmanian metric defined as $\langle \eta, \xi\rangle_M = \mathrm{tr}\left(M^{-1}\eta M^{-1}\xi\right)$. We obtain the relation

$$\pi\imath p\frac{\nabla h(M)}{2\sqrt{h(M)}} = \frac{1}{p}M^2 + \sum_{j=1}^p \int_{\xi_j}^{\lambda_j}\sqrt{\frac{1}{m_{\tilde{\mu}_p}(x)}}\mathrm{sym}\left(M\hat{C}_2(M\hat{C}_2 - xI_p)^{-2}M\right)dx$$

where $\mathrm{sym}(A) = \frac{1}{2}(A + A^{\mathsf{T}})$ is the symmetric part of $A \in \mathbb{R}^{p\times p}$. We can write the latter as:

$$\nabla h(M) = 2\sqrt{h(M)}\left[\mathrm{sym}\left(V\Lambda_\nabla V^{-1}\right) + \frac{1}{p}M^2\right]$$

where V is the orthogonal matrix of the eigenvectors of $M\hat{C}_2$ and Λ_∇ is the diagonal matrix with

$$[\Lambda_\nabla]_{kk} = \frac{1}{\pi p}\sum_{j\neq k}\int_{\xi_j}^{\lambda_j}\sqrt{\frac{1}{m_{\tilde{\mu}_p}(x)}}\frac{1}{(\lambda_k - x)^2}dx + \frac{1}{\pi p}\sum_{j\neq k}\int_{\xi_k}^{\lambda_k}\sqrt{\frac{1}{m_{\tilde{\mu}_p}(x)}}\frac{1}{(\lambda_j - x)^2}dx.$$

Algorithm 1 Proposed estimation algorithm.

Require Positive definite initialization $M = M_0$.

Repeat $M \leftarrow M^{\frac{1}{2}} \exp\left(-tM^{-\frac{1}{2}} \nabla h(M) M^{-\frac{1}{2}}\right) M^{\frac{1}{2}}$ with t either fixed or optimized by backtracking line search.

Until Convergence.

Return M.

This finally entails the gradient descent Algorithm 1.

Figure 2 depicts the results of the algorithm. There is displayed the Wasserstein distance $D_W(C, \cdot)$ between a matrix C having four distinct eigenvalues of equal multiplicity (precisely, $\nu_p = \frac{1}{4}(\delta_{.1} + \delta_3 + \delta_4 + \delta_5)$) and various estimators of C: the sample covariance matrix (SCM), the state-of-the-art "non-linear shrinkage" estimators QuEST1 [16] (based on a Frobenius distance minimization) and QuEST2 [17] (based on a Stein loss minimization), and the result of the gradient descent approach proposed in this section. For fair comparison, the iterative QuEST1, QuEST2 and our proposed method are all initialized at M_0 the linear shrinkage estimator from [18]. Note that our proposed choice of C is particularly suited to mimick an "optimal transport" problem of displacing the eigenvalues of M_0 to the discrete four positions of the eigenvalues of C.

In addition to the computational simplicity of our gradient-descent approach with respect to the QuEST estimators (see the numerical method details in [19]), the figure demonstrates significant gains brought by our proposed approach for large values of p/n, where the SCM particularly fails.

V. Concluding Remarks

Interestingly, while the Fisher distance or Kullbach-Liebler divergence, which depend on logarithms of *inverse* of covariance matrices, are understandably difficult to estimate in the $n_1, n_2 < p$ regime (see [10] for advanced discussions on this matter), the Wasserstein distance should not be confronted with this limitation. Yet, the invertibility of C_1, C_2 and the request for $c_1, c_2 \in (0, 1)$ (i.e., $p < n_1, n_2$) from Assumption 1 are fundamental to our proofs. Precisely, the variable changes exploited in the proof of Theorem 1 to reach a contour Γ_ν correctly surrounding $\mathrm{supp}(\nu_p)$ from a contour Γ surrounding $\mathrm{supp}(\mu_p)$ are not satisfying if $c_1 > 1$ or $c_2 > 1$. These surprising difficulties need clarification.

