# ROBUST COVARIANCE ESTIMATION AND LINEAR SHRINKAGE IN THE LARGE DIMENSIONAL REGIME

*Romain Couillet*[1] *and Matthew McKay*[2].

[1]Telecommunication Department, Supélec, France.
[2]Hong-Kong University of Science and Technology, Hong Kong.

## ABSTRACT

The article studies two regularized robust estimators of scatter matrices proposed in parallel in [1] and [2], based on Tyler's robust M-estimator [3] and on Ledoit and Wolf's shrinkage covariance matrix estimator [4]. These hybrid estimators convey robustness to outliers or impulsive samples and small sample size adequacy to the classical sample covariance matrix estimator. We consider here the case of i.i.d. elliptical zero mean samples in the regime where both sample and population sizes are large. We prove that the above estimators behave similar to well-understood random matrix models, which allows us to derive optimal shrinkage strategies to estimate the population scatter matrix, largely improving existing methods.

## I. INTRODUCTION

Many scientific domains customarily deal with (possibly small) sets of large dimensional data samples from which statistical inference is performed. That is, the number $n$ of independent data samples $x_1,\ldots,x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) may not be large compared to the size $N$ of the population, suggesting that the empirical sample covariance matrix $\bar{C}_N = \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x})^*$, $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, is a poor estimate for $C_N = \mathrm{E}[(x_1-\mathrm{E}[x_1])(x_1-\mathrm{E}[x_1])^*]$. Several solutions have been proposed to work around this problem. If the end application is not to retrieve $C_N$ but some metric of it, recent works on random matrix theory showed that replacing $C_N$ in the metric by $\bar{C}_N$ often leads to a biased estimate of the metric, but that this estimate can be corrected by an improved estimation of the metric itself via the samples $x_1,\ldots,x_n$ [5]. However, when the object under interest is $C_N$ itself and $N \simeq n$, there is little hope to retrieve any consistent estimate of $C_N$. A popular alternative proposed originally in [4] is to "shrink" $\bar{C}_N$, i.e., consider instead $\bar{C}_N(\rho) = (1-\rho)\bar{C}_N + \rho I_N$ for an appropriate $\rho \in [0,1]$ that minimizes the average distance $\mathrm{E}[\mathrm{tr}((\bar{C}_N(\rho)-C_N)^2)]$. The interest of $\rho$ here is to give more or less weight to $\bar{C}_N$ depending on the relevance of the $n$ samples, so that in particular $\rho$ is better chosen close to zero when $n$ is large and close to one when $n$ is small.

In addition to the scarcity of samples, outliers may be present among the set of samples. These outliers, if not correctly handled, may further corrupt the statistical inference and in particular the estimation of $C_N$. The field of robust estimation intends to deal with this problem [6] by proposing estimators that have the joint capability to naturally attenuate the effect of outliers as well as to handle samples of an impulsive nature, e.g., elliptically distributed data. A common denominator of such estimators is their belonging to the class of M-estimators, therefore taking the form of the solution to an implicit equation. This poses important problems of analysis in small $N,n$ dimensions, resulting mostly in only asymptotic results in the regime $N$ fixed and $n \to \infty$. Nonetheless, recent works based on random matrix theory have shown that a certain family of such robust covariance matrix estimators asymptotically behave as $N,n \to \infty$ and $N/n \to c \in (0,1)$ similar to classical random matrices taking explicit forms [7], [8].

In the present article, we study two hybrid robust shrinkage covariance matrix estimates $\hat{C}_N(\rho)$ (hereafter referred to as the Abramovich–Pascal estimate) and $\check{C}_N(\rho)$ (hereafter referred to as the Chen estimate) proposed in parallel in [9], [2] and in [1], respectively. Both matrices are empirically built upon Tyler's M-estimate [3] and upon the Ledoit–Wolf shrinkage estimator [4]. This allows for an improved degree of freedom for approximating the population covariance matrix and importantly allows for $N > n$, which Maronna's and Tyler's estimators do not. In [2] and [1], $\hat{C}_N(\rho)$ and $\check{C}_N(\rho)$ were proved to be well-defined as the unique solutions to their defining fixed-point matrices. However, little is known of their performance as estimators of $C_N$ in the regime $N \simeq n$ of interest here. Some progress in this direction was made in [1] but this work does not manage to solve the optimal shrinkage problem consisting of finding $\rho$ such that $\mathrm{E}[\mathrm{tr}((\check{C}_N(\rho)-C_N)^2)]$ is minimized and resorts to solving an approximate problem instead.

The present article fills the gap by showing that, as $N,n \to \infty$ with $N/n \to c \in (0,\infty)$, both $\hat{C}_N(\rho)$ and $\check{C}_N(\rho)$ asymptotically behave similar to well-known random matrix models and are in fact asymptotically equivalent. This

Couillet's work is supported by the ERC MORE EC–120133. McKay's work is supported by the Hong Kong Research Grants Council under grant number 616713. The article is part of an extended article to appear in Elsevier Journal of Multivariate Analysis.

result is then used to derive an optimal shrinkage strategy for both estimators that, similar to [4], minimizes the square Frobenius norm metric.

