A CONCENTRATION OF MEASURE PERSPECTIVE TO ROBUST STATISTICS

Cosme Louart

GIPSA-lab, University Grenoble-Alpes
CEA LIST

Romain Couillet

GIPSA-lab, University Grenoble-Alpes
CentraleSupélec, University ParisSaclay

ABSTRACT

We provide promising mathematical considerations for the study of robust scatter matrices in the regime where the data number and dimension are large. Chiefly, we present a new realistic model for data with an assumption inspired from the concentration of measure phenomenon. Our technical contribution is to provide a deterministic equivalent for the robust scatter matrix (i) under relaxed assumptions when compared to the robust statistics literature and (ii) with an original proof based on the introduction of a new semi-metric. This brings simultaneously a new methodological approach to robust statistics analysis and a wider application spectrum to more realistic large dimensional data models.

***Index Terms*—** Scatter matrix, concentration of measure, random matrices, fixed point theorem.

1. INTRODUCTION

Given a set of n data vectors $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p\times n}$ arising from one or several distributions, a classical statistical question is to estimate the population covariance matrix $\mathbb{E}[\frac{1}{n}XX^T]$. This is generally, if not almost always, empirically performed via the sample covariance matrix $\frac{1}{n}XX^T$.

However, there exist cases (presence of outliers, spurious or missing entries, etc.) where the data have a diverging sample covariance due to a heavy tail behavior, resulting in a weak population covariance estimation. To solve this issue one classically exploits, in place of the sample covariance, the so-called *robust scatter matrix* ([1], [2] and [3]) $\hat{C}$ solution of the equation:

$$\hat{C} = \frac{1}{n}\sum_{i=1}^{n} u\left(\frac{1}{n}x_i^T\left(\hat{C}+\gamma I_p\right)^{-1}x_i\right)x_i x_i^T$$

where $\gamma > 0$ is here to ensure the invertibility of $\hat{C}+\gamma I_p$ when $p > n$, and $u : \mathbb{R}^+ \to \mathbb{R}^+$ is in general a decreasing function meant to mitigate (or cut out) the outliers among the dataset. The study is generally conducted under the elliptical model assumption relying on the decomposition $x_i = \sqrt{\tau_i}C^{\frac{1}{2}}z_i$ where $C \in \mathbb{R}^{p\times p}$ is the sought-for scatter matrix and z_i is uniformly distributed on the sphere and independent of τ_i.

Following the approach initiated in [4], we suggest here to relax the hypothesis on the random $w_i = C^{\frac{1}{2}}z_i$ to assume that they belong to the class of "concentrated random vectors" – to visualize this class, for now, just keep in mind that this class contains any 1-Lipschitz transformation of a Gaussian vector. In particular the realistic images built from generative adversarial neural networks (GAN), or more generally any output of neural networks fed in by random Gaussian vectors, belong to this class by construction but we tend to think that even real images and most of the data types commonly explored by machine learning techniques verify this assumption. Yet, to still account for outliers in the data model, as a first extension of the concentrated vector model, we investigate here the case $x_i = \sqrt{\tau_i}w_i$ with w_i concentrated and τ_i heavy-tailed and independent of w_i. It is a very general model in that it allows the entries of the w_i's (that carry the information) to have complex dependence.

A major novelty of the article lies in the introduction of a new semi-metric d_s defined as $d_s(x,y) = |x-y|/\sqrt{xy}$ for $x, y \in \mathbb{R}^+$. Assuming that the mapping u is 1-Lipschitz for this semi-metric (along with some bounding conditions), our main contribution is to prove the "concentration" of the spectral density (i.e., the eigenvalue distribution) of the robust estimator and devise a deterministic limiting measure conditioned on τ.

2. THE CONCENTRATION OF MEASURE FRAMEWORK

We note $\mathcal{M}_{p,n}$ (or simply $\mathcal{M}_p$ if $p = n$) the set of matrices of size $p \times n$ that can be endowed with three possible norms:

- the Frobenius norm $\|M\|_F = \sqrt{\mathrm{Tr}(MM^T)}$
- the operator norm $\|M\| = \sup_{\|x\|\leq 1}\|Mx\|$, where $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$, for $x \in \mathbb{R}^n$
- the nuclear norm $\|M\|_1 = \mathrm{Tr}(\sqrt{MM^T})$ (it is the dual norm of $\|\cdot\|$ for the canonical scalar product on $\mathcal{M}_{p,n}$).

