Random Matrix Improved Covariance Estimation for a Large Class of Metrics

Malik Tiomoko, Florent Bouchard, Guillaume Ginolhac, Romain Couillet

Random Matrix Improved Covariance Estimation for a Large Class of Metrics

Malik Tiomoko [1 2] **Florent Bouchard** [2] **Guillaume Ginholac** [3] **Romain Couillet** [2 1]

Abstract

Relying on recent advances in statistical estimation of covariance distances based on random matrix theory, this article proposes an improved covariance and precision matrix estimation for a wide family of metrics. The method is shown to largely outperform the sample covariance matrix estimate and to compete with state-of-the-art methods, while at the same time being computationally simpler. Applications to linear and quadratic discriminant analyses also demonstrate significant gains, therefore suggesting practical interest to statistical machine learning.

1. Introduction

Covariance and precision matrix estimation is a fundamental and simply posed, yet still largely considered, key problems of statistical data analysis, with countless applications in statistical inference. In machine learning, it is notably at the core of elementary methods as linear (LDA) and quadratic discriminant analysis (QDA) (McLachlan, 2004).

Estimation of the covariance matrix $C \in \mathbb{R}^{p\times p}$ based on n independent (say zero mean) samples $x_1, \ldots, x_n \in \mathbb{R}^p$ is conventionally performed using the sample covariance matrix (SCM) $\hat{C} \equiv \frac{1}{n}\sum_{i=1}^n x_i x_i^\mathsf{T}$ (and its inverse using $\hat{C}^{-1}$). The estimate is however only consistent for $n \gg p$ and only invertible for $n \geq p$. Treating the important practical cases where $n \sim p$ and even $n \ll p$ has recently spurred a series of parallel lines of research. These directions rely either on structural constraints, such as "toeplitzification" procedures for Toeplitz covariance models (particularly convenient for time series) (Bickel et al., 2008; Wu & Pourahmadi, 2009; Vinogradova et al., 2015), on sparse constraints with LASSO and graphical LASSO-based approaches (Friedman et al., 2008) or, more interestingly for the present article, on exploiting the statistical independence in the entries of the vectors x_i.

(Ledoit & Wolf, 2004) proposes to linearly "shrink" $\hat{C}$ as $\hat{C}(\rho) \equiv \rho I_p + \sqrt{1-\rho^2}\hat{C}$ for $\rho > 0$ chosen to minimize the expected Frobenius distance $\mathrm{E}[\|C - \hat{C}(\rho)\|_F]$ in the asymptotic $p, n \to \infty$ limit with $p/n \to c > 0$. Basic results from random matrix theory (RMT) are used here to estimate ρ consistently. This procedure is simple and quite flexible and has been generalized in various directions (e.g., in (Couillet & McKay, 2014) with a robust statistics approach). However, the method only applies a naive homothetic map to each $\lambda_i(\hat{C})$ of $\hat{C}$ in order to better estimate $\lambda_i(C)$. A strong hope to recover a better approximation of the $\lambda_i(C)$'s then arose from (Silverstein & Bai, 1995; Silverstein & Choi, 1995) that provide a random matrix result relating directly the limiting eigenvalue distributions of C and $\hat{C}$. Unfortunately, while estimating the $\lambda_i(\hat{C})$'s from the $\lambda_i(C)$'s is somewhat immediate, estimating the $\lambda_i(C)$'s backward from the $\lambda_i(\hat{C})$'s is a difficult task. (El Karoui et al., 2008) first proposed an optimization algorithm to numerically solve this problem, however with little success as the method is quite unstable and has rarely been efficiently reproduced. (Mestre, 2008) later offered a powerful idea, based on contour integral, to consistently estimate linear functionals $\frac{1}{n}\sum_{i=1}^n f(\lambda_i(C))$ from the $\lambda_i(\hat{C})$'s. But f is constrained to be very smooth (complex analytic) which prevents the estimation of the individual $\lambda_i(C)$'s. Recently, Ledoit and Wolf took over the work of El Karoui, which they engineered to obtain a more efficient numerical method, named QuEST (Ledoit & Wolf, 2015). Rather than inverting the Bai–Silverstein equations, the authors also proposed, with the same approach, to estimate the $\lambda_i(C)$'s by minimizing a Frobenius norm distance (Ledoit & Wolf, 2015) (named QuEST1 in the present article) or a Stein loss (Ledoit et al., 2018) (QuEST2 here).

These methods, although more stable than El Karoui's initial approach, however suffer several shortcomings: (i) they are still algorithmically involved as they rely on a series of fine-tuned optimization schemes, and (ii) they are only adaptable to few error metrics (Frobenius, Stein).

Inspired by Mestre's approach and the recent work (Couil-

[1]CentraleSupélec, University ParisSaclay, France [2]GIPSA-lab, University Grenoble-Alpes, France [3]LISTIC, University Savoie Mont-Blanc, France. Correspondence to: Malik Tiomoko <malik.tiomoko@gipsa-lab.grenoble-inp.fr>, Florent Bouchard <florent.bouchard@gipsa-lab.grenoble-inp.fr>, Guillaume Ginholac <guillaume.ginholac@univ-smb.fr>, Romain Couillet <romain.couillet@gipsa-lab.grenoble-inp.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

let et al., 2018), this article proposes a different procedure consisting in (i) writing C as the solution to $\mathrm{argmin}_{M \succ 0} \delta(M, C)$ for a wide range of metrics δ (Fisher, Batthacharyya, Stein's loss, Wasserstein, etc.), (ii) based on (Couillet et al., 2018), using the fact that $\delta(M, C) - \hat{\delta}(M, X) \to 0$ for some consistent estimator $\hat{\delta}$, valid for all deterministic M and samples $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$ having zero mean and covariance C, and (iii) proceeding to a gradient descent on $\hat{\delta}$ rather than on the unknown δ itself. With appropriate adaptations, the estimation of C^{-1} is similarly proposed by solving instead $\mathrm{argmin}_{M \succ 0} \delta(M, C^{-1})$.

While only theoretically valid for matrices M independent of X, and thus only in the first steps of the gradient descent, the proposed method has several advantages: (i) it is easy to implement and technically simpler than QuEST, (ii) it is adaptable to a large family of distances and divergences, and, most importantly, (iii) simulations suggest that it systematically outperforms the SCM and is competitive with, if not better than, QuEST.

The remainder of the article is organized as follows. Section 2 introduces preliminary notions and concepts on which are hinged our proposed algorithms, thereafter described in Section 3. Section 4 provides experimental validations and applications, including an improved version of LDA/QDA based on the proposed enhanced estimates.