## II. Main Results

We start by introducing the main assumptions of the data model under study. Let $x_1, \ldots, x_n \in \mathbb{C}^N$ (or $\mathbb{R}^N$) be $n$ sample vectors characterized as follows.

*Assumption 1:* Denoting $c_N = N/n$, $c_N \to c \in (0, \infty)$ as $N \to \infty$.

*Assumption 2:* The vectors $x_1, \ldots, x_n \in \mathbb{C}^N$ are independent with

a. $x_i = \sqrt{\tau_i} A_N y_i$, $y_i \in \mathbb{C}^{\bar{N}}$, $\bar{N} \geq N$, random zero mean unitarily invariant with norm $\|y_i\|^2 = \bar{N}$, $A_N \in \mathbb{C}^{N \times \bar{N}}$ deterministic, and $\tau_1, \ldots, \tau_n$ a collection of positive scalars. We shall denote $z_i = A_N y_i$.

b. $C_N \triangleq A_N A_N^*$ nonnegative definite, $\frac{1}{N} \operatorname{tr} C_N = 1$, and $\limsup_N \|C_N\| < \infty$ in spectral norm.

c. $\nu_N \triangleq \frac{1}{N} \sum_{i=1}^N \boldsymbol{\delta}_{\lambda_i(C_N)} \to \nu$ weakly with $\nu \neq \boldsymbol{\delta}_0$.

With this definition, the distribution of the vectors $x_i$ contains in particular the class of elliptical distributions. Note that, since $y_i$ is zero mean unitarily invariant with norm $\bar{N}$, $y_i = \sqrt{\bar{N}} \frac{\tilde{y}_i}{\|\tilde{y}_i\|}$ with $\tilde{y}_i \in \mathbb{C}^{\bar{N}}$ standard Gaussian. We can now introduce our main results.

*Theorem 1 (Abramovich–Pascal Estimate):* Let Assumptions 1 and 2 hold. For $\varepsilon \in (0, \min\{1, c^{-1}\})$, define $\hat{\mathcal{R}}_\varepsilon = [\varepsilon + \max\{0, 1 - c^{-1}\}, 1]$. For each $\rho \in (\max\{0, 1 - c_N^{-1}\}, 1]$, let $\hat{C}_N(\rho)$ be the unique solution to

$$\hat{C}_N(\rho) = (1-\rho) \frac{1}{n} \sum_{i=1}^n \frac{x_i x_i^*}{\frac{1}{N} x_i^* \hat{C}_N(\rho)^{-1} x_i} + \rho I_N.$$

Then, as $N \to \infty$,

$$\sup_{\rho \in \hat{\mathcal{R}}_\varepsilon} \left\| \hat{C}_N(\rho) - \hat{S}_N(\rho) \right\| \xrightarrow{\text{a.s.}} 0$$

where

$$\hat{S}_N(\rho) = \frac{1}{\hat{\gamma}(\rho)} \frac{1-\rho}{1-(1-\rho)c} \frac{1}{n} \sum_{i=1}^n z_i z_i^* + \rho I_N$$

and $\hat{\gamma}(\rho)$ is the unique positive solution to the equation in $\hat{\gamma}$

$$1 = \int \frac{t}{\hat{\gamma}\rho + (1-\rho)t} \nu(dt).$$

The function $\rho \mapsto \hat{\gamma}(\rho)$ thus defined is continuous on $(0, 1]$.

*Proof:* The proof can be found in [10, Section 5.1]. ■

*Theorem 2 (Chen Estimate):* Let Assumptions 1 and 2 hold. For $\varepsilon \in (0, 1)$, define $\check{\mathcal{R}}_\varepsilon = [\varepsilon, 1]$. For each $\rho \in (0, 1]$, let $\check{C}_N(\rho)$ be the unique solution to

$$\check{C}_N(\rho) = \frac{\check{B}_N(\rho)}{\frac{1}{N} \operatorname{tr} \check{B}_N(\rho)}$$

where

$$\check{B}_N(\rho) = (1-\rho) \frac{1}{n} \sum_{i=1}^n \frac{x_i x_i^*}{\frac{1}{N} x_i^* \check{C}_N(\rho)^{-1} x_i} + \rho I_N.$$

Then, as $N \to \infty$,

$$\sup_{\rho \in \check{\mathcal{R}}_\varepsilon} \left\| \check{C}_N(\rho) - \check{S}_N(\rho) \right\| \xrightarrow{\text{a.s.}} 0$$

where

$$\check{S}_N(\rho) = \frac{1-\rho}{1-\rho+T_\rho} \frac{1}{n} \sum_{i=1}^n z_i z_i^* + \frac{T_\rho}{1-\rho+T_\rho} I_N$$

in which $T_\rho = \rho \check{\gamma}(\rho) F(\check{\gamma}(\rho); \rho)$ with, for all $x > 0$,

$$F(x; \rho) = \frac{\rho - c(1-\rho)}{2} + \sqrt{\left( \frac{\rho - c(1-\rho)}{2} \right)^2 + \frac{1-\rho}{x}}$$

and $\check{\gamma}(\rho)$ is the unique positive solution to the equation in $\check{\gamma}$

$$1 = \int \frac{t}{\check{\gamma}\rho + \frac{1-\rho}{(1-\rho)c + F(\check{\gamma}; \rho)} t} \nu(dt).$$

The function $\rho \mapsto \check{\gamma}(\rho)$ thus defined is continuous on $(0, 1]$.