We note $\mathcal{D}_n \subset \mathcal{M}_n$ the set of diagonal matrices and $\mathcal{D}_n^+$ the set of positive diagonal matrices.

Consider $w_1, \ldots, w_n \in \mathbb{R}^p$ seen as n independent drawings of one out of k possible distributions $\mu_1, \ldots \mu_k$ with respective covariances $C_1, \ldots, C_k$. We place ourselves under the classical quasi-asymptotic random matrix setting where p and n are large and of the same order.

Assumption 1. *$p = O(n)$, $n = O(p)$ and $k = O(1)$.*

The first two points of the assumption can be understood as the existence of a positive ratio $c > 0$ such that $\frac{p}{n} \to c$ as $n \to \infty$. In fact, p and n need not be too large as it has repeatedly been shown in practice that the estimators provided by random matrix theory are already quite accurate for p, n of order 10. Then, the number of classes k needs to be at least ten times smaller than p and n.

Let us denote $W = [w_1, \ldots, w_n] \in \mathcal{M}_{n,p}$ for which we wish to devise a concentration hypothesis. Although seemingly artificial, the hypothesis presented here is very general and efficient (as recalled in the introduction and proven in [5]) – when satisfied, we call those random vectors *concentrated vectors*. The important point is that the subsequent concentration inequality is independent of p and n.

Assumption 2. *There exist two constants $C, c > 0$ (such that $C, c = O(1)$) such that, for any 1-Lipschitz function $f : \mathcal{M}_{p,n} \to \mathbb{R}$,*

$$\forall t > 0 : \mathbb{P}\left(|f(W) - \mathbb{E}[f(W)]| \geq t\right) \leq Ce^{-(t/c)^2}.$$

In particular (see [6]), the assumption holds if the vectors w_i are:

- Gaussian vectors with bounded norm covariance,
- uniformly distributed on the sphere $\mathcal{S}^{p-1}$ (or on the ball $\mathcal{B}^p = \{x \in \mathbb{R}^p, \|x\| \leq 1\}$)
- any 1-Lipschitz transformation of the upper cases.

This last item allows one to easily construct a wide range of concentrated vectors, among which random vectors with intricate dependence between their entries (such as randomly generated GAN images). We need a supplementary hypothesis on W to control the non-centered sample covariance matrix $S = \frac{1}{n}WW^T$:

Assumption 3. $\|\mathbb{E}[w_i]\| = O(\sqrt{p})$.

The large dimensional behavior of the spectral measure $\mu_S = \frac{1}{p}\sum_{\lambda \in \mathrm{Sp}(S)} \delta_\lambda$ is classically determined by the Stieltjes transform

$$m_S : t \in \mathcal{C} \setminus \mathrm{Sp}(S) \longmapsto \int_{\mathbb{R}^+} \frac{1}{t - z} d\mu_S(t),$$

where $\mathrm{Sp}(S)$ is the set of eigenvalues of S ([7]). For $z > 0$, we further introduce the resolvent $R_z = (\frac{1}{n}WW^T + zI_p)^{-1}$ that satisfies $m_S(z) = \frac{1}{p}\operatorname{Tr}(R_{-z})$. We showed in [5], under our hypotheses, that R_z is concentrated in the sense of Assumption 2 and that $\|\mathbb{E}[W] - \tilde{R}_z\| = O(\frac{1}{\sqrt{n}})$ for $\tilde{R}_z \in \mathcal{M}_p$ a deterministic matrix that only depends on the (non-centered) population covariance matrices $C_1, \ldots, C_k$ (a more general result is given in Theorem 1 below). Thus, μ_S also converges to a distribution defined only by $C_1, \ldots, C_k$, i.e., as if $w_1, \ldots w_n$ were Gaussian vectors.

