Large-Dimensional Behavior of Regularized Maronna's M-Estimators of Covariance Matrices

Nicolas Auguin, David Morales-Jimenez, Matthew R. Mckay, Romain Couillet

Large-dimensional behavior of regularized Maronna's M-estimators of covariance matrices

N. Auguin*, D. Morales-Jimenez*, M. R. McKay*, R. Couillet†

***Abstract*—Robust estimators of large covariance matrices are considered, comprising regularized (linear shrinkage) modifications of Maronna's classical M-estimators. These estimators provide robustness to outliers, while simultaneously being well-defined when the number of samples does not exceed the number of variables. By applying tools from random matrix theory, we characterize the asymptotic performance of such estimators when the numbers of samples and variables grow large together. In particular, our results show that, when outliers are absent, many estimators of the regularized-Maronna type share the same asymptotic performance, and for these estimators we present a data-driven method for choosing the asymptotically optimal regularization parameter with respect to a quadratic loss. Robustness in the presence of outliers is then studied: in the non-regularized case, a large-dimensional robustness metric is proposed, and explicitly computed for two particular types of estimators, exhibiting interesting differences depending on the underlying contamination model. The impact of outliers in regularized estimators is then studied, with remarkable differences with respect to the non-regularized case, leading to new practical insights on the choice of particular estimators.**

I. Introduction

Covariance or scatter matrix estimation is a fundamental problem in statistical signal processing [1, 2], with applications ranging from wireless communications [3] to financial engineering [4] and biology [5]. Historically, the sample covariance matrix (SCM) $\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^\dagger$, where $\mathbf{y}_1,\cdots,\mathbf{y}_n \in \mathbb{C}^N$ are zero-mean data samples, has been a particularly appealing choice among possible estimators. The SCM is known to be the maximum likelihood estimator (MLE) of the covariance matrix when the $\mathbf{y}_i$ are independent, identically distributed zero-mean Gaussian observations, and its simple structure makes it easy to implement. Nonetheless, the SCM is known to suffer from three major drawbacks: first, it is not resilient to outliers nor samples of impulsive nature; second, it is a poor estimate of the true covariance matrix whenever the number of samples n and the number of variables N are of similar order; lastly, it is not invertible for $n < N$. The sensitivity to outliers is a particularly important issue in radar-related applications [6, 7], where the background noise usually follows a heavy-tailed distribution, often modelled as a complex elliptical distribution [8, 9]. In such cases, the MLE of the covariance matrix is no longer the SCM. On the other hand, data scarcity is a relevant issue in an ever-growing number of signal processing applications where n and N are generally of similar order, possibly with $n < N$ [4, 5, 10, 11]. New improved covariance estimators are needed to account for both potential data anomalies and high-dimensional scenarios.

In order to harness the effect of outliers and thus provide a better inference of the true covariance matrix, robust estimators known as M-estimators have been designed [9, 12–14]. Their structure is non-trivial, involving matrix fixed-point equations, and their analysis challenging. Nonetheless, significant progress towards understanding these estimators has been made in large-dimensional settings [15–19], motivated by the increasing number of applications where N, n are both large and comparable. Salient messages of these works are: (i) outliers or impulsive data can be handled by these estimators, if appropriately designed (the choice of the specific form of the estimator is important to handle different types of outliers) [19]; (ii) in the absence of outliers, robust M-estimators essentially behave as the SCM and, therefore, are still subject to the data scarcity issue [16].

To alleviate the issue of scarce data, regularized versions of the SCM have originally been proposed [20, 21]. Such estimators consist of a linear combination of the SCM and a shrinkage target (often the identity matrix), which guarantees their invertibility, and often provides a large improvement in accuracy over the SCM when N and n are of the same order. Nevertheless, the regularized SCM (RSCM) inherits the sensitivity of the SCM to outliers/heavy-tailed data. To alleviate both the data scarcity issue and the sensitivity to data anomalies, regularized M-estimators have been proposed [1, 2, 15, 22]. Such estimators are similar in spirit to the RSCM, in that they consist of a combination of a robust M-estimator and a shrinkage target. However, unlike the RSCM, but similar to the estimators studied in [16, 19], these estimators are only defined implicitly as the solution to a matrix fixed-point equation, which makes their analysis particularly challenging.

In this article, we propose to study these robust regularized estimators in the double-asymptotic regime where N and n grow large together. Building upon recent works [17, 23], we will make use of random matrix theory tools to understand the yet-unknown asymptotic behavior of these estimators and subsequently to establish

*N. Auguin, D. Morales-Jimenez, M. R. McKay are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. E-mail: nicolas.auguin@connect.ust.hk, {eedmorales, m.mckay}@ust.hk.

†R. Couillet is with CentraleSupélec and Université Paris-Saclay, Gif-sur-Yvette, France. E-mail: romain.couillet@centralesupelec.fr

N. Auguin, D. Morales-Jimenez and M. R. McKay were supported by the Hong Kong RGC General Research Fund under grant numbers 16206914 and 16203315. R. Couillet's work was supported by the ANR Project RMT4GRAPH (ANR-14-CE28-0006).

design principles aimed at choosing appropriate estimators in different scenarios. In order to do so, we will first study the behavior of these regularized M-estimators in an outlier-free scenario. In this setting, we will show that, upon optimally choosing the regularization parameter, most M-estimators perform asymptotically the same, meaning that the form of the underlying M-estimator does not impact the performance of its regularized counterpart in clean data. Second, we will investigate the effect of the introduction of outliers in the data, under different contamination models. Initial insights were obtained in [19] for non-regularized estimators, focusing on the weights given by the M-estimator to outlying and legitimate data. However, the current study, by proposing an intuitive measure of robustness, takes a more formal approach to qualify the robustness of these estimators. In particular, we will demonstrate which form of M-estimators is preferable given a certain contamination model, first in the non-regularized setting, and then for regularized estimators.

Notation: $||\mathbf{A}||$, $||\mathbf{A}||_F$ and $\operatorname{Tr}\mathbf{A}$ denote the spectral norm, the Frobenius norm and the trace of the matrix $\mathbf{A}$, respectively. The superscript $(\cdot)^\dagger$ stands for Hermitian transpose. Thereafter, we will use $\lambda_1(\mathbf{A}) \leq \cdots \leq \lambda_N(\mathbf{A})$ to denote the ordered eigenvalues of the square matrix $\mathbf{A}$. The statement $\mathbf{A} \succ 0$ (resp. $\succeq 0$) means that the symmetric matrix $\mathbf{A}$ is positive definite (resp. positive semi-definite). The arrow $\xrightarrow{\text{a.s.}}$ designates almost sure convergence, while δ_x denotes the Dirac measure at point x.

II. Review of the large dimensional behavior of non-regularized M-estimators

A. General form of non-regularized M-estimators

In the non-regularized case, robust M-estimators of covariance matrices are defined as the solution (when it exists) to the equation in $\mathbf{Z}$ [13]

$$\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{n} u\left(\frac{1}{N}\mathbf{y}_i^\dagger \mathbf{Z}^{-1}\mathbf{y}_i\right)\mathbf{y}_i\mathbf{y}_i^\dagger, \tag{1}$$

where $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n] \in \mathbb{C}^{N\times n}$ represents the data matrix, and where u satisfies the following properties:

- u is a nonnegative, nonincreasing, bounded, and continuous function on $\mathbb{R}^+$,
- $\phi : x \mapsto xu(x)$ is increasing and bounded, with $\phi_\infty \triangleq \lim_{x\to\infty}\phi(x) > 1$.

If well-defined, the solution of (1) can be obtained via an iterative procedure (see, for example, [16, 24]). Intuitively, the i-th data sample is given a weight $u(\frac{1}{N}\mathbf{y}_i^\dagger\mathbf{Z}^{-1}\mathbf{y}_i)$, which should be smaller for outlying samples than for legitimate ones. The choice of the u function determines the degree of robustness of the M-estimator. As a rule of thumb, the larger ϕ_∞, the more robust the underlying M-estimator to potential extreme outliers [13]. However, such increased robustness is usually achieved at the expense of accuracy [9].

A related and commonly-used estimator is Tyler's [14], which is associated with the unbounded function $u(x) = 1/x$. We remark that, for such u function, the existence of a solution to (1) depends on the sample (see, e.g., [14, 22, 25]). To avoid this issue, we here focus on a wider class of estimators with bounded u functions (as prescribed above). Examples of practical interest, which we study in some detail, include

$$u_{\text{M-Tyler}}(x) \triangleq K\frac{1+t}{t+x} \tag{2}$$

$$u_{\text{M-Huber}}(x) \triangleq K\min\left\{1, \frac{1+t}{t+x}\right\}, \tag{3}$$

for some $t, K > 0$. For a specific t, $u_{\text{M-Tyler}}$ is known to be the MLE of the true covariance matrix when the $\mathbf{y}_i$ are independent, zero-mean, multivariate Student vectors [19], whereas $u_{\text{M-Huber}}$ refers to a modified form of the so-called Huber estimator [26]. Observe that for these functions, $\phi_\infty = K(1+t)$, such that the robustness of the associated M-estimator to extreme outliers is controlled by both t and the scale factor K. In what follows, with a slight abuse of terminology, we will refer to these estimators as "Tyler's" and "Huber's" estimators, respectively.

B. Asymptotic equivalent form under outlier-free data model

Assume now the following "outlier free" data model: let $\mathbf{y}_i$ be N-dimensional data samples, drawn from $\mathbf{y}_i = \mathbf{C}_N^{1/2}\mathbf{x}_i$, where $\mathbf{C}_N \in \mathbb{C}^{N\times N} \succ 0$ is deterministic and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are random vectors, the entries of which are independent with zero mean, unit variance and finite $(8+\sigma)$-th order moment (for some $\sigma > 0$). With this model, we now recall the main result from [16].

Theorem 1. *[16] Assume that $c_N \triangleq N/n \to c \in (0,1)$ as $N, n \to \infty$. Further assume that $0 < \liminf_N\{\lambda_1(\mathbf{C}_N)\} \leq \limsup_N\{\lambda_N(\mathbf{C}_N)\} < \infty$. Then, denoting by $\hat{\mathbf{C}}_N$ a solution to (1), we have*

$$\left\|\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N\right\| \xrightarrow{\text{a.s.}} 0,$$

where $\hat{\mathbf{S}}_N \triangleq \frac{1}{\phi^{-1}(1)}\frac{1}{n}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^\dagger$.

This shows that, up to a multiplying constant, regardless of the choice of u, Maronna's M-estimators behave (asymptotically) like the SCM. As such, in the absence of outliers, no information is lost.