*Proof:* The proof is available in [10, Section 5.2]. ■

Theorem 1 and Theorem 2 show that, as $N, n \to \infty$ with $N/n \to c$, the matrices $\hat{C}_N(\rho)$ and $\check{C}_N(\rho)$, defined as the non-trivial solution of fixed-point equations, behave similar to matrices $\hat{S}_N(\rho)$ and $\check{S}_N(\rho)$, respectively, whose characterization is well-known and much simpler than that of $\hat{C}_N(\rho)$ and $\check{C}_N(\rho)$ themselves. Indeed, $\hat{S}_N(\rho)$ and $\check{S}_N(\rho)$ are random matrices of the sample covariance matrix type thoroughly studied in e.g., [11].

Technically speaking, the proof of both Theorem 1 and Theorem 2 unfold from the same technique as developed in [8]. However, while the proof of Theorem 1 comes with no major additional difficulty compared to [8], due to the scale normalization imposed in the definition of $\check{C}_N(\rho)$, the proof of Theorem 2 requires a more elaborate approach than used in [8]. The readers are invited to find the detailed proofs in [10]. Another difference to previous works lies here in that, unlike Maronna's estimator that only attenuates the effect of the scale parameters $\tau_i$, the proposed Tyler-based estimators discard this effect altogether. Also, the technical study of Maronna's estimator can be made under the assumption that $C_N = I_N$ (from a natural variable change) while here, because of the regularization term $\rho I_N$, $C_N$ does intervene in an intricate manner in the results.

As a side remark, it is shown in [2] that for each $N, n$ fixed with $n \geq N + 1$, $\hat{C}_N(\rho) \to \hat{C}_N(0)$ as $\rho \to 0$ with $\hat{C}_N(0)$ defined (almost surely) as one of the (uncountably many) solutions to

$$\hat{C}_N(0) = \frac{1}{n} \sum_{i=1}^n \frac{x_i x_i^*}{\frac{1}{N} x_i^* \hat{C}_N(0)^{-1} x_i}. \tag{1}$$

In the regime where $N, n \to \infty$ and $N/n \to c$, this result is difficult to generalize as it is challenging to handle the limit $\|\hat{C}_N(\rho_N) - \hat{S}_N(\rho_N)\|$ for a sequence $\{\rho_N\}_{N=1}^{\infty}$ with $\rho_N \to 0$. The requirement that $\rho_N \to \rho_0 > 0$ on any such sequence is indeed at the core of the proof of Theorem 1. This explains why the set $\hat{\mathcal{R}}_\varepsilon$ in Theorem 1 excludes the region $[0, \varepsilon)$. Similar arguments hold for $\check{C}_N(\rho)$.

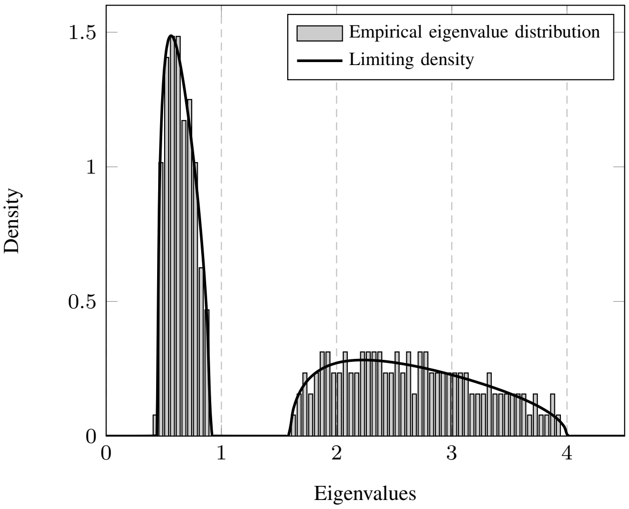

**Fig. 1**. Histogram of the eigenvalues of $\hat{C}_N$ (Abramovich–Pascal type) for $n = 2048$, $N = 256$, $C_N = \frac{1}{3}\operatorname{diag}(I_{128}, 5I_{128})$, $\rho = 0.2$, versus limiting eigenvalue distribution.