Here, however, to extend our results, we are interested in the second order statistics of non concentrated random vectors $X = W\tau^{\frac{1}{2}}$ where $\tau = \mathrm{Diag}(\tau_1, \ldots, \tau_n)$ is a positive random diagonal matrix independent of W and possibly non concentrated in the sense of Assumption 2 (imagine that the τ_i are Student or Cauchy random variables). To study this new setting, we introduce, for a given $z > 0$ and $D \in \mathcal{D}_n^+$, the (respectively random and deterministic) resolvent matrices

$$R_z(D) = \left(\frac{1}{n}WDW^T + zI_p\right)^{-1}$$
$$\tilde{R}_z(D) = \left(\frac{1}{n}\sum_{i=1}^n D_i C_{k(i)} + zI_p\right)^{-1}$$

where $k(i)$ is the class of datum i. We will extensively use the inequalities (in the set of symmetric matrices):

$$\frac{I_p}{z + \frac{1}{n}\|W\|^2} \leq R_z \leq \frac{I_p}{z} \quad \text{and} \quad \frac{I_p}{z + K_C} \leq \tilde{R}_z \leq \frac{I_p}{z} \tag{1}$$

where $K_C = \sup(\|C_a\|)_{1 \leq a \leq k}$. Given $D \in \mathcal{D}_k$, we note $D^{(n)} = \mathrm{Diag}(D_{k(i)}, 1 \leq i \leq n)$, that can be used to adapt the results of [5] and design a deterministic equivalent for $R_z(D)$. The proof of the next proposition relies on tools presented in Section 3 it is thus reported later in the paper (after Theorem 2).

Proposition 1. *For any $D \in \mathcal{D}_n^+$, the fixed point equation*

$$\Lambda = \mathrm{Diag}\left(\frac{1}{n}\operatorname{Tr}\left(C_a \tilde{R}_z\left(\frac{D}{1 + D\Lambda^{(n)}}\right)\right), 1 \leq a \leq k\right)$$

admits a unique solution that we note $\Lambda_z(D)$.

Theorem 1. *Given $D \in \mathcal{D}_n^+$ such that $\|D\| = O(1)$ and a matrix $A \in \mathcal{M}_p$ such that $\|A\|_1 = O(1)$, there exist two constants $C, c = O(1)$ such that:*

$$\mathbb{P}\left(\left|\operatorname{Tr}\left(A\left(R_z(D) - \tilde{R}_z\right)\right)\right| \geq t\right) \leq Ce^{-nt^2/c},$$

with the shortcut notation $\tilde{R}_z = \tilde{R}_z\left(\frac{D}{1 + D\Lambda_z^{(n)}(D)}\right)$.

In particular, since $\|\frac{1}{p}I_p\|_1 = 1 = O(1)$, Theorem 1 provides the concentration of the Stieltjes transform $m_{S(D)}(z) = \frac{1}{p}\operatorname{Tr}(R_z(D))$ of the spectral measure of $S(D) = \frac{1}{n}WDW^T$.

This result cannot be employed directly by replacing D with τ since the assumption $\|\tau\| = O(1)$ would ruin our setting. Instead, as mentioned in the introduction, we look into $\hat{C}$, the robust estimator of scatter for data $X = W\tau^{\frac{1}{2}}$, that solves:

$$\hat{C} = \frac{1}{n}\sum_{i=1}^n u\left(\frac{1}{n}x_i^T(\hat{C} + \gamma I_p)^{-1}x_i\right)x_i x_i^T \tag{2}$$

where $\gamma > 0$, $X = [x_1, \ldots, x_n]$ and u is a well chosen function whose properties are described next.

In the remaining section, we successively prove the existence and uniqueness of $\hat{C}$, and then the convergence properties of its eigenvalue distribution.

3. EXISTENCE AND UNIQUENESS OF $\hat{C}$, AND A NEW SEMI-METRIC

The set $\mathcal{D}_n^+$ can be considered an extension of $\mathbb{R}^+$ since elements behaving mostly like reals. Its commutativity allows us to employ the fractional notation:

$$\frac{D'}{D} = D'D^{-1} = D^{-1}D', \text{ for } D, D' \in \mathcal{D}_n^+.$$

Besides, the spectral norm on $\mathcal{D}_n^+$ is equal to the infinite norm as we have $\|D\| = \|\mathrm{Diag}(D_i, 1 \leq i \leq n)\| = \sup_{1 \leq i \leq n} D_i$.