Reproducibility. Matlab codes for the proposed estimation algorithms are available as supplementary materials and are based on Manopt, a Matlab toolbox for optimization on manifolds (Boumal et al., 2014).

2. Preliminaries

Let n random vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ be of the form $x_i = C^{\frac{1}{2}} z_i$ for some positive definite matrix $C \in \mathbb{R}^{p \times p}$ and $z_1, \ldots, z_n \in \mathbb{R}^p$ independent random vectors of independent entries with $\mathrm{E}[[z_i]_j] = 0$ and $\mathrm{E}[|[z_i]_j|^2] = 1$. We further assume the following large dimensional regime for n and p.

Assumption 1 (Growth Rate)**.** *As* $n \to \infty$, $p/n \to c \in (0, 1)$ *and* $\limsup_p \max\{\|C^{-1}\|, \|C\|\} < \infty$ *for* $\|\cdot\|$ *the matrix operator norm.*

Our objective is to estimate C and C^{-1} based on $x_1, \ldots, x_n$ under the above large p, n regime. For simplicity of exposition and readability, we mostly focus on the estimation of C and more briefly discuss that of C^{-1}.

Our approach relies on the following elementary idea:

$$C \equiv \mathrm{argmin}_{M \succ 0} \delta(M, C)$$

where, for some function f,

$$\delta(M, C) \equiv \frac{1}{p} \sum_{i=1}^{p} f(\lambda_i(M^{-1}C)) \tag{1}$$

is a divergence (possibly a squared distance $\delta = d^2$) between the positive definite matrices M and C, depending only on the eigenvalues of $M^{-1}C$. Among divergences satisfying this condition, we find the *natural Riemannian distance* d^2_{R} (Bhatia, 2009), which corresponds to the Fisher metric for the multivariate normal distribution (Skovgaard, 1984); the *Battacharyya distance* d^2_{B} (Sra, 2013), which is close to the natural Riemannian distance while numerically less expensive; the *Kullback-Leibler divergence* δ_{KL}, linked to the likelihood and studied for example in (Moakher, 2012); the *Rényi divergence* $\delta_{\alpha\mathrm{R}}$ for Gaussian x_i's (Van Erven & Harremos, 2014); etc.[1] Table 1 reports the explicit values of f for these divergences.

Since $\delta(M, C)$ is not accessible as C is unknown, our approach exploits an estimator for $\delta(M, C)$ which is consistent in the large n, p regime of Assumption 1.

To this end, our technical arguments are fundamentally based on random matrix theory, and notably rely on the so-called *Stieljes transform* of eigenvalue distributions. For an arbitrary real-supported probability measure θ, the Stieltjes transform $m_\theta : \mathbb{C} \setminus \mathrm{supp}(\theta) \to \mathbb{C}$ is defined as

$$m_\theta(z) = \int \frac{\theta(dt)}{t - z}.$$

The key interest of the Stieltjes transform in this article is that it allows one to relate the distributions of the eigenvalues of C and $\hat{C}$ as $p, n \to \infty$ (Silverstein & Bai, 1995). More specifically here, for arbitrary deterministic matrices M, Stieltjes transform relations connect the empirical spectral (i.e., eigenvalue) distribution ν_p of $M^{-1}C$ to the empirical spectral distribution μ_p of $M^{-1}\hat{C}$ (Couillet et al., 2018), defined as

$$\mu_p \equiv \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(M^{-1}\hat{C})} \text{ and } \nu_p \equiv \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(M^{-1}C)}.$$

The connecting argument goes as follows: first, from Cauchy's integral formula (stating that $f(t) = \frac{1}{2\pi i} \oint_\Gamma f(z)/(t-z)dz$ for Γ a complex contour enclosing t), the metric $\delta(M, C)$ in (1) relates to the Stieltjes transform $m_{\nu_p}(z; M)$ through

$$\delta(M, C) = \frac{1}{2\pi i} \oint_\Gamma f(z) m_{\nu_p}(z; M) dz \tag{2}$$

[1] The Frobenius distance does not fall into this setting but has already largely been investigated and optimized.

Divergences	$f(z)$
d_{R}^2	$\log^2(z)$
d_{B}^2	$-\frac{1}{4}\log(z)+\frac{1}{2}\log(1+z)-\frac{1}{2}\log(2)$
δ_{KL}	$\frac{1}{2}z-\frac{1}{2}\log(z)-\frac{1}{2}$
$\delta_{\alpha\mathrm{R}}$	$\frac{-1}{2(\alpha-1)}\log(\alpha+(1-\alpha)z)+\frac{1}{2}\log(z)$

Table 1. Distances d and divergences δ, and their corresponding $f(z)$ functions.

for $\Gamma \subset \mathbb{C}$ a (positively oriented) contour surrounding the eigenvalues of $M^{-1}C$. The notation $m_{\nu_p}(z;M)$ reminds the dependence of $m_{\nu_p}(z)$ in the matrix M.

Then, by exploiting the relation between the Stieltjes transforms μ_{ν_p} and m_{μ_p}, it is shown in (Couillet et al., 2018) that, under Assumption 1, for all deterministic M of bounded operator norm,

$$\delta(M,C)-\hat{\delta}(M,X)\to 0 \tag{3}$$

almost surely, where $X=[x_1,\ldots,x_n]$ and

$$\hat{\delta}(M,X)\equiv\frac{1}{2\pi\imath c}\oint_{\hat{\Gamma}} G\left(-m_{\tilde{\mu}_p}(z;M)\right)dz \tag{4}$$

with G such that $G'(z)\equiv g(z)=f(1/z)$, $\hat{\Gamma}$ a contour surrounding the support of the almost sure limiting eigenvalue distribution of $M^{-1}\hat{C}$ and $\tilde{\mu}_p=\frac{p}{n}\mu_p+(1-\frac{p}{n})\delta_0$ (and thus $m_{\tilde{\mu}_p}(z)=cm_{\mu_p}(z)+(1-\frac{p}{n})/z$). Note that, by the linearity of G in (4), it is sufficient in practice to evaluate $\hat{\delta}(M,X)$ for elementary functions (such as $f(z)=z$, $f(z)=\log(z)$, etc.) in order to cover most distances and metrics of interest (see again Table 1). Table 2 reports the values of G for such atomic functions f.