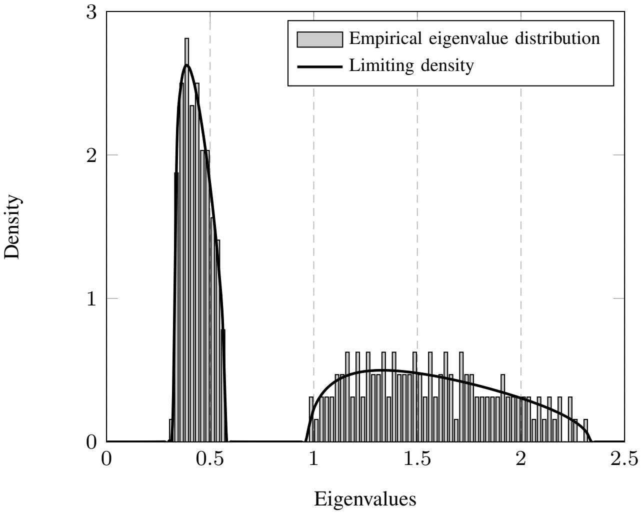

**Fig. 2**. Histogram of the eigenvalues of $\check{C}_N$ (Chen type) for $n = 2048$, $N = 256$, $C_N = \frac{1}{3}\operatorname{diag}(I_{128}, 5I_{128})$, $\rho = 0.2$, versus limiting eigenvalue distribution.

Figure 1 and Figure 2 depict the histogram of the eigenvalues of $\hat{C}_N(\rho)$ and $\check{C}_N(\rho)$ for $\rho = 0.2$, $N = 256$, $n = 2048$, $C_N = \operatorname{diag}(I_{128}, 5I_{128})$, versus the limiting distributions of the eigenvalues of $\hat{S}_N(\rho)$ and $\check{S}_N(\rho)$ for $c = 1/8$, respectively.

A corollary of Theorem 1 and Theorem 2 is the joint convergence (over both $\rho$ and the eigenvalue index) of the individual eigenvalues of $\hat{C}_N(\rho)$ to those of $\hat{S}_N(\rho)$ and of the individual eigenvalues of $\check{C}_N(\rho)$ to those of $\check{S}_N(\rho)$, as well as the joint convergence over $\rho$ of the moments of the empirical spectral distributions of $\hat{C}_N(\rho)$ and $\check{C}_N(\rho)$. These joint convergence properties are fundamental in problems of optimization of the parameter $\rho$ as discussed in Section III.

*Corollary 1 (Joint convergence properties):* Under the settings of Theorem 1 and Theorem 2,

$$\sup_{\rho \in \hat{\mathcal{R}}_\varepsilon} \max_{1 \le i \le n} \left| \lambda_i(\hat{C}_N(\rho)) - \lambda_i(\hat{S}_N(\rho)) \right| \xrightarrow{\text{a.s.}} 0$$

$$\sup_{\rho \in \check{\mathcal{R}}_\varepsilon} \max_{1 \le i \le n} \left| \lambda_i(\check{C}_N(\rho)) - \lambda_i(\check{S}_N(\rho)) \right| \xrightarrow{\text{a.s.}} 0.$$

This result implies

$$\limsup_N \sup_{\rho \in \hat{\mathcal{R}}_\varepsilon} \|\hat{C}_N(\rho)\| < \infty$$

$$\limsup_N \sup_{\rho \in \check{\mathcal{R}}_\varepsilon} \|\check{C}_N(\rho)\| < \infty.$$

almost surely. This in turn induces that, for each $\ell \in \mathbb{N}$,

$$\sup_{\rho \in \hat{\mathcal{R}}_\varepsilon} \left| \frac{1}{N} \operatorname{tr} \left( \hat{C}_N(\rho)^\ell \right) - M_{\hat{\mu}_\rho, \ell} \right| \xrightarrow{\text{a.s.}} 0$$

$$\sup_{\rho \in \check{\mathcal{R}}_\varepsilon} \left| \frac{1}{N} \operatorname{tr} \left( \check{C}_N(\rho)^\ell \right) - M_{\check{\mu}_\rho, \ell} \right| \xrightarrow{\text{a.s.}} 0$$

where we recall that $M_{\mu,\ell} = \int t^\ell \mu(dt) \in (0, \infty]$ for any probability measure $\mu$ with support in $\mathbb{R}^+$; in particular, $M_{\hat{\mu}_\rho,1} = \frac{1}{\hat{\gamma}(\rho)} \frac{1-\rho}{1-(1-\rho)c} + \rho$ and $M_{\check{\mu}_\rho,1} = 1$.

*Proof:* The proof is found in [10, Section 5.3]. ■

## III. APPLICATION TO OPTIMAL SHRINKAGE

We now apply Theorems 1 and 2 to the problem of optimal linear shrinkage, originally considered in [4] for the simpler sample covariance matrix model. The optimal linear shrinkage problem consists in choosing $\rho$ to be such that a certain distance metric between $\hat{C}_N(\rho)$ (or $\check{C}_N(\rho)$) and $C_N$ is minimized, therefore allowing for a more appropriate estimation of $C_N$ via $\hat{C}_N(\rho)$ or $\check{C}_N(\rho)$. The metric selected here is the squared Frobenius norm of the difference between the (possibly scaled) robust estimators and $C_N$, which has the advantage of being a widespread matrix distance (e.g., as considered in [4]) and a metric amenable to mathematical analysis.[1] In [1], the authors studied this problem in the specific case of $\check{C}_N(\rho)$ but did not find an expression for

