Large Dimensional Analysis of Robust M-Estimators of Covariance With Outliers
David Morales-Jimenez, Romain Couillet, Matthew Mckay

To cite this version:
David Morales-Jimenez, Romain Couillet, Matthew Mckay. Large Dimensional Analysis of Robust M-Estimators of Covariance With Outliers. IEEE Transactions on Signal Processing, Institute of Electrical and Electronics Engineers, 2015, 63 (21), pp.5784 -5797. 10.1109/TSP.2015.2460225. hal-01242442

HAL Id: hal-01242442
https://hal.archives-ouvertes.fr/hal-01242442
Submitted on 12 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract—A large dimensional characterization of robust M-estimators of covariance (or scatter) is provided under the assumption that the dataset comprises independent (essentially Gaussian) legitimate samples as well as arbitrary deterministic samples, referred to as outliers. Building upon recent random matrix advances in the area of robust statistics, we specifically show that the so-called Maronna M-estimator of scatter asymptotically behaves similar to well-known random matrices when the population and sample sizes grow together to infinity. The introduction of outliers leads the robust estimator to behave asymptotically as the weighted sum of the sample outer products, with a constant weight for all legitimate samples and different weights for the outliers. A fine analysis of this structure reveals importantly that the propensity of the M-estimator to attenuate (or enhance) the impact of outliers is mostly dictated by the alignment of the outliers with the inverse population covariance matrix of the legitimate samples. Thus, robust M-estimators can bring substantial benefits over more simplistic estimators such as the per-sample normalized version of the sample covariance matrix, which is not capable of differentiating the outlying samples. The analysis shows that, within the class of Maronna’s estimators of scatter, the Huber estimator is most favorable for rejecting outliers. On the contrary, estimators more similar to Tyler’s scale invariant estimator (often preferred in the literature) run the risk of inadvertently enhancing some outliers.

Index Terms—Robust statistics, M-estimation, outliers.

I. INTRODUCTION

The growing momentum of big data applications along with the recent advances in large dimensional random matrix theory have raised much interest for problems in statistics and signal processing under the assumption of large but similar population dimension $N$ and sample size $n$. Due to the intrinsic complexity of large dimensional random matrix theory, as compared to classical statistics where $N$ is fixed and $n \to \infty$, most of the classical applications were concerned with sample covariance matrix (SCM) based methods (as in e.g., [1, 2] for source detection or [3] for subspace estimation). Only recently have other random matrix structures started to be explored which are adequate to deal with more advanced statistical problems; see for instance [4] on Toeplitz random matrix structures, or [5] on kernel random matrices. Of particular interest is the structure of robust M-estimators of covariance (or scatter), which have very recently come to a better understanding in the large dimensional regime and is the focus of the present work.

The field of robust M-estimation, born with the early works of Huber [6], roughly consists in improving classical Gaussian maximum-likelihood estimators, such as the sample mean or SCM, into estimators that (unlike the classical estimators) are resilient to both the possibly heavy-tailed nature of the observed data or the presence of outliers in the dataset. Assuming observation data of known zero mean, robust estimators of the population covariance matrix, referred to as robust M-estimators of scatter, were proposed successively in [6] for data composed of a majority of independent Gaussian samples and a few outliers and then in [7] and [8] for elliptically distributed or arbitrary scaled Gaussian data.

But the analysis for each given $N,n$ of the aforementioned robust estimators of scatter, which often take the form of solutions of implicit equations, is in general intractable. In a series of recent works [9–12] (see also [13, 14] for applications), this limitation was alleviated by considering the random matrix regime where both $N,n$ are large and commensurable. These works have shown that in this regime several classes of robust estimators of scatter (Maronna, Tyler, and regularized Tyler) behave similar to simpler and explicit random matrix models, which are fully understandable via (now standard) random matrix methods. Nonetheless, all these works were pursued under the assumption that the input data are independent and follow a zero-mean elliptical distribution. One of the salient outcomes of these works is that, under elliptical inputs, the Tyler and regularized Tyler estimators asymptotically behave similar to the SCM of the normalized data henceforth referred to as the normalized SCM, and therefore do not provide any apparent gain in robustness versus simpler sample covariance estimators.

This fact, however, fundamentally disregards the important role of robust estimators as arbitrary outlier rejectors. In the present work, we shall consider data comprising both legitimate data (that are essentially independent Gaussian samples) and a certain (a priori unknown) amount of arbitrary deterministic outliers. Focusing our attention specifically to the (larger) class of Maronna’s M-estimators of scatter, similar to all of the aforementioned works and following the approach in [9], we will show that in this setting the robust estimator of scatter behaves similar for large $N,n$ to an explicit and easily understood random matrix. But it will appear, unlike in

\[1\] This being valid up to second-order fluctuations [1].
[2], that this random matrix no longer behaves similar to
the normalized SCM. Our main finding is that, under suitable
conditions, the robust estimator of scatter manages to attenuate
(to some extent) the impact of the deterministic outliers,
which the SCM (or normalized SCM) may not be capable of.
Calling $C_N$ the population covariance matrix of the legitimate
data, $a_i \in \mathbb{C}^N$ the $i$-th outlier, and assuming the number
of outliers is small compared to $n$, it will be demonstrated
that the rejection power of the robust estimator of scatter is
monotonically related to the quadratic form $a_i^\dagger C_N^{-1} a_i$. This
shows that, if $C_N$ is (invertible but) essentially of low rank,
$a_i^\dagger C_N^{-1} a_i$ can take large values and thus $a_i$ is likely to be
suppressed. If $a_i^\dagger C_N^{-1} a_i$ is quite small instead, an inverse effect
of outlier enhancement may appear that needs be controlled
by an appropriate choice of estimator within Maronna’s class.
We shall show that such an estimator should resemble the
original Huber estimator from [4] and substantially differ from
the Tyler estimator.

In the remainder of the article, we provide a rigorous state-
ment of our main results. The problem at hand is discussed
in Section III and our main results introduced in Section III
all proofs being deferred to the appendices. Special attention
will then be made on the analytically tractable cases where t he
all proofs being deferred to the appendices. Special attention
will then be made on the analytically tractable cases where t he
will be considered the legiti-
mate data, while $a_1, \ldots, a_{n_o}$ are deterministic unknown
outliers. It is important to note at this point that all estimators
of $C_N$ considered in the following are invariant to column
permutations in $Y$ so that we can freely assume the first
columns of $Y$ to be the legitimate data and the last columns
to be the outliers. Note also that we consider here a more
general setting than Gaussian legitimate data as we merely
request the $x_i$’s to have independent normalized entries with
some bounded moment condition.

Although $a_1, \ldots, a_{n_o}$ are arbitrary, for technical reasons
we shall need the following control.

**Assumption 1.** $\limsup_n \left\| \frac{1}{n} \sum_{i=1}^{n_o} C_N^{-1/2} a_i a_i^\dagger C_N^{-1/2} \right\| < \infty$.

Note that, if $\limsup_n \varepsilon_n/n < \infty$, Assumption I reduces to
$\limsup_n \max_{1 \leq i \leq n_o} \frac{1}{n} a_i^\dagger C_N^{-1} a_i < \infty$.

If one were aware of the presence and position of out-
liers in the dataset, then the natural estimator for $C_N$ (up
renormalization by $1 - \varepsilon_n$) would read $\frac{1}{n} Y^\dagger Y^{o\dagger}$ with
$Y^o = [y_1, \ldots, y_{(1 - \varepsilon_n)n}]$; this estimator, which we shall refer to
as the Oracle estimator (hence the “$o$” superscript), merely
consists in a SCM with discarded outliers. For lack of knowing
the outliers presence and positions, the immediate alternative
for $C_N$ is the SCM, which reads here $\frac{1}{n} Y Y^\dagger$. If
one is only interested in estimating any scaled version of
$C_N$, then, to mitigate the negative impact of outliers
with arbitrarily large norm, a simple robust procedure consists in
estimating $C_N$ via the normalized SCM $\frac{1}{n} Y_n Y_n^{\dagger}$, where
$Y_n \triangleq Y \text{diag}(\sqrt{\frac{n}{n}} Y^\dagger Y)^{-\frac{1}{2}}$. This matrix has the advantage of
avoiding arbitrarily large biases in the estimation of $C_N$. How-
ever, being only based on a per-data norm control, $\frac{1}{n} Y_n Y_n^{\dagger}$ does not take into account the fact that outliers can also be
detected if they significantly differ, not just in norm, from
the majority of the data. The robust estimators of scatter,
introduced by Huber [6] and later studied by Maronna [7],
were precisely designed for this purpose. Our objective here
is to finely understand this outlier identification and mitigation
procedure by means of a large random matrix analysis.

To be able to define a robust $M$-estimator of scatter in
the sense of Maronna under the presence of arbitrary outlier vectors,
a constraint must be set on $\varepsilon_n$ and $N$. In particular,
as $n$ grows large, we shall require that $n(1 - \varepsilon_n)/N$ (and not
only $n/N$) be always beyond one.

**Assumption 2 (Growth rate).** As $n \to \infty \varepsilon_n \to \varepsilon \in (0, 1)$
and $\varepsilon_n \triangleq \frac{N}{n} \to c$ with $0 < c < 1 - \varepsilon$.

We then define Maronna’s $M$-estimator of scatter $\hat{C}_N$ as a
solution, when it exists, to the equation in $Z$

$$Z = \frac{1}{n} \sum_{i=1}^{(1 - \varepsilon_n)n} u \left( \frac{1}{N} y_i^\dagger Z^{-1} y_i \right) y_i y_i^\dagger + \frac{1}{n} \sum_{i=1}^{\varepsilon_n n} u \left( \frac{1}{N} a_i a_i^\dagger \right) a_i a_i^\dagger,$$

(1)

where $u : [0, \infty) \to (0, \infty)$ is continuous, non-increasing, and
such that $\phi(x) \triangleq xu(x)$ is increasing with $\lim_{x \to \infty} \phi(x) \triangleq \phi_\infty$ and $(1 - \varepsilon)^{-1} < \phi_\infty < c^{-1}$. Note that the latter
assumption on $\phi_\infty$ is equivalent to that in [9] with a slight
modification accounting for the presence of outliers.
A standard choice for the function $u$ is $u = u_S$, where, for some $t > 0$,
$$
u_S(x) \triangleq \frac{1 + t}{t + x} \quad \text{(2)}$$
which, for an appropriate $t$, turns $\hat{C}_N$ into the maximum-likelihood estimator of $C_N$ when the columns of $Y$ are independent multivariate Student vectors (hence the superscript “$S$”). As $t \to 0$, $\hat{C}_N$ converges to one version of the so-called Tyler estimator [8], as shown in [13]. We shall however restrict our study here to Maronna’s class of estimators. Of particular interest in the present work is another function $u$, which we shall (somewhat abusively) refer to as Huber’s estimator function $u_H$, defined, for some $t > 0$, as
$$
u_H(x) \triangleq \max \left\{ 1, \frac{1 + t}{t + x} \right\}. \quad \text{(3)}$$
This function has the particularity of being constant for all $x \leq 1$, which will be later seen as an important property.