Our main idea is to estimate C by minimizing the approximation $\hat{\delta}(M,X)$ of $\delta(M,C)$ over M. However, it is important to note that, as discussed in (Couillet et al., 2018), the random quantity $\hat{\delta}(M,X)$ may be negative with nonzero probability. As such, minimizing $\hat{\delta}(M,X)$ over M may lead to negative solutions. Our proposed estimation method therefore consists in approximating C by the solution to the optimization problem

$$\mathrm{argmin}_{M\succ 0}\{h_X(M)\},\ \text{where } h_X(M)\equiv(\hat{\delta}(M,X))^2. \tag{5}$$

3. Methodology and Main Results

3.1. Estimation Method

We solve (5) via a gradient descent algorithm in the Riemannian manifold S_n^{++} of positive definite $n\times n$ matrices.

To evaluate the gradient $\nabla h_X(M)$ of h_X at $M\in S_n^{++}$, recall that on S_n^{++} the differential $\mathrm{D}h_X(M)[\xi]$ of the functional $h_X: S_n^{++}\to\mathbb{R}^+$, at position $M\in S_n^{++}$ and in the direction of $\xi\in S_n$ (the Riemannian manifold of symmetric $n\times n$ matrices), is given by (Absil et al., 2009)

$$\mathrm{D}h_X(M)[\xi]=\langle\nabla h_X(M),\xi\rangle_M^{S_n^{++}}$$

where $\langle\cdot,\cdot\rangle_\cdot^{S_n^{++}}$ is the Riemannian metric defined through

$$\langle\eta,\xi\rangle_M^{S_n^{++}}=\mathrm{tr}\left(M^{-1}\eta M^{-1}\xi\right).$$

Differentiating $\hat{\delta}^2(M,X)$ at M in the direction ξ yields:

$$\begin{aligned}&\mathrm{D}h_X(M)[\xi]\\&=\frac{-\hat{\delta}(M,X)}{\pi\imath c}\oint_\Gamma g(-m_{\tilde{\mu_p}}(z,M))\mathrm{D}m_{\tilde{\mu_p}}(z,M)[\xi]dz.\end{aligned}$$

By using the fact that

$$\begin{aligned}&\mathrm{D}m_{\tilde{\mu_p}}(z,M)[\xi]\\&=\frac{c}{p}\mathrm{Dtr}\left(\left[M^{-1}\hat{C}-zI_p\right]^{-1}\right)[\xi]\\&=\frac{c}{p}\mathrm{tr}\left(M^{-1}\hat{C}\left[M^{-1}\hat{C}-zI_p\right]^{-2}M^{-1}\xi\right)\\&=\left\langle\frac{c}{p}\mathrm{sym}\left(\hat{C}\left[M^{-1}\hat{C}-zI_p\right]^{-2}\right),\xi\right\rangle_M^{S_n^{++}}\end{aligned}$$

where $\mathrm{sym}(A)=\frac{1}{2}(A+A^\mathsf{T})$ is the symmetric part of $A\in\mathbb{R}^{p\times p}$, we retrieve the gradient of $h_X(M)$ as

$$\begin{aligned}&-\imath\pi p\frac{\nabla h_X(M)}{\hat{\delta}(M,X)}\\&=\oint_{\hat{\Gamma}} g\left(-m_{\tilde{\mu}_p}(z;M)\right)\mathrm{sym}\left(\hat{C}(M^{-1}\hat{C}-zI_p)^{-2}\right)dz\end{aligned} \tag{6}$$

(recall that the right-hand side still depends on X implicitly through $\tilde{\mu}_p$ and $\hat{C}$).

$f(z)$	$G(z)$
$\log^2(z)$	$z\left(\log^2(z)-2\log(z)+2\right)$
$\log(z)$	$-z\log(z)+z$
$\log(1+sz)$	$s\log(s+z)+z\log\left(\frac{s+z}{z}\right)$
z	$\log(z)$

$f(z)$	$F(z)$
$\log^2(z)$	$z\left(\log^2(z)-2\log(z)+2\right)$
$\log(z)$	$z\log(z)-z$
$\log(1+sz)$	$\left(\frac{1}{s}+z\right)\log(1+sz)-z$
z	$\frac{1}{2}z^2$

Table 2. Values of $G(z)$ and $F(z)$ for "atomic" $f(z)$ functions used in most distances and divergences under study; here $s>0$ and $z\in\mathbb{C}$.

Once ∇h_X estimated, every gradient descent step in S_n^{++} corresponds to a small displacement on the geodesic starting at M and towards $-\nabla h_X(M)$, defined as the curve

$$\begin{aligned}\mathbb{R}_+ &\to S_n^{++} \\ t &\mapsto M^{\frac{1}{2}} \exp\left(-tM^{-\frac{1}{2}}\nabla h_X(M)M^{-\frac{1}{2}}\right) M^{\frac{1}{2}}\end{aligned}$$

where, for $A = U\Lambda U^\mathsf{T} \in S_n^{++}$ in its spectral decomposition, $\exp(A) \equiv U\exp(\Lambda)U^\mathsf{T}$ (with exp understood here applied entry-wise on the diagonal elements of Λ).

That is, letting $M_0, M_1, \ldots$ and $t_0, t_1, \ldots$ be the successive iterates and step sizes of the gradient descent, we have, for some given initialization $M_0 \in S_n^{++}$,

$$M_{k+1} = M_k^{\frac{1}{2}} \exp\left(-t_k M_k^{-\frac{1}{2}}\nabla h_X(M_k)M_k^{-\frac{1}{2}}\right) M_k^{\frac{1}{2}}. \tag{7}$$

Our proposed method is summarized as Algorithm 1.

Algorithm 1 Proposed estimation algorithm.

Require $M_0 \in C_n^{++}$.

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

Until Convergence.

Return M.

We conclude this section by an important remark on the fundamental limitations of the proposed algorithm.

Remark 1 (Approximation of $\delta(M_k, C)$ by $\hat{\delta}(M_k, X)$)**.** It is fundamental to understand the result from (Couillet et al., 2018) at the heart of the proposed method. There, it is precisely shown that, for every deterministic sequence of matrices $\{M^{(p)},\ p = 1, 2, \ldots\}$ and $\{C^{(p)},\ p = 1, 2, \ldots\}$, with $M^{(p)}, C^{(p)} \in \mathbb{R}^{p\times p}$ and $\max(\|C^{(p)}\|, \|M^{(p)}\|) < K$ for some constant K independent of p, we have that, for $X^{(p)} = [x_1^{(p)}, \ldots, x_n^{(p)}]$ with $x_i^{(p)} = C^{(p)\frac{1}{2}} z_i^{(p)}$ and $z_i^{(p)}$ i.i.d. vectors of i.i.d. zero mean and unit variance entries,

$$\delta(M^{(p)}, C^{(p)}) - \hat{\delta}(M^{(p)}, X^{(p)}) \to 0$$

almost surely as $n, p \to \infty$ and $p/n \to c \in (0, 1)$. This result seems to suggest that $\hat{\delta}(M_k, X)$ in our algorithm is a good approximation for the sought for $\delta(M_k, C)$. This however only holds true so long that M_k *is independent of* X which clearly does not stand when proceeding to successive gradient descent steps in the direction of $\nabla h_X(M)$ which depends explicitly on X. As such, while initializations with, say, $M_0 = I_p$, allow for a close approximation of $\delta(M_k, C)$ in the very first steps of the descent, for larger values of k, the descent is likely to drive the optimization in less accurate directions.