Definition 1. *For $D, D' \in \mathcal{D}_n^+$, we introduce the semi-metric:*

$$d_s(D, D') = \left\|\frac{D - D'}{\sqrt{DD'}}\right\|.$$

Definition 2. *The class of 1-Lipschitz functions for the semi-metric d_s is called the* stable class*, denoted $\mathcal{S}(\mathcal{D}_n^+)$.*

When on $\mathbb{R}^+$, one can find a very simple characterization of the stable class that is lost on $\mathcal{D}_n^+$:

Property 2. *A function $f : \mathbb{R}^+ \to \mathbb{R}^+$ is stable if and only if $x \mapsto \frac{f(x)}{x}$ is nonincreasing and $x \mapsto xf(x)$ is nondecreasing.*

Sketch of Proof. You can let y tend to x in the inequality:

$$\left|\frac{f(x) - f(y)}{x - y}\right| \leq \sqrt{\frac{f(x)f(y)}{xy}}$$

to obtain the two inequalities $-f(x) \leq xf'(x) \leq f(x)$ that provide the monotonicity of $x \mapsto xf(x)$ and $x \mapsto \frac{f(x)}{x}$. □

The qualifier "stable" indicates the existence of a set of stability properties that we give below.

Property 3. *Given $A \in \mathcal{M}_{p,n}$ with stricly positive entries and $f, g \in \mathcal{S}(\mathcal{D}_n^+)$:*

$$Af \in \mathcal{S}(\mathcal{D}_n^+), \quad \frac{1}{f} \in \mathcal{S}(\mathcal{D}_n^+), \quad \textit{and} \quad f \circ g \in \mathcal{S}(\mathcal{D}_n^+).$$

The example of stable function that interests us particularly is given by Proposition 1.

Proposition 4. *The function*

$$\tilde{I} \; : \; D \longmapsto \mathrm{Diag}\left(\frac{1}{n}\,\mathrm{Tr}\left(C_a \tilde{R}_z(D)\right),\; 1 \le a \le k\right),$$

is stable and its Lipschitz parameter is bounded with

$$\sup\left\{\lambda_{\tilde{I}}(D), D \in \mathcal{D}_n^+, \right\} \quad \textit{where} \quad \lambda_{\tilde{I}}(D) \equiv \left\| 1 - \gamma \tilde{R}_z(D) \right\|$$

It can be shown that the semi-metric space $(\mathcal{D}_n^+, d_s)$ is complete. We thus prove a result quite similar to Picard's fixed point theorem that is at the center of our study.

Theorem 2. *Let $f : \mathcal{D}_n^+ \to \mathcal{D}_n^+$, bounded from above and below, and contracting for the semi-metric d_s. Then there exists a unique $D^* \in \mathcal{D}_n^+$ satisfying $D^* = f(D^*)$.*

This result together with Proposition 4 entices Proposition 1.

Proof of Proposition 1. For any $D \in \mathcal{D}_n^+$, $f_D : \Lambda \mapsto \frac{D}{1+D\Lambda^{(n)}}$ is stable and consequently, $\tilde{I} \circ f$ – unlike $\tilde{I}$ – satisfies the hypothesis of Theorem 2. Indeed, it is contracting thanks to (1):

$$\sup\{\lambda_{\tilde{I}}(f_D(\Lambda)), D \in \mathcal{D}_n^+\} \le \frac{\|D\|\, K_C}{\gamma + \|D\| K_C} < 1,$$

and it is bounded from above and below respectively by $\frac{pK_C}{n\gamma}$ and $\frac{\inf(\frac{1}{n}\,\mathrm{Tr}(C_a))_{1\le a\le k}}{\gamma+\|D\|K_C}$. □

With the same idea, we can set the existence and uniqueness of $\hat{C}$ as defined in (2) thanks to the next assumption:

Assumption 4. *The mapping $u : \mathbb{R}^+ \to \mathbb{R}^+$ is in the stable class $\mathcal{S}(\mathbb{R}^+)$ and is bounded from above.*

