IMPROVED ESTIMATION OF THE DISTANCE BETWEEN COVARIANCE MATRICES

Malik Tiomoko[1]*, Romain Couillet*[2,1]*, Eric Moisan*[2] *and Steeve Zozor*[2]

[1]CentraleSupélec, University of Paris–Saclay and [2]GIPSA-lab, University of Grenoble–Alpes.

ABSTRACT

A wide range of machine learning and signal processing applications involve data discrimination through statistical covariances. A broad family of covariance matrix metrics, among which the Frobenius, Fisher, Bhattacharyya distances, as well as the Kullback-Leibler or Rényi divergences, are regularly exploited. However, not being directly accessible, these metrics are assessed through *empirical* sample covariances, which we shall show here may lead to dramatically erroneous estimates in large dimensional data.

In this article, based on advanced random matrix considerations, we provide a novel generic consistent estimate for such covariance matrix distances and divergences. While theoretically developed for both large and numerous data, practical simulations demonstrate its large performance gains over the standard approach even for very small dimensions. A particular emphasis is made on the Fisher information metric and a concrete application to covariance-based spectral clustering is investigated.

Index Terms— Covariance distance, random matrix theory, Fisher information metric.

1. INTRODUCTION

Similarities between covariance matrices are objects of interest for many engineering applications, among which the machine learning problem of classifying data from covariance-based features (for instance, pattern detection and classification in synthetic aperture radar and hyperspectral imaging [1], or in EEG datasets [2]), dimensionality reduction [3], portfolio-optimization and asset clustering in finance [4], etc. Depending on context and application, various metrics are available in the literature to compare semi-definite positive matrices (the Frobenius norm, the Fisher Information metric [5], the Bhattacharyya distance [6], the Rényi or Kullback-Leibler divergence, etc.). Many of these metrics (here all of the aforementioned except the Frobenius distance) can be written under a common functional form involving the distribution of the eigenvalues of $C_1^{-1}C_2$, where C_1 and C_2 are the two matrices to be compared. Based on a simple law-of-large-numbers argument, these metrics are commonly estimated from a simple replacement of the genuine $p \times p$-dimensional matrices C_1 and C_2 by their sample covariance estimates $\hat{C}_1$ and $\hat{C}_2$. Such estimates, as shall be shown next, are however bound to sometimes extremely severe errors, particularly when the respective numbers n_1 and n_2 of samples to estimate C_1 and C_2 are not large compared to p. This scenario is however frequently met in practice (short-time brain activity scans with high resolution EEG, large number of shortly-stationary assets in finance, high-resolution hyperspectral imaging, etc.) and therefore induces possibly weak data processing performances.

To tackle these problems, in this paper, we introduce a new estimate for a broad family of covariance matrix metrics, with an exemplary emphasis on the Fisher information metric. More results involving other metrics, along with advanced technical details and comments, are reported in the extended article [**?**]. This estimate is n_1, n_2, p*-consistent* in the sense that it (almost surely) converges to the sought-for metric as n_1, n_2, p grow simultaneously large. Yet, simulation results will show that even for very small dimensional settings, the proposed method largely outperforms the traditional sample covariance "plug-in" estimator.

Technically speaking, our results rely on the following approach. We express a generic form of the metric under study under the form of a complex integral involving the *Stieltjes transform* of the (population) eigenvalue distribution of $C_1^{-1}C_2$. As the latter distribution is not accessible, we then link it to the (sample) eigenvalue distribution of $\hat{C}_1^{-1}\hat{C}_2$, through a functional equation relating the Stieltjes transforms of population and empirical eigenvalue distribution. This results, through an appropriate change of variable, to a complex integral involving only the eigenvalues of $\hat{C}_1^{-1}\hat{C}_2$, which may finally be evaluated using complex analysis techniques.

This approach is notably inspired by the seminal work of Mestre [7] (see also [8]) where functional estimates of the eigenvalue distribution of a single covariance matrix C is performed similarly from the corresponding eigenvalue distribution of the sample estimate $\hat{C}$. Aside from the more involved statistical model $\hat{C}_1^{-1}\hat{C}_2$, the originality of the present work mostly lies in that the family of metrics involve non-smooth complex functionals (in particular logarithms) that result in more advanced technical considerations from real and complex analysis than in [7].