III. MAIN RESULT

From the problem setting, Assumption [2] and [16, Thm. 2.3], it is easily seen that, with probability one, the solution of (1) is unique for all large $n$ and thus $\hat{C}_N$ is unequivocally defined. In the same spirit as in [19, 10] (and with similar notations), our first objective is to find an explicit tight approximation of the implicitly defined $\hat{C}_N$ in the regime where $N, n \to \infty$ as per Assumption [2]. Our main result unfolds as follows.

**Theorem 1** (Asymptotic Behavior). Let Assumptions [7, 12] hold and let $\hat{C}_N$ the solution to (1) (unique for all large $n$, with probability one). Then, as $n \to \infty$,
$$
\left\| \hat{C}_N - \hat{S}_N \right\| \overset{a.s.}{\to} 0
$$
where
$$
\hat{S}_N \triangleq v(\gamma_n) \frac{1}{n} \sum_{i=1}^{n} (1 - \varepsilon_n) y_i y_i^\top + \frac{1}{n} \sum_{i=1}^{n} v(\alpha_{i,n}) a_i a_i^\top
$$
with $v(x) = u \left( (g^{-1})(x) \right)$, $g(x) = x/(1 - c\phi(x))$, and $(\gamma_n, \alpha_{1,n}, \ldots, \alpha_{\varepsilon_n,n})$ the solution to
$$
\gamma_n = \frac{1}{N} \text{tr} C_N \left( \frac{1 - \varepsilon_n v(\gamma_n)}{1 + cv(\gamma_n)\gamma_n} C_N + \frac{1}{n} \sum_{j=1}^{n} v(\alpha_{j,n}) a_j a_j^\top \right)^{-1},
$$
$$
\alpha_{i,n} = \frac{1}{N} \frac{1}{a_i^\top} \left( \frac{1 - \varepsilon_n v(\gamma_n)}{1 + cv(\gamma_n)\gamma_n} C_N + \frac{1}{n} \sum_{j \neq i} v(\alpha_{j,n}) a_j a_j^\top \right)^{-1} a_i \quad \text{for } i = 1, \ldots, \varepsilon_n. \quad \text{(4)}
$$

As opposed to Maronna’s class of estimators, Tyler estimator is only defined up to a constant factor; thus it estimates $C_N$ up to a scale parameter.

**Remark 1** (Function $v$). The function $v$ defined in Theorem [7] was already introduced in [2] and, through $g$, the assumption that $\phi(x) < e^{-1}$. It has essentially the same general properties as $u$ in that it is continuous, non-increasing and such that $v(x) \equiv x e(x)$ is increasing and bounded with \( \lim_{x \to \infty} v(x) = \psi_\infty = \phi_\infty/(1 - c\phi_\infty) \).

**Remark 2** (Relation to previous results). Taking $\varepsilon_n = 0$, Theorem [7] reduces to the result obtained in [18] and [12], i.e., $\hat{S}_N = v(\gamma_n) y_n y_n^\top$. In this case, (4) reduces to
$$
\gamma_n = \frac{1 + cv(\gamma_n)}{v(\gamma_n)\gamma_n} \frac{1}{\gamma_n}, \quad \text{which, after basic algebra, entails } \gamma_n = \frac{1}{1 - \phi^{-1}(1) - c} \text{ and } v(\gamma_n) = 1/\phi^{-1}(1).
$$

Theorem [11] allows us to transfer many properties of the implicit matrix $\hat{C}_N$ into the more tractable matrix $\hat{S}_N$, the random matrix structure of which is well known and has been studied as early as in [19]. The structure of $\hat{S}_N$ is particularly interesting as it mostly consists of two terms: the sum of outer products of the legitimate data scaled by a constant factor $v(\gamma_n)$ along with a per-sample weighted sum of the outer products of the outlying data. Therefore, as one would expect, $\hat{C}_N$ sets a specific emphasis (either small or large) on each outlying data sample while maintaining all legitimate data under constant weight. We expect here that, as opposed to the SCM that provides no control on the data or to the normalized SCM that merely normalizes the outliers, $\hat{C}_N$ will appropriately ensure a reduction of the outlier impact by letting $v(\alpha_{i,n})$ be quite small compared with $v(\gamma_n)$, especially if $\varepsilon_n$ is small.

An immediate corollary of Theorem [11] concerns the large $N$ eigenvalue distribution of $\hat{C}_N$ and reads as follows.

**Corollary 1** (Spectral Distribution). Define the empirical spectral distribution $F_N(\gamma_n)(x) = \frac{1}{N} \sum_{i=1}^{N} 1_{\lambda_i(\hat{C}_N) \leq x}$ for $x \in \mathbb{R}$. Then, under the setting of Theorem [11],
$$
F_N(\gamma_n)(x) - F_N(x) \overset{a.s.}{\to} 0
$$
almost surely as $n \to \infty$, where $F_N(x)$ is a real distribution function with density defined via its Stieltjes transform $m_N(z)$ (i.e., $m_N(z) = \frac{1}{N} \sum_{i=1}^{N} v(\lambda_i(\hat{C}_N))$) for all $z \in \mathbb{C}^+$ by
$$
m_N(z) = \frac{1}{N} \text{tr} \left( \frac{1 - \varepsilon_n v(\gamma_n)}{1 + cv(\gamma_n)\gamma_n} C_N + A_N - z I_N \right)^{-1},
$$
with $A_N \triangleq \frac{1}{N} \sum_{i=1}^{n} v(\alpha_{i,n}) a_i a_i^\top$ and $e_N(z)$ the unique solution in $\mathbb{C}^+$ of the equation
$$
e_N(z) = \frac{v(\gamma_n)}{n} \text{tr} C_N \left( \frac{1 - \varepsilon_n v(\gamma_n)}{1 + cv(\gamma_n)\gamma_n} C_N + A_N - z I_N \right)^{-1}.
$$

In the appendix, it is importantly shown that $\lim \sup_N \left\| \hat{C}_N \right\| < \infty$ a.s. (as a result of $\lim \sup_N \left\| \hat{S}_N \right\| < \infty$).

**Remark 3** (Estimation error). Recall that any distribution function $F$ is uniquely defined by its Stieltjes transform $m(z)$ by the fact that, for all continuity points $a, b$ of $F$,
$$
F(b) - F(a) = \lim_{y \downarrow 0} \int_{a}^{b} S[m(t + iy)] dt.
$$
as). This implies that $F^C_N$ and $F_N$ have compact supports and are fully determined by their respective moments $M^C_{N,k} \triangleq \int t^k dF^C_N(t)$ and $M_{N,k} \triangleq \int t^k dF_N(t)$, $k = 1, 2, \ldots$, which satisfy $M^C_{N,k} - M_{N,k} \xrightarrow{a.s.} 0$ (by the dominated convergence theorem). While $F_N$ is defined via its deterministic but implicit Stieltjes transform, the $M_{N,k}$ can be retrieved explicitly using successive derivatives of the moment generating formula (for $|z| < 1/\sup(\text{Supp}(F_N))$)

$$m_N(1/z) = -\sum_{k=0}^{\infty} z^{k+1} M_{N,k}.$$ 

Precisely, we obtain here the following result.

**Corollary 2 (Moments).** For $F_N$ defined in Corollary 1 letting $M_{N,p} \triangleq \int t^p dF_N(t)$, $p = 1, 2, \ldots$,

$$M_{N,p} = (-1)^p \frac{1}{p!} \frac{1}{N} \text{tr } T_p$$

where $T_p$ is obtained from the following recursive formulas

$$T_{p+1} = -\sum_{i=0}^{p} T_{p-i} A_N T_i + \sum_{i=0}^{p} \sum_{j=0}^{i} \binom{i}{j} T_{p-i} Q_{i-j+1} T_j$$

$$Q_{p+1} = (p+1)f_p(1-\varepsilon)v(\gamma_n)C_N$$

$$f_{p+1} = \frac{p}{2} \sum_{i=0}^{p} \sum_{j=0}^{i} \binom{i}{j} (-1)^{p-i} \beta_{p-i}$$

$$\beta_{p+1} = v(\gamma_n) \frac{1}{n} \text{tr } C_N T_{p+1},$$

with initial values $T_0 = I_N$, $f_0 = -1$, $\beta_0 = v(\gamma_n) \frac{1}{n} \text{tr } C_N$. In particular,

$$M_{N,1} = \frac{1}{N} \text{tr } [A_N + (1-\varepsilon)v(\gamma_n)C_N]$$

$$M_{N,2} = \frac{1}{N} \text{tr } [A_N^2 + 2(1-\varepsilon)v(\gamma_n)C_N A_N + (1-\varepsilon^2)v^2(\gamma_n)C_N^2 + \left[ \frac{1}{n} \text{tr } C_N \right] (1-\varepsilon)v^2(\gamma_n)C_N],$$

Albeit having characterized the random matrix $\hat{S}_N$, which approximates the behavior of $C_N$ for large $N$, it is quite challenging to gain a good intuitive understanding of the weight structure as the expression relating $\gamma_n$ to the $\alpha_{i,n}$’s is still implicit (while being deterministic). To get more insight on the properties of $\hat{S}_N$, we shall successively consider two specific scenarios that simplify the system.