Property 2 and the stability rules given by Property 3 to which we can add the stability with the maximum and the minimum (only for mappings defined on $\mathbb{R}^+$) give us a full range of basic operations to construct possible u's. For instance, one can consider $u : t \mapsto \min(t+1, \frac{1}{t+1})$. Although u is defined on $\mathbb{R}^+$, we can apply it on diagonal matrices entry-wise and thus introduce the shortcut notation $u^\tau : D \mapsto \tau u(\tau D)$. Let us also denote:

$$I : D \in \mathcal{D}_n^+ \mapsto \mathrm{Diag}\left(\frac{1}{n} w_i^T R_\gamma(D) w_i, 1 \le i \le n\right) \in \mathcal{D}_n^+$$

Proposition 5. *The fixed point equation*

$$\Delta = I(u^\tau(\Delta)) \tag{3}$$

admits a unique solution $\Delta \in \mathcal{D}_n^+$.

The proof is the same as for Proposition 1, the mapping I could be seen as $\tilde{I}$ where the matrices $C_{k(i)}$ would be replaced by $w_i w_i^T$ and z by γ. The superior bound on u^τ, allows us to conclude the proof the same way as the bound $\|f_D(\Lambda)\| \le \|D\|$ authorized the application of Theorem 2 on $\tilde{I} \circ f$.

We can finally set $\hat{C} = \frac{1}{n} X u(\Delta) X^T$ to retrieve the existence and uniqueness of $\hat{C}$.

4. CONCENTRATION AND ESTIMATE OF $\hat{C}$

In this section we prove our main result: the convergence (concentration) of the eigenvalue distribution of $\hat{C}$ under the conditions of Assumptions 1-4, along with extra conditions to be introduced.

If we try to show directly the concentration of Δ, we rapidly face the dependence between Δ and x_i. As a workaround, we employ the Schur identities. For any $D \in \mathcal{D}_n^+$, denote for simplicity $R(D) = R_\gamma(D)$ and, for $i \in \{1, \ldots, n\}$, denote $W_{-i} = [w_1, \ldots, w_{i-1}, 0, w_{i+1}, \ldots, w_n]$ and $R_{-i}(D) = (W_{-i} D W_{-i}^T + \gamma I_p)^{-1}$, the resolvent $R(D)$ deprived of the contribution of w_i. We then have:

$$R(D) w_i = \frac{R_{-i}(D) w_i}{1 + \frac{D_i}{n} w_i^T R_{-i}(D) w_i}. \tag{4}$$

Let us set $\Delta^- \equiv I^-(u^\tau(\Delta))$ where I^- is the stable mapping:

$$I^- : D \mapsto \mathrm{Diag}\left(\frac{1}{n} w_i^T R_{-i}(D) w_i, 1 \le i \le n\right)$$

(the stability is a similar result to Proposition 4). We have from (4) the identity:

$$\tau\Delta = \frac{1}{\frac{1}{\tau\Delta^-} + u(\tau\Delta)}.$$

We are naturally led to introducing the function $\eta : \mathbb{R}^+ \to \mathbb{R}^+$ satisfying $\tau\Delta = \eta(\tau\Delta^-)$ defined thanks to the next proposition which is again an application of Theorem 2 with the convenient 1-Lipschitz mapping of $(\mathbb{R}^+, d_s)$.

Proposition 6. *Given $x \in \mathbb{R}^+$, there is a unique $\eta(x) \in \mathbb{R}^+$ such that $\eta(x) = \frac{1}{\frac{1}{x} + u(\eta(x))}$. In addition, $\eta \in \mathcal{S}(\mathbb{R}^+)$.*

We can deduce from the fixed point equation (3) satisfied by Δ a fixed point equation satisfied by Δ^- that could be taken as a new definition for Δ^- (here $u_\eta^\tau : D \mapsto \tau\, u \circ \eta(\tau D)$)

$$\Delta^- = I^-(u_\eta^\tau(\Delta^-))$$

To design a deterministic equivalent of Δ^-, we then inspire from a corollary of Theorem 1 that sets for any $D \in \mathcal{D}_n^+$ such that $\|D\| = O(1)$ the existence of two constants $C, c = O(1)$ such that:

$$\mathbb{P}\left(\left|I^-(D) - \Lambda_\gamma(D)_{k(i)}\right| \ge t\right) \le C e^{-nt^2/c}, \tag{5}$$