However, Theorem 1 excludes the "under-sampled" case $N \geq n$. Regularized versions of Maronna's M-estimators have been proposed to alleviate this issue, in most cases considering regularized versions of Tyler's estimator ($u(x) = 1/x$) [1, 2, 15, 25], the behavior of which has been studied in [17, 18]. Recently, a regularized M-estimator which accounts for a wider class of u functions has been introduced in [22], but its large-dimensional behavior remains unknown. We address this in the next section. Moreover, note that Theorem 1 does not tell us anything about the behavior of different estimators, associated with different u functions, in the presence of outlying or contaminating data. While

progress to better understand the effect of outliers was recently made in [19], their study focused on non-regularized estimators. In this work, a new measure to characterize the robustness of different M-estimators will be proposed, allowing us to study both non-regularized and regularized estimators (Sections IV and V).

III. Regularized M-estimators: Large dimensional analysis and calibration

A. General form of regularized M-estimators

We consider the class of regularized M-estimators introduced in [22], and given as the unique solution to

$$\mathbf{Z} = (1-\rho)\frac{1}{n}\sum_{i=1}^{n} u\left(\frac{1}{N}\mathbf{y}_i^\dagger \mathbf{Z}^{-1}\mathbf{y}_i\right)\mathbf{y}_i\mathbf{y}_i^\dagger + \rho\mathbf{I}_N \quad , \quad (4)$$

where $\rho \in (0,1]$ is a regularization (or shrinkage) parameter, and where $\mathbf{I}_N$ denotes the identity matrix. The introduction of a regularization parameter allows for a solution to exist when $N > n$. The structure of (4) strongly resembles that of the RSCM, defined as

$$\mathbf{R}(\beta) \triangleq (1-\beta)\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^\dagger + \beta\mathbf{I}_N, \quad (5)$$

where $\beta \in [0,1]$, also referred to as linear shrinkage estimator [27], linear combination estimator [28], diagonal loading [21], or ridge regression [29]. Regularized M-estimators are robust versions of the RSCM.

B. Asymptotic equivalent form under outlier-free data model

Based on recent random matrix theory results, we now characterize the asymptotic behavior of these M-estimators. Under the same data model as that of Section II, we answer the basic question of whether (and to what extent) different regularized estimators, associated with different u functions, are asymptotically equivalent. We need the following assumption on the growth regime and the underlying covariance matrix $\mathbf{C}_N$:

Assumption 1.

a. $c_N \triangleq N/n \to c \in (0,\infty)$ as $N, n \to \infty$.
b. $\limsup_N \{\lambda_N(\mathbf{C}_N)\} < \infty$.
c. $\nu_n \triangleq \frac{1}{N}\sum_{i=1}^{N}\delta_{\lambda_i(\mathbf{C}_N)}$ satisfies $\nu_n \to \nu$ weakly with $\nu \neq \delta_0$ almost everywhere.

Assumption 1 slightly differs from the assumptions of Theorem 1. In particular, the introduction of a regularization parameter now allows $c \geq 1$. Likewise, $\mathbf{C}_N$ is now only required to be positive semidefinite.

For each $\rho \in (0,1]$, we denote by $\hat{\mathbf{C}}_N(\rho)$ the unique solution to (4). We first characterize its behavior in the large n, N regime. To this end, we need the following assumption:

Assumption 2. $\phi_\infty = \lim_{x\to\infty}\phi(x) < \frac{1}{c}$.

We now introduce an additional function, which will be useful in characterizing a matrix equivalent to $\hat{\mathbf{C}}_N(\rho)$.

Definition. Let Assumption 2 hold. Define $v : [0,\infty) \to [u(0), 0)$ as $v(x) = u(g^{-1}(x))$ where g^{-1} denotes the inverse function of $g(x) = \frac{x}{1-(1-\rho)c\phi(x)}$, which maps $[0,\infty)$ onto $[0,\infty)$.

The function v is continuous, non-increasing and onto. We remark that Assumption 2 guarantees that g (and thus v) is properly defined[1]. Note that, importantly, ϕ_∞ does not have to be lower bounded by 1, as opposed to the non-regularized setting. With this in hand, we have the following theorem:

Theorem 2. *Define $\mathcal{I}$ a compact set included in $(0,1]$. Let $\hat{\mathbf{C}}_N(\rho)$ be the unique solution to (4). Then, as $N, n \to \infty$, under Assumptions 1-2,*

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

where

$$\hat{\mathbf{S}}_N(\rho) \triangleq (1-\rho)v(\gamma)\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^\dagger + \rho\mathbf{I}_N,$$

with γ the unique positive solution to the equation

$$\gamma = \frac{1}{N}\operatorname{Tr}\left[\mathbf{C}_N\left((1-\rho)\frac{v(\gamma)}{1+c(1-\rho)v(\gamma)\gamma}\mathbf{C}_N + \rho\mathbf{I}_N\right)^{-1}\right]. \quad (6)$$

Furthermore, the function $\rho \mapsto \gamma(\rho)$ is bounded, continuous on $(0,\infty]$ and greater than zero.

The proof of Theorem 2 (as well as that of the other technical results in this section) is provided in Appendix A.

Remark 1. *The uniform convergence in Theorem 2 will be important for finding the optimal regularization parameter of a given estimator. As a matter of fact, the set $\mathcal{I}$, required to establish such uniform convergence, can be taken as $[0,1]$ in the over-sampled case (provided that $\liminf_N\{\lambda_1(\mathbf{C}_N)\} > 0$).*

Theorem 2 shows that, for every u function, the estimator $\hat{\mathbf{C}}_N(\rho)$ asymptotically behaves (uniformly on $\rho \in \mathcal{I}$) like the RSCM, with weights $\{(1-\rho)v(\gamma), \rho\}$ in lieu of the parameters $\{1-\beta, \beta\}$ in (5). Importantly, the relative weight given to the SCM depends on the underlying u function, which entails that, for a fixed ρ, two different estimators may have different asymptotic behaviors. However, while this is indeed the case, in the following it will be shown that, upon properly choosing the regularization parameter, all regularized M-estimators share the same, optimal asymptotic performance, at least with respect to a quadratic loss.

C. Optimized regularization and asymptotic equivalence of different regularized M-estimators

First, we will demonstrate that any trace-normalized regularized M-estimator is in fact asymptotically equivalent to the RSCM, up to a simple transformation of the regularization parameter. The result is as follows:

[1]Assumption 2 could in fact be relaxed by considering instead the inequality $(1-\rho)\phi_\infty < 1/c$, which therefore enforces a constraint on the choice of both the u function (through ϕ_∞) and the regularization parameter ρ. The proof of Theorem 2 (provided in Appendix) considers this more general case. Nevertheless, for simplicity of exposition, we will avoid this technical aspect in the core of the paper.

Proposition 1. *Let Assumptions 1-2 hold. For each $\rho \in (0,1]$, the parameter $\underline{\rho} \in (0,1]$ defined as*

$$\underline{\rho} \triangleq \frac{\rho}{(1-\rho)v(\gamma)+\rho} \tag{7}$$

is such that

$$\frac{\hat{\mathbf{S}}_N(\rho)}{\frac{1}{N}\operatorname{Tr}\hat{\mathbf{S}}_N(\rho)} = \frac{\mathbf{R}(\underline{\rho})}{\frac{1}{N}\operatorname{Tr}\mathbf{R}(\underline{\rho})}, \tag{8}$$

where we recall that $\mathbf{R}(\underline{\rho}) = (1-\underline{\rho})\frac{1}{n}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^\dagger + \underline{\rho}\mathbf{I}_N$.

Reciprocally, for each $\underline{\rho} \in (0,1]$, there exists a solution $\rho \in (0,1]$ to the equation (7) for which equality (8) holds.

Proposition 1 implies that, in the absence of outliers, any (trace-normalized) estimator $\hat{\mathbf{S}}_N(\rho)$ is equal to a trace-normalized RSCM estimator with a regularization parameter $\underline{\rho}$ depending on ρ and on the underlying u function (through v). From Theorem 2, it then follows that the estimator $\hat{\mathbf{C}}_N(\rho)$ asymptotically behaves like the RSCM estimator with parameter $\underline{\rho}$.

Thanks to Proposition 1, we may thus look for optimal asymptotic choices of ρ. Given an estimator $\hat{\mathbf{B}}_N$ of $\mathbf{C}_N$, define the quadratic loss of the associated trace-normalized estimator as:

$$\mathcal{L}\left(\frac{\hat{\mathbf{B}}_N}{\frac{1}{N}\operatorname{Tr}\hat{\mathbf{B}}_N}, \frac{\mathbf{C}_N}{\frac{1}{N}\operatorname{Tr}\mathbf{C}_N}\right) \triangleq \frac{1}{N}\left\|\frac{\hat{\mathbf{B}}_N}{\frac{1}{N}\operatorname{Tr}\hat{\mathbf{B}}_N} - \frac{\mathbf{C}_N}{\frac{1}{N}\operatorname{Tr}\mathbf{C}_N}\right\|_F^2.$$

We then have the following proposition:

Proposition 2. (Optimal regularization)

Let Assumptions 1 and 2 hold. Define

$$\mathcal{L}^\star \triangleq c\frac{M_2-1}{c+M_2-1}M_1^2$$
$$\rho^\star \triangleq \frac{c}{c+M_2-1},$$

where $M_1 \triangleq \int t\nu(dt)$ and $M_2 \triangleq \int t^2\nu(dt)$. Then,

$$\inf_{\rho\in\mathcal{I}} \mathcal{L}\left(\frac{\hat{\mathbf{C}}_N(\rho)}{\frac{1}{N}\operatorname{Tr}\hat{\mathbf{C}}_N(\rho)}, \frac{\mathbf{C}_N}{\frac{1}{N}\operatorname{Tr}\mathbf{C}_N}\right) \xrightarrow{\text{a.s.}} \mathcal{L}^\star.$$

Furthermore, for $\hat{\rho}^\star$ a solution to $\frac{\hat{\rho}^\star}{(1-\hat{\rho}^\star)v(\gamma)+\hat{\rho}^\star} = \rho^\star$,

$$\mathcal{L}\left(\frac{\hat{\mathbf{C}}_N(\hat{\rho}^\star)}{\frac{1}{N}\operatorname{Tr}\hat{\mathbf{C}}_N(\hat{\rho}^\star)}, \frac{\mathbf{C}_N}{\frac{1}{N}\operatorname{Tr}\mathbf{C}_N}\right) \xrightarrow{\text{a.s.}} \mathcal{L}^\star.$$

(Optimal regularization parameter estimate)

The solution $\hat{\rho}_N \in \mathcal{I}$ to

$$\frac{\hat{\rho}_N}{\frac{1}{N}\operatorname{Tr}\hat{\mathbf{C}}_N(\hat{\rho}_N)} = \frac{c_N}{\frac{1}{N}\operatorname{Tr}\left[\left(\frac{1}{n}\sum_{i=1}^n \frac{\mathbf{y}_i\mathbf{y}_i^\dagger}{\frac{1}{N}\|\mathbf{y}_i\|^2}\right)^2\right] - 1},$$

satisfies

$$\hat{\rho}_N \xrightarrow{\text{a.s.}} \hat{\rho}^\star$$
$$\mathcal{L}\left(\frac{\hat{\mathbf{C}}_N(\hat{\rho}_N)}{\frac{1}{N}\operatorname{Tr}\hat{\mathbf{C}}_N(\hat{\rho}_N)}, \frac{\mathbf{C}_N}{\frac{1}{N}\operatorname{Tr}\mathbf{C}_N}\right) \xrightarrow{\text{a.s.}} \mathcal{L}^\star.$$

Proposition 2 states that, irrespective of the choice of u, there exists some ρ for which the quadratic loss of the corresponding regularized M-estimator is minimal, this minimum being the same as the minimum achieved by an optimally-regularized RSCM. The last result of Proposition 2 provides a simple way to estimate this optimal parameter.