[1] Alternative metrics (such as the geodesic distance on the cone of nonnegative definite matrices) can be similarly considered. The appropriate choice of such a metric heavily depends on the ultimate problem to optimize.

the optimal theoretical $\rho$ due to the involved structure of $\check{C}_N(\rho)$ for all finite $N, n$ and therefore resorted to solving an approximate problem, the solution of which is denoted here $\check{\rho}_O$. Instead, we show that for large $N, n$ values the optimal $\rho$ under study converges to a limiting value $\check{\rho}^\star$ that takes an extremely simple explicit expression and a similar result holds for $\hat{C}_N(\rho)$ for which an equivalent optimal $\hat{\rho}^\star$ is defined.

Our first result is a lemma of fundamental importance which demonstrates that, up to a change in the variable $\rho$, $\hat{S}_N(\rho)/M_{\hat{\mu}_\rho,1}$ and $\check{S}_N(\rho)$ (constructed from the samples $x_1, \ldots, x_n$) are completely equivalent to the original Ledoit–Wolf linear shrinkage model for the (non observable) samples $z_1, \ldots, z_n$.

*Lemma 1 (Model Equivalence):* For each $\rho \in (0, 1]$, there exist unique $\hat{\rho} \in (\max\{0, 1 - c^{-1}\}, 1]$ and $\check{\rho} \in (0, 1]$ such that

$$\frac{\hat{S}_N(\hat{\rho})}{M_{\hat{\mu}_{\hat{\rho}},1}} = \check{S}_N(\check{\rho}) = (1-\rho)\frac{1}{n}\sum_{i=1}^n z_i z_i^* + \rho I_N.$$

Besides, the maps $(0, 1] \to (\max\{0, 1 - c^{-1}\}, 1]$, $\rho \mapsto \hat{\rho}$ and $(0, 1] \to (0, 1]$, $\rho \mapsto \check{\rho}$ thus defined are continuously increasing and onto.

*Proof:* The proof is found in [10, Section 5.4]. ■

Along with Theorem 1 and Theorem 2, Lemma 1 indicates that, up to a change in the variable $\rho$, $\hat{C}_N(\rho)$ and $\check{C}_N(\rho)$ can be somewhat viewed as asymptotically equivalent (but there is no saying whether they can be claimed equivalent for all finite $N, n$). As such, thanks to Lemma 1, we now show that the optimal shrinkage parameters $\rho$ for both $\hat{C}_N(\rho)/(\frac{1}{N}\operatorname{tr}\hat{C}_N(\rho))$ and $\check{C}_N(\rho)$ lead to the same asymptotic performance, which corresponds to the asymptotically optimal Ledoit–Wolf linear shrinkage performance but for the vectors $z_1, \ldots, z_n$.

*Proposition 1 (Optimal Shrinkage):* For each $\rho \in (0, 1]$, define[2]

$$\hat{D}_N(\rho) = \frac{1}{N}\operatorname{tr}\left(\left(\frac{\hat{C}_N(\rho)}{\frac{1}{N}\operatorname{tr}\hat{C}_N(\rho)} - C_N\right)^2\right)$$

$$\check{D}_N(\rho) = \frac{1}{N}\operatorname{tr}\left(\left(\check{C}_N(\rho) - C_N\right)^2\right).$$

Also denote $D^\star = c\frac{M_{\nu,2}-1}{c+M_{\nu,2}-1}$, $\rho^\star = \frac{c}{c+M_{\nu,2}-1}$, and $\hat{\rho}^\star \in (\max\{0, 1 - c^{-1}\}, 1]$, $\check{\rho}^\star \in (0, 1]$ the unique solutions to

$$\frac{\hat{\rho}^\star}{\frac{1}{\hat{\gamma}(\hat{\rho}^\star)}\frac{1-\hat{\rho}^\star}{1-(1-\hat{\rho}^\star)c} + \hat{\rho}^\star} = \frac{T_{\check{\rho}^\star}}{1 - \check{\rho}^\star + T_{\check{\rho}^\star}} = \rho^\star.$$

Then, letting $\varepsilon < \min(\hat{\rho}^\star - \max\{0, 1 - c^{-1}\}, \check{\rho}^\star)$, under the setting of Theorem 1 and Theorem 2,

$$\inf_{\rho \in \hat{\mathcal{R}}_\varepsilon} \hat{D}_N(\rho) \xrightarrow{\text{a.s.}} D^\star, \quad \inf_{\rho \in \check{\mathcal{R}}_\varepsilon} \check{D}_N(\rho) \xrightarrow{\text{a.s.}} D^\star$$

and

$$\hat{D}_N(\hat{\rho}^\star) \xrightarrow{\text{a.s.}} D^\star, \quad \check{D}_N(\check{\rho}^\star) \xrightarrow{\text{a.s.}} D^\star.$$