Thanks again to Theorem 2 we therefore define the deterministic equivalent to be of Δ^- as the unique deterministic (conditionally on τ) diagonal matrix $\tilde{\Delta} \in \mathcal{D}_n^+$ satisfying:

$$\tilde{\Delta} = \Lambda_\gamma(u_\eta^\tau(\tilde{\Delta}^{(n)})). \tag{6}$$

To set the concentration bounds, we need to bound asymptotically (i.e. when p, n tend to ∞) from above and below the diagonal matrix $u_\eta^\tau(\tilde{\Delta})$ (in particular to be able to employ Theorem 1 in the case $D = u_\eta^\tau(\tilde{\Delta})$). We need for that two last assumptions, the first one allows us to bound inferiorly $u_\eta^\tau(\Delta)$.

Assumption 5. $\left\|\frac{1}{\tau}\right\| = \frac{1}{\tau_{\min}} = O(1)$ *and* $\lim_{x\to 0} u(x) > 0$.

Remark 1. *Assumption 5 is very weak because for any pair of independent random matrices $(W, \tau) \in \mathcal{M}_{p,n} \times \mathcal{D}_n^+$ such that W satisfies Assumption 2, one can show that the couple (W', τ') defined with:*

$$\tau' = \mathrm{Diag}\left(\max(\tau_i, 1), 1 \leq i \leq n\right), \quad W' = W \frac{\tau^{\frac{1}{2}}}{\tau'^{\frac{1}{2}}}$$

is such that W' satisfies Assumption 2 and τ' satisfies Assumption 5. Therefore, our proposed model for "real data" is the same with or without Assumption 5.

Assumption 6. *$x \mapsto xu(x)$ is bounded by $\kappa_u < 1$.*

We know from the definition of η that $\forall x > 0$, $\frac{\eta(x)}{x} + \eta(x)u(\eta(x) = 1$, thus, since $x \mapsto xu(x)$ is nondecreasing, $\frac{\eta(x)}{x} \geq 1 - \kappa_u > 0$. Now, given $i \in \{1, \ldots, n\}$:

$$u_\eta^\tau(\tilde{\Delta}^{(n)})_i = \tau_i u(\eta(\tau_i \tilde{\Delta}_{k(i)})) \leq \frac{\tau_i \kappa_u}{\eta(\tau_i \tilde{\Delta}_{k(i)})} \leq \frac{\kappa_u / \tilde{\Delta}_{k(i)}}{1 - \kappa_u} = O(1)$$

thanks to Assumption 5. Note that $\frac{\kappa_u}{1-\kappa_u} = O(1)$ because we implicitly assume that u is independent of n and p (as γ).

Theorem 3. *There exist two constants $C, c > 0$ such that $C, c = O(1)$ and $\forall \varepsilon \in (0, 1)$:*

$$\mathbb{P}\left(\left\|\Delta^- - \tilde{\Delta}\right\| \geq \varepsilon\right) \leq Ce^{-\sqrt{\frac{n}{c \log n}}\varepsilon}.$$

Sketch of proof. The idea is to decompose $\Delta^- - \tilde{\Delta}$ into $\Delta^- - I^-(u_\eta^\tau(\tilde{\Delta})) + I^-(u_\eta^\tau(\tilde{\Delta})) - \tilde{\Delta}$ and to employ the contracting character of $I^- \circ u_\eta^\tau$ (recall that $\Delta^- = I^-(u_\eta^\tau(\Delta^-))$):