In the following, we validate these theoretical findings through simulation. Let $[\mathbf{C}_N]_{ij} = .9^{|i-j|}$, and consider the u functions specified in (2) and (3), with $K = 1/c_N$ and $t = 0.1$. For $\rho \in (0,1]$, Fig. 1 depicts the expected quadratic loss $\mathcal{L}$ associated with the solution $\hat{\mathbf{C}}_N(\rho)$ of (4) and that associated with the RSCM (line curves), along with the expected quadratic loss associated with the random equivalent $\hat{\mathbf{S}}_N(\rho)$ of Tyler's and Huber's estimators (marker).

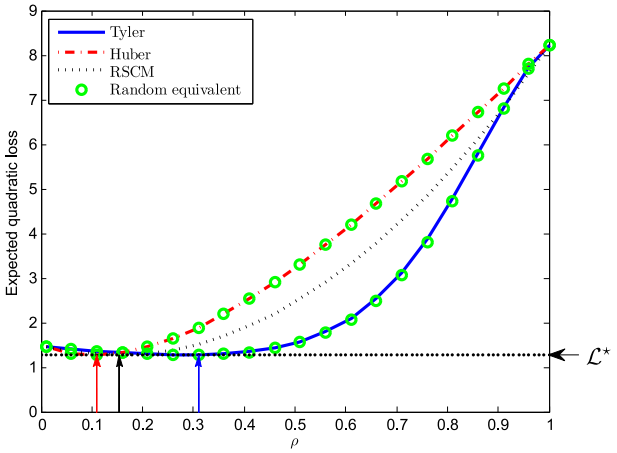


Fig. 1: Expected quadratic loss of different estimators as ρ varies, for $N = 150$, $n = 100$ ($c_N = 3/2$), and $[\mathbf{C}_N]_{ij} = .9^{|i-j|}$, averaged over 100 realizations. Arrows indicate the estimated optimal regularization parameters for the considered estimators, while $\mathcal{L}^\star$ indicates the asymptotic, minimal quadratic loss.

For both u functions and all $\rho \in (0,1]$, there is a close match between the quadratic loss of $\hat{\mathbf{C}}_N(\rho)$ and that of $\hat{\mathbf{S}}_N(\rho)$. This shows the accuracy of the (asymptotic) equivalence of $\hat{\mathbf{C}}_N(\rho)$ and $\hat{\mathbf{S}}_N(\rho)$ described in Theorem 2. As suggested by our analysis, while the estimators associated with different u functions have different performances for a given ρ, they have the same performance when ρ is optimized, with a quadratic loss approaching $\mathcal{L}^\star$ for N large. Furthermore, the optimal regularization parameter for a given u function is accurately estimated, as shown by the arrows in Fig. 1.

IV. Large-dimensional robustness: non-regularized case

In this section, we turn to the case where the data is contaminated by random outliers and study the robustness of M-estimators for distinct u functions. Some initial insight has been previously provided in [19] for the non-regularized case. Specifically, that study focused on the comparison of the weights given by the estimator to outlying and legitimate samples. Albeit insightful, the analysis in [19]

did not directly assess robustness, understood as the impact of outliers on the estimator's performance. Here we propose a different approach to analyze robustness, by introducing and evaluating a robustness metric which measures the bias induced by data contamination.

We start by studying non-regularized estimators (thereby excluding the case $c_N \geq 1$), which are technically easier to handle. This will provide insight on the capabilities of different M-estimators to harness outlying samples. Then, in the following section, we will investigate how this study translates to the regularized case. The proofs of the technical results in this section are provided in Appendix B.

A. Asymptotic equivalent form under outlier data model

We focus on a particular type of contamination model where outlying samples follow a distribution different from that of the legitimate samples. Similar to [19], the data matrix $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_{(1-\epsilon_n)n}, \mathbf{a}_1, \cdots, \mathbf{a}_{\epsilon_n n}] \in \mathbb{C}^N$ is constructed with the first $(1-\epsilon_n)n$ data samples $(\mathbf{y}_1, \cdots, \mathbf{y}_{(1-\epsilon_n)n})$ being the legitimate data, and following the same distribution as in Sections II and III (that is, they verify $\mathbf{y}_i = \mathbf{C}_N^{1/2}\mathbf{x}_i$). The remaining $\epsilon_n n$ "contaminating" data samples $(\mathbf{a}_1, \cdots, \mathbf{a}_{\epsilon_n n})$ are assumed to be random, independent of the $\mathbf{y}_i$, with $\mathbf{a}_i = \mathbf{D}_N^{1/2}\mathbf{x}'_i$, where $\mathbf{D}_N \in \mathbb{C}^{N\times N}$ is deterministic positive definite and $\mathbf{x}'_1, \cdots, \mathbf{x}'_{\epsilon_n n}$ are independent random vectors with i.i.d. zero mean, unit variance, and finite $(8+\eta)$-th order moment entries, for some $\eta > 0$.

To characterize the asymptotic behavior of M-estimators for this data model, we require the following assumptions on the growth regime and on the underlying covariance matrices $\mathbf{C}_N$ and $\mathbf{D}_N$:

Assumption 3.

a. $\epsilon_n \to \epsilon \in [0,1)$ and $c_N \to c \in (0,1)$ as $N, n \to \infty$.
b. $0 < \liminf_N \{\lambda_1(\mathbf{C}_N)\} \leq \limsup_N \{\lambda_N(\mathbf{C}_N)\} < \infty$.
c. $\limsup_N \left\|\mathbf{D}_N \mathbf{C}_N^{-1}\right\| < \infty$.

Let us consider a function u with now $1 < \phi_\infty < \frac{1}{c}$. For such a u function, the equation in $\mathbf{Z}$

$$\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{(1-\epsilon_n)n} u\left(\frac{1}{N}\mathbf{y}_i^\dagger \mathbf{Z}^{-1}\mathbf{y}_i\right)\mathbf{y}_i\mathbf{y}_i^\dagger + \frac{1}{n}\sum_{i=1}^{\epsilon_n n} u\left(\frac{1}{N}\mathbf{a}_i^\dagger \mathbf{Z}^{-1}\mathbf{a}_i\right)\mathbf{a}_i\mathbf{a}_i^\dagger \tag{9}$$

has a unique solution [13], hereafter referred to as $\hat{\mathbf{C}}_N^{\epsilon_n}$.

In this setting, we have the following result:

Theorem 3. *Let Assumption 3 hold and let $\hat{\mathbf{C}}_N^{\epsilon}$ be the unique solution to (9). Then, as $N, n \to \infty$,*

$$\left\|\hat{\mathbf{C}}_N^{\epsilon_n} - \hat{\mathbf{S}}_N^{\epsilon_n}\right\| \xrightarrow{\text{a.s.}} 0$$

where

$$\hat{\mathbf{S}}_N^{\epsilon_n} \triangleq v(\gamma^{\epsilon_n})\frac{1}{n}\sum_{i=1}^{(1-\epsilon_n)n}\mathbf{y}_i\mathbf{y}_i^\dagger + v(\alpha^{\epsilon_n})\frac{1}{n}\sum_{i=1}^{\epsilon_n n}\mathbf{a}_i\mathbf{a}_i^\dagger,$$

with γ^{ϵ_n} and α^{ϵ_n} the unique positive solutions to:

$$\gamma^{\epsilon_n} = \frac{1}{N}\operatorname{Tr}\mathbf{C}_N\mathbf{B}_N^{-1}$$
$$\alpha^{\epsilon_n} = \frac{1}{N}\operatorname{Tr}\mathbf{D}_N\mathbf{B}_N^{-1},$$

with

$$\mathbf{B}_N \triangleq \left(\frac{(1-\epsilon_n)v(\gamma^{\epsilon_n})}{1+cv(\gamma^{\epsilon_n})\gamma^{\epsilon_n}}\mathbf{C}_N + \frac{\epsilon_n v(\alpha^{\epsilon_n})}{1+cv(\alpha^{\epsilon_n})\alpha^{\epsilon_n}}\mathbf{D}_N\right).$$

Theorem 3 shows that $\hat{\mathbf{C}}_N^{\epsilon}$ behaves similar to a weighted SCM, with the legitimate samples weighted by $v(\gamma^{\epsilon_n})$, and the outlying samples by $v(\alpha^{\epsilon_n})$. This result generalizes [19, Corollary 3] to allow for $\epsilon \in [0,1)$, without the constraint $(1-\epsilon)^{-1} < \phi_\infty < \frac{1}{c}$ (along with $c < 1-\epsilon$).

A scenario which will be of particular interest in the following concerns the case where there is a vanishingly small proportion of outliers. This occurs when $\epsilon_n = O(1/n^\mu)$ for some $0 < \mu \leq 1$, in which case $\epsilon_n \to \epsilon = 0$. For this scenario, the weights given to the legitimate and outlying data are

$$\gamma^0 \triangleq \lim_{n\to\infty}\gamma^{\epsilon_n} = \frac{\phi^{-1}(1)}{1-c} \tag{10}$$
$$\alpha^0 \triangleq \lim_{n\to\infty}\alpha^{\epsilon_n} = \gamma^0\frac{1}{N}\operatorname{Tr}(\mathbf{C}_N^{-1}\mathbf{D}_N), \tag{11}$$

respectively.

In the following, we exploit the form of $\hat{\mathbf{S}}_N^{\epsilon_n}$ to characterize the effect of random outliers on the estimator $\hat{\mathbf{C}}_N^{\epsilon_n}$.