Remark 1 is in fact not surprising. Indeed, finding the minimum of $\delta(M, C)$ over $M \succ 0$ would result in finding C, which cannot be achieved for unconstrained matrices C and for non vanishing values of p/n. Figure 1 provides a typical evolution of the distance $\delta(M_k, C)$ versus its approximation $\hat{\delta}(M_k, X)$ at the successive steps $k = 1, 2, \ldots$ of Algorithm 1, initialized at $M_0 = I_p$. As expected, the difference $|\hat{\delta}(M_k, X) - \delta(M_k, C)|$, initially small (at $k = 1$, $\hat{\delta}(I_p, X) \simeq \delta(I_p, C)$), increases with k, until the gradient vanishes and the divergence $\delta(M_k, C)$ converges.

3.2. Practical Implementation

In order to best capture the essence of Algorithm 1, as well as its various directions of simplification and practical fast implementation, a set of important remarks are in order.

First note that the generic computation of M_{k+1} in Equation (7) may be numerically expensive, unless M_k and $\nabla h_X(M)$ share the same eigenvectors. In this case, letting $M_k = U\Omega_k U^\mathsf{T}$ and $\nabla_X h(M_k) = U\Delta_k U^\mathsf{T}$, we have the recursion

$$[\omega_{k+1}]_i = [\omega_k]_i \exp\left(-t\frac{[\delta_k]_i}{[\omega_k]_i}\right) \tag{8}$$

where $\delta_k = \mathrm{diag}(\Delta_k)$ and $\omega_k = \mathrm{diag}(\Omega_k)$.

In particular, if $M_0 = \alpha I_n + \sqrt{1-\alpha^2}\hat{C}$ is a linear shrinkage for some $\alpha \in [0, 1]$, we immediately find that, for all $k \geq 0$,

$$M_k = \hat{U}\Omega_k\hat{U}^\mathsf{T}$$

where $\hat{U} \in \mathbb{R}^{p\times p}$ are the eigenvectors of $\hat{C}$ (i.e., in its spectral decomposition, $\hat{C} = \hat{U}\hat{\Lambda}\hat{U}^\mathsf{T}$) and Ω_k is recursively defined through (8).

This shows that, initialized as such, the ultimate limiting estimate M_∞ (i.e., the limit of M_k) of C shares the same eigenvectors as $\hat{C}$, and thus reduces to a "non-linear shrinkage" procedure, similar to that of (Ledoit & Wolf, 2015). Extensive simulations in fact suggest that, if initialized randomly (say with M_0 a random Wishart matrix), after a few iterates, the eigenvectors of M_k do converge to those of $\hat{C}$ (see further discussions in Section 4). As such, for computational ease, we suggest to initialize the algorithm with $M_0 = I_p$ or with M_0 a linear shrinkage of $\hat{C}$.

A further direction of simplification of Algorithm 1 relates to the fact that, for generic values of M_0 (notably having eigenvectors different from those of $\hat{C}$), Equation (7) is computationally expensive to evaluate. A second-order simplification for small t is often used in practice (Jeuris

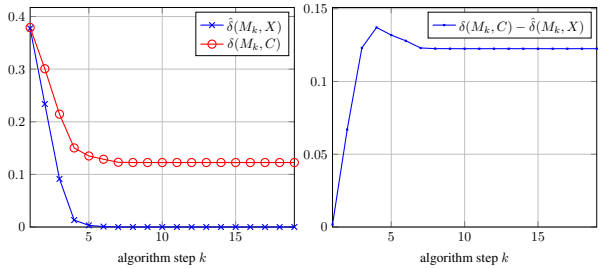

Figure 1. (left) Evolution of the Fisher distance $\delta(M_k, C)$ versus $\hat{\delta}(M_k, X)$ for $k = 1, 2, \ldots$, initialized to $M_0 = I_p$,. (right) Evolution of $\delta(M_k, C) - \hat{\delta}(M_k, X)$.

et al., 2012), as follows

$$M_{k+1} = M_k - t\nabla h_X(M_k) + \frac{t^2}{2}\nabla h_X(M_k) M_k^{-1} \nabla h_X(M_k) + O(t^3).$$

Simulations with this approximation suggest almost no difference in either the number of steps until convergence or accuracy of the solution.

3.3. Estimation of C^{-1}

In our framework, estimating C^{-1} rather than C can be performed by minimizing $\delta(M, C^{-1})$ instead of $\delta(M, C)$. In this case, under Assumption 1, (3) now becomes

$$\delta(M, C^{-1}) - \hat{\delta}^{\text{inv}}(M, X) \to 0$$

almost surely, for every deterministic M of bounded operator norm and $X = [x_1, \ldots, x_n]$, where

$$\hat{\delta}^{\text{inv}}(M, X) \equiv \frac{1}{2\pi\imath c} \oint_{\hat{\Gamma}} F\left(-m_{\tilde{\mu}_p^{\text{inv}}}(z; M)\right) dz$$

for F such that $F'(z) \equiv f(z)$, $\hat{\Gamma}$ a contour surrounding the support of the almost sure limiting eigenvalue distribution of $M\hat{C}$ and $\tilde{\mu}_p^{\text{inv}} = \frac{p}{n}\mu_p^{\text{inv}} + (1 - \frac{p}{n})\delta_0$, where $\mu_p^{\text{inv}} \equiv \frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i(M\hat{C})}$. The cost function to minimize under this setting is now given by $h^{\text{inv}}(M) \equiv (\hat{\delta}^{\text{inv}}(M, X))^2$ with gradient $\nabla h_X^{\text{inv}}(M)$ satisfying

$$\imath\pi p \frac{\nabla h_X^{\text{inv}}(M)}{\hat{\delta}^{\text{inv}}(M, X)} = \oint_{\hat{\Gamma}} f\left(-m_{\tilde{\mu}_p^{\text{inv}}}(z; M)\right) \text{sym}\left(M\hat{C}(M\hat{C} - zI_p)^{-2}M\right) dz.$$

With these amendments, Algorithm 1 can be adapted to the estimation of C^{-1}. Table 2 provides the values of F for the atomic functions f of interest.