### IV. FINITELY MANY OUTLIERS SCENARIO

Let us first assume that $\varepsilon_n = K$ is maintained constant as $n \to \infty$ (thus $\varepsilon = 0$). Recall that, in this scenario, Assumption A can be replaced by the sufficient condition

$$\limsup_{n \to \infty} \max_{1 \leq l \leq n} \frac{1}{N} \text{tr } C_N^{-l} \alpha_i < \infty.$$ 

In the appendix, it is shown that $\gamma_n$ cannot grow unbounded with $n$. As such, by a rank-one perturbation argument iterated $K$ times, see e.g., [19], Lemma 2.6], we find that

$$\gamma_n = 1 + \frac{c v(\gamma_n) \gamma_n}{v(\gamma_n)} = O(1/N)$$

which ensures by Remark 2 that

$$\gamma_n = \frac{\phi^{-1}(1)}{1-c} + O(1/N).$$

We shall denote next $\gamma \triangleq \frac{\phi^{-1}(1)}{1-c}$ (and thus $v(\gamma) = 1/\phi^{-1}(1)$). Then we obtain that

$$\hat{S}_N - \hat{S}_N^* \xrightarrow{a.s.} 0$$

with

$$\hat{S}_N = v(\gamma) \left[ \frac{1}{n} \sum_{i=1}^{n-K} y_i^T y_i^\dagger + \frac{K}{n} \sum_{i=1}^{K} v(\alpha_{i,n}^*) \alpha_i \alpha_i^\dagger \right]^{-1} a_1,$$

where $\alpha_{i,n}^*$ are the unique positive solutions to

$$\alpha_{i,n}^* = \frac{1}{N} a_1^\dagger \left( \gamma^{-1} C_N + \frac{1}{n} \sum_{j \neq i} v(\alpha_{j,n}^*) \alpha_j \alpha_j^\dagger \right)^{-1} a_1.$$

As such, when the number $K$ of outliers is fixed, the common weight $v(\gamma_n)$ becomes independent of the vectors $\alpha_i$’s (even if they are of arbitrarily large norm) while the individual weights $v(\alpha_{i,n})$ eventually solve a system of $K$ equations involving the $\alpha_i$’s and $C_N$.

A more specific case lies in the scenario where $a_1 = \ldots = a_K$. There, $\alpha_{1,n} = \ldots = \alpha_{K,n}$ and the $K$ equations above reduce to a single one reading

$$\alpha_{1,n} = \frac{1}{N} a_1^\dagger \left( \gamma^{-1} C_N + \frac{K-1}{n} v(\gamma_n) a_1 \alpha_1^\dagger \right)^{-1} a_1$$

which, using $a_1^\dagger (A + ta a_1)^{-1} = a_1 A^{-1}/(1 + ta a_1^{-1})$ for invertible $A$, simplifies as

$$\alpha_{1,n} = \gamma \frac{1}{N} a_1^\dagger C_N^{-1} a_1$$

or equivalently

$$\alpha_{1,n} = \frac{1}{N} a_1^\dagger C_N^{-1} a_1.$$

Since the right-hand side is positive, so should be the left-hand side, which may then be seen as an increasing function of $\alpha_{1,n}$. Thus, since $\gamma$ depends neither on $C_N$ nor $a_1$, it comes that $\alpha_{1,n}^*$ is an increasing function of $\frac{1}{N} a_1^\dagger C_N^{-1} a_1$. Moreover, $\alpha_{1,n}^* < \psi^{-1}(1/(cn(K-1)))$ and thus converges to zero as $K$ grows large. When $K = 1$, and thus the outlier is now isolated, this reduces to

$$\alpha_{1,n} = \frac{1}{N} a_1^\dagger C_N^{-1} a_1.$$

This short calculus leads to two important remarks. First, for $K = 1$, $\hat{C}_N$ asymptotically allocates a weight $v(\gamma)$ to the legitimate data and a weight $v(\gamma^*) a_1^\dagger C_N^{-1} a_1$ for the outlier. As a consequence, by the non-increasing property of $v$, the effect of the outlier will be (for most choices of the $v$ function) attenuated if $\frac{1}{N} a_1^\dagger C_N^{-1} a_1 \geq 1$ but will be increased if $\frac{1}{N} a_1^\dagger C_N^{-1} a_1 \leq 1$. As such, the robust estimator of scatter will tend to mitigate the effect of outliers $a_1$ having either large norm or, more interestingly, having strong alignment to the weakest eigenmodes of $C_N$. In particular, note that when
Thus should be taken constant for all $x \leq \gamma$, or equivalently, $u(x)$ should be taken constant for $x \leq \phi^{-1}(1)$. A natural choice is the Huber estimator $u = \psi_{1.345}$ introduced in (3).

The second remark is a slightly more surprising outcome. Indeed, despite $n$ being potentially extremely large, the presence of (already few) $K > 1$ identical outliers drives $\hat{S}_N$ (and thus $\hat{C}_N$) to allocate large weights $v(\alpha_i, n)$ (since $\alpha_i, n$ is small) to these outliers, therefore seemingly contradicting the very purpose of the robust estimator. This seems to indicate that $\hat{C}_N$ has the propensity to put forward both large quantities of data with similar distribution as well as rather small quantities of vectors with strong pairwise alignment, while more naturally rejecting isolated outliers.

In terms of large dimensional spectral distribution and moments, the scenario of finitely many outliers is asymptotically equivalent to the outlier-free scenario. This can be observed from a rank-one perturbation argument along with $\varepsilon_n \to 0$ applied to Corollaries 1 and 2. A similar reasoning would hold for the normalized SCM. However, the matrices $C_N$ and $\hat{Y}^n \hat{Y}^n$ themselves experience a (maximum) rank-$K$ perturbation which can severely compromise the estimation of $C_N$, along the previous argumentation lines.

Figure 1 displays an artificially generated scenario where a single outlier $a_1$ of norm $\|a_1\|^2 = 1$ produces a large value for $\|a_1\|^2 C_N^{-1} a_1$ (= 14.50), thus entailing a strong attenuation by $\hat{C}_N$. The terms $a_1$ and $C_N$ were made such that the SCM and normalized SCM have the same asymptotic eigenvalues and produce an isolated eigenvalue (around 1.00), with $\|a_1\|^2 = 1$, $a_1 = (a_{11}^R, 0, 0, ..., 0)^T$, $a_{11}^R \in \mathbb{R}^{10}$, $\alpha_{11} = \sqrt{10}$, and $\gamma_{11} = 0$, so that $\|a_1\|^2 = 1$. $\hat{C}_N$ is a real distribution of Corollary 3 we shall use the subscript “R” standing for “random outliers scenario”.

Corollary 3 (Random Outliers). Let Assumption 2 hold with $\varepsilon > 0$ and let $a_{11}, \ldots, a_{\varepsilon_n, n}$ be random independent of the $y_i$’s with $a_i \sim N_1$, where $D_N \in C^{N \times N}$ is deterministic Hermitian positive definite and $x_1, \ldots, x_\varepsilon_{n}, n$ are independent random vectors with i.i.d. zero mean, unit variance, and finite $(8 + \eta)$-th order moment entries, for some $\eta > 0$. Let us further assume that $\lim_{n \to \infty} \|D_N C_N^{-1} \| < \infty$. Then, as $n \to \infty$,

$$\|\hat{C}_N - \hat{S}_N^R\| \overset{a.s.}{\to} 0$$

where

$$\hat{S}_N^R \& v \gamma_n, n \frac{1}{n} \sum_{i=1}^{\varepsilon_n, n} y_i y_i^\dagger + v \alpha_n, n \frac{1}{n} \sum_{i=1}^{\varepsilon_n, n} a_i a_i^\dagger$$

with $\alpha_n$ and $\gamma_n$ the unique positive solutions to

$$\gamma_n = \frac{1}{N} \text{tr} C_N \left( (1 - \varepsilon) v \gamma_n \right)^{-1} + \frac{\varepsilon v \alpha_n, n}{1 + \varepsilon v \alpha_n, n}$$

$$\alpha_n = \frac{1}{N} \text{tr} D_N \left( (1 - \varepsilon) v \gamma_n \right)^{-1} + \frac{\varepsilon v \alpha_n, n}{1 + \varepsilon v \alpha_n, n}$$

In particular, for $F_N^C(x)$ as defined in Corollary 1

$$F_N^C(x) - F_N(x) \Rightarrow 0$$

almost surely as $n \to \infty$, where $F_N^C(x)$ is a real distribution function with density, defined via its Stieltjes transform

$$m_N^R(z) = \frac{1}{N} \text{tr} E_N^{-1}$$

$$E_N = (1 - \varepsilon) v \gamma_n C_N + \frac{\varepsilon v \alpha_n}{1 + \varepsilon N, x(z)} D_N$$

V. RANDOM OUTLIERS SCENARIO

Assuming $a_1, \ldots, a_{\varepsilon_n, n}$ to be independent with zero mean and covariance $D_N \neq C_N$ provides a rather immediate corollary of Theorem 1 given below. In the results to come, to differentiate between the conditions of Theorem 1 and those
the other to the outlying data. In the defining equations for $\gamma_n^{R}$ and $\alpha_n^{R}$, an interesting symmetrical interplay arises between the weights applied to the legitimate and the outlying data, which are only differentiatied by $\varepsilon$. In particular, if $\varepsilon > 1/2$, the $a_i$'s will be considered legitimate (being in majority) and the $y_i$'s become outliers.

Despite the symmetrical form of the equations defining $\gamma_n^{R}$ and $\alpha_n^{R}$, it remains difficult to extract general insight on these quantities. Thus, again, it is interesting to study the regime where $\varepsilon \to 0$. In this case, $\gamma_n^{R} \to \gamma = \phi^{-1}(1)/(1-c)$, and

$$\alpha_n^{R} \to \gamma \frac{1}{N} \text{tr} D_N C_N^{-1}.$$  

As such, the factor dictating the outlier mitigation strength of $\hat{C}_N$ is now $\frac{1}{N} \text{tr} D_N C_N^{-1}$. Similar to before, when larger than one, the impact of the outliers will be reduced but these might be enhanced when smaller than one. Interestingly, if $\frac{1}{N} \text{tr} D_N = \frac{1}{N} \text{tr} C_N = 1$ (say), both legitimate and outlier samples have similar norm for all large $N$. As such, under this scenario, the SCM $\frac{1}{N} Y Y^\dagger$ or its normalized version $\frac{1}{N} Y^\dagger Y$ behave asymptotically equivalently, neither of which being capable of differentiating between legitimate and outlier data. On the contrary, $\hat{C}_N$ is capable of reducing the impact of the outliers as long as $\frac{1}{N} \text{tr} D_N C_N^{-1} > 1$. Note here again that $\hat{C}_N$ must be sufficiently distinct from $I_N$, which would otherwise entail $\frac{1}{N} \text{tr} D_N C_N^{-1} \approx 1$ and thus $\hat{C}_N$ would be indifferent to outliers. Also, similar to previously, $u$ must be well chosen to avoid enhancing the outlier effect if $\frac{1}{N} \text{tr} D_N C_N^{-1} < 1$ (so in particular it is advised that $u$ be similar to $u_H$).