Moreover, letting $\hat{\rho}_N$ and $\check{\rho}_N$ be random variables such that $\hat{\rho}_N \xrightarrow{\text{a.s.}} \hat{\rho}^\star$ and $\check{\rho}_N \xrightarrow{\text{a.s.}} \check{\rho}^\star$,

$$\hat{D}_N(\hat{\rho}_N) \xrightarrow{\text{a.s.}} D^\star, \quad \check{D}_N(\check{\rho}_N) \xrightarrow{\text{a.s.}} D^\star.$$

*Proof:* The proof is provided in [10, Section 5.5]. ■

The last part of Proposition 1 states that, if consistent estimates $\hat{\rho}_N$ and $\check{\rho}_N$ of $\hat{\rho}^\star$ and $\check{\rho}^\star$ exist, then they have optimal shrinkage performance in the large $N, n$ limit. Such estimates may of course be defined in multiple ways. We present below a simple example based on $\hat{C}_N(\rho)$ and $\check{C}_N(\rho)$.

*Proposition 2 (Optimal Shrinkage Estimate):* Under the setting of Proposition 1, let $\hat{\rho}_N \in (\max\{0, 1 - c^{-1}\}, 1]$ and $\check{\rho}_N \in (0, 1]$ be solutions (not necessarily unique) to

$$\frac{\hat{\rho}_N}{\frac{1}{N}\operatorname{tr}\hat{C}_N(\hat{\rho}_N)} = \frac{c_N}{\frac{1}{N}\operatorname{tr}\left[\left(\frac{1}{n}\sum_{i=1}^n \frac{x_i x_i^*}{\frac{1}{N}\|x_i\|^2}\right)^2\right] - 1}$$

$$\frac{G(\check{\rho}_N)}{1 - \check{\rho}_N + G(\check{\rho}_N)} = \frac{c_N}{\frac{1}{N}\operatorname{tr}\left[\left(\frac{1}{n}\sum_{i=1}^n \frac{x_i x_i^*}{\frac{1}{N}\|x_i\|^2}\right)^2\right] - 1}$$

where

$$G(\rho) = \rho\frac{1}{n}\sum_{i=1}^n \frac{x_i^* \check{C}_N(\rho)^{-1} x_i}{\|x_i\|^2}.$$

and defined arbitrarily when no such solutions exist. Then

$$\hat{\rho}_N \xrightarrow{\text{a.s.}} \hat{\rho}^\star \text{ and } \check{\rho}_N \xrightarrow{\text{a.s.}} \check{\rho}^\star$$

so that

$$\hat{D}_N(\hat{\rho}_N) \xrightarrow{\text{a.s.}} D^\star \text{ and } \check{D}_N(\check{\rho}_N) \xrightarrow{\text{a.s.}} D^\star.$$

*Proof:* The proof can be found in [10, Section 5.6]. ■

Figure 3 illustrates the performance in terms of the metric $\check{D}_N$ of the empirical shrinkage coefficient $\check{\rho}_N$ introduced in Proposition 2 versus the optimal value $\inf_{\rho \in (0,1]}\{\check{D}_N(\rho)\}$, averaged over 10 000 Monte Carlo simulations. We also present in this graph the almost sure limiting value $D^\star$ of both $\check{D}_N(\check{\rho}_N)$ and $\inf_{\rho \in \check{\mathcal{R}}_\varepsilon}\{\check{D}_N(\rho)\}$ for some sufficiently small $\varepsilon$, as well as $\check{D}_N(\check{\rho}_O)$ of $\check{\rho}_O$ defined in [1, Equation (12)] as the minimizing solution of $\mathrm{E}[\frac{1}{N}\operatorname{tr}(\check{C}_O(\rho) - C_N)^2]$ with $\check{C}_O(\rho)$ the so-called "clairvoyant estimator"

$$\check{C}_O(\rho) = (1-\rho)\frac{1}{n}\sum_{i=1}^n \frac{x_i x_i^*}{\frac{1}{N}x_i^* C_N^{-1} x_i} + \rho I_N.$$

We consider in this graph $N = 32$ constant, $n \in \{2^k, k = 1, \ldots, 7\}$, and $C_N = [C_N]_{i,j=1}^N$ with $[C_N]_{ij} = r^{|i-j|}$, $r = 0.7$, which is the same setting as considered in [1, Section 4].

It appears in Figure 3 that a significant improvement is brought by $\check{\rho}_N$ over $\check{\rho}_O$, especially for small $n$,

[2] Recall that, for $A$ Hermitian, $\frac{1}{N}\operatorname{tr}(A^2) = \frac{1}{N}\operatorname{tr}(AA^*) = \frac{1}{N}\|A\|_F^2$ with $\|\cdot\|_F$ the Frobenius norm for matrices.

which translates the poor quality of $\check{C}_O(\rho)$ as an approximation of $\check{C}_N(\rho)$ for large values of $c_N$ (obviously linked to $\frac{1}{N}x_i^* C_N^{-1} x_i$ being then a bad approximation for $\frac{1}{N}x_i^* \check{C}_N(\rho)^{-1} x_i$). Another important remark is that, even for so small values of $N, n$, $\inf_{\rho\in(0,1]} \check{D}_N(\rho)$ is extremely close to the limiting optimal, suggesting here that the limiting results of Proposition 1 are already met for small practical values. The approximation $\check{\rho}_N$ of $\check{\rho}^\star$, translated here through $\check{D}_N(\check{\rho}_N)$, also demonstrates good practical performance at small values of $N, n$.