3.4. Application to Explicit Metrics

Algorithm 1 is very versatile as it merely consists in a gradient descent method for various metrics f through adaptable definitions of the function $h_X(M) = \hat{\delta}(M, X)^2$ and its resulting gradient. Yet, because of the integral form assumed by the gradient (Equation (6)), a possibly computationally involved complex integration needs to be numerically performed at each gradient descent step.

In this section, we specify closed-form expressions for the gradient for the atomic f functions of Table 2 (which is enough to cover the list of divergences in Table 1).

3.4.1. Estimation of C

Let us denote

$$\nabla h_X(M) \equiv 2\hat{\delta}(M, X) \cdot \text{sym}\left(\hat{C} \cdot V \Lambda_\nabla V^{-1}\right)$$

where V are the eigenvectors of $M^{-1}\hat{C}$ and we will determine Λ_∇ for each function f. Again, we recall from the discussion in Section 3.2, that $V = \hat{U}$ the eigenvectors of $\hat{C}$ if M shares the same eigenvectors as $\hat{C}$ (which thus avoids evaluating the eigenvectors V of $M_k^{-1}\hat{C}$ at each step k of the algorithm).

For readability in the following, let us denote $\lambda_i \equiv \lambda_i(M^{-1}\hat{C})$, $i \in \{1, \ldots, p\}$, the eigenvalues of the matrix $M^{-1}\hat{C}$ and $\xi_1, \ldots, \xi_p$ the eigenvalues of

$$\Lambda - \frac{1}{n}\sqrt{\lambda}\sqrt{\lambda}^\mathsf{T}$$

with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and $\lambda = (\lambda_1, \ldots, \lambda_p)^\mathsf{T}$. Finally, for $s > 0$, let $\kappa_s \in (-1/(s(1 - p/n)), 0)$ be the unique negative number t solution of the equation (see (Couillet et al., 2018) for details)

$$m_{\tilde{\mu}_p}(t) = -s.$$

With these notations at hand, following the derivations in (Couillet et al., 2018) (detailed in supplementary material), we have the following determinations for Λ_∇.

Proposition 1 (Case $f(t) = t$)**.** For $f(t) = t$,

$$[\Lambda_\nabla]_{kk} = -\frac{1}{c} + \frac{1}{p}\sum_{i=1}^p \frac{1}{m'_{\tilde{\mu}_p}(\xi_i)\,(\lambda_k - \xi_i)^2}$$

with $m'_{\tilde{\mu}_p}$ the derivative of $m_{\tilde{\mu}_p}$.

Proposition 2 (Case $f(t) = \log(t)$)**.** For $f(t) = \log(t)$,

$$[\Lambda_\nabla]_{kk} = \frac{-1}{p\lambda_k}.$$

Proposition 3 (Case $f(t) = \log(1 + st)$)**.** For $s > 0$ and $f(t) = \log(1 + st)$,

$$[\Lambda_\nabla]_{kk} = \frac{-1}{p(\lambda_k - \kappa_s)}.$$

Proposition 4 (Case $f(t) = \log^2(t)$)**.** For $f(t) = \log^2(t)$,

$$[\Lambda_\nabla]_{kk} = \frac{2}{p}\log(\lambda_k)\left[\sum_{i=1}^{p}\frac{1}{\lambda_k-\xi_i} - \sum_{\substack{i=1\\i\neq k}}^{p}\frac{1}{\lambda_k-\lambda_i} - \frac{1}{\lambda_k}\right] - \frac{2}{p}\sum_{i=1}^{p}\frac{\log(\xi_i)}{\lambda_k-\xi_i} + \frac{2}{p}\sum_{\substack{i=1\\i\neq k}}^{p}\frac{\log(\lambda_i)}{\lambda_k-\lambda_i} - \frac{2-2\log(1-c)}{p\lambda_k}.$$

These results are mostly achieved by residue calculus for entire analytic functions f or by exploiting more advanced complex integration methods (in particular branch-cut methods) for more challenging functions (involving logarithms in particular).

Combining these formulas provides an analytical expression for the gradient of all aforementioned divergences and square distances, for the estimation of C.

3.4.2. Estimation of C^{-1}

Similarly, for the problem of estimating C^{-1}, recalling Remark 3.3, we may denote

$$\nabla h_X^{\text{inv}}(M) \equiv 2\hat{\delta}^{\text{inv}}(M,X)\cdot \text{sym}\left(M\cdot V_{\text{inv}}\Lambda_\nabla^{\text{inv}}V_{\text{inv}}^{-1}\right)$$

with V_{inv} the eigenvectors of $M\hat{C}$.

We redefine in this section $\lambda_i \equiv \lambda_i(M\hat{C})$, and $\xi_1,\ldots,\xi_p$ the eigenvalues of $\Lambda - \frac{1}{n}\sqrt{\lambda}\sqrt{\lambda}^\mathsf{T}$ with $\Lambda = \text{diag}(\lambda_1,\ldots,\lambda_p)$ and $\lambda = (\lambda_1,\ldots,\lambda_p)^\mathsf{T}$. Again, for $s>0$, let $\kappa_s < 0$ be the only negative real number t solution of

$$m_{\tilde{\mu}_p^{\text{inv}}}(t) = -\frac{1}{s}.$$

With the same approach as in the previous section, we here obtain the following values for $\Lambda_\nabla^{\text{inv}}$.

Proposition 5 (Case $f(t) = t$)**.** For $f(t) = t$,

$$[\Lambda_\nabla^{\text{inv}}]_{kk} = -\frac{1-c}{p\lambda_k}.$$

Proposition 6 (Case $f(t) = \log(t)$)**.** For $f(t) = \log(t)$,

$$[\Lambda_\nabla^{\text{inv}}]_{kk} = -1$$

Proposition 7 (Case $f(t) = \log(1+st)$)**.** For $s>0$ and $f(t) = \log(1+st)$,

$$[\Lambda_\nabla^{\text{inv}}]_{kk} = \frac{\lambda_k}{\lambda_k-\kappa_s} - 1$$

Proposition 8 (Case $f(t) = \log^2(t)$)**.** For $f(t) = \log^2(t)$,

$$[\Lambda_\nabla^{\text{inv}}]_{kk} = -\frac{2}{p}\log(\lambda_k)\left[\sum_{\substack{i=1\\i\neq k}}^{p}\frac{\lambda_k}{\lambda_k-\xi_i} - \sum_{\substack{i=1\\i\neq k}}^{p}\frac{\lambda_k}{\lambda_k-\lambda_i} - 1\right] + \frac{2}{p}\sum_{i=1}^{p}\frac{\lambda_k\log(\xi_i)}{\lambda_k-\xi_i} - \frac{2}{p}\sum_{\substack{i=1\\i\neq k}}^{p}\frac{\lambda_k\log(\lambda_i)}{\lambda_k-\lambda_i} + \frac{2}{p} - \frac{2}{p}\log(1-c).$$

4. Experimental Results

This section introduces experimental results on the direct application of our proposed method to the estimation of C and C^{-1} as well as on its use as a plug-in estimator in more advanced procedures, here in the scope of linear and quadratic discriminant analyses (LDA/QDA).