Figure 4 depicts the previous observations in terms of the deterministic equivalent spectral distributions: $F_N^{R}$ of $\hat{C}_N$, $F_{SCM}^{R}$ of $\frac{1}{N} Y Y^\dagger$ (or $F_{NN}^{R}$ of $\frac{1}{N} Y^\dagger Y$ which satisfies $F_{SCM}^{R} = F_{NN}^{R}$ here), and $F_{oracle}^{R}$ of the outlier-free oracle estimator $\frac{1}{N} Y^\dagger Y$; we take here $C_N$ and $D_N$ to ensure $\frac{1}{N} \text{tr} D_N C_N^{-1}$ large and $\varepsilon$ is taken small. The sought-for distribution that would optimally discard all outliers is the oracle distribution and, thus, highly robust estimators are expected to have a similar distribution. Figure 4 confirms that this is indeed the case of $\hat{C}_N$ which shows a close tail behavior but is slightly mismatched in the main distribution lobe. On the contrary, the SCM (normalized or not) shows a strong decay in the main lobe and a non matching tail. The associated theoretical values of $\gamma_n^{R}$ and $\alpha_n^{R}$ for $\varepsilon_n = .05$ are here

---

**Figure 2.** Density of $F_N^{R}$ versus histogram of $E[F_N^{C_N}]$ for $C_N$ with $(|C_N|_{ij} = .2 |i-j|, D_N = I_N, \varepsilon = .05$, and $u(z) = (t+1)/(t+x)$ where $t = .1$.

**Figure 3.** Mean and standard deviation (error bars) of $||\hat{C}_N - \hat{S}_N^{R}||/||\hat{C}_N||$ for $\varepsilon_n = .25$, $|C_N|_{ij} = .2 |i-j|, D_N = 2 |i-j|$, $\varepsilon_n = .05$, and $u(z) = (1+t)/(t+x)$, with $t = .1$.

For $z \in \mathbb{C}^+$ and $(e_{N,1}(z), e_{N,2}(z))$ the unique solution in $(\mathbb{C}^+)^2$ to

$$e_{N,1}(z) = \frac{u(\gamma_n^{R})}{n} \text{tr} C_N (E_N - z I_N)^{-1},$$

$$e_{N,2}(z) = \frac{u(\alpha_n^{R})}{n} \text{tr} D_N (E_N - z I_N)^{-1}.$$
Fig. 4. Density of the approximate (deterministic) spectral distributions for the outlier-free oracle ($F^\text{oracle}_N$), the SCM or normalized SCM ($F^\text{SCM}_N = F^\text{oracle}_N$), and $\hat{C}_N (FN)$, with $u = u_{1H}$ with parameter $t = .1$, $[C_N]_{ij} = .9^{i-j}$, $D_N = I_N$, $N = 100$, $c = .2$, and $\varepsilon = .05$.

$v_H(\gamma_n^R) \approx 1.00$, $v_H(\alpha_n^R) \approx 1.219$, while in the limit $\varepsilon_n \to 0$, these values become $v_H(\gamma_n^R) \to 1$ and $v_H(\alpha_n^R) \to .1179$.

As it appears from Figure 4 that the tail of the various estimator distributions may be strongly affected by a weak outlier control, it is interesting to investigate the impact on $M^\text{oracle}_N$.

Under the setting of Corollary 1, letting $M^R_N = \int t^p dF^R_N(t)$, we have

$$M^R_N = \frac{(-1)^p}{p!} \frac{1}{N} \text{tr} T^R_p,$$

where $T^R_p$ is obtained recursively as

$$T^R_{p+1} = \sum_{i=0}^p \sum_{j=0}^i \begin{pmatrix} p \atop i \end{pmatrix} \begin{pmatrix} p \atop j \end{pmatrix} T^R_{p-i} Q^R_{i-j+1} T^R_j,$$

$$Q^R_{p+1} = (p+1) \left[ (1-\varepsilon) f_{1,p} R_1 + \varepsilon f_{2,p} R_2 \right],$$

$$f_{k,p+1} = \sum_{i=0}^p \sum_{j=0}^i \begin{pmatrix} p \atop i \end{pmatrix} \begin{pmatrix} p \atop j \end{pmatrix} (p-i+1) f_{k,j} f_{k,i-j} \beta_{k,p-i},$$

$$\beta_{k,p+1} = \frac{1}{n} \text{tr} R_k T^R_{p+1},$$

with initial values $T^R_0 = I_N$, $f_{k,0} = -1$, $\beta_{k,0} = \frac{1}{n} \text{tr} R_k$, and

with $R_1 = \nu(\gamma_n^R) C_N$, $R_2 = \nu(\alpha_n^R) D_N$. In particular,

$$M^R_{N,1} = \frac{1}{N} \text{tr} \left[ \varepsilon \nu(\alpha_n^R) D_N + (1-\varepsilon) \nu(\gamma_n^R) C_N \right].$$

$$M^R_{N,2} = \frac{1}{N} \text{tr} \left[ (\varepsilon \nu(\alpha_n^R) D_N + (1-\varepsilon) \nu(\gamma_n^R) C_N)^2 \right.$$

$$+ \varepsilon^2 \nu(\alpha_n^R)^2 D_N \left[ \frac{1}{n} \text{tr} D_N \right] + (1-\varepsilon)^2 \nu(\gamma_n^R)^2 C_N \left[ \frac{1}{n} \text{tr} C_N \right].$$

As expected, $\hat{C}_N$ induces a bias in the mean. For fair comparison with the normalized SCM, which estimates $C_N$ up to a scale constant, let us define the normalized moments $\bar{M}_{N,p}$ and define similarly $\bar{M}_{N,p}^R$ as well as $\bar{M}_{N,p}^\text{oracle}$ for the SCM, $\bar{M}_{N,p}^\text{SCM}$ for the normalized SCM, and $\bar{M}_{N,p}^\text{oracle}$ for the oracle estimator. Under the same setting as in Figure 4 we provide in the table of Figure 5 the successive normalized moments and relative error compared to $M^\text{oracle}_N$.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$M^\text{oracle}_N$</th>
<th>$\bar{M}_{N,p}^\text{oracle}$ (error)</th>
<th>$\bar{M}_{N,p}^\text{SCM}$ (error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>9.28</td>
<td>9.18 (1.1%)</td>
<td>8.53 (8.2%)</td>
</tr>
<tr>
<td>3</td>
<td>129</td>
<td>126 (1.8%)</td>
<td>112 (13%)</td>
</tr>
<tr>
<td>4</td>
<td>1993</td>
<td>1945 (2.4%)</td>
<td>1600 (17%)</td>
</tr>
</tbody>
</table>

Fig. 5. Normalized moments $\bar{M}_{N,p}^R$, $\bar{M}_{N,p}^\text{SCM}$, versus $\bar{M}_{N,p}^\text{oracle}$ and relative error $| - \bar{M}_{N,p}^\text{oracle} | / \bar{M}_{N,p}^\text{oracle}$. Random outliers, $N = 100$, $c = .2$, $[C_N]_{ij} = .9^{i-j}$, $D_N = I_N$, $\varepsilon = .05$, $u = u_{1H}$, $t = .1$.

VI. DISCUSSION AND CONCLUDING REMARKS

Our study of the robust estimator $\hat{C}_N$ in the large random matrix regime has already led to several interesting conclusions, which we shall more thoroughly address in this section.

Most investigations of robust estimators of scatter focus on the more tractable case where the samples (i.e., the columns of $Y$) are independent with identical elliptical distribution. The recent results of [9, 10] have revealed that, as $u(x)$ gets close to the Tyler $1/x$ function, in the large random matrix regime, $\hat{C}_N$ tends to behave similar to the normalized SCM defined in (3). This conclusion was quite pessimistic as it suggested no real improvement of $\hat{C}_N$ over simplistic alternative robust methods. In the concluding remarks of [10, Section 4], the authors anticipated a change of behavior of $\hat{C}_N$ versus the normalized SCM for deterministic outlier data. This was revealed here both in Section [IV] and in Section [V] where it is made clear that, unlike the normalized SCM, the robust estimators of scatter smartly detect the outliers, essentially by evaluating and comparing the quadratic forms $y^t C_N^{-1} y$ for each column vector $y$ of $Y$. Larger $y^t C_N^{-1} y$ imply more attenuation of $y$ within the observed samples. However, an incidental consequence of this behavior of $\hat{C}_N$ is that small values of $y^t C_N^{-1} y$ enhance the effect of $y$ even though it might not comply with the legitimate sample distribution, thus increasing the probability of inducing false alarms. This has led us to conclude that the function $u$ should be adequately tuned to avoid such a phenomenon. Another consequence is that matrices $\hat{C}_N$ with legitimate data of covariance $C_N$ close to the identity will have very poor outlier rejection properties.

When the outliers are few, the empirical spectral measure $F_{\hat{C}_N}$ of $\hat{C}_N$ is asymptotically the same as that of the SCM,
normalized SCM, and oracle estimators. As such, if one’s interest is on functionals of the eigenvalues of $C_N$, such as moments, and only few outliers are expected, sophisticated robust estimators come to no avail. This being said, the outliers may naturally engender extra isolated eigenvalues (only finitely many) in the spectrum of $\hat{2}YY^\dagger$ which $\hat{C}_N$ might suitably remove while the normalized SCM may not (recall Figure 1). For subspace detection and estimation applications, where the information often lies in the eigenvectors of isolated eigenvalues, discarding such outlying information is critical and thus robust estimators may bring important performance gains. For instance, applications in finance and biostatistics (where data are often assumed to contain outliers) heavily rely on isolated eigenvalue-eigenvector pairs, see e.g., [20, 21]. The experimenter must however keep in mind that, according to our analysis, $\hat{C}_N$ is most effective at automatically suppressing isolated outliers (the less of these relative to the legitimate samples the better) and loses discriminatory power as the outliers approach one another.

The observation made in Section VIII that the distribution (in particular through its first order moments) $F^{\dagger}_N$ is much closer to the oracle estimator than would the (normalized or not) SCM be leads to some interesting applications when it comes to designing improved estimators for $C_N$ that both account for the fact that $n$ is not large compared to $N$ and for the fact that the observed data are prone to outliers. Such investigations were successively made in [22] for the finite $N, n$ regime and later in [10] for the large $N, n$ regime where hybrid Ledoit–Wolf [23] and Tyler [8] estimators were proposed that improve the estimation of $C_N$ by providing an extra degree of freedom (a regularization parameter) which is selected so to maximize the expected Frobenius norm between $C_N$ and the estimator under study. Since the Frobenius norm is nothing but a functional of second order moments, the observation made in the table of Figure 5 strongly suggests that the Ledoit–Wolf estimator alone (being based on the SCM) would be quite sensitive to deterministic outliers while the estimators studied in [10, 22], which are essentially of a similar class as $\hat{C}_N$, would be much more resilient to such outliers.

When the number of outliers is much larger, even in the random outlier scenario studied in Section VIII very little can be said. However, we noticed an interesting symmetry in the equations defining the weights $\gamma_n^R$ and $\delta_n^R$ of Corollary 3 which reveals that the asymptotic proportion $\varepsilon$ of outliers versus $1 - \varepsilon$ of legitimate data could tip for $\varepsilon > .5$ towards letting the outliers be considered as the truly legitimate data.