Couillet's work is supported by the UGA IDEX GSTATS DataScience Chair and by the ANR RMT4GRAPH (ANR-14-CE28-0006) project.

The remainder of the article is organized as follows. In Section 2, we introduce the main model and assumptions. In Section 3, our main result providing the generic consistent estimate in complex integral form is introduced. Section 4 then proposes an application of this result to the Fisher information metric, which are then simulated and compared to the traditional method in Section 5.

2. MODEL AND ASSUMPTIONS

For $a \in \{1,2\}$, we consider n_a vectors $x_1^{(a)},\dots,x_{n_a}^{(a)}$ independent and identically distributed, with $x_i^{(a)} = C_a^{\frac{1}{2}}\tilde{x}_i^{(a)}$ where $\tilde{x}_i^{(a)}$ has zero mean, unit variance and finite fourth order moment entries. In addition, we shall make the following growth rate 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.*

The object under investigation in this article is any metric of (C_1, C_2) expressible under the form:

$$D(C_1,C_2) = \frac{1}{p}\sum_{i=1}^{n} f\left(\lambda_i(C_1^{-1}C_2)\right)$$

with $\lambda_i(X)$ the i-th eigenvalue of X and $f : R \to R$ a given function. For instance, letting $f(t) = \log^2(t)$, $D(C_1,C_2)$ is the Fisher distance between C_1 and C_2 (more in Section 4); for $f(t) = \frac{1}{2}t - \frac{1}{2} + \frac{1}{2}\log(t)$, $D(C_1,C_2)$ is the Kullback-Leibler divergence; for $f(t) = \frac{1}{2}\log(1+t) - \frac{1}{4}\log(t) - \frac{1}{2}\log(2)$, $D(C_1,C_2)$ is the Bhattacharyya distance, etc.

Our main theoretical result, rooted in random matrix techniques, involves the *Stieltjes transform* of probability measures defined, for a measure θ as $m_\theta : C \setminus \mathrm{supp}(\theta) \to C$,

$$m_\theta(z) = \int \frac{d\theta(\lambda)}{\lambda - z}.$$

Specifically we will relate the Stieltjes transform of the population and empirical eigenvalue distributions:

$$\mu_p = \frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i(\hat{C}_1^{-1}\hat{C}_2)}, \quad \nu_p = \frac{1}{p}\sum_{i=1}^{p}\delta_{\lambda_i(C_1^{-1}C_2)}.$$

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

3. MAIN RESULT

Our main result is as follows:

Theorem 1. *Let $f :\to$ be analytic on a (positively oriented) contour $\Gamma \subset \{z \in C, \mathcal{R}[z] > 0\}$ surrounding $\cup_{p=1}^{\infty}\mathrm{supp}(\mu_p)$. Then,*

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

with $\varphi_p(z) = z(1 + c_1 z m_{\mu_p}(z))$, $\psi_p(z) = 1 - c_2 - c_2 z m_{\mu_p}(z)$.

The result of Theorem 1 has the strong advantage to be flexible to any smooth function f over $\{z \in C, \mathcal{R}[z] > 0\}$, so in particular to $f(z) = \log^k(z)$ or $f(z) = \log^k(1+\alpha z)$, which commonly appear in covariance distances and divergences. The constraint $c_2 < 1$ is however mandatory and cannot be relaxed, unless f is analytic on all (which fails for logarithm functions); see [**?**] for details.

Before getting to the proof, note that the formulation of Theorem 1 exhibits two important quantities, the functions φ_p and ψ_p, which both relate to the eigenvalue distribution of $\hat{C}_1^{-1}\hat{C}_2$ respectively through c_1 and c_2; each function therefore emphasizes the impact of the restricted number of data with respect to the dimension p.

We subsequently provide a sketch of proof of Theorem 1.

Sketch of proof. The main observation arises from Cauchy's integral formula by which:

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

with Γ_ν a contour surrounding the support of ν_p.