4.1. Validation on synthetic data

In this first section, we provide a series of simulations on the estimation of C and C^{-1} based on the Fisher distance and for several examples of genuine matrices C. Similar results and conclusions were obtained for the other metrics discussed above (the KL divergence and the square Battacharyya distance especially) which are thus not explicitly reported here. The interested reader can refer to the code provided by the authors for self experimentation as supplementary material.

Our preference for the Fisher distance for fair comparisons lies in the fact that it is the "natural" Riemannian distance to compare covariance matrices in S_n^{++}, therefore in entire agreement with the proposed estimation strategy through gradient descents in S_n^{++}. Besides, for this specific case, Theorem 4 in (Smith, 2005) establishes an exact and very straightforward formula for the Cramer-Rao bound (CRB) on *unbiased* estimators of C. Although the compared estimators of C are likely all biased and that M_0 initializations may by chance bring additional information disrupting a formally fair CRB comparison, the CRB at least provides an indicator of relevance of the estimators.

The examples of covariance matrix C under consideration in the following are:
(i) [Wishart] a random (p-dimensional) standard Wishart matrix with $2p$ degrees of freedom,
(ii) [Toeplitz a] the Toeplitz matrix defined by $C_{ij} = a^{|i-j|}$,
(iii) [Discrete] a matrix C with uniform eigenvector distribution and eigenvalues equal to .1, 1, 3, 4 each with multiplicity $p/4$.

Figure 2 (for the estimation of C) and Figure 3 (for C^{-1}) report comparative performances on the aforementioned C

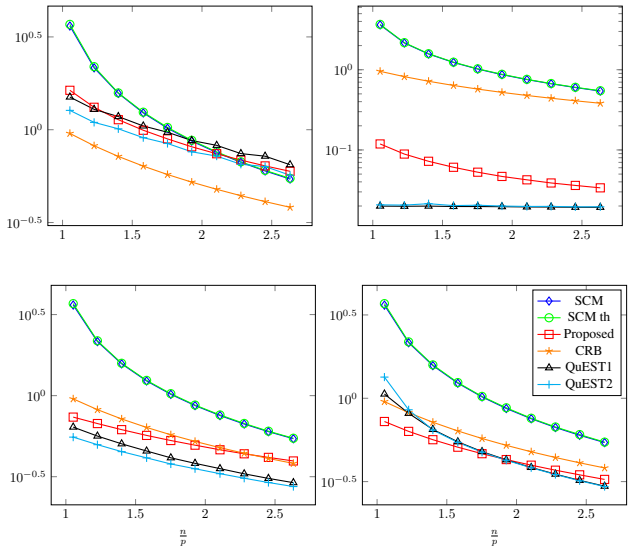

Figure 2. Fisher distance of estimates of C, initialized at linear-shrinkage. From top-left to bottom-right: Wishart, Toeplitz 0.1, Toeplitz 0.9, Discrete. "SCM th" defined in Remark 2. Averaged over 100 random realizations of X, $p = 200$.

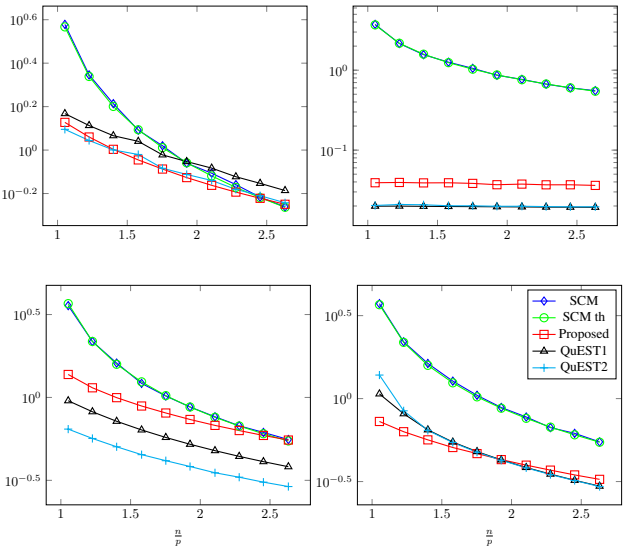

Figure 3. Fisher distance of estimates of C^{-1}, initialized at linear-shrinkage. From top-left to bottom-right: Wishart, Toeplitz 0.1, Toeplitz 0.9, Discrete. Averaged over 100 random realizations of X, $p = 200$.

matrices for the SCM, QuEST1, QuEST2, and our proposed estimator, the latter three being initialized at M_0 the shrinkage estimation from (Ledoit & Wolf, 2004) (consistent with the choice made for QuEST1, QuEST2 in (Ledoit & Wolf, 2015; Ledoit et al., 2018)). In the figures, "SCM th" refers to the asymptotic analytical approximation of $\delta(C, \hat{C})$ as defined in Remark 2. It is observed that, while the SCM never reaches the unbiased CRB, in many cases the QuESTx estimators and our proposed method overtake the CRB, sometimes significantly so. The Wishart matrix case seems more challenging from this perspective. In terms of performances, both our proposed method and QuESTx perform competitively and systematically better than the sample covariance matrix.

Remark 2 (Consistent estimator for $\delta(C, \hat{C})$)**.** With the same technical tools from (Couillet et al., 2018), it is straightforward to estimate the distance $\delta(C, \hat{C})$. Indeed, $\delta(C, \hat{C}) = \frac{1}{2\pi i} \oint_\Gamma f(z) m_\gamma(z)$ for $m_\gamma(z)$ the Stieljes transform of the eigenvalue distribution of $C^{-1}\hat{C}$; the limiting distribution of the latter is the popular Marcenko-Pastur law (Marčenko & Pastur, 1967), the expression of which is well known. The estimate is denoted "SCM th" in Figures 2–3. The observed perfect match between limiting theory and practice confirms the consistency of the random matrix approach even for not too large p, n.