In summary, the present study provides a first step towards a better understanding of the behavior of (classical) robust estimators of scatter against arbitrary outliers. Our findings underline several key aspects of such estimators of profound practical relevance, such as the importance of the population covariance matrix $C_N$ of the legitimate data in the rejection power of the estimator, as well as the risks inherent to using weight functions $u$ of the Tyler type. Nonetheless, this study remains at the theoretical level of the estimator itself and does not consider the implications when used as a plug-in estimator in detection or estimation methods. Whether these methods are based on local information (isolated eigenvalue, specific eigenvectors, etc.) or global information (functional of the eigenvalues, projections on large subspaces, etc.) about $C_N$ will entail significant differences in the way $\hat{C}_N$, through the weight function $u$, must be tailored. Such considerations are left to future investigations.

**Appendix A**

**Proof of Theorem 1**

The main technical difficulty of the article lies in the proof of Theorem 1 which extends the methods developed in [2] to multiple sample types. The present section is dedicated to this proof. Some auxiliary random matrix results will be then listed in Appendix B while Appendix C will deal with the (rather immediate) proof of Corollary 2.

The proof of Theorem 1 is divided in two parts. First, we show that the system of fixed-point equations (4) admits a unique vector solution and that such solution is bounded as $n \to \infty$. This then defines unequivocally the matrix $\hat{S}_N$. We then show in a second part that $\|\hat{C}_N - \hat{S}_N\| \to \infty$.

**A. Existence, uniqueness, boundedness of the solution to (4)**

To prove existence and uniqueness, we use the framework of standard interference functions (24).

**Definition 1.** A function $h = (h_0, \ldots, h_s) : \mathbb{R}_{+}^{1+s} \to \mathbb{R}_{+}^{1+s}$ is a standard interference function if it satisfies the conditions:

1) **Positivity:** if $q_0, \ldots, q_s \geq 0$, then $h_i(q_0, \ldots, q_s) > 0$ for all $i$.

2) **Monotonicity:** if $q_i \geq q'_i$, $\ldots$, $q_s \geq q'_s$ then, for all $i$, $h_i(q_0, \ldots, q_s) \geq h_i(q'_0, \ldots, q'_s)$.

3) **Scalability:** for all $\delta > 1$ and all $i$, $\delta h_i(q_0, \ldots, q_s) > h_i(\delta q_0, \ldots, \delta q_s)$.

By (24, Thm. 2), if $h$ is a standard interference function for which there exists $(q_0, \ldots, q_s)$ such that $q_i \geq h_i(q_0, \ldots, q_s)$ for all $i$, then the system of equations $q_i = h_i(q_0, \ldots, q_s)$, $i = 0, \ldots, s$, has a unique solution.

Define $h \triangleq (h_0, \ldots, h_{s,n}) : \mathbb{R}_{+}^{1+s} \to \mathbb{R}_{+}^{1+s}$ with

$$h_0(q_0, \ldots, q_{s,n}) = \frac{1}{N} \text{tr} \left( \frac{(1 - \varepsilon)v(q_0)}{1 + cv(q_0)q_0} C_N + \frac{1}{n} \sum_{j=1}^{s} v(q_j)a_j^*a_j \right)^{-1}$$

$$h_i(q_0, \ldots, q_{s,n}) = \frac{1}{N} a_i^* \left( \frac{(1 - \varepsilon)v(q_0)}{1 + cv(q_0)q_0} C_N + \frac{1}{n} \sum_{j \neq i} v(q_j) a_j a_j^* \right)^{-1} a_i$$

for $i = 1, \ldots, s, n$. Let us prove that $h$ meets the conditions of Definition 1 and that, for $i = 0, \ldots, s, n$, $h_i(q_0, \ldots, q_{s,n}) \leq q_i$ for some $(q_0, \ldots, q_{s,n})$, which will then prove existence and uniqueness.

From Assumption 1 and the fact that $v$ is bounded, we clearly have $h_i > 0$ for all $i$. To show monotonicity, let us
first define
\[ B_N(q_0, \ldots, q_{\varepsilon n}) = \frac{(1 - \varepsilon)v(q_0)}{1 + cv(q_0)} C_N + \frac{1}{n} \sum_{j=1}^{\varepsilon n} v(q_j) a_j a_j^\dagger \]
and take \( q_0, \ldots, q_{\varepsilon n} \) and \( q'_0, \ldots, q'_{\varepsilon n} \) such that \( q_i \geq q'_i \) for all \( i \). Then, since \( v \) is non-increasing and \( \psi(x) = xv(x) \) is increasing,
\[ B_N(q_0, \ldots, q_{\varepsilon n}) \preceq B_N(q'_0, \ldots, q'_{\varepsilon n}). \]
From (17) Cor. 7.7.4, this implies
\[ (B_N(q_0, \ldots, q_{\varepsilon n}))^{-1} \geq (B_N(q'_0, \ldots, q'_{\varepsilon n}))^{-1} \]
from which \( h_0(q_0, \ldots, q_{\varepsilon n}) \geq h_0(q'_0, \ldots, q'_{\varepsilon n}) \). By the same arguments, \( h_1(q_0, \ldots, q_{\varepsilon n}) \geq h_1(q'_0, \ldots, q'_{\varepsilon n}) \) for \( i = 1, \ldots, \varepsilon n \), thus proving the monotonicity of \( h \). Finally, to show scalability, let us rewrite \( h_0 \) as
\[ h_0(q_0, \ldots, q_{\varepsilon n}) = \frac{1}{N} \text{tr} \, C_N \left( \left( 1 - \varepsilon \right) \frac{\Theta(q_0)}{q_0} C_N + \frac{1}{n} \sum_{j=1}^{\varepsilon n} \frac{\psi(q_j)}{q_j} a_j a_j^\dagger \right)^{-1} \]
where \( \Theta(x) = \frac{\psi(x)}{1 + cv(x)} \). Since \( \psi(x) \) is increasing, so is \( \Theta(x) \) and, for any \( \delta > 1 \),
\[ h_0(\delta q_0, \ldots, \delta q_{\varepsilon n}) = \frac{\delta}{N} \text{tr} \, C_N \left( \left( 1 - \varepsilon \right) \frac{\Theta(q_0)}{q_0} C_N + \frac{1}{n} \sum_{j=1}^{\varepsilon n} \frac{\psi(q_j)}{q_j} a_j a_j^\dagger \right)^{-1} < \delta h_0(q_0, \ldots, q_{\varepsilon n}). \]
We show similarly \( h_i(\delta q_0, \ldots, \delta q_{\varepsilon n}) < \delta h_i(q_0, \ldots, q_{\varepsilon n}) \) for \( i = 1, \ldots, \varepsilon n \), thus proving the scalability of \( h \).

Thus, \( h \) is a standard interference function and it remains to show that \( h_i(q_0, \ldots, q_{\varepsilon n}) \leq q_i \) for some \( (q_0, \ldots, q_{\varepsilon n}) \) and for all \( i \). For \( i = 0 \),
\[ h_0(q_0, \ldots, q_{\varepsilon n}) = \frac{1}{N} \text{tr} \, C_N \left( B_N(q_0, \ldots, q_{\varepsilon n}) \right)^{-1} \]
where
\[ B_N(q_0, \ldots, q_{\varepsilon n}) \geq \frac{(1 - \varepsilon)v(q_0)}{1 + cv(q_0)} C_N \]
and thus, by definition of \( \psi \),
\[ h_0(q_0, \ldots, q_{\varepsilon n}) \leq \frac{1 + cv(q_0)}{(1 - \varepsilon)\psi(q_0)} q_0. \]
As a consequence, we need to find some \( q_0 \) for which \( \frac{1 + cv(q_0)}{1 - \varepsilon \psi(q_0)} \leq 1 \) or, equivalently, \( \psi(q_0) \geq \frac{1}{1 - \varepsilon} \). Such a choice of \( q_0 \) is always possible since \( \psi \) is increasing on \([0, \infty)\) with image \([0, \psi_\infty)\) where \( \frac{1}{1 - \varepsilon} \leq \psi_\infty \) (this unfolds from \( \phi_\infty > \frac{1}{1 - \varepsilon} \)). Therefore, for any \( q_0 \) such that \( \frac{1}{1 - \varepsilon} \leq \psi(q_0) < \psi_\infty \), we have \( h_0(q_0, \ldots, q_{\varepsilon n}) \leq q_0 \). Take for instance \( q_0 = \psi^{-1}(\frac{1}{1 - \varepsilon}) \) and consider now the functions \( h_i, i = 1, \ldots, \varepsilon n \) for which, using [25] Lemma 10 and similar arguments as above,
\[ h_i(q_0, \ldots, q_{\varepsilon n}) \leq q_0 \left( \frac{1 + cv(q_0)}{(1 - \varepsilon)\psi(q_0)} \right) \frac{1}{N} a_i a_i^\dagger C_N^{-1} \]
\[ = q_0 \frac{1}{N} a_i a_i^\dagger C_N^{-1} \preceq w_i. \]
Therefore, taking \( q_i = w_i \) for \( i = 1, \ldots, \varepsilon n \), we also have \( h_i(q_0, \ldots, q_{\varepsilon n}) \leq q_i \). Altogether, we have shown that the function \( h \) satisfies the conditions of [24] Thm. 2 implying that there exists a unique solution to (4). As such, \( \hat{S}_N \) as introduced in the statement of Theorem [1] is well-defined.

We now turn our focus to the boundedness of the solution to (4). From (6) and (7), along with Assumption [1] we immediately have that \( (\gamma_n, \alpha_{1,n}, \ldots, \alpha_{n,n}) \) is uniformly bounded in \( n \), i.e., \( \limsup_n \gamma_n < \infty \) and \( \limsup_n \max_{1 \leq i \leq n} \alpha_{i,n} < \infty \). Furthermore, \( \gamma_n \) can be shown to be uniformly away from zero as follows. By monotonicity of the \( h \) function, \( h_0(q_0, \ldots, q_{\varepsilon n}) \geq h_0(0, 0, 0) \), i.e.,
\[ h_0(q_0, \ldots, q_{\varepsilon n}) \geq \frac{1}{v(0)} \frac{1}{N} \text{tr} \, C_N^{-1} \geq \frac{1}{v(0)} \frac{1}{\|H_N\|}, \]
where the matrix \( H_N \) is defined as
\[ H_N \triangleq \frac{1}{(1 - \varepsilon)} I_N + \frac{1}{N} \sum_{j=1}^{\varepsilon n} C_N^{-1/2} a_j a_j^\dagger C_N^{-1/2}. \]
By Assumption [1] we have \( \limsup_n \|H_N\| < \infty \) and, consequently, \( \min_n \gamma_n > 0 \).