The next step is to link the Stieltjes transform $m_{\nu_p}(z)$ to the Stieltjes transform $m_{\mu_p}(z)$. The latter being a random quantity, we first resort to an asymptotic estimation. A first important remark is that, since only the eigenvalues of $\hat{C}_1^{-1}\hat{C}_2$ are involved, the problem is unchanged if $x_i^{(1)}$ had identity covariance while $x_i^{(2)}$ had covariance $C_1^{-\frac{1}{2}}C_2C_1^{-\frac{1}{2}}$. With this change of notation, $\hat{C}_1^{-1}\hat{C}_2$ is the product of two independent sample covariance matrices and, following the seminal works of Bai and Silverstein [9], one may successively relate the (almost sure) *limiting eigenvalue distribution* μ of $\hat{C}_1^{-1}\hat{C}_2$ to the (almost sure) limiting eigenvalue distribution ζ_2 of $\hat{C}_2$ (by conditioning on the latter) before next relating ζ_2 to the eigenvalue distribution ν_p (which can be asked to coincide its limit) of $C_1^{-\frac{1}{2}}C_2C_1^{-\frac{1}{2}}$. All calculus made, we obtain the coupled fundamental equations:

$$m_\mu(z) = z(1 + c_1 z m_\mu(z))m_{\zeta_2}\left(z(1 + c_1 z m_\mu(z))\right) \qquad (2)$$

$$m_\nu\left(\frac{z}{1 - c_2 - c_2 z m_{\zeta_2}(z)}\right) = m_{\zeta_2}(z)\left(1 - c_2 - c_2 z m_{\zeta_2}(z)\right). \qquad (3)$$

Successively plugging (2)–(3) in (1) in two successive appropriate changes of variables, we obtain an exact equality for Theorem 1 with μ_p replaced by μ, and Γ the pre-image of Γ_ν by φ/ψ. Since $\mu_p \to \mu$ uniformly on the bounded Γ, the result unfolds. A remaining non-trivial hidden difficulty though is to ensure that there does indeed exist a Γ such that

$(\varphi/\psi)(\Gamma)$ leads to a valid contour Γ_ν. In [**?**], we show that this is in general only possible if $c_2 < 1$ ($c_1 < 1$ is mandatory for the existence of μ_p). □

4. APPLICATION: THE FISHER DISTANCE

Theorem 2 requires to evaluate numerically a complex integral. This may be inefficient and particularly inaccurate as it numerically depends on the chosen contour and integration step size. Besides, an integral formula leaves little room to interpretation. In this section, letting $f(t) = \log^2(t)$, we establish a closed form expression for Theorem 2 adapted to the Fisher distance D_{F}. Indeed, the latter is defined through [10]

$$D_F(C_1, C_2)^2 = \frac{1}{p}\sum_{i=1}^{p} \log^2\left(\lambda_i(C_1^{-1}C_2)\right) = \int \log^2(t)\nu_p(dt).$$

For this distance, we have the following corollary of Theorem 1.

Theorem 2. *For $\lambda = (\lambda_1, \ldots, \lambda_p)^\mathsf{T}$ (ordered) and $\Lambda = \mathrm{diag}(\lambda)$, let $\eta \in R^p$ be the (ordered) eigenvalues of $\Lambda + \frac{1}{n_1 - p}\sqrt{\lambda}\sqrt{\lambda}^\mathsf{T}$ and $\zeta \in R^p$ the (ordered) eigenvalues of $\Lambda - \frac{1}{n_2}\sqrt{\lambda}\sqrt{\lambda}^\mathsf{T}$. Then, under Assumption 1, we have*

$$\int \log^2(t)\nu_p(dt) - \left[\frac{1}{p}\sum_{i=1}^{p}\log^2((1-c_1)\lambda_i) - \frac{2}{p}\left(\Delta_\zeta^\eta\right)^\mathsf{T} N 1_p \right.$$
$$+ 2\frac{c_1 + c_2 - c_1c_2}{c_1c_2}\left\{\left(\Delta_\zeta^\eta\right)^\mathsf{T} M\left(\Delta_\lambda^\eta\right) + \left(\Delta_\lambda^\eta\right)^\mathsf{T} r\right\}$$
$$\left. -2\frac{1-c_2}{c_2}\left\{\frac{1}{2}\log^2((1-c_1)(1-c_2)) + \left(\Delta_\zeta^\eta\right)^\mathsf{T} r\right\}\right] \xrightarrow{\text{a.s.}} 0$$