4.2. Application to LDA/QDA

As pointed out in the introduction, the estimation of the covariance and inverse covariance matrices of random vectors are at the core of a wide range of applications in statistics, machine learning and signal processing. As a basic illustrative example, we focus here on linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). Both exploit estimates covariance matrices of the data or their inverse in order to perform the classification.

Suppose $x_1^{(1)}, \ldots, x_{n_1}^{(1)} \sim N(\mu_1, C_1)$ and $x_1^{(2)}, \ldots, x_{n_2}^{(2)} \sim N(\mu_2, C_2)$ are two sets of random independent p-dimensional training vectors forming two classes of a Gaussian mixture. The objective of LDA and QDA is to estimate the probability for an arbitrary random vector x to belong to either class by replacing the genuine means μ_a and covariances C_a by sample estimates, with in the case of LDA the underlying (possibly erroneous) assumption that $C_1 = C_2$. Defining C as $C \equiv \frac{n_1}{n_1+n_2} C_1 + \frac{n_2}{n_1+n_2} C_2$, the classification rules for LDA and QDA for data point x depend on the signs of the respective quantities:

$$\delta_x^{\text{LDA}} = (\hat{\mu}_1 - \hat{\mu}_2)^\mathsf{T} \check{C}^{-1} x + \frac{1}{2}\mu_2^\mathsf{T} \check{C}^{-1} \mu_2 - \frac{1}{2}\hat{\mu}_1^\mathsf{T} \check{C}^{-1} \hat{\mu}_1$$

$$\delta_x^{\text{QDA}} = \frac{1}{2} x^\mathsf{T} \left(\check{C}_2^{-1} - \check{C}_1^{-1} \right) x + \left(\hat{\mu}_1^\mathsf{T} \check{C}_1^{-1} - \hat{\mu}_2^\mathsf{T} \check{C}_2^{-1} \right) x + \frac{1}{2}\hat{\mu}_2^\mathsf{T} \check{C}_2^{-1} \hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1^\mathsf{T} \check{C}_1^{-1} \hat{\mu}_1 + \frac{1}{2} \log\det \frac{\check{C}_1^{-1}}{\check{C}_2^{-1}} - \log \frac{n_2}{n_1}$$

where $\hat{\mu}_a \equiv \frac{1}{n_a} \sum_{i=1}^{n_a} x_i^{(a)}$ is the sample estimate of μ_a and $\check{C}_a^{-1}$ are some estimate of C_a^{-1}, while $\check{C} \equiv \frac{n_1}{n_1+n_2} \check{C}_1 + \frac{n_2}{n_1+n_2} \check{C}_2$ with $\check{C}_a$ the estimation of C_a. As such, in the following simulations, LDA will exclusively exploit estimations of C_1 and C_2 (before inverting their estimated average), while QDA will focus on estimating directly the inverses C_1^{-1} and C_2^{-1}.

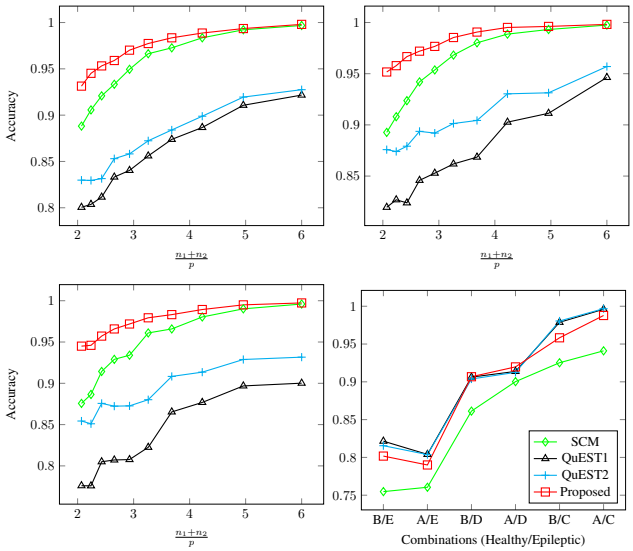

Figure 4. Mean accuracy obtained over 10 realizations of LDA classification. From left to right and top to bottom: C_1 and C_2 are respectively Wishart/Wishart (independent), Wishart/Toeplitz-0.2, Toeplitz-0.2/Toeplitz-0.4, and real application to EEG data.

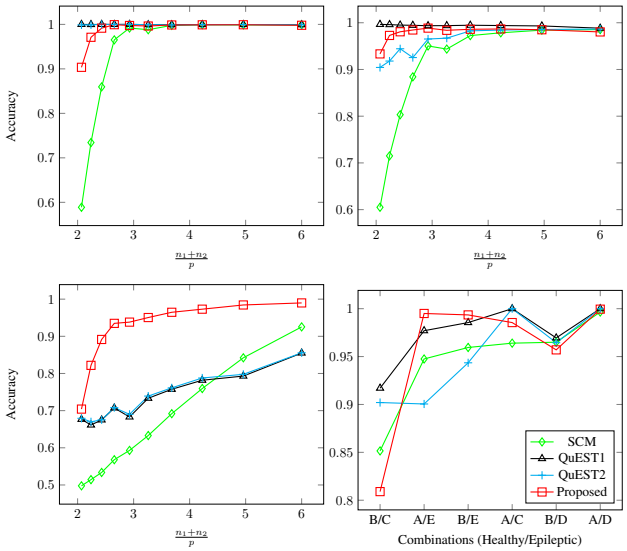

Figure 5. Mean accuracy obtained over 10 realizations of QDA classification. From left to right and top to bottom: C_1 and C_2 are respectively Wishart/Wishart (independent), Wishart/Toeplitz-0.2, Toeplitz-0.2/Toeplitz-0.4, and real application to EEG data.

The first three displays in Figures 4 and 5 compare the accuracies of the LDA/QDA algorithms for C_1 and C_2 chosen among Wishart and Toeplitz matrices, and for $\mu_2 = \mu_1 + \frac{80}{p}$ for the LDA and $\mu_2 = \mu_1 + \frac{1}{p}$ for the QDA settings (in order to avoid trivial classification). The bottom right displays are applications to real EEG data extracted from the dataset in (Andrzejak et al., 2001). The dataset contains five subsets (denoted A-E). Sets A and B were collected from healthy volunteers while C, D, E were collected from epileptic patients. The graph presents all combinations of binary classes between healthy volunteers and epilectic subjects (e.g., A/E for subsets A and E). There we observe that, for most considered settings, our proposed algorithm almost systematically outperforms competing methods, with QuEST1 and QuEST2 exhibiting a much less stable behavior and particularly weak performances in all synthetic scenarios.