B. Robustness analysis

Let $\hat{\mathbf{C}}_N^0$ be the solution to (1), and $\hat{\mathbf{C}}_N^{\epsilon_n}$ the solution to (9). We propose the following metric, termed *measure of influence*, to assess the robustness of a given estimator to an ϵ-contamination of the data:

Definition 1. *For $\epsilon_n \to \epsilon \in [0,1)$, the measure of influence $\mathrm{MI}(\epsilon_n)$ is given by*

$$\mathrm{MI}(\epsilon_n) \triangleq \left\|\mathbb{E}\left[\frac{\hat{\mathbf{C}}_N^0}{\frac{1}{N}\operatorname{Tr}\hat{\mathbf{C}}_N^0} - \frac{\hat{\mathbf{C}}_N^{\epsilon_n}}{\frac{1}{N}\operatorname{Tr}\hat{\mathbf{C}}_N^{\epsilon_n}}\right]\right\|.$$

For simplicity, we assume hereafter that $\frac{1}{N}\operatorname{Tr}\mathbf{C}_N = \frac{1}{N}\operatorname{Tr}\mathbf{D}_N = 1$ for all N. From Theorems 1 and 3, we have the following:

Corollary 1. *As $N, n \to \infty$,*

$$\mathrm{MI}(\epsilon_n) - \overline{\mathrm{MI}}(\epsilon_n) \to 0,$$

where

$$\overline{\mathrm{MI}}(\epsilon_n) = \frac{\epsilon_n v(\alpha^{\epsilon_n})}{(1-\epsilon_n)v(\gamma^{\epsilon_n}) + \epsilon_n v(\alpha^{\epsilon_n})}\left\|\mathbf{C}_N - \mathbf{D}_N\right\|. \tag{12}$$

Note that $\lim_{\epsilon_n\to 0}\overline{\mathrm{MI}}(\epsilon_n) = \overline{\mathrm{MI}}(0) = 0$, as expected. The result (12) shows that $\overline{\mathrm{MI}}$ is globally influenced by $\|\mathbf{C}_N - \mathbf{D}_N\|$, which is also an intuitive result, since it

suggests that the more "different" $\mathbf{D}_N$ is from $\mathbf{C}_N$, the higher the influence of the outliers on the estimator. To get clearer insight on the effect of a small proportion of outliers, assuming $\epsilon_n = O(1/n^\mu)$ for some $0 < \mu \leq 1$, we compute

$$\overline{\mathrm{IMI}} \triangleq \lim_{n\to\infty} \frac{1}{\epsilon_n} \overline{\mathrm{MI}}(\epsilon_n), \tag{13}$$

which we will refer to as the infinitesimal measure of influence (IMI).

From (12) and (13),

$$\overline{\mathrm{IMI}} = \frac{v(\alpha^0)}{v(\gamma^0)} \left\| \mathbf{C}_N - \mathbf{D}_N \right\|, \tag{14}$$

with γ^0, α^0 given in (10) and (11), respectively.

For particular u functions, these general results reduce to even simpler forms: for example, for u functions such that $\phi^{-1}(1) = 1$ (such as $u_{\mathrm{M-Tyler}} = \frac{1+t}{t+x}$ or $u_{\mathrm{M-Huber}} = \min\{1, \frac{1+t}{t+x}\}$), which entails $v(\gamma^0) = 1$, (14) further yields

$$\overline{\mathrm{IMI}} = v\left(\frac{\frac{1}{N}\operatorname{Tr} \mathbf{C}_N^{-1}\mathbf{D}_N}{1-c} \right) \left\| \mathbf{C}_N - \mathbf{D}_N \right\|.$$

Further, considering t small, the IMI associated with $u_{\mathrm{M-Tyler}}$ and $u_{\mathrm{M-Huber}}$ can be approximated as

$$\overline{\mathrm{IMI}}_{\mathrm{M-Tyler}} \simeq \frac{1+t}{t + \frac{1}{N}\operatorname{Tr} \mathbf{C}_N^{-1}\mathbf{D}_N} \left\| \mathbf{C}_N - \mathbf{D}_N \right\| \tag{15}$$

and

$$\overline{\mathrm{IMI}}_{\mathrm{M-Huber}} \simeq \begin{cases} \left\| \mathbf{C}_N - \mathbf{D}_N \right\| & \text{if } \frac{1}{N}\operatorname{Tr} \mathbf{C}_N^{-1}\mathbf{D}_N \leq 1 \\ \overline{\mathrm{IMI}}_{\mathrm{M-Tyler}} & \text{if } \frac{1}{N}\operatorname{Tr} \mathbf{C}_N^{-1}\mathbf{D}_N > 1 \end{cases}. \tag{16}$$

Hence, when $\frac{1}{N}\operatorname{Tr} \mathbf{C}_N^{-1}\mathbf{D}_N \leq 1$, $\overline{\mathrm{IMI}}_{\mathrm{M-Huber}} \leq \overline{\mathrm{IMI}}_{\mathrm{M-Tyler}}$, which shows that the influence of an infinitesimal fraction of outliers is higher for Tyler's estimator than for Huber's. In contrast, when $\frac{1}{N}\operatorname{Tr} \mathbf{C}_N^{-1}\mathbf{D}_N > 1$, both Huber's and Tyler's estimators exhibit the same IMI.

For comparison, the measure of influence of the SCM can be written as

$$\overline{\mathrm{MI}}_{\mathrm{SCM}}(\epsilon_n) = \epsilon_n \left\| \mathbf{C}_N - \mathbf{D}_N \right\|,$$

which is linear in ϵ_n. It follows immediately that

$$\overline{\mathrm{IMI}}_{\mathrm{SCM}} = \left\| \mathbf{C}_N - \mathbf{D}_N \right\|. \tag{17}$$

The fact that $\overline{\mathrm{IMI}}_{\mathrm{SCM}}$ is bounded may seem surprising, since it is known that a single arbitrary outlier can arbitrarily bias the SCM [30], however we recall that the current model focuses on a particular random outlier scenario. From (12), the SCM is more affected than given M-estimators by the introduction of outliers if and only if

$$\overline{\mathrm{MI}}(\epsilon_n) \leq \overline{\mathrm{MI}}_{\mathrm{SCM}}(\epsilon_n) \Leftrightarrow v(\alpha^{\epsilon_n}) \leq v(\gamma^{\epsilon_n}).$$

This further legitimizes the study in [19], which focused on these weights to assess the robustness of a given M-estimator. However, in the regularized case it will be shown that the relationship between the relative weights and robustness is more complex (see Subsection V-B).

Fig. 2 depicts the measure of influence $\overline{\mathrm{MI}}(\epsilon_n)$ for different u functions, as the proportion ϵ_n of outlying samples increases. For every u function ($u_{\mathrm{M-Tyler}}$ or $u_{\mathrm{M-Huber}}$), we take $t = 0.1$. In addition, we show the measure of influence of the SCM, as well as the linear approximation $\epsilon_n \mapsto \epsilon_n \overline{\mathrm{IMI}}$ (computed using (15), (16) and (17)) of the measure of influence in the neighborhood of $\epsilon = 0$. We first set $[\mathbf{C}_N]_{ij} = .9^{|i-j|}$ and $[\mathbf{D}_N]_{ij} = .2^{|i-j|}$ (such that $\frac{1}{N}\operatorname{Tr} \mathbf{C}_N^{-1}\mathbf{D}_N > 1$), and then swap the roles of $\mathbf{C}_N$ and $\mathbf{D}_N$ (such that $\frac{1}{N}\operatorname{Tr} \mathbf{C}_N^{-1}\mathbf{D}_N < 1$).

In the case where $\frac{1}{N}\operatorname{Tr} \mathbf{C}_N^{-1}\mathbf{D}_N > 1$, Fig. 2 confirms that the measure of influence of both Tyler's and Huber's estimators is lower than that of the SCM, as corroborated by the fact that $\overline{\mathrm{IMI}}_{\mathrm{M-Tyler}}, \overline{\mathrm{IMI}}_{\mathrm{M-Huber}} < \left\| \mathbf{C}_N - \mathbf{D}_N \right\| = \overline{\mathrm{IMI}}_{\mathrm{SCM}}$ (see (15), (16)). This shows that in the case where $\mathbf{C}_N$ is more "structured" than $\mathbf{D}_N$, the considered M-estimators are more robust to the introduction of outliers than the SCM. Furthermore, both Tyler's and Huber's estimators exhibit the same robustness for small ϵ_n. However, in the opposite case where $\frac{1}{N}\operatorname{Tr} \mathbf{C}_N^{-1}\mathbf{D}_N < 1$, Tyler's estimator is much less robust than both Huber's estimator and the SCM, which are both equally robust (for small ϵ_n). Since both $\mathbf{C}_N$ and $\mathbf{D}_N$ are unknown in practice, it suggests that choosing Huber's estimator is preferable over Tyler's.

V. Large dimensional robustness: Regularized case

We now turn to the regularized case, which, in particular, allows $c \geq 1$. The proofs of the technical results in this section are provided in Appendix C.

A. Asymptotic equivalent form under outlier data model

To facilitate the robustness study of regularized M-estimators, we start by analyzing the large-dimensional behavior of these estimators in the presence of outliers.

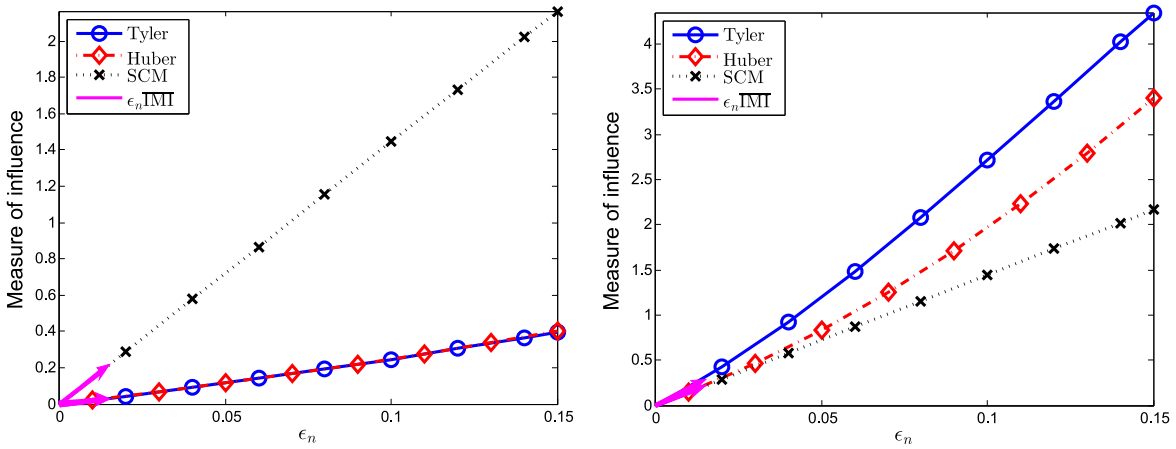
For $\rho \in \mathcal{R} = (\rho_0, 1]$, where $\rho_0 = \max\{0, 1 - \frac{1}{c\phi_\infty}\}$, we define the regularized estimator $\hat{\mathbf{C}}_N^{\epsilon}(\rho)$ associated with the function u as the unique solution to the equation in $\mathbf{Z}$:

$$\begin{aligned} \mathbf{Z} = (1-\rho)\frac{1}{n}\sum_{i=1}^{(1-\epsilon_n)n} u\left(\frac{1}{N}\mathbf{y}_i^\dagger \mathbf{Z}^{-1}\mathbf{y}_i\right)\mathbf{y}_i\mathbf{y}_i^\dagger \\ + (1-\rho)\frac{1}{n}\sum_{i=1}^{\epsilon_n n} u\left(\frac{1}{N}\mathbf{a}_i^\dagger \mathbf{Z}^{-1}\mathbf{a}_i\right)\mathbf{a}_i\mathbf{a}_i^\dagger + \rho\mathbf{I}_N. \end{aligned} \tag{18}$$