where we defined Δ_a^b the vector with $(\Delta_a^b)_i = b_i - a_i$ and, for $i, j \in \{1, \ldots, p\}$,

$$M_{ij} = \begin{cases} \frac{\frac{\lambda_i}{\lambda_j} - 1 - \log\left(\frac{\lambda_i}{\lambda_j}\right)}{(\lambda_i - \lambda_j)^2} & , i \neq j \\ \frac{1}{2\lambda_i^2} & , i = j \end{cases},$$

$$N_{ij} = \begin{cases} \frac{\log\left(\frac{\lambda_i}{\lambda_j}\right)}{\lambda_i - \lambda_j} & , i \neq j \\ \frac{1}{\lambda_i} & , i = j. \end{cases} \text{ and } r_i = \frac{\log((1-c_1)\lambda_i)}{\lambda_i}.$$

A few remarks are in order before proving this result. First, despite the seemingly involved formulation of Theorem 2, the latter shows that the proposed estimate is somewhat related to the classical one (through the leading term $1/p\sum_i \log^2((1-c_1)\lambda_i)$ to which a non trivial bias term is added. It is also worth noting that, if C_1 were perfectly known, then Theorem 2 holds by taking the limit where $c_1 \to 0$; this enlarges the practical perspective of the estimator in applications where one aims at evaluating the distance of an unknown covariance matrix to a reference point (for instance when tracking the centroid of multiple covariance matrices [11]).

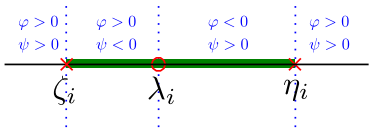

Fig. 1. Analysis of φ_p and ψ_p close to the real axis. A branch cut appears where $\varphi_p/\psi_p < 0$.

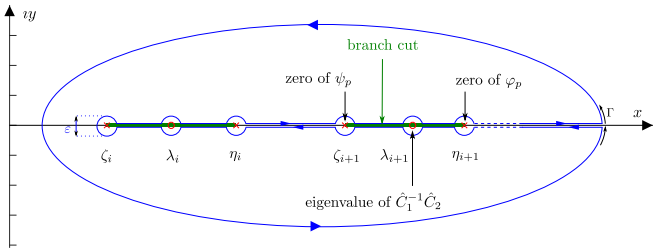

Fig. 2. Integration contour and the logarithm branch cuts.

Sketch of proof. For functions f analytical inside the contour Γ, the evaluation of the complex integral in Theorem 1 follows from a mere residue theorem. For the Fisher distance, this would hold if $\log^2(\varphi_p(z)/\psi_p(z))$ were analytical inside Γ. This is however not the case, as a precise analysis of the latter reveals branch cuts in the complex logarithm arising in the real segments $[\zeta_i, \eta_i]$, $i = 1, \ldots, p$, with ζ_i the zeros of $\psi_p(z)$ and η_i those of $\varphi_p(z)$; see Figure 1 for a visualization.

To evaluate the complex integral, we thus resort to a contour deformation carefully avoiding branch cuts and singularities as depicted in Figure 2 (see e.g., [12] for details on complex analysis integration methods). Of utmost interest here are the resulting (limiting) real integrals right above and under branch cuts. These involve real integrals of the type $\int_c^d \frac{\log(x-a)}{x-b}dx$, for $a, b, c, d > 0$, which do not have a closed form in general but can be written as function of *dilogarithms* $\mathrm{Li}_2(x) = -\int_0^x \frac{\log(1-u)}{u}du$ [13]. Using functional relations of the latter (e.g., to relate $\mathrm{Li}_2(x)$ to $\mathrm{Li}_2(1/x)$), the summation of all real integrals then simplifies to reach a compact closed-form formula. This formula however involves the time-consuming evaluation of $\mathcal{O}(p^2)$ values for Li_2. Observing that $\lambda_i = \zeta_i + \mathcal{O}(\frac{1}{p})$ and $\eta_i = \lambda_i + \mathcal{O}(\frac{1}{p})$, a Taylor expansion of the latter finally brings the desired expression. □

5. NUMERICAL EXPERIMENTS

In this section, we compare our estimate of the Fisher distance derived in Theorem 2 to the classical "plug-in" estimator $\frac{1}{p}\sum_{i=1}^p \log^2(\lambda_i)$.