B. Convergence of \( \hat{C}_N - \hat{S}_N \)

Having proved that \( \hat{S}_N \) is well defined, we now turn to the core of the proof of Theorem [1]. The outline of the proof follows tightly that of [3] Thm. 2 but for a model that is (i) simpler in its assuming the legitimate data to be essentially Gaussian instead of elliptical, but (ii) made more complex due to the deterministic addition of the vectors \( a_1, \ldots, a_n \). Our way to deal with (ii) is by controlling in parallel the quantities asymptotically approximated by \( \gamma_n \) and those asymptotically approximated by \( \alpha_{i,n} \). Since some parts of the proof mirror closely those in [3] Thm. 2], we shall mainly focus on the significantly differing aspects.

First note that we can assume \( C_N = I_N \) by studying \( C_N^{-1/2} \hat{C}_N C_N^{-1/2} \) instead of \( \hat{C}_N \), in which case we have \( C_N^{-1/2} a_i \) in place of the original \( a_i \). This can be seen from (1), the implicit equation solved by \( \hat{C}_N \). Hence, from now on we assume \( C_N = I_N \) without loss of generality. Using the definition \( v(x) \triangleq u(q^{-1}(x)) \) with \( q^{-1}(x) = x/(1 - c \phi(x)) \), and following the same steps as in [3], let us write
\[ \hat{C}_N = \frac{1}{n} \sum_{i=1}^{\varepsilon n} v(d_i) x_i x_i^\dagger + \frac{1}{n} \sum_{i=1}^{\varepsilon n} v(b_i) a_i a_i^\dagger \]
with \( d_i \triangleq \frac{1}{n} \sum_{i=1}^{\varepsilon n} C_N^{-1} x_i x_i^\dagger \) and \( b_i \triangleq \frac{1}{n} \sum_{i=1}^{\varepsilon n} C_N^{-1/2} a_i a_i^\dagger \). Further define
\[ e_i \triangleq \frac{v(d_i)}{v(\gamma_n)} \quad f_i \triangleq \frac{v(b_i)}{v(\alpha_{i,n})}, \]
with \( \gamma_n \) and \( \alpha_{i,n} \) as in the statement of Theorem [1] but for
\[ C_N = I_N, \text{i.e., } \gamma_n \text{ and } \alpha_{i,n} \text{ are the positive solutions to} \]
\[
\gamma_n = \frac{1}{N} \text{tr} \left( \frac{(1-e)v(\gamma_n)}{1 + cv(\gamma_n)} \right)^{-1} I_N + \frac{e_n}{n} \sum_{j=1}^{e_n} v(\alpha_{j,n}) a_j a_j^{\dagger} \right) \]
\[
\alpha_{i,n} = \frac{1}{N} a_i^{\dagger} \left( \frac{1-e}{1 + cu(\gamma_n)} \right)^{-1} I_N + \frac{e_n}{n} \sum_{j \neq i} v(\alpha_{j,n}) a_j a_j^{\dagger} \right) a_i. \]

The core of the proof is to show that
\[
\max_{1 \leq i \leq (1-\epsilon_n)n} |e_i - 1| \to 0 \quad \text{and} \quad \max_{1 \leq i \leq \epsilon_n n} |f_i - 1| \to 0.
\]

Let us first relabel \( e_i \) and \( f_i \) such that \( e_1 \leq \cdots \leq e_{(1-\epsilon_n)n} \) and \( f_1 \leq \cdots \leq f_{\epsilon_n n} \) and denote \( \delta_n = \max(e_{(1-\epsilon_n)n}, f_{\epsilon_n n}) \). For any \( i = 1, \ldots, (1-\epsilon_n)n \), we have
\[
v \left( \frac{1}{N} x_i^{\dagger} \left( \frac{1}{n} \sum_{j \neq i} v(d_j) x_j x_j^{\dagger} + \frac{e_n}{n} \sum_{j=1}^{e_n} v(b_j) a_j a_j^{\dagger} \right) \right) x_i \]
\[
eq v \left( \frac{1}{N} x_i^{\dagger} \left( \frac{1}{n} \sum_{j \neq i} v(\gamma_n) x_j x_j^{\dagger} + \frac{e_n}{n} \sum_{j=1}^{e_n} v(\alpha_{j,n}) a_j a_j^{\dagger} \right) \right) x_i \leq v(\gamma_n)
\]
where we used \( v(d_j) = v(\gamma_n) \), \( v(b_j) = v(\alpha_{j,n}) \), and the inequality arises from \( e_j, f_j \leq \delta_n \), from \( v \) being non-increasing, and from \[17, \text{Cor. 7.7.4}.\] For readability, let
\[
F_{N, (i)} = \frac{1}{n} \sum_{j \neq i} v(\gamma_n) x_j x_j^{\dagger} + \frac{e_n}{n} \sum_{j=1}^{e_n} v(\alpha_{j,n}) a_j a_j^{\dagger}.
\]

From the random matrix result, Lemma 1 of Appendix B
\[
\max_{1 \leq i \leq (1-\epsilon_n)n} \left| \frac{1}{N} x_i^{\dagger} F_{N, (i)}^{-1} x_i - \gamma_n \right| \to 0.
\]

Thus, for \( \zeta > 0 \), with probability one, we have for all large \( n \)
\[
eq e_{(1-\epsilon_n)n} \leq v \left( \frac{1}{N} (\gamma_n - \zeta) \right) \quad \text{and} \quad v \left( \frac{1}{N} (\gamma_n - \zeta) \right) \to 0.
\]

We can proceed similarly to bound \( f_i \) from above as
\[
f_i \leq v \left( \frac{1}{N} a_i^{\dagger} G_{N, (i)}^{-1} a_i \right) \quad \text{for any } i = 1, \ldots, \epsilon_n n, \text{ with}
\]
\[
G_{N, (i)} = \frac{1}{n} \sum_{j=1}^{(1-\epsilon_n)n} v(\gamma_n) x_j x_j^{\dagger} + \frac{e_n}{n} \sum_{j=1}^{e_n} v(\alpha_{j,n}) a_j a_j^{\dagger}
\]
and we now use Lemma 2 in Appendix B which states
\[
\max_{1 \leq i \leq \epsilon_n n} \left| \frac{1}{N} a_i^{\dagger} G_{N, (i)}^{-1} a_i - \alpha_{i,n} \right| \to 0.
\]

Therefore, for the same \( \zeta > 0 \) and for all large \( n \) a.s.,
\[
f_{\epsilon_n n} \leq v \left( \frac{1}{N} (\alpha_{i,n} - \zeta) \right) \quad \text{and} \quad v \left( \frac{1}{N} (\alpha_{i,n} - \zeta) \right) \to 0.
\]

We now consider separately the subsequence of \( n \) over which \( e_{(1-\epsilon_n)n} \geq f_{\epsilon_n n} \) and that over which \( e_{(1-\epsilon_n)n} < f_{\epsilon_n n} \) (these subsequences may be empty or finite).

Subsequence \( e_{(1-\epsilon_n)n} \geq f_{\epsilon_n n} \). On this subsequence, \[10\]

becomes
\[
eq e_{(1-\epsilon)n} \leq v \left( \frac{1}{N} (\gamma_n - \zeta) \right) v(\gamma_n)
\]
or alternatively, since \( e_{(1-\epsilon_n)n} \) is positive,
\[
1 \leq \frac{\psi \left( \frac{\gamma_n}{\psi(\gamma_n)} \right)}{\psi(\gamma_n)} v(\gamma_n)
\]

We want to prove that, for any \( \ell > 0 \), \( e_{(1-\epsilon)n} \leq 1 + \ell \) for all large \( n \) a.s. Let us assume the opposite, i.e., \( e_{(1-\epsilon)n} \geq 1 + \ell \) infinitely often, and let us restrict ourselves to a (further) subsequence where this always holds. Then,
\[
1 \leq \frac{\psi \left( \frac{\gamma_n}{\psi(\gamma_n)} \right)}{\psi(\gamma_n)} v(\gamma_n)
\]

From the uniform boundedness of \( \gamma_n \) away from zero and infinity (see Appendix A-A), considering yet a further subsequence over which \( \gamma_n \to 0 \), we obtain in the limit
\[
\psi(\gamma_n) \left( 1 - \frac{\zeta}{\gamma_n} \right) \leq \psi \left( \frac{\gamma_n}{\psi(\gamma_n)} \right) \left( 1 - \frac{\zeta}{\gamma_n} \right).
\]

This being valid for each \( \zeta > 0 \), a contradiction is raised in the limit \( \zeta \to 0 \). Therefore, either the subsequence over which \( e_{(1-\epsilon)n} \geq f_{\epsilon_n n} \) is finite or \( e_{(1-\epsilon)n} \leq 1 + \ell \) for all large \( n \) a.s. Assuming the former, then \( e_{(1-\epsilon)n} < f_{\epsilon_n n} \) for all large \( n \), which is considered next.

Subsequence \( e_{(1-\epsilon_n)n} < f_{\epsilon_n n} \). On this subsequence, \[11\]

becomes
\[
f_{\epsilon_n n} \leq v \left( \frac{1}{f_{\epsilon_n n}} (\alpha_{i,n} - \zeta) \right) \quad \text{for all large } n \text{ a.s.}
\]

Again, we wish to prove that with, say, the same \( \ell > 0 \) as above, \( f_{\epsilon_n n} \leq 1 + \ell \) for all large \( n \) a.s. Consider first the case \( \lim \inf_n \alpha_{i,n,n} = 0 \) and restrict ourselves to those converging subsequences over which \( \alpha_{i,n,n} \to 0 \). In the limit, \( v(\alpha_{i,n,n}) \to v(0) \) so that, for any \( \theta > 0 \) and for \( n \) large enough, \( v(\alpha_{i,n,n}) > v(0) - \theta \). This, along with \( v(1/f_{\epsilon_n n}(\alpha_{i,n,n} - \zeta)) \leq v(0) \) gives \( f_n \leq v(0)/(v(0) - \theta) \) for all large \( n \) implying that, for any \( \ell > 0 \), \( f_n \leq 1 + \ell \) for all large \( n \) a.s. Consider now the rest of subsequences for which \( \lim \inf_n \alpha_{i,n,n} > 0 \) and rewrite \[12\] as
\[
1 \leq \frac{\psi \left( \frac{\alpha_{i,n,n}}{\psi(\alpha_{i,n,n})} \right)}{\psi(\alpha_{i,n,n})} v(\alpha_{i,n,n})
\]

As above for \( e_{(1-\epsilon_n)n} \), we assume \( f_{\epsilon_n n} > 1 + \ell \) infinitely often, and restrict ourselves to a further subsequence where this holds for all \( n \). Then,
\[
1 \leq \frac{\psi \left( \frac{\alpha_{i,n,n}}{\psi(\alpha_{i,n,n})} \right)}{\psi(\alpha_{i,n,n})} v(\alpha_{i,n,n})
\]
From the boundedness of \(\alpha_{n,n} \) (see Appendix A-A), we can take a converging (further) subsequence over which \(\alpha_{n,n} \to \alpha_0 > 0\). In the limit,

\[
\psi(\alpha_0) \left(1 - \frac{\xi}{\alpha_0}\right) \leq \psi \left(\frac{\alpha_0}{1 + \ell}\right)
\]

which is contradictory for sufficiently small \(\xi\). Thus, necessarily \(f_{\varepsilon_n,n} \leq 1 + \ell\) for all large \(n\) a.s., unless we have \(\epsilon(1-\varepsilon_n)n \geq f_{\varepsilon_n,n}\) in which case, as shown above, \(f_{\varepsilon_n,n} \leq \epsilon(1-\varepsilon_n)n \leq 1 + \ell\) for all large \(n\) a.s.