Remark 2. *If we assume that $\phi_\infty < \frac{1}{c}$, then the range of admissible ρ is $\mathcal{R} = (0, 1]$. Furthermore, if $c < 1$, we can in fact take $\mathcal{R} = [0, 1]$. In the following, we assume that $\phi_\infty < \frac{1}{c}$. Note that, similar to the outlier-free scenario, the introduction of a regularization parameter allows us to relax the assumption of $\phi_\infty > 1$.*

Theorem 4. *Assume the same contaminated data model as in Theorem 3. Let Assumptions 1-2 hold and let $\hat{\mathbf{C}}_N^{\epsilon_n}(\rho)$ be the unique solution to (18). Then, as $N, n \to \infty$, for all $\rho \in \mathcal{R}$,*

$$\left\| \hat{\mathbf{C}}_N^{\epsilon_n}(\rho) - \hat{\mathbf{S}}_N^{\epsilon_n}(\rho) \right\| \xrightarrow{\text{a.s.}} 0$$



(a) $[\mathbf{C}_N]_{ij} = .9^{|i-j|}$ and $[\mathbf{D}_N]_{ij} = .2^{|i-j|}$. (b) $[\mathbf{C}_N]_{ij} = .2^{|i-j|}$ and $[\mathbf{D}_N]_{ij} = .9^{|i-j|}$.

Fig. 2: Measure of influence for $\epsilon_n \in [0, 0.15]$, in the non-regularized case for $N = 50$, $n = 200$ ($c_N = 1/4$).

where

$$\hat{\mathbf{S}}_N^{\epsilon_n}(\rho) \triangleq (1-\rho)v(\gamma^{\epsilon_n})\frac{1}{n}\sum_{i=1}^{(1-\epsilon_n)n}\mathbf{y}_i\mathbf{y}_i^\dagger + (1-\rho)v(\alpha^{\epsilon_n})\frac{1}{n}\sum_{i=1}^{\epsilon_n n}\mathbf{a}_i\mathbf{a}_i^\dagger + \rho\mathbf{I}_N,$$

with γ^{ϵ_n} and α^{ϵ_n} the unique positive solutions to:

$$\gamma^{\epsilon_n} = \frac{1}{N}\operatorname{Tr}\mathbf{C}_N\mathbf{B}_N^{-1} \qquad (19)$$
$$\alpha^{\epsilon_n} = \frac{1}{N}\operatorname{Tr}\mathbf{D}_N\mathbf{B}_N^{-1},$$

with

$$\mathbf{B}_N \triangleq (1-\rho)\frac{(1-\epsilon_n)v(\gamma^{\epsilon_n})}{1+(1-\rho)cv(\gamma^{\epsilon_n})\gamma^{\epsilon_n}}\mathbf{C}_N + (1-\rho)\frac{\epsilon_n v(\alpha^{\epsilon_n})}{1+(1-\rho)cv(\alpha^{\epsilon_n})\alpha^{\epsilon_n}}\mathbf{D}_N + \rho\mathbf{I}_N.$$

Remark 3. *In the case $\epsilon_n = 0$ (no outliers), Theorem 4 reduces to Theorem 2, while in the case $\rho = 0$ (if $c < 1$), it reduces to Theorem 3.*

B. Robustness analysis

Similar to the non-regularized case, we next make use of $\hat{\mathbf{S}}_N^{\epsilon_n}(\rho)$ to study the robustness of $\hat{\mathbf{C}}_N^{\epsilon_n}(\rho)$. Importantly, introducing a regularization parameter entails an additional variable to consider when studying the robustness of M-estimators.

We denote by $\hat{\mathbf{C}}_N^0(\rho)$ the solution to (18) for a given $\rho \in \mathcal{R}$ when there are no outliers. Similar to the non-regularized case, we define

$$\mathrm{MI}(\rho, \epsilon_n) \triangleq \left\| \mathbb{E}\left[\frac{\hat{\mathbf{C}}_N^0(\rho)}{\frac{1}{N}\operatorname{Tr}\hat{\mathbf{C}}_N^0(\rho)} - \frac{\hat{\mathbf{C}}_N^{\epsilon_n}(\rho)}{\frac{1}{N}\operatorname{Tr}\hat{\mathbf{C}}_N^{\epsilon_n}(\rho)} \right] \right\|.$$

By Theorem 4, we have the following corollary:

Corollary 2. *Let the same assumptions as in Theorem 4 hold. Then,*

$$\forall \rho \in \mathcal{R}, \quad \mathrm{MI}(\rho, \epsilon_n) - \overline{\mathrm{MI}}(\rho, \epsilon_n) \to 0,$$

with

$$\overline{\mathrm{MI}}(\rho, \epsilon_n) = \left\| \frac{U(\epsilon_n, \rho)}{V(\epsilon_n, \rho)} \right\|,$$

with $U(\epsilon_n, \rho), V(\epsilon_n, \rho)$ defined as

$$\begin{aligned} U(\epsilon_n, \rho) &= \rho(1-\rho)((1-\epsilon_n)v(\gamma^{\epsilon_n}) - v(\gamma^0))(\mathbf{C}_N - \mathbf{I}_N) \\ &\quad + \rho(1-\rho)\epsilon_n v(\alpha^{\epsilon_n})(\mathbf{D}_N - \mathbf{I}_N) \\ &\quad + (1-\rho)^2\epsilon_n v(\gamma^0)v(\alpha^{\epsilon_n})(\mathbf{D}_N - \mathbf{C}_N) \\ V(\epsilon_n, \rho) &= ((1-\rho)(1-\epsilon_n)v(\gamma^{\epsilon_n}) \\ &\quad + (1-\rho)\epsilon_n v(\alpha^{\epsilon_n}) + \rho)((1-\rho)v(\gamma^0) + \rho). \end{aligned}$$

Unlike in the non-regularized case, the form of $\overline{\mathrm{MI}}(\rho, \epsilon_n)$ renders the analysis difficult in general. For the specific case $\rho = 1$ however, $\overline{\mathrm{MI}}(1, \epsilon_n) = 0$ for all ϵ_n. This is intuitive, and reflects the fact that the more we regularize an estimator, the more robust it becomes (eventually, it boils down to taking $\hat{\mathbf{C}}_N = \mathbf{I}_N$). This extreme regularization, however, leads to a significant bias, and is therefore not desirable.

In the following, we focus on the infinitesimal measure of influence associated with $\mathrm{MI}(\rho, \epsilon_n)$, which is defined in a similar way as for the non-regularized case: assume $\epsilon_n = O(1/n^\mu)$ for some $0 < \mu \leq 1$ ($\epsilon_n \to \epsilon = 0$). Then, for v smooth enough[2], and $\rho \in \mathcal{R}$,

$$\overline{\mathrm{IMI}}(\rho) \triangleq \lim_{n\to\infty} \frac{1}{\epsilon_n}\overline{\mathrm{MI}}(\rho, \epsilon_n).$$

Corollary 2 allows us to compute $\overline{\mathrm{IMI}}(\rho)$ explicitly in the particular case where $\frac{1}{N}\operatorname{Tr}\mathbf{C}_N = \frac{1}{N}\operatorname{Tr}\mathbf{D}_N = 1$ for all N. This is given as follows:

Corollary 3. *Let the same assumptions as in Theorem 4 hold. If $\gamma \mapsto v(\gamma)$ is differentiable in the neighborhood of*

[2]Precise details are provided in Appendix C4.

$\gamma^0 = \lim_{n\to\infty} \gamma^{\epsilon_n}$,

$$\overline{\mathrm{IMI}}(\rho) = \frac{1}{((1-\rho)v(\gamma^0)+\rho)^2} \left\|\mathbf{G}(\rho)\right\|, \qquad (20)$$

where

$$\begin{aligned}\mathbf{G}(\rho) &= \rho(1-\rho)[v(\alpha^0)(\mathbf{D}_N - \mathbf{I}_N) - v(\gamma^0)(\mathbf{C}_N - \mathbf{I}_N)] \\ &\quad + (1-\rho)^2 v(\gamma^0)v(\alpha^0)(\mathbf{D}_N - \mathbf{C}_N) \\ &\quad + \rho(1-\rho)\left.\frac{dv}{d\epsilon_n}\right|_{\epsilon_n=0}(\mathbf{C}_N - \mathbf{I}_N),\end{aligned}$$

and where

$$\begin{aligned}\alpha^0 &\triangleq \lim_{n\to\infty} \alpha^{\epsilon_n} \\ &= \frac{1}{N}\operatorname{Tr}\mathbf{D}_N \left(\frac{(1-\rho)v(\gamma^0)}{1+(1-\rho)c\gamma^0 v(\gamma^0)}\mathbf{C}_N + \rho\mathbf{I}_N\right)^{-1}.\end{aligned}$$

Details on how to evaluate $\left.\frac{dv}{d\epsilon}\right|_{\epsilon=0}$ are given in Appendix C4.

While the intricate expression $\mathbf{G}(\rho)$ does not yield simple analytical insight for an arbitrary regularization parameter $\rho \in \mathcal{R}$, it can still be leveraged to numerically assess the robustness of regularized estimators, as we show below. As a given ρ plays an a priori different role for distinct M-estimators, a direct comparison of $\overline{\mathrm{MI}}(\rho, \epsilon_n)$ or $\overline{\mathrm{IMI}}(\rho)$ (for fixed ρ) for different estimators is not meaningful. However, using Proposition 2, we can choose $\rho = \rho^\star$ such that, in the absence of outliers, a given estimator's quadratic loss is minimal. This allows us to meaningfully compare how robust these estimators are to a small proportion of outliers.

For our subsequent numerical studies, we will focus on the scenario where $\mathbf{C}_N$ is more structured than $\mathbf{D}_N$. In the alternative case ($\mathbf{D}_N$ more structured than $\mathbf{C}_N$), the differences between Huber, Tyler, and the RSCM are marginal (at least for small ϵ). In Fig. 3, we compute the measure of influence $\overline{\mathrm{MI}}(\rho^\star, \epsilon_n)$ of the RSCM and of the estimators associated with $u_{\mathrm{M-Tyler}}$ and $u_{\mathrm{M-Huber}}$ (with $K = 1/c_N$), as ϵ_n increases. We also plot the linear approximation $\epsilon_n \mapsto \epsilon_n \overline{\mathrm{IMI}}(\rho^\star)$ (computed using (15), (16) and (17)) of $\overline{\mathrm{MI}}$ in the neighborhood of $\epsilon_n = 0$. We observe that the MI of Tyler's estimator is lower than that of Huber's estimator. This differs from the non-regularized case, where Tyler's and Huber's IMI were shown to be the same. This suggests that "less-correlated" outlying samples have a greater negative impact on regularized Huber's estimator, as compared with Tyler's estimator. It also appears that $\epsilon_n \overline{\mathrm{IMI}}(\rho^\star)$ is a fairly good approximation of $\overline{\mathrm{MI}}(\rho^\star, \epsilon_n)$ for small ϵ_n, which shows the interest of Corollary 3.