We first report in Table 5 the genuine versus estimated values of the Fisher distance on a synthetic setting (details in caption). A first surprising observation is that the plug-in estimator is extremely unfit to large values of p/n_1, p/n_2,

bringing up to 500% error figures for $n_1 = 2p$; the proposed estimator is instead resilient to large p. Possibly more surprisingly, while Theorems 1–2 provably hold for asymptotically large p, n_1, n_2, our estimator already outperforms the standard approach for $p = 2$. This may be explained by the fact that the proposed approach essentially exploits randomness both from the size *and* the number of the dataset, with accuracies provably of order $\mathcal{O}(1/\sqrt{pn})$ thereby already reaching accurate values for not too large p (note that this in particular implies central limit theorems and thus convergence speed quadratically faster than in the large-n_1, n_2 alone setting).

p	$D_{\mathrm{F}}(C_1, C_2)$	Classical	Proposed
2	0.0980	0.1002	0.0973
4	0.1456	0.1520	0.1461
8	0.1694	0.1820	0.1703
16	0.1812	0.2081	0.1845
32	0.1872	0.2363	0.1886
64	0.1901	**0.2892**	0.1920
128	0.1916	**0.3955**	0.1934
256	0.1924	**0.6338**	0.1942
512	0.1927	**1.2715**	0.1953

(error $>$ 50%) **(error $>$ 100%)** **(error $>$ 500%)**

Table 1. Proposed versus classical estimator for the Fisher distance between C_1 and C_2 with $[C_1^{-\frac{1}{2}} C_2 C_1^{-\frac{1}{2}}]_{ij} = .3^{|i-j|}$, $x_i^{(a)} \sim \mathcal{N}(0, C_a)$; $n_1 = 1024$ and $n_2 = 2048$ for different values of p. Averaged over 10 000 trials.

In a second experiment, we perform the unsupervised classification, based on kernel spectral clustering [14], of $m = 200$ independent generations of zero mean Gaussian samples $X_i \in R^{p \times n_i}$, half with covariance C_1 and half with covariance C_2. The chosen kernel $K \in R^{m \times m}$ is defined by $K_{ij} = \exp(-\frac{1}{2}\hat{D}_{\mathrm{F}}(X_i, X_j)^2)$ with $\hat{D}_{\mathrm{F}}$ either the classical or our proposed estimator. A first interesting surprising outcome is that, for all tested (rather large) values of p, m and $n_1 = \ldots = n_m$, spectral clustering based on either algorithm perform equally well with often extremely similar eigenvectors; this can be understood by the fact that *non-trivial clustering tasks* occur for very similar C_1 and C_2 matrices for which the standard estimator of D_{F} is *almost equally biased* on all data realizations, thereby solely inducing a constant shift for the entries of K (thus loosely interfering with the informative eigenvectors).

However, this observation breaks down for differing values of n_i. Figure 3 (top) displays a scatter plot of the two leading eigenvectors of K for the same setting as above, but now with n_i chosen uniformly at random in $[2p, 4p]$, $p = 128$, $m = 200$, $C_1 = I_p$, $[C_2]_{ij} = .05^{|i-j|}$. There, we observe both a large spread of the eigenvectors for the classical estimator and a smaller inter-class spacing, suggesting poor clustering performance, as opposed to the well-centered eigenvectors achieved by the proposed estimator. In a possibly more realistic setting, Figure 3 (bottom) displays the result obtained for $n_1 = \ldots = n_{m-1} = 512$ and $n_m = 256$ (simulating a data retrieval failure for one observation). As expected, the classical estimator brings an isolated outlier in the eigenvector scatter plot, but more surprisingly, this very outlier also starkly contaminates the resolution power of the rest of the data; this effect is exacerbated when adding more outliers and is likely due to a competing effect between the outliers and the genuine clusters to "drive" the dominant eigenvectors.