Altogether, we necessarily have

\[
\max \{e(1-\varepsilon_n)n, f_{\varepsilon_n,n}\} \leq 1 + \ell
\]

for all large \(n\) a.s. All the same, by reverting the inequalities, we prove that, for all large \(n\) a.s.,

\[
\min \{e_1, f_1\} \geq 1 - \ell
\]

and therefore, altogether,

\[
\max \limits_{1 \leq i \leq (1-\ell)n} |e_i - 1| \leq \ell
\]
\[
\max \limits_{1 \leq i \leq n} |f_i - 1| \leq \ell
\]

for all large \(n\) a.s., which eventually proves (8) and (9) by taking a countable sequence of \(\varepsilon_n\) going to zero. This establishes the main result, from which Theorem (1) unfolds. Specifically, from (8)–(9) and by uniform boundedness of \(\gamma_n\) and \(\alpha_{i,n}\),

\[
\max \limits_{1 \leq i \leq (1-\varepsilon_n)n} |v(d_i) - v(\gamma_n)| \xrightarrow{a.s.} 0
\]
\[
\max \limits_{1 \leq i \leq n} |v(b_i) - v(\alpha_{i,n})| \xrightarrow{a.s.} 0.
\]

Thus, for any \(\ell > 0\) and for all large \(n\) a.s.

\[
(1 - \ell)\hat{S}_N \leq \hat{C}_N \leq (1 + \ell)\hat{S}_N
\]

and, therefore \(\|\hat{C}_N - \hat{S}_N\| \leq 2\ell \|\hat{S}_N\|\). Using the triangle inequality and the fact that \(v\) is non-increasing, we have

\[
\|\hat{C}_N - \hat{S}_N\|
\]
\[
\leq 2\ell v(0) \left(\frac{1}{n} \sum_{i=1}^{(1-\varepsilon_n)n} x_i x_i^\top \right) + \frac{1}{n} \sum_{i=1}^{\varepsilon_n,n} a_i a_i^\top \right).
\]

From (26) and Assumption 2 \(\|\frac{1}{n} \sum_{i=1}^{(1-\varepsilon_n)n} x_i x_i^\top \| < 4(1 - \varepsilon)\) for all large \(n\) a.s. and, from Assumption 2 \(\lim \sup \|n \sum_{i=1}^{\varepsilon_n,n} a_i a_i^\top \| < \infty\). Then, since \(\ell\) is arbitrarily small, \(\|\hat{C}_N - \hat{S}_N\|\) tends to zero a.s. as \(n \to \infty\), which concludes the proof of Theorem 1. For \(C_N \neq I\) we simply need to show \(\|C_N^{1/2}(\hat{C}_N - \hat{S}_N)C_N^{1/2}\| \xrightarrow{a.s.} 0\), which follows from \(\|C_N^{1/2}(\hat{C}_N - \hat{S}_N)C_N^{1/2}\| \leq \|C_N\|\|\hat{C}_N - \hat{S}_N\|\) since, by assumption, \(\lim \sup \|C_N\| \leq 1\).

For the random outliers scenario, Assumption 2 holds a.s. by virtue of (26), provided that \(\lim \sup \|D_N C_N^{-1}\| \leq 1\). Then, the proof of Corollary 3 follows from applying standard random matrix arguments to the model of \(\hat{S}_N\) in Theorem 2 considered now as a random matrix in both \(y_i\) and \(a_i\). The result may be straightforwardly obtained from, e.g., [23, Thm. 1] (see Appendix B for similar applications).

### Appendix B

#### RANDOM MATRIX RESULTS

In this section we list several intermediate results needed in Appendix A.

**Lemma 1.** Let Assumptions 0 hold. Define

\[
F_N \triangleq \frac{1}{n} \sum_{j=1}^{(1-\varepsilon_n)n} v(\gamma_n) x_j x_j^\top + \frac{1}{n} \sum_{j=1}^{\varepsilon_n,n} v(\alpha_{j,n}) a_j a_j^\top
\]

and \(F_{N,(i)} = F_N - \frac{1}{n} v(\gamma_n) x_i x_i^\top\), with \(\gamma_n\) and \(\alpha_{j,n}\) given in Theorem 1. Then, as \(n \to \infty\),

\[
\max \limits_{1 \leq i \leq (1-\varepsilon_n)n} \left|\frac{1}{n} x_i F_{N,(i)} x_i - \gamma_n\right| \xrightarrow{a.s.} 0.
\]

**Proof:** We first need to establish a result on \(\lambda_1(\gamma_n)\), for which we know that \(\lambda_1(F_{N,(i)}) \geq \lambda_1(v(\gamma_n) x_i x_i^\top)\). Then, [18, Lemma 1] along with Assumption 2 and the boundedness of \(\gamma_n\) show that there exists \(\xi > 0\) such that, for all large \(n\) a.s.,

\[
\min \limits_{1 \leq i \leq (1-\varepsilon_n)n} \lambda_1(\gamma_n) > \xi.
\]

With this acquired, the outline of the proof is divided into two main steps. We first prove that \(\max_{1 \leq i \leq (1-\varepsilon_n)n} \|v(\gamma_n) x_i x_i^\top - \frac{1}{n} v F_{N} x_i^\top x_i\| \xrightarrow{a.s.} 0\) using quadratic form-close-to-the trace and rank-one perturbation arguments. Then, using [23, Thm. 1], we show that \(\frac{1}{n} \tr F_{N}^{-1} x_i x_i^\top \xrightarrow{a.s.} 0\).

The triangle inequality allows us to write

\[
\left|\frac{1}{n} x_i F_{N,(i)} x_i - \frac{1}{n} \tr F_{N}\right| \leq \left|\frac{1}{n} x_i F_{N,(i)} x_i - \frac{1}{n} \tr F_{N}\right| + \left|\frac{1}{n} \tr F_{N,(i)} - \frac{1}{n} \tr F_{N}\right|.
\]

Let us bound the two terms on the right hand side of (14). Denote by \(E_x\) the expectation with respect to \(x_i\) (i.e., conditionally on \(F_{N,(i)}\)) and \(K_x \triangleq 1_{\lambda_1(\gamma_n) > \xi}\) with \(\xi\) defined in (13). For the first term, we can apply [27, Lemma B.26] (since \(x_i\) is independent of \(\kappa_1^{1/p} F_{N,(i)}^{1/p}\)), so that for \(p > 2\),

\[
E_x \left[\kappa_i \left|\frac{1}{n} x_i F_{N,(i)} x_i - \frac{1}{n} \tr F_{N,(i)}^{1/p}\right|^p\right] \leq \frac{\kappa_i K_x}{N^{p/2}} \left[\frac{1}{N} \tr \left(F_{N,(i)}^{1/p}\right)^2\right]^{p/2} + \frac{\nu_2 p}{N^{p/2}} \tr F_{N,(i)}^{1/p}.
\]

For some constant \(K_p\) depending only on \(p\), with \(\nu_2\) any value such that \(E[|x_{ij}|^2] \leq \nu_2\). Using \(\frac{1}{N} \tr B^k \leq \left(\frac{1}{N} \tr B\right)^k\) for \(B \in \mathbb{C}^{n \times n}\) nonnegative definite and \(k \geq 1\) leads to

\[
E_x \left[\kappa_i \left|\frac{1}{n} x_i F_{N,(i)} x_i - \frac{1}{n} \tr F_{N,(i)}^{1/p}\right|^p\right] \leq \frac{\kappa_i K_p}{N^{p/2}} \left(\nu_2^2/p + \nu_2 p\right) \left[\frac{1}{N} \tr F_{N,(i)}^{1/p}\right]^{p/2} \leq \frac{K_p}{\xi^{p/2} N^{p/2}} \left(\nu_2^2/p + \nu_2 p\right) \left[\frac{1}{N} \tr F_{N,(i)}^{1/p}\right]^{p/2}.
\]
where for the second inequality we have used $\text{tr} B \leq ||B||$ for $B \in \mathbb{C}^{N \times N}$ nonnegative definite and the fact that $\kappa_i ||F_{N, (i)}^{-1}|| < \xi^{-1}$, which holds from the definition of $\kappa_i$. The bound (16) being irrespective of $F_{N, (i)}$, we can now take the expectation over $F_{N, (i)}$ to obtain

$$E \left[ \kappa_i \left| \frac{1}{N} x_i^T F_{N, (i)}^{-1} x_i - \frac{1}{N} \text{tr} F_{N, (i)}^{-1} \right|^p \right] = O \left( \frac{1}{N^{p/2}} \right).$$

(17)

For the second term in (14), we can write $F_{N, (i)} = (F_{N, (i)} - \frac{1}{N} I_N) + \frac{1}{N} I_N$ with $F_{N, (i)} - \frac{1}{N} I_N > 0$ and we have from [19] Lemma 2.6 (rank-one perturbation lemma)

$$E \left[ \kappa_i \left| \frac{1}{N} \text{tr} F_{N, (i)}^{-1} - \frac{1}{N} \text{tr} F_N^{-1} \right|^p \right] \leq \frac{1}{N^p} (2 \xi^{p/2}).$$

(18)

From (14), we can now use Hölder’s inequality and the bounds (17)–(18) to obtain

$$E \left[ \kappa_i \left| \frac{1}{N} x_i^T F_{N, (i)}^{-1} x_i - \frac{1}{N} \text{tr} F_N^{-1} \right|^p \right] = O \left( \frac{1}{N^{p/2}} \right).$$