So far, we have only considered two possible values of c_N: $c_N = 1/4$ (non-regularized case, Fig. 2) and $c_N = 3/2$ (regularized case, Fig. 3). To connect our results in the regularized and non-regularized scenarios, we now evaluate $\overline{\mathrm{IMI}}(\rho^\star)$ for various c_N. Such experiment shall shed light on the effect of the aspect ratio c_N on the robustness of different estimators. We consider again $u_{\mathrm{M-Tyler}}$ and $u_{\mathrm{M-Huber}}$, but now with $K = \min\{1, \frac{1}{c_N}\}$, such that Assumption 2 is verified; note that for $c_N \leq 1$, we retrieve the setting of

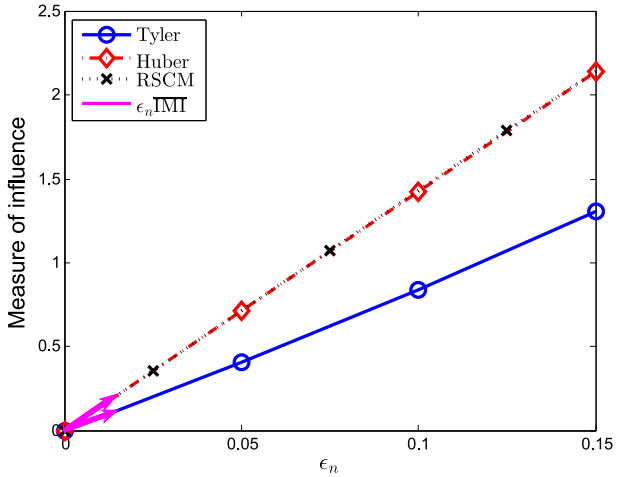


Fig. 3: Measure of influence for $\epsilon_n \in [0, 0.15]$, in the regularized case ($N = 150$, $n = 100$, such that $c_N = 3/2$). The IMI is computed at the optimal regularization parameter (assuming clean data) that minimizes the quadratic loss of the estimator. $[\mathbf{C}_N]_{ij} = .9^{|i-j|}$ and $[\mathbf{D}_N]_{ij} = .2^{|i-j|}$.

Fig. 2. Results are reported in Fig. 4. It appears that the IMI of a regularized estimator varies with c_N in a non-trivial manner. Indeed, different c_N call for different amounts of regularization (through $\rho^\star$), which in turn lead to substantial differences in robustness. Nonetheless, when $c_N \to 0$, the IMI of a given estimator tends to its non-regularized counterpart (indicated by arrows). This is a natural result, since in such case the need for regularization vanishes. We also notice that Tyler's estimator shows better robustness than all other estimators for nearly all values of c_N.

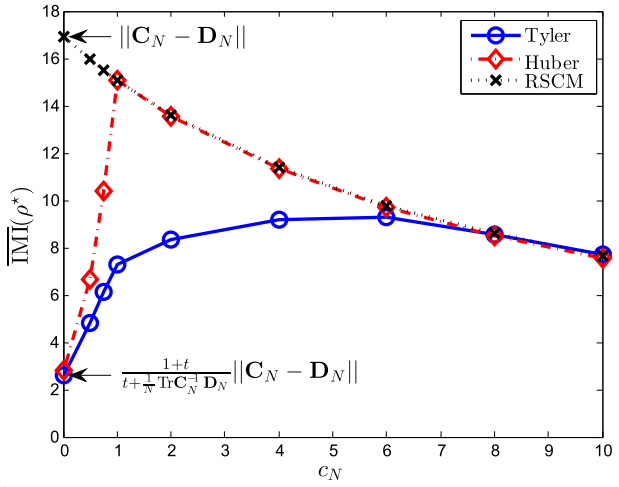


Fig. 4: Infinitesimal measure of influence vs. the aspect ratio c_N. The IMI is computed at the optimal regularization parameter (assuming clean data) that minimizes the quadratic loss of the estimator. Arrows indicate the IMI in the non-regularized case ($\rho = 0$). $[\mathbf{C}_N]_{ij} = .9^{|i-j|}$ and $[\mathbf{D}_N]_{ij} = .2^{|i-j|}$.

VI. Discussion and concluding remarks

In summary, we have shown that, in the absence of outliers, regularized M-estimators are asymptotically equivalent RSCM estimators and that, when optimally regularized, they all attain the same performance as the optimal RSCM, at least with respect to the quadratic loss. We proposed an intuitive metric to assess the robustness of different estimators when random outliers are introduced. In particular, it was shown in the non-regularized case that Huber's estimator is generally preferable over Tyler's, while, in the regularized case, the opposite is true.

The comparatively different behaviour in regularized and non-regularized settings evidences the substantial (and non trivial) effect of regularization on the robustness of M-estimators. This point is further emphasized in Fig. 5, where we plot $\overline{\mathrm{IMI}}(\rho)$ for $\rho \in (0,1]$. Interestingly, the IMI of

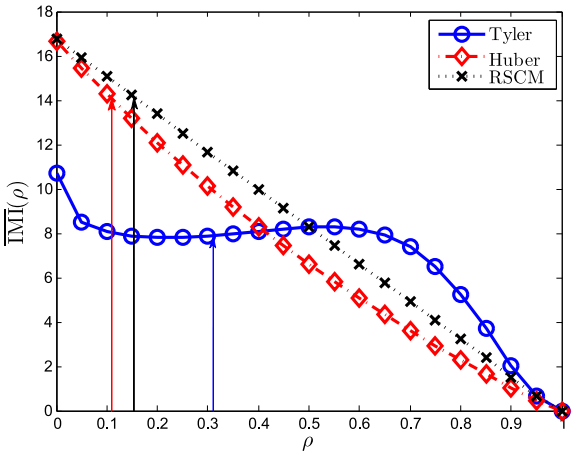


Fig. 5: Infinitesimal measure of influence for $\rho \in (0,1]$ ($c_N = 3/2$). Arrows indicate the (oracle) optimal shrinkage parameter of the considered estimators in the absence of outliers. $[\mathbf{C}_N]_{ij} = .9^{|i-j|}$ and $[\mathbf{D}_N]_{ij} = .2^{|i-j|}$.

Tyler's estimator is somewhat less sensitive in ρ than that of Huber's estimator or that of the RSCM (at least in the region where $\rho \approx \rho^\star$, indicated by arrows). This suggests that, in the presence of outliers, a small variation in the estimation of the optimal regularization parameter can have a different impact on the robustness of different estimators. Therefore, properly choosing the regularization parameter is crucial, both in terms of performance and robustness. These observations call for the need of a careful estimation of this optimal regularization parameter in the presence of outliers.

Another important problem relates to the fact that, while the proposed (clean-data-optimal) choice of regularization parameter proves to be a practical solution for a vanishing proportion of outliers, this choice would become more suboptimal under a more substantial non-vanishing proportion of outliers. In such scenarios, and in particular if some a-priori knowledge on the outliers could be exploited, different choices of regularization and/or estimators would be advisable. These problems will be investigated in future work.

Appendix

A. Proofs of results in Section III

1) Theorem 2: We will first start by proving the existence of $\hat{\mathbf{S}}_N(\rho)$, before turning to the continuity of $\rho \mapsto \gamma(\rho)$. Finally, we will show the uniform convergence of the spectral norm of $\hat{\mathbf{C}}_N(\rho) - \hat{\mathbf{S}}_N(\rho)$. The structure of the proof mirrors that of [17, Theorem 1], but non-trivial modifications are necessary to generalize the result to the wider class of u functions considered in this work. In particular, we have to make extensive use of the properties of the function $v = u(g^{-1}(x))$, where we recall that $g(x) = \frac{x}{1-(1-\rho)c\phi(x)}$.

Let us first prove the existence and uniqueness of $\hat{\mathbf{S}}_N(\rho)$. Notice that γ (as defined in (6)) can be rewritten as the solution to the fixed-point equation

$$\int \frac{t}{(1-\rho)\phi(g^{-1}(\gamma))t + \rho\gamma}\nu(dt) = 1. \tag{21}$$

Notice that the left-hand side of (21) is a decreasing function of γ (recall that ϕ is a increasing function and that $\rho > 0$). Furthermore, it has limits ∞ as $\gamma \to 0$ (since $\phi(g^{-1}(0)) = 0$ and $\nu \neq \delta_0$ a.e.) and zero as $\gamma \to \infty$. This proves the existence and uniqueness of γ, from which the existence and uniqueness of $\hat{\mathbf{S}}_N(\rho)$ unfold.

Now, let us turn to the continuity of $\rho \mapsto \gamma(\rho)$. Consider a given compact set $I \in \mathcal{I}$, where $\mathcal{I}$ is the set of compact sets included in $(0,1]$. Take $\rho_1, \rho \in I$ and $\gamma_1 = \gamma(\rho_1)$, $\gamma = \gamma(\rho)$. We can then write

$$\int \frac{t}{(1-\rho)\phi(g^{-1}(\gamma))t + \rho\gamma}\nu(dt) - \int \frac{t}{(1-\rho_1)\phi(g^{-1}(\gamma_1))t + \rho_1\gamma_1}\nu(dt) = 0,$$

which, after some algebra, leads to (22) (see top of next page). By Assumption 1.c, the support of ν is bounded by $\limsup_N ||\mathbf{C}_N|| < \infty$. In particular, recalling that $0 \leq \phi(x) \leq \phi_\infty$, from (21) we necessarily have that $\rho\gamma \leq \limsup_N ||\mathbf{C}_N||$. It follows that the above integrals are uniformly bounded on ρ in a neighborhood of $\rho_1 \leq \rho$. Taking the limit $\rho \to \rho_1$, we then have (23) (see top of next page). As g^{-1} and ϕ are increasing, $\phi(g^{-1}(\gamma_1)) - \phi(g^{-1}(\gamma))$ and $\gamma_1 - \gamma$ have the same sign. As the above integrals are uniformly bounded on a neighborhood of ρ_1, we have $\gamma_1 - \gamma \to 0$, from which we conclude that $\rho \mapsto \gamma(\rho)$ is continuous on $\mathcal{I}$.