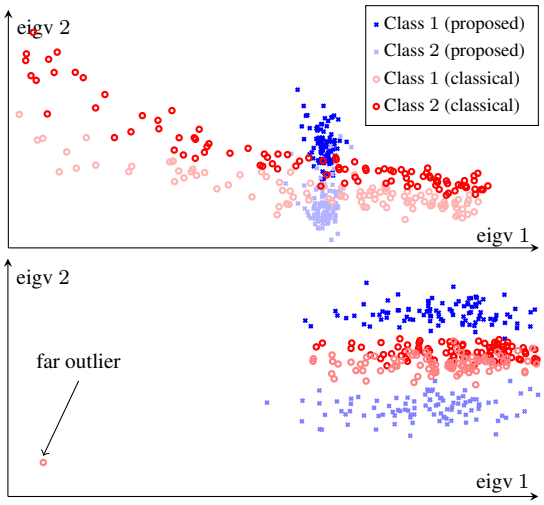

Fig. 3. First and second eigenvectors of K for the traditional estimator (red circles) versus the proposed one (blue crosses); (top) random number of snapshots n_i; (bottom) $n_1 = \ldots = n_{m-1} = 512$ and $n_m = 256$.

6. CONCLUDING REMARKS

The present study has revealed a strong lack of consistency for the traditional "plug-in" covariance matrix-distance (and divergence) estimators, when the data dimension p is not small. This is particularly dramatic as p and the number of snapshots n are close. We provided a consistent solution to recover consistency, exploiting random matrix tools. Yet, our proposed estimator still suffers from the need $n > p$, which may not be met in practice; further investigations and more elaborate tools to cover the case $p > n$ are needed.

Importantly, by exploiting both randomness in p and n, our estimator converges as fast as $\mathcal{O}(1/\sqrt{pn})$, but a more precise central limit analysis is required to exactly assess confidence intervals, which is yet another avenue of research.

But the real strength and robustness of the proposed estimator will only be demonstrated when applied to real (non Gaussian) datasets and more exotic applications. Brain signal processing (or human-machine interaction) and radar imaging (SAR or hyperspectral) are both interesting application candidates that shall be investigated in the future.

7. REFERENCES

[1] Chein-I Chang, *Hyperspectral imaging: techniques for spectral detection and classification*, vol. 1, Springer Science & Business Media, 2003.

[2] Jonas Richiardi, Sophie Achard, Horst Bunke, and Dimitri Van De Ville, "Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 58–70, 2013.

[3] Kevin Michael Carter, "Dimensionality reduction on statistical manifolds.," 2009.

[4] Vincenzo Tola, Fabrizio Lillo, Mauro Gallegati, and Rosario N Mantegna, "Cluster analysis for portfolio optimization," *Journal of Economic Dynamics and Control*, vol. 32, no. 1, pp. 235–258, 2008.

[5] Sueli IR Costa, Sandra A Santos, and João E Strapasson, "Fisher information distance: a geometrical reading," *Discrete Applied Mathematics*, vol. 197, pp. 59–69, 2015.

[6] Anil Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[7] X. Mestre, "Improved estimation of eigenvalues of covariance matrices and their associated subspaces using their sample estimates," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 5113–5129, Nov. 2008.

[8] R. Couillet, J. W. Silverstein, and M. Debbah, "Eigen-Inference for Energy Estimation of Multiple Sources," *IEEE Trans. Inf. Theory*, 2011, To appear.

[9] 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.

[10] Colin Atkinson and Ann FS Mitchell, "Rao's distance measure," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 345–365, 1981.

[11] Marco Congedo, Alexandre Barachant, and Anton Andreev, "A new generation of brain-computer interface based on riemannian geometry," *arXiv preprint arXiv:1310.8115*, 2013.

[12] Steven G Krantz, *Handbook of complex variables*, Springer Science & Business Media, 2012.

[13] Don Zagier, "The dilogarithm function," in *Frontiers in number theory, physics, and geometry II*, pp. 3–65. Springer, 2007.

[14] Romain Couillet, Florent Benaych-Georges, et al., "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.