We additionally mention that we produced similar curves for $\hat{C}_N(\rho)$ in place of $\check{C}_N(\rho)$ which happened to show virtually the same performance as the equivalents curves for $\check{C}_N(\rho)$. This is of course expected (with exact match) for $\inf_{\rho\in(0,1]} \hat{D}_N(\rho)$ which, up to the region $[0, \varepsilon)$, matches $\inf_{\rho\in(0,1]} \check{D}_N(\rho)$ for large enough $N, n$, and similarly for $\hat{D}_N(\hat{\rho}_N)$ since $\hat{\rho}_N$ was designed symmetrically to $\check{\rho}_N$.

Associated to Figure 3 is Figure 4 which provides the shrinkage parameter values, optimal and approximated, for both the Abramovich–Pascal and Chen estimates, along with the clairvoyant $\check{\rho}_O$ of [1]. Recall that the $\hat{(\cdot)}$ values must only be compared to one another, and similarly for the $\check{(\cdot)}$ values (so in particular $\check{\rho}_O$ only compares against the $\check{(\cdot)}$ values). It appears here that $\check{\rho}_O$ is a rather poor estimate for $\operatorname{argmin}_{\rho\in(0,1]} \check{D}_N(\rho)$ for a large range of values of $n$. It tends in particular to systematically overestimate the weight to be put on the sample covariance matrix.

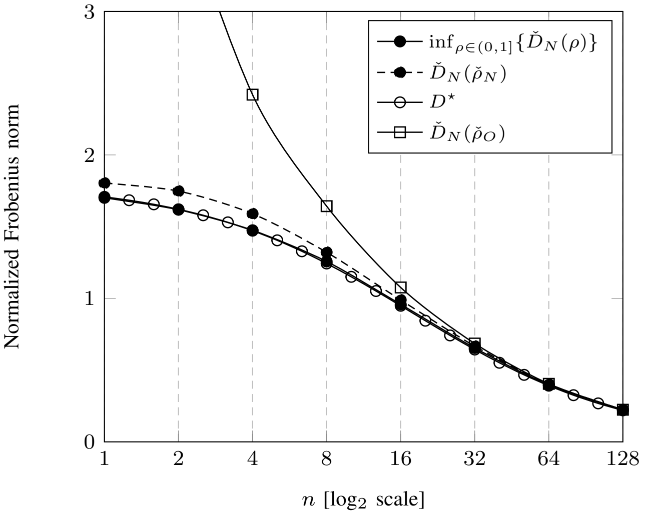

**Fig. 3**. Performance of optimal shrinkage averaged over 10 000 Monte Carlo simulations, for $N = 32$, various values of $n$, $[C_N]_{ij} = r^{|i-j|}$ with $r = 0.7$; $\check{\rho}_N$ is given in Proposition 2; $\check{\rho}_O$ is the clairvoyant estimator proposed in [1, Equation (12)]; $D^\star$ taken with $c = N/n$.

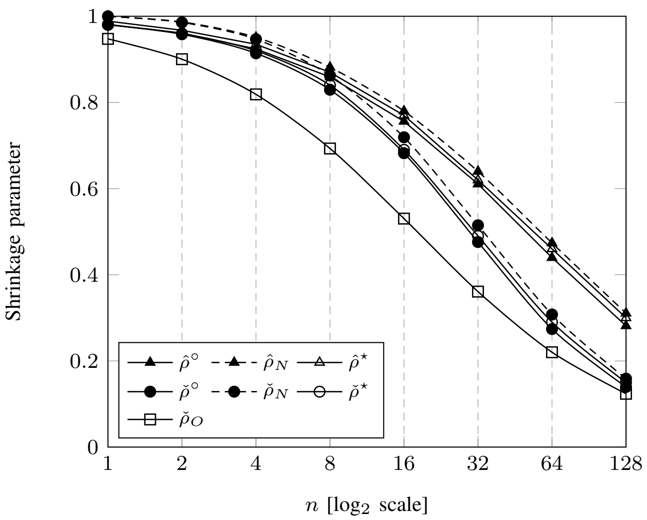

**Fig. 4**. Shrinkage parameter $\rho$ averaged over 10 000 Monte Carlo simulations, for $N = 32$, various values of $n$, $[C_N]_{ij} = r^{|i-j|}$ with $r = 0.7$; $\hat{\rho}_N$ and $\check{\rho}_N$ given in Proposition 2; $\check{\rho}_O$ is the clairvoyant estimator proposed in [1, Equation (12)]; $\rho^\star$, $\hat{\rho}^\star$, and $\check{\rho}^\star$ taken with $c = N/n$; $\hat{\rho}^\circ = \operatorname{argmin}_{\{\rho\in(\max\{0,1-c_N^{-1}\},1]\}}\{\hat{D}_N(\rho)\}$ and $\check{\rho}^\circ = \operatorname{argmin}_{\{\rho\in(0,1]\}}\{\check{D}_N(\rho)\}$.