Now, we will show the uniform convergence of the spectral norm of $\hat{\mathbf{C}}_N(\rho) - \hat{\mathbf{S}}_N(\rho)$. Let us fix $\rho \in I$, and denote $\hat{\mathbf{C}}_{(i)}(\rho) = \hat{\mathbf{C}}_N(\rho) - (1-\rho)\frac{1}{n}v\left(\frac{1}{N}\mathbf{y}_i^\dagger\hat{\mathbf{C}}_N^{-1}(\rho)\mathbf{y}_i\right)\mathbf{y}_i\mathbf{y}_i^\dagger$. After some algebra, we can rewrite

$$\hat{\mathbf{C}}_N(\rho) = (1-\rho)\frac{1}{n}\sum_{i=1}^n v(d_i(\rho))\mathbf{y}_i\mathbf{y}_i^\dagger + \rho\mathbf{I}_N,$$

where, for $i \in \{1, \cdots, n\}$, $d_i(\rho) \triangleq \frac{1}{N}\mathbf{y}_i^\dagger\hat{\mathbf{C}}_{(i)}^{-1}(\rho)\mathbf{y}_i$. Without loss of generality, we can assume that $d_1(\rho) \leq \cdots \leq d_n(\rho)$. Then, using the fact that v is non-increasing, and the fact that $\mathbf{A} \succeq \mathbf{B} \Rightarrow \mathbf{B}^{-1} \succeq \mathbf{A}^{-1}$ for positive Hermitian matrices

$$(\gamma_1-\gamma)\rho_1+\gamma(\rho_1-\rho)-\left((1-\rho)\phi(g^{-1}(\gamma))-(1-\rho_1)\phi(g^{-1}(\gamma_1))\right)\frac{\int\frac{t^2\nu(dt)}{((1-\rho)\phi(g^{-1}(\gamma))t+\rho\gamma)((1-\rho_1)\phi(g^{-1}(\gamma_1))t+\rho_1\gamma_1)}}{\int\frac{t\nu(dt)}{((1-\rho)\phi(g^{-1}(\gamma))t+\rho\gamma)((1-\rho_1)\phi(g^{-1}(\gamma_1))t+\rho_1\gamma_1)}}=0 \tag{22}$$

$$(\gamma_1-\gamma)\rho_1+(1-\rho_1)\left(\phi(g^{-1}(\gamma_1))-\phi(g^{-1}(\gamma))\right)\frac{\int\frac{t^2\nu(dt)}{((1-\rho_1)\phi(g^{-1}(\gamma))t+\rho\gamma)((1-\rho_1)\phi(g^{-1}(\gamma_1))t+\rho_1\gamma_1)\nu(dt)}}{\int\frac{t\nu(dt)}{((1-\rho_1)\phi(g^{-1}(\gamma))t+\rho_1\gamma)((1-\rho_1)\phi(g^{-1}(\gamma_1))t+\rho_1\gamma_1)}}\to 0 \tag{23}$$

$\mathbf{A}$ and $\mathbf{B}$, we have

$$\begin{aligned}d_n(\rho)&=\tfrac{1}{N}\mathbf{y}_n^\dagger\left((1-\rho)\tfrac{1}{n}\textstyle\sum_{j=1}^{n-1}v(d_j(\rho))\mathbf{y}_j\mathbf{y}_j^\dagger+\rho\mathbf{I}_N\right)^{-1}\mathbf{y}_n\\&\leq\tfrac{1}{N}\mathbf{y}_n^\dagger\left((1-\rho)\tfrac{1}{n}\textstyle\sum_{j=1}^{n-1}v(d_n(\rho))\mathbf{y}_j\mathbf{y}_j^\dagger+\rho\mathbf{I}_N\right)^{-1}\mathbf{y}_n.\end{aligned}$$

Since $\mathbf{y}_n\neq 0$ with probability 1, we then have

$$\mathbf{y}_n^\dagger\left((1-\rho)\frac{1}{n}\sum_{j=1}^{n-1}d_n(\rho)v(d_n(\rho))\mathbf{y}_j\mathbf{y}_j^\dagger+\rho d_n(\rho)\mathbf{I}_N\right)^{-1}\mathbf{y}_n\geq N. \tag{24}$$

Similarly,

$$\mathbf{y}_1^\dagger\left((1-\rho)\frac{1}{n}\sum_{j=2}^{n}d_1(\rho)v(d_1(\rho))\mathbf{y}_j\mathbf{y}_j^\dagger+\rho d_1(\rho)\mathbf{I}_N\right)^{-1}\mathbf{y}_1\leq N.$$

We want to show that:

$$\sup_{\rho\in\mathcal{I}}\max_{1\leq i\leq n}|d_i(\rho)-\gamma(\rho)|\xrightarrow{\text{a.s.}}0.$$

This will be proven by a contradiction argument: assume there exists a sequence $\{\rho_n\}_{n=1}^\infty$ over which $d_n(\rho_n)>\gamma(\rho_n)+l$ infinitely often, for some $l>0$ fixed. Let us consider a subsequence of $\{\rho_n\}_{n=1}^\infty$ such that $\rho_n\to\rho_1$ (since $\{\rho_n\}_{n=1}^\infty$ is bounded, such subsequence exists by the Bolzano-Weierstrass theorem). On this subsequence, (24) gives us (25) (see top of next page). Assume for now that $\rho_1\neq 1$. Rewriting $xv(x)=\psi(x)$, we have (26) (see top of next page), where, for $x>0$, $\delta(x)$ is the unique positive solution to the equation

$$\delta(x)=\int\frac{t}{-x+\frac{t}{1+c\delta(x)}}\nu(dt). \tag{27}$$

The convergence above follows from random matrix tools exposed in the proof of [17, Theorem 1]. Define $(l,e)\mapsto h(l,e)$ as

$$h(l,e)\triangleq\int\frac{t}{(\gamma(\rho_1)+l)\rho_1 e+\frac{te}{\frac{1}{(1-\rho_1)\psi(\gamma(\rho_1)+l)}+ce}}\nu(dt),$$

which is clearly decreasing in both l and e. Using (26) and (27) and a little algebra, we have that $h(l,e^+)=1$ for all $l>0$. Furthermore, from the definition of $\gamma(\rho_1)$, we also have that $h(0,1)=1$. Therefore, $h(0,1)=h(l,e^+)=1$ for all $l>0$. Along with the fact that $e\mapsto h(\cdot,e)$ and $l\mapsto h(l,\cdot)$ are both decreasing functions, we then necessarily have $e^+<1$. But this is in contradiction with $e_n\geq 1$ from (25).

Assume now that $\rho_1=1$. Since $\frac{1}{N}||\mathbf{y}_n||^2\xrightarrow{\text{a.s.}}M_{\nu,1}<\infty$, $\limsup_n||\frac{1}{n}\sum_{i=1}^n\mathbf{y}_i\mathbf{y}_i^\dagger||<\infty$ a.s. (from Assumption 1.b. and [31]), and $\gamma(1)=M_{\nu,1}$, from the definition of e_n we have:

$$e_n\xrightarrow{\text{a.s.}}\frac{M_{\nu,1}}{M_{\nu,1}+l}<1,$$

which is again a contradiction.

It follows that for all large n, there is no sequence of ρ_n such that $d_n(\rho)>\gamma(\rho)+l$ infinitely often. Consequently, $d_n(\rho)\leq\gamma(\rho)+l$ for all large n a.s., uniformly on $\rho\in I$. We can apply the same strategy to prove that $d_1(\rho)$ is greater than $\gamma(\rho)-l$ for all large n uniformly on $\rho\in I$. As this is true for arbitrary $l>0$, we then have $\sup_{\rho\in I}\max_{1\leq i\leq n}|d_i(\rho)-\gamma(\rho)|\xrightarrow{\text{a.s.}}0$. By continuity of v, we also have $\sup_{\rho\in I}\max_{1\leq i\leq n}|v(d_i(\rho))-v(\gamma(\rho))|\xrightarrow{\text{a.s.}}0$. It follows that

$$\begin{aligned}&\sup_{\rho\in I}\left|\left|\hat{\mathbf{C}}_N(\rho)-\hat{\mathbf{S}}_N(\rho)\right|\right|\\&\leq\left|\left|\frac{1}{n}\sum_{i=1}^n\mathbf{y}_i\mathbf{y}_i^\dagger\right|\right|\sup_{\rho\in I}\max_{1\leq i\leq n}(1-\rho)|v(d_i)-v(\gamma)|\xrightarrow{\text{a.s.}}0,\end{aligned}$$

where we used the fact that $\limsup_n\left|\left|\frac{1}{n}\sum_{i=1}^n\mathbf{y}_i\mathbf{y}_i^\dagger\right|\right|<\infty$ a.s., as above.

2) Proposition 1: Since $\rho\mapsto v(\gamma)$ is non-negative, it is clear that $\underline{\rho}$ is indeed in $(0,1]$. Then, for a couple $(\rho,\underline{\rho})$ satisfying (7), the (relative) weights given to the SCM $\frac{1}{n}\sum_{i=1}^n\mathbf{y}_i\mathbf{y}_i^\dagger$ and the shrinkage target $\mathbf{I}_N$ are the same for $\hat{\mathbf{S}}_N(\rho)$ and $\mathbf{R}(\underline{\rho})$. After trace-normalization, the first result of Proposition 1 follows. Now, since $\rho\mapsto v(\gamma)$ is continuous and bounded (from Theorem 2), it follows that $F:(0,1]\to(0,1]$ is continuous and onto, from which the second result of Proposition 1 unfolds.

3) Proposition 2: The proof of Proposition 2 makes use of the asymptotic equivalence of $\hat{\mathbf{C}}_N(\rho)$ with $\hat{\mathbf{S}}_N(\rho)$ (as given in Theorem 2) and the equivalence and mapping between $\hat{\mathbf{S}}_N(\rho)$ and the RSCM (as given in Proposition 1). It is known that the RSCM can be optimized with respect to the Frobenius norm, with the corresponding optimal

$$1 \leq \frac{1}{N}\mathbf{y}_n^\dagger \left((1-\rho_n)\frac{1}{n}\sum_{j=1}^{n-1}(\gamma(\rho_n)+l)v(\gamma(\rho_n)+l)\mathbf{y}_j\mathbf{y}_j^\dagger + \rho_n(\gamma(\rho_n)+l)\mathbf{I}_N \right)^{-1} \mathbf{y}_n \triangleq e_n. \tag{25}$$

$$\begin{aligned} e_n &= \frac{1}{(1-\rho_n)\psi(\gamma(\rho_n)+l)}\frac{1}{N}\mathbf{y}_n^\dagger \left(\frac{1}{n}\sum_{j=1}^{n-1}\mathbf{y}_j\mathbf{y}_j^\dagger + \rho_n\frac{\gamma(\rho_n)+l}{(1-\rho_n)\psi(\gamma(\rho_n)+l)}\mathbf{I}_N \right)^{-1}\mathbf{y}_n \\ &\xrightarrow{\text{a.s.}} \frac{1}{(1-\rho_1)\psi(\gamma(\rho_1+l))}\delta\left(-(\gamma(\rho_1)+l)\rho_1\frac{1}{(1-\rho_1)\psi(\gamma(\rho_1)+l)} \right) \triangleq e^+ \end{aligned} \tag{26}$$

regularization parameter $\rho^\star$ given in Proposition 2 (see, e.g., [17, 27]). With (7), it follows that for $\hat{\rho}^\star$ a solution to $\frac{\hat{\rho}^\star}{(1-\hat{\rho}^\star)v(\gamma)+\hat{\rho}^\star} = \rho^\star$, the associated estimator $\hat{\mathbf{C}}_N(\hat{\rho}^\star)$ will have (asymptotically) minimal quadratic loss. Similarly to [17, Proposition 2], the second part of Proposition 2 provides a consistent estimate $\hat{\rho}^\star$ based on a possible estimate of $\rho^\star$, the optimal regularization parameter for the RSCM.