(19)

Then, we have that

$$\Pr \left[ \max_{1 \leq i \leq (1-\varepsilon_n)n} \kappa_i^{1/p} \left| \frac{1}{N} x_i^T F_{N, (i)}^{-1} x_i - \frac{1}{N} \text{tr} F_N^{-1} \right| > \zeta \right] \leq \sum_{i=1}^{(1-\varepsilon_n)n} \Pr \left[ \kappa_i^{1/p} \left| \frac{1}{N} x_i^T F_{N, (i)}^{-1} x_i - \frac{1}{N} \text{tr} F_N^{-1} \right| > \zeta \right] \leq \frac{(1-\varepsilon_n)n}{\xi^p} E \left[ \kappa_i \left| \frac{1}{N} x_i^T F_{N, (i)}^{-1} x_i - \frac{1}{N} \text{tr} F_N^{-1} \right|^p \right] = O \left( \frac{1}{N^{p/2}} \right).$$

where we have used (in order) Boole’s inequality, Markov’s inequality, and (19). Recall from (13) that the entries of $x_i$ are required to have finite $2\beta$-th order moment and that, by our initial assumption, $E[|x_{ij}|^{8+\gamma}] < \infty$ for some $\gamma > 0$. Then, taking $p > 4$, the Borel Cantelli lemma along with the fact that $\min_{1 \leq i \leq (1-\varepsilon_n)n} \kappa_i \xrightarrow{a.s.} 1$ ensure

$$\max_{1 \leq i \leq (1-\varepsilon_n)n} \left| \frac{1}{N} x_i^T F_{N, (i)}^{-1} x_i - \frac{1}{N} \text{tr} F_N^{-1} \right| \xrightarrow{a.s.} 0.$$  

(20)

It remains to show that $\gamma_n$ is a deterministic equivalent for $\frac{1}{N} \text{tr} F_N^{-1}$. From (13) and the fact that any subtraction of a nonnegative definite matrix cannot increase the smallest eigenvalue, we have that $\lambda_1 (F_N) > \xi$ for all large $n$ a.s. Then, we can write $F_N = (F_N - \frac{1}{N} I_N) + \frac{1}{N} I_N$ with $\lim_{n \to \infty} \lambda_1 (F_N - \frac{1}{N} I_N) > 0$ a.s. and we are in position to apply [22] Thm. 1 which ensures

$$\frac{1}{N} \text{tr} F_N^{-1} - \frac{1}{N} \text{tr} \left( \frac{1-\varepsilon}{1+\varepsilon} v(\gamma_n) \right) \left( \frac{1}{1+\varepsilon} I_N + A_N \right)^{-1} \xrightarrow{a.s.} 0$$

where $A_N = \frac{1}{n} \sum_{j=1}^{n} v(\alpha_{j,n}) \alpha_j \alpha_j^\dagger$ and $e_N$ is the unique positive solution to

$$e_N = c_n v(\gamma_n) \frac{1}{N} \text{tr} \left( \frac{1-\varepsilon}{1+\varepsilon} v(\gamma_n) \right) \left( \frac{1}{1+\varepsilon} I_N + A_N \right)^{-1}.$$  

According to the definition of $\gamma_n$, $c_N = c_n v(\gamma_n)\gamma_n$ with $\gamma_n$ the solution to

$$\gamma_n = \frac{1}{N} \text{tr} \left( \frac{1-\varepsilon}{1+\varepsilon} v(\gamma_n) \right) \left( \frac{1}{1+\varepsilon} I_N + A_N \right)^{-1}$$

which has been proven to be unique. Altogether,

$$\left| \frac{1}{N} \text{tr} F_N^{-1} - \gamma_n \right| \xrightarrow{a.s.} 0.$$  

(21)

Combining (20) and (21) concludes the proof. ■

**Lemma 2.** Let Assumptions [22] hold and define

$$G_{N, (i)} \equiv \frac{1}{n} \sum_{j=1}^{n} v(\gamma_n) x_j x_j^\dagger + \frac{1}{n} \sum_{j \neq i} v(\alpha_{j,n}) \alpha_j \alpha_j^\dagger$$

with $\gamma_n$ and $\alpha_{j,n}$ defined as in Theorem 7. Then, as $n \to \infty,$

$$\max_{1 \leq i \leq (1-\varepsilon_n)n} \left| \frac{1}{N} \text{tr} G_{N, (i)}^{-1} a_i - \alpha_{i,n} \right| \xrightarrow{a.s.} 0.$$  

**Proof:** Since $\lambda_1 (G_{N, (i)}) \geq \lambda_1 (v(\gamma_n) \xi^{1/n} \sum_{j=1}^{n} x_j x_j^\dagger),$ we can use [18] Lemma 1 along with Assumption [2] and the uniform boundedness of $\gamma_n$ to show that there exists $\xi > 0$ such that, for all large $n$ a.s.

$$\min_{1 \leq i \leq (1-\varepsilon_n)n} \lambda_1 (G_{N, (i)}) > \xi.$$  

Denote $\kappa_i \equiv 1 (\lambda_1 (G_{N, (i)}) > \xi).$ Using similar derivations as for [22] Lemma 3 adapted to the present model, we have

$$E \left[ \kappa_i \left| \frac{1}{N} a_i^\dagger G_{N, (i)}^{-1} a_i - \alpha_{i,n} \right|^p \right] = O \left( \frac{1}{N^{p/2}} \right).$$  

(22)

Then

$$\Pr \left[ \max_{1 \leq i \leq (1-\varepsilon_n)n} \kappa_i^{1/p} \left| \frac{1}{N} a_i^\dagger G_{N, (i)}^{-1} a_i - \alpha_{i,n} \right| > \zeta \right]$$

$$\leq \sum_{i=1}^{(1-\varepsilon_n)n} \Pr \left[ \kappa_i^{1/p} \left| \frac{1}{N} a_i^\dagger G_{N, (i)}^{-1} a_i - \alpha_{i,n} \right| > \zeta \right]$$

$$\leq \frac{\varepsilon_n n}{\xi^{p/2}} E \left[ \kappa_i \left| \frac{1}{N} a_i^\dagger G_{N, (i)}^{-1} a_i - \alpha_{i,n} \right|^p \right]$$

$$= O \left( \frac{1}{N^{p/2}} \right)$$

where we used (in order) Boole’s inequality, Markov’s inequality, and (22). Taking $p > 4$, the Borel Cantelli lemma ensures

$$\max_{1 \leq i \leq (1-\varepsilon_n)n} \kappa_i^{1/p} \left| \frac{1}{N} a_i^\dagger G_{N, (i)}^{-1} a_i - \alpha_{i,n} \right| \xrightarrow{a.s.} 0$$

which then proves Lemma 2 using $\min_{1 \leq i \leq (1-\varepsilon_n)n} \kappa_i \xrightarrow{a.s.} 1.$ ■

**APPENDIX C**

**ASYMPTOTIC MOMENTS**

In this last appendix, we derive the moments of the deterministic equivalents studied in [22]. We provide in full the generic result, which may be used for independent purposes. We first recall [22] Thm. 1.

**Theorem 2** (Wagner et al., [22]). Let $Y \in \mathbb{C}^{N \times n}$ have independent columns $y_i = H_i x_i$, where $x_i \in \mathbb{C}^N$, has
i.i.d. entries of zero mean, variance $1/n$, and $4 + \eta$ moment of order $O(1/n^{2+\eta})$, and $H_i \in \mathbb{C}^{N \times N}$ such that $R_i \triangleq HH_i^H$ has uniformly bounded spectral norm over $n, N$. Let also $A_N \in \mathbb{C}^{N \times N}$ be Hermitian non-negative and denote $F_N = YY^H + A_N$. Then, as $N, N_1, \ldots, N_p$, and $n$ grow large with ratios $c_i = N_i/n$, and $c_0 = N/n$ satisfying $0 < \lim inf_n c_i \leq \lim sup_n c_i < \infty$ for $0 \leq i \leq n$, we have

$$
\frac{1}{n} \text{tr} \left( F_N - zI_N \right)^{-1} - m_N(z) \xrightarrow{\text{as}n} 0
$$

with

$$
m_N(z) = \frac{1}{n} \text{tr} \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + e_{N,i}(z)} R_i + A_N - zI_N \right)^{-1}
$$

where $e_{N,1}(z), \ldots, e_{N,n}(z)$ form the unique solution of

$$
e_{N,j}(z) = \frac{1}{n} \text{tr} R_j \left( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + e_{N,i}(z)} R_i + A_N - zI_N \right)^{-1}
$$

such that all $e_{N,j}(z)$ are Stieltjes transforms of a non-negative finite measure on $\mathbb{R}^+$.

From Theorem 2, the distribution function $F_N$ with Stieltjes transform $m_N(z)$ is a deterministic equivalent for the eigenvalue distribution of $F_N$. We next describe the successive moments of the distribution function of $F_N$ generalizes the asymptotic moment results in [29], valid only for $A_N = 0$.

**Theorem 3.** Let $F_N$ be the distribution function associated with the Stieltjes transform $m_N(z)$, and denote $M_{N,0}, M_{N,1}, \ldots$ the successive moments of the distribution function of $F_N$, i.e., $M_{N,p} \triangleq \int x^p dF_N$. Then,

$$
M_{N,p} = \frac{(-1)^p}{p!} \frac{1}{N} \text{tr} T_p
$$

with $T_0, T_1, \ldots$ defined recursively from

$$
T_{p+1} = -\frac{p}{n} T_{p-1} A_N T_1 + \sum_{i=0}^{p} \sum_{j=0}^{i} \left( \frac{p}{i} \left( \frac{i}{j} \right) T_{p-i} Q_{i-j+1} T_j \right)
$$

$$
Q_{p+1} = \frac{p + 1}{n} \sum_{k=1}^{n} f_{k,p} R_k
$$

$$
f_{k,p} = \sum_{i=0}^{p} \sum_{j=0}^{i} \left( \frac{p}{i} \left( \frac{i}{j} \right) (p - i + 1) f_{j,k} f_{i-j-1,k} \beta_{k,p-i} \right)
$$

$$
\beta_{k,p+1} = \frac{1}{n} \text{tr} [R_k T_{p+1}]
$$

and $T_0 = I_N, f_{k,0} = -1, \beta_{k,0} = \frac{1}{n} \text{tr} R_k$ for $k \in \{1, \ldots, n\}$.

**Proof:** Follows the same steps as the proof of [29, Thm. 2] with proper modifications to account for $A_N \neq 0$.

**REFERENCES**