## IV. CONCLUDING REMARKS

The article shows that, in the large dimensional random matrix regime, the Abramovich–Pascal and Chen estimators for elliptical samples $x_1, \ldots, x_n$ are (up to a variable change) asymptotically equivalent, so that both can be used interchangeably. They are also equivalent to the classical Ledoit–Wolf estimator for the samples $z_1, \ldots, z_n$ or, as can be easily verified, for the samples $\sqrt{N}x_1/\|x_1\|, \ldots, \sqrt{N}x_n/\|x_n\|$. This means that for elliptical samples, at least as far as first order convergence is concerned, the Abramovich–Pascal and Chen estimators perform similar to a normalized version of Ledoit–Wolf.

Recalling that robust estimation theory aims in particular at handling sample sets corrupted by outliers, the performance of the Abramovich–Pascal and Chen estimators given in this paper (not considering outliers) can be seen as a base reference for the "clean data" scenario which paves the way for future work in more advanced scenarios. In the presence of outliers, it is expected that the Abramovich–Pascal and Chen estimates exhibit robustness properties that the normalized Ledoit–Wolf scheme does not possess by appropriately weighting good versus outlying data. The study of this scenario is currently under investigation. Also, the extension of this work to second order analysis, e.g., to central limit theorems on linear statistics of the robust estimators, is a direction of future work that will allow to handle more precisely the gain of robust versus non-robust schemes in the not-too-large dimensional regime.

In terms of applications, Proposition 2 allows for the design of covariance matrix estimators, with minimal Frobenius distance to the population covariance matrix for impulsive i.i.d. samples but in the absence of outliers, and having robustness properties in the presence of outliers. This is fundamental to those scientific fields where the covariance matrix is the object of central interest. More generally though, Theorems 1 and 2 can be used to design optimal covariance matrix estimators under other metrics than the Frobenius norm. This is in particular the case in applications to finance where a possible target consists in the minimization of the risk induced by portfolios built upon such covariance matrix estimates, see e.g., [12], [13], [14]. The possibility to let the number of samples be less than the population size (as opposed to robust estimators of the Maronna-type [15]) is also of interest to applications where optimal shrinkage is not a target but where robustness is fundamental, such as array processing with impulsive noise (e.g., multi-antenna radar) where direction-of-arrival estimates are sought for (see e.g., [16], [7]). These considerations are also left to future work.

## V. REFERENCES

[1] Y. Chen, A. Wiesel, and A. O. Hero, "Robust shrinkage estimation of high-dimensional covariance matrices," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4097–4107, 2011.

[2] F. Pascal, Y. Chitour, and Y. Quek, "Generalized robust shrinkage estimator – Application to STAP data," *Submitted for publication*, 2013. [Online]. Available: http://arxiv.org/pdf/1311.6567

[3] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *The Annals of Statistics*, vol. 15, no. 1, pp. 234–251, 1987.

[4] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004.

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

[6] R. A. Maronna, D. R. Martin, and J. V. Yohai, *Robust Statistics: Theory and Methods*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, 2006.

[7] R. Couillet, F. Pascal, and J. W. Silverstein, "Robust Estimates of Covariance Matrices in the Large Dimensional Regime," *IEEE Transactions on Information Theory*, 2013. [Online]. Available: http://arxiv.org/abs/1204.5320

[8] ——, "The random matrix regime of Maronna's M-estimator with elliptically distributed samples," *Journal of Multivariate Analysis*, 2013. [Online]. Available: http://arxiv.org/abs/1311.7034

[9] Y. I. Abramovich and N. K. Spencer, "Diagonally loaded normalised sample matrix inversion (lnsmi) for outlier-resistant adaptive filtering," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, vol. 3. IEEE, 2007, pp. III–1105.

[10] R. Couillet and M. McKay, "Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators," *to appear in Journal of Multivariate Analysis*, 2013.

[11] J. W. Silverstein and S. Choi, "Analysis of the limiting spectral distribution of large dimensional random matrices," *Journal of Multivariate Analysis*, vol. 54, no. 2, pp. 295–309, 1995.

[12] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, vol. 10, no. 5, pp. 603–621, 2003.

[13] F. Rubio, X. Mestre, and D. P. Palomar, "Performance analysis and optimal selection of large minimum variance portfolios under estimation risk," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 4, pp. 337–350, 2012.

[14] J. Yu, M. R. McKay, and F. Rubio, "Minimum variance portfolio optimisation with high frequency data: A robust approach based on random matrix theory," *Submitted for publication*, 2013.

[15] R. A. Maronna, "Robust M-estimators of multivariate location and scatter," *The annals of statistics*, pp. 51–67, 1976.

[16] X. Mestre and M. Lagunas, "Modified subspace algorithms for DoA estimation with large arrays," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 598–614, Feb. 2008.