5. Discussion and Concluding Remarks

Based on elementary yet powerful contour integration techniques and random matrix theory, we have proposed in this work a systematic framework for the estimation of covariance and precision matrices. Unlike alternative state-of-the-art techniques that attempt to invert the fundamental Bai–Silverstein equations (Silverstein & Bai, 1995), our proposed method relies on a basic gradient descent approach in S_n^{++} that, in addition to performing competitively (if not better), is computationally simpler.

While restricted to metrics depending on the eigenvalues of products of covariance matrices, our approach may be flexibly adapted to further matrix divergences solely depending on eigenvalue relations. The same framework can notably be applied to the Wasserstein distance between zero-mean Gaussian laws. A reservation nonetheless remains on the need for $n > p$ in settings involving inverse matrices that, when not met, does not allow to relate the sought-for eigenvalues to the (undefined) inverse sample covariance matrices. (Couillet et al., 2018) shows that this problem can be partially avoided for some divergences (not for the Fisher distance though). A systematic treatment of the $n < p$ case is however lacking.

Our approach also suffers from profound theoretical limitations that need be properly addressed: the fact that $\hat{\delta}(M, X)$ only estimates the sought-for $\delta(M, C)$ for M independent of X poses a formal problem when implemented in the gradient descent approach. This needs be tackled: (i) either by estimating the introduced bias so to estimate the loss incurred or, better, (ii) by accounting for the dependence to provide a further estimator $\hat{\hat{\delta}}(M(X), X)$ of $\delta(M(X), C)$ for all *X-dependent* matrices $M(X)$ following a specific form. Notably, given that, along the gradient descent initialized at $M_0 = \hat{U} D_0 \hat{U}^\mathsf{T}$, with $\hat{U}$ the eigenvectors of $\hat{C}$ and D_0 some diagonal matrix, all subsequent M_k matrices share the same profile (i.e., $M_k = \hat{U} D_k \hat{U}^\mathsf{T}$ for some diagonal D_k), a first improvement would consist in estimating consistently $\delta(\hat{U} D \hat{U}^\mathsf{T}, C)$ for deterministic diagonal matrices D.

The stability of the eigenvectors when initialized at $M_0 = \hat{U} D_0 \hat{U}^\mathsf{T}$, with $\hat{U}$ the eigenvectors of $\hat{C}$, turns our algorithm into a "non-linear shrinkage" method, as called by (Ledoit & Wolf, 2015). Parallel simulations also suggest that arbitrary initializations M_0 which *do not* follow this structure tend to still lead to solutions converging to the eigenspace of $\hat{C}$. However, this might be a consequence of the inconsistency of $\hat{\delta}(M, X)$ for X-dependent matrices M: it is expected that more consistent estimators might avoid this problem of *"eigenvector of $\hat{C}$" attraction*, thereby likely leading to improved estimations of the eigenvectors of C.

We conclude by emphasizing that modern large dimensional statistics have lately realized that substituting large covariance matrices by their sample estimators (or even by improved covariance estimators) is in general a weak approach, and that one should rather focus on estimating some ultimate functional (e.g., the result of a statistical test) involving the covariance (see, e.g., (Mestre & Lagunas, 2008) in array processing or (Yang et al., 2015) in statistical finance). It is to be noted that our proposed approach is consistent with these considerations as various functionals of C can be obtained from Equation (2), from which similar derivations can be performed.

Acknowledgement

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

References

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., and Elger, C. E. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907, 2001.

Bhatia, R. *Positive definite matrices*. Princeton University Press, 2009.

Bickel, P. J., Levina, E., et al. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36 (1):199–227, 2008.

Boumal, N., Mishra, B., Absil, P.-A., and Sepulchre, R. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014. URL `http://www.manopt.org`.

Couillet, R. and McKay, M. Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. *Journal of Multivariate Analysis*, 131:99–120, 2014.

Couillet, R., Tiomoko, M., Zozor, S., and Moisan, E. Random matrix-improved estimation of covariance matrix distances. *arXiv preprint arXiv:1810.04534*, 2018.

El Karoui, N. et al. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, 36(6):2757–2790, 2008.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Jeuris, B., Vandebril, R., and Vandereycken, B. A survey and comparison of contemporary algorithms for computing the matrix geometric mean. *Electronic Transactions on Numerical Analysis*, 39(ARTICLE), 2012.

Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Ledoit, O. and Wolf, M. Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. *Journal of Multivariate Analysis*, 139: 360–384, 2015.

Ledoit, O., Wolf, M., et al. Optimal estimation of a large-dimensional covariance matrix under steins loss. *Bernoulli*, 24(4B):3791–3832, 2018.

Marčenko, V. A. and Pastur, L. A. Distributions of eigenvalues for some sets of random matrices. *Math USSR-Sbornik*, 1(4):457–483, April 1967.

McLachlan, G. *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons, 2004.

Mestre, X. On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices. 56(11):5353–5368, November 2008.

Mestre, X. and Lagunas, M. Modified Subspace Algorithms for DoA Estimation With Large Arrays. 56(2): 598–614, February 2008.

Moakher, M. Divergence measures and means of symmetric positive-definite matrices. In *New Developments in the Visualization and Processing of Tensor Fields*, pp. 307–321. Springer, 2012.

Silverstein, J. W. and Bai, Z. D. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2): 175–192, 1995.

Silverstein, J. W. and Choi, S. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

Skovgaard, L. T. A Riemannian geometry of the multivariate normal model. *Scandinavian Journal of Statistics*, pp. 211–223, 1984.

Smith, S. T. Covariance, subspace, and intrinsic crame/spl acute/r-rao bounds. *IEEE Transactions on Signal Processing*, 53(5):1610–1630, 2005.

Sra, S. Positive definite matrices and the s-divergence. *arXiv preprint arXiv:1110.1773*, 2013.

Van Erven, T. and Harremos, P. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Vinogradova, J., Couillet, R., and Hachem, W. Estimation of toeplitz covariance matrices in large dimensional regime with application to source detection. *IEEE Trans. Signal Processing*, 63(18):4903–4913, 2015.

Wu, W. B. and Pourahmadi, M. Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, pp. 1755–1768, 2009.

Yang, L., Couillet, R., and McKay, M. R. Minimum variance portfolio optimization in the spiked covariance model. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on*, pp. 13–16. IEEE, 2015.