Another point of interest lies in the comparative advantage of exploiting a particular covariance matrix distance in specific scenarios. For instance, it may seem that ill-conditioned matrices should be more tolerated by Wasserstein distance estimators than by Fisher distance estimators. Yet, this aspect is not obvious in our proofs and also deserves more insights.

Acknowledgement

We thank Pedro Rodrigues for helpful discussions and references on optimal transport.

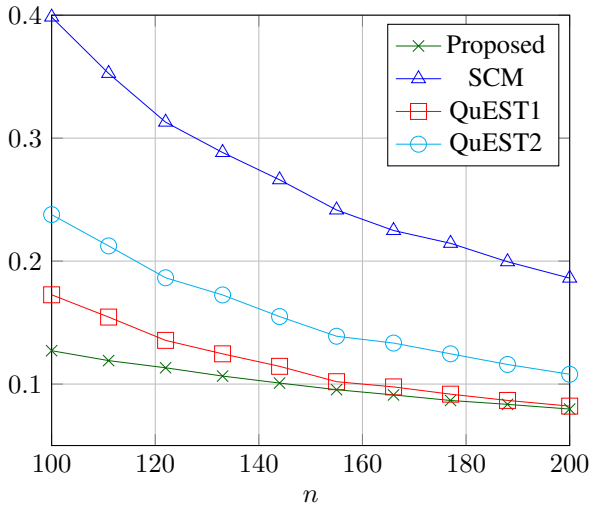

Fig. 2. Wasserstein distance $D_W(C, \cdot)$ between C with $\nu_p = \frac{1}{4}(\delta_{.1} + \delta_3 + \delta_4 + \delta_5)$ and (green) our proposed estimator, (blue) the sample covariance matrix, (red) and (light blue) the QuEST estimators proposed in [17], [16]; for $p = 100$ and varying number of samples n averaged over 10 realizations.

References

[1] Gaspard Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[2] Leonid V Kantorovich, "On the translocation of masses," in *Dokl. Akad. Nauk. USSR (NS)*, 1942, vol. 37, pp. 199–201.

[3] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, no. 2, pp. 99–121, 2000.

[4] Zhengyu Su, Yalin Wang, Rui Shi, Wei Zeng, Jian Sun, Feng Luo, and Xianfeng Gu, "Optimal mass transport for shape matching and comparison," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 11, pp. 2246–2259, 2015.

[5] Kangyu Ni, Xavier Bresson, Tony Chan, and Selim Esedoglu, "Local histogram based segmentation using the wasserstein distance," *International journal of computer vision*, vol. 84, no. 1, pp. 97–111, 2009.

[6] Marco Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in neural information processing systems*, 2013, pp. 2292–2300.

[7] Xavier Mestre, "On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices," *IEEE Transactions on Signal Processing*, vol. 56, no. 11, pp. 5353–5368, 2008.

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

[9] Gabriel Peyré and Marco Cuturi, "Computational optimal transport," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

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

[11] 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.

[12] Joel N Franklin, *Matrix theory*, Courier Corporation, 2012.

[13] EB Saff and AD Snider, "Fundamentals of complex analysis with applications to engineering and science," 2003.

[14] 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.

[15] P-A Absil, Robert Mahony, and Rodolphe Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, 2009.

[16] Olivier Ledoit and Michael Wolf, "Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions," *Journal of Multivariate Analysis*, vol. 139, pp. 360–384, 2015.

[17] Olivier Ledoit, Michael Wolf, et al., "Optimal estimation of a large-dimensional covariance matrix under stein's loss," *Bernoulli*, vol. 24, no. 4B, pp. 3791–3832, 2018.

[18] Olivier Ledoit and Michael Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of multivariate analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[19] Olivier Ledoit and Michael Wolf, "Numerical implementation of the quest function," *Computational Statistics & Data Analysis*, vol. 115, pp. 199–223, 2017.