$$\begin{aligned}\left\|\frac{\Delta^- - \tilde{\Delta}}{\sqrt{I^-(u_\eta^\tau(\tilde{\Delta}))\Delta^-}}\right\| &\leq \left\|\frac{\Delta^- - I^-(u_\eta^\tau(\tilde{\Delta}))}{\sqrt{I^-(u_\eta^\tau(\tilde{\Delta}))\Delta^-}}\right\| + \left\|\frac{I^-(u_\eta^\tau(\tilde{\Delta})) - \tilde{\Delta}}{\sqrt{I^-(u_\eta^\tau(\tilde{\Delta}))\Delta^-}}\right\| \\ &\leq \lambda_{I^-}\left\|\frac{\Delta^- - \tilde{\Delta}}{\sqrt{\tilde{\Delta}\Delta^-}}\right\| + \left\|\frac{I^-(u_\eta^\tau(\tilde{\Delta})) - \tilde{\Delta}}{\sqrt{I^-(u_\eta^\tau(\tilde{\Delta}))\Delta^-}}\right\|,\end{aligned}$$

with $\lambda_{I^-} = \lambda_{I^-}(\Delta^-) < 1$ (like the parameter $\lambda_{\tilde{I}}(f_D(\Lambda))$ that appeared in the proof of Proposition 1). Since we know from (5) that $\tilde{\Delta} = \Lambda_\gamma(u_\eta^\tau(\tilde{\Delta}))$ is close to $I^-(u_\eta^\tau(\tilde{\Delta}))$, we can derive from this last inequality:

$$\left\|\frac{\tilde{D} - D^-}{\sqrt{I_-(u_\eta^\tau(\tilde{D}))D^-}}\right\| \leq \frac{1 + C}{1 - \lambda}\left\|\frac{I_-(u_\eta^\tau(\tilde{D})) - \tilde{D}}{\sqrt{I_-(u_\eta^\tau(\tilde{D}))D^-}}\right\|.$$

and we conclude again thanks to (5) (and the lower and upper bounds on $u_\eta^\tau(\tilde{D})$ following from Assumptions 5 and 6). □

Recalling that $\hat{C} = \frac{1}{n} W u_\eta^\tau(\Delta^-) W^T$ and letting $\hat{R}_z = R_z(u^\tau(\Delta))$, the resolvent of $\hat{C}$, we now understand the spectral measure $\mu_{\hat{C}}$ of $\hat{C}$ thanks to the next proposition (which is a simple adaptation of Theorem 1):

Corollary 1. *For any $A \in \mathcal{M}_p$ such that $\|A\|_1 = O(1)$, there exist two constants $C, c = O(1)$ such that $\forall \varepsilon \in (0, 1)$:*

$$\mathbb{P}\left(\left|\mathrm{Tr}\left(A\left(\hat{R}_z - \tilde{R}_z^{u_\eta^\tau(\tilde{D})}\right)\right)\right| \geq \varepsilon\right) \leq Ce^{-\sqrt{\frac{n}{c \log n}}\varepsilon}$$

with the short notation $\tilde{R}_z^{u_\eta^\tau(\tilde{D})} = \tilde{R}_z\left(\frac{u_\eta^\tau(\tilde{D})}{1 + u_\eta^\tau(\tilde{D})\Lambda_z^{(n)}(u_\eta^\tau(\tilde{D}))}\right)$.

If we note $m_{\hat{C}}$ the Stieltjes transform of $\mu_{\hat{C}}$, we have the identity $\forall z > 0$:

$$m_{\hat{C}}(z) = \frac{1}{p}\mathrm{Tr}\left(\hat{R}_z\right) = \frac{1}{p}\mathrm{Tr}\left(\left(Wu^\tau(\Delta)W^T + zI_p\right)^{-1}\right).$$

Therefore, as depicted in Figure 1, asymptotically, the spectral measure of $\hat{C}$ is close to a distribution with the deterministic (conditionally on τ) Stieltjes transform $\tilde{m}_{\hat{C}}(z) = \frac{1}{p}\mathrm{Tr}(\tilde{R}_z^{u_\eta^\tau(\tilde{D})})$. This last property induces a lot of inferences on the matrix $\hat{C}$ and, in turn, on the robust kernel matrix $K = \frac{1}{n} u^\tau(\Delta)^{\frac{1}{2}} X^T X u^\tau(\Delta)^{\frac{1}{2}}$ fully studied in our companion article [8].

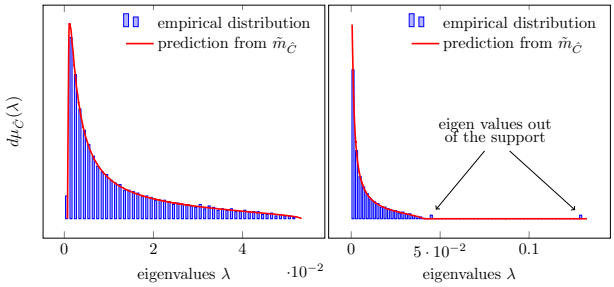

Fig. 1: Histogram of the eigenvalues of the matrix $\hat{C}$ without the mass at 0. We took $X = W\tau^{\frac{1}{2}} \in \mathcal{M}_{p,n}$ with $p = 3000$, $n = 1000$. To avoid a purely Gaussian scenario, we took $W = W' - \mathbb{E}[W']$ with $W' = f(TZ)$, $f = \mathrm{ReLu} : t \mapsto \max(0, t)$, $T = \mathrm{Toeplitz}(0.5^i, 0 \leq i \leq p-1)$ and $Z \in \mathcal{M}_{p,n}$ with independent standard Gaussian entries. To compute Δ, we took $\gamma = 0.1$ and $\tau_i = \max(|\tau_i'|, 0.1)$ where τ_i' is a student random variable with parameter 1. On the left $u = \min(5, 0.1/t)$ and on the right $u = \min(0.005, 0.1/t)$; on this last case, there is a saturation effect on Δ ($\#\{i, u_\eta^\tau(\Delta_i) = 0.005\tau_i\} = 600$) causing a little spreading of the eigen values out of their asymptotic support. This is made possible because the assumption of independence of u towards p and n is not respected ($0.005 \sim \frac{1}{n}$). The predictions are computed thanks to an estimation of the population covariance of W made on $n = 50\,000$ samples.

5. CONCLUDING REMARKS

The model $X = W\tau^{\frac{1}{2}}$ proposed here distinguishes the information – held in W – from a disturbing term – τ – whose effect must be mitigated to efficiently recover the behavior of W. Such a model is not relevant in practice since one has access to W easily dividing the columns of X by their norm. A model commonly found in array processing is rather $X = W\tau^{\frac{1}{2}} + A$ where $A \in \mathcal{M}_{p,n}$ is a deterministic matrix such that $\|A\|_F \leq \sqrt{n}$. Hopefully, in that case, the matrix X writes $X = W'\tau'$ where W' and τ' verify the same assumption as W and τ; consequently, our results are still valid. The main difference though is that the naive vector-wise normalization approach becomes inappropriate and one is therefore led to cleverly choose the stable mapping u that helps recovering A. It is in particular shown in a machine learning classification context in our companion paper [8] that a good choice of u for classification issues can be $u : t \mapsto \frac{1+\alpha}{1+\alpha t}$ for some appaorpriate $\alpha > 0$.

One could also object that the right diagonal action of τ in the model is somewhat limited to match a sufficiently large range of applications.The introduction of a diagonal term $\tau' \in \mathcal{D}_p^+$ that would disrupt the matrix W from the left or a "variance profile" multiplying the entries of W are possible directions of practical extension of the present model, for which the tools of the concentration of measure framework are still valid.

6. REFERENCES

[1] Peter Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics, Vol 35*, pp. 73–101, 1964.

[2] Ricardo Maronna, "Robust m-estimators of multivariate location and scatter," *The Annals of Statistics, Vol 4*, pp. 51–76, 1976.

[3] David Tyler, "A distribution-free m-estimator of multivariate scatter," *The Annals of Statistics, Vol 15*, p. 234–251, 1987.

[4] Mohamed El Amine Seddik, Mohamed Tamaazousti, and Romain Couillet, "Kernel random matrices of large concentrated data : the example of gan-generated images," *ICASSP'19*, 2019.

[5] Cosme Louart and Romain Couillet, "Concentration of measure and large random matrices with an application to sample covariance matrices," *submitted*, 2019.

[6] Michel Ledoux, *The Concentration of Measure Phenomenon*, Mathematical Surveys and Monographs, Number 89, 2001.

[7] Zhidong Bai and Jack W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, Springer Series in Statistics, 2010.

[8] Romain Couillet, "High dimensional robust classification: a random matrix analysis," *submitted to CAMSAP*, 2019.