B. Proofs of the results in Section IV

1) Theorem 3: The convergence of the spectral norm of $\hat{\mathbf{C}}_N - \hat{\mathbf{S}}_N$ unfolds from the proof of [19, Theorem 1]. However, the proof of the existence and uniqueness of $\hat{\mathbf{S}}_N$ for ϵ arbitrary requires additional arguments. To proceed, we make use of the standard interference function framework [32]. Define the real-valued functions $h_i : [0,\infty) \to [0,\infty), (q_0,q_1) \mapsto h_i(q_0,q_1)$ (with $i=0,1$) as:

$$\begin{aligned} h_0(q_0,q_1) &= \frac{1}{N}\operatorname{Tr}\mathbf{C}_N\left((1-\epsilon)\frac{f(q_0)}{q_0}\mathbf{C}_N + \epsilon\frac{f(q_1)}{q_1}\mathbf{D}_N\right)^{-1} \\ h_1(q_0,q_1) &= \frac{1}{N}\operatorname{Tr}\mathbf{D}_N\left((1-\epsilon)\frac{f(q_0)}{q_0}\mathbf{C}_N + \epsilon\frac{f(q_1)}{q_1}\mathbf{D}_N\right)^{-1}, \end{aligned}$$

where $f(x) \triangleq \frac{xv(x)}{1+cxv(x)}$, and where we dropped the subscript n of ϵ_n for readability. Thus defined, f is onto from $[0,\infty)$ to $[0,\phi_\infty)$, where we recall that $\phi_\infty > 1$. It can be easily verified that h_0, h_1 are standard interference functions[3] (see, for example, [19] for details). According to [32, Theorem 2], if there exist some $q_0, q_1 > 0$ such that $h_0(q_0,q_1) \leq q_0$ and $h_1(q_0,q_1) \leq q_1$, then the system of fixed-point equations $h_0(q_0,q_1) = q_0$, $h_1(q_0,q_1) = q_1$ admits a unique solution $\{q_0,q_1\}$. It therefore remains to find q_0 and q_1 that satisfy $h_0(q_0,q_1) \leq q_0$ and $h_1(q_0,q_1) \leq q_1$. It is in fact sufficient to show that there exist $q_0, q_1 \geq f^{-1}(1)$ such that

$$\begin{aligned} h_0'(q_0,q_1) &\leq q_0 \\ h_1'(q_0,q_1) &\leq q_1, \end{aligned}$$

where

$$\begin{aligned} h_0'(q_0,q_1) &\triangleq \frac{1}{N}\operatorname{Tr}\mathbf{C}_N\left((1-\epsilon)\frac{1}{q_0}\mathbf{C}_N + \epsilon\frac{1}{q_1}\mathbf{D}_N\right)^{-1} \\ &\geq h_0(q_0,q_1) \\ h_1'(q_0,q_1) &\triangleq \frac{1}{N}\operatorname{Tr}\mathbf{D}_N\left((1-\epsilon)\frac{1}{q_0}\mathbf{C}_N + \epsilon\frac{1}{q_1}\mathbf{D}_N\right)^{-1} \\ &\geq h_1(q_0,q_1). \end{aligned}$$

Consider two cases depending on whether $\epsilon \in \{0,1\}$ or not. If $\epsilon = 0$, then

$$\begin{aligned} h_0'(q_0,q_1) &= q_0 \\ h_1'(q_0,q_1) &= \frac{1}{N}\operatorname{Tr}\mathbf{D}_N\mathbf{C}_N^{-1}q_0. \end{aligned}$$

Taking $q_1 = \mathbf{D}_N\mathbf{C}_N^{-1}q_0$, we have $h_0'(q_0,q_1) \leq q_i$ for $i = 0,1$. It remains to choose q_0 such that $\min\{q_0,q_1\} \geq f^{-1}(1)$ (which is always possible), and the proof is done. Similarly, if $\epsilon = 1$, it suffices to take $q_0 = \mathbf{C}_N\mathbf{D}_N^{-1}q_1$, with q_1 chosen such that $\min\{q_0,q_1\} \geq f^{-1}(1)$.

Assume now that $\epsilon \in (0,1)$. Let us define $\alpha \triangleq \frac{q_1}{q_0} > 0$. We can rewrite:

$$\begin{aligned} h_0'(q_0,q_1) &= q_0\frac{1}{N}\operatorname{Tr}\left((1-\epsilon)\mathbf{I}_N + \frac{\epsilon}{\alpha}\mathbf{C}_N^{-1}\mathbf{D}_N\right)^{-1} \\ h_1'(q_0,q_1) &= \frac{q_1}{\alpha}\frac{1}{N}\operatorname{Tr}\mathbf{C}_N^{-1}\mathbf{D}_N\left((1-\epsilon)\mathbf{I}_N + \frac{\epsilon}{\alpha}\mathbf{C}_N^{-1}\mathbf{D}_N\right)^{-1}. \end{aligned}$$

Finding q_0, q_1 such that $h_i' \leq q_i$ is then equivalent to finding α such that

$$\frac{1}{N}\operatorname{Tr}\left((1-\epsilon)\mathbf{I}_N + \epsilon\frac{1}{\alpha}\mathbf{C}_N^{-1}\mathbf{D}_N\right)^{-1} \leq 1 \tag{28}$$

$$\frac{1}{\alpha}\frac{1}{N}\operatorname{Tr}\mathbf{C}_N^{-1}\mathbf{D}_N\left((1-\epsilon)\mathbf{I}_N + \epsilon\frac{1}{\alpha}\mathbf{C}_N^{-1}\mathbf{D}_N\right)^{-1} \leq 1. \tag{29}$$

By applying Lemma 1 (see below) with $\mathbf{A} = \frac{1}{\alpha}\mathbf{C}_N^{-1}\mathbf{D}_N$, we can show that

$$\begin{aligned} &\frac{1}{\alpha}\frac{1}{N}\operatorname{Tr}\mathbf{C}_N^{-1}\mathbf{D}_N\left((1-\epsilon)\mathbf{I}_N + \epsilon\frac{1}{\alpha}\mathbf{C}_N^{-1}\mathbf{D}_N\right)^{-1} \leq 1 \\ &\quad\Leftrightarrow \frac{1}{N}\operatorname{Tr}\left((1-\epsilon)\mathbf{I}_N + \epsilon\frac{1}{\alpha}\mathbf{C}_N^{-1}\mathbf{D}_N\right)^{-1} \geq 1. \end{aligned}$$

[3] In particular, they should verify conditions of positivity, monotonocity and scalability.

Combined with (28) and (29), it follows that q_0, q_1 verify $h_i' \leq q_i$ if and only if

$$\frac{1}{N}\operatorname{Tr}\left((1-\epsilon)\mathbf{I}_N + \epsilon\frac{1}{\alpha}\mathbf{C}_N^{-1}\mathbf{D}_N\right)^{-1} = 1.$$

Denote by $a_i > 0$ the i-th eigenvalue of $\mathbf{C}^{-1}\mathbf{D}$. We then have:

$$\begin{aligned}
&\frac{1}{N}\operatorname{Tr}\left((1-\epsilon)\mathbf{I}_N + \epsilon\frac{1}{\alpha}\mathbf{C}_N^{-1}\mathbf{D}_N\right)^{-1} = 1 \\
&\Leftrightarrow \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1-\epsilon+\epsilon\frac{a_i}{\alpha}} = 1 \\
&\Leftrightarrow \frac{1}{N}\alpha\sum_{i=1}^{N}\prod_{j\neq i}((1-\epsilon)\alpha+\epsilon a_j) = \prod_{i=1}^{N}((1-\epsilon)\alpha+\epsilon a_i),
\end{aligned}$$

where the last equality comes from putting all the terms in the sum on the same denominator, and multiplying by $\alpha^N \neq 0$. Finding an eligible α therefore boils down to finding whether the polynomial in α appearing in the last equation has positive roots. Notice now that the leading coefficient of this N-order polynomial is $b_N = \epsilon(1-\epsilon)^{N-1} > 0$, while the constant is $b_0 = -\epsilon^N\prod_{i=1}^N a_i < 0$. As $b_N \times b_0 < 0$, it follows that this polynomial admits (at least) one positive root (by applying the intermediate value theorem). Call α_0 such a root. Choosing q_0 such that $\min\{q_0, q_0\alpha_0\} \geq f^{-1}(1)$, q_0 and $q_1 = q_0\alpha_0$ will be such that $h_i' \leq q_i$, and therefore such that $h_i \leq q_i$. The existence and uniqueness of γ^ϵ and α, as given in Theorem 3, unfold.

Lemma 1. *For* $\mathbf{A}$ *an invertible matrix and* $\alpha > 0$, $\epsilon \in (0,1)$, *we have the following equivalence:*

$$\begin{aligned}
&\frac{1}{N}\operatorname{Tr}\mathbf{A}\left((1-\epsilon)\mathbf{I}_N + \epsilon\mathbf{A}\right)^{-1} \leq 1 \Leftrightarrow \\
&\quad\frac{1}{N}\operatorname{Tr}\left((1-\epsilon)\mathbf{I}_N + \epsilon\mathbf{A}\right)^{-1} \geq 1.
\end{aligned}$$

2) Corollary 1: This is a direct consequence of Theorems 1 and 3, by writing $\overline{\mathrm{MI}}(\epsilon_n) = \left\|\mathbb{E}\left[\frac{\hat{\mathbf{S}}_N^0}{L^0} - \frac{\hat{\mathbf{S}}_N^{\epsilon_n}}{L^\epsilon}\right]\right\|$, where $L_N^{\epsilon_n} \triangleq v(\gamma^{\epsilon_n})(1-\epsilon_n) + v(\alpha^{\epsilon_n})\epsilon_n$, and using the fact that $\frac{1}{N}\operatorname{Tr}\hat{\mathbf{C}}_N^{\epsilon_n} - L_N^{\epsilon_n} \xrightarrow{\text{a.s.}} 0$ and $\frac{1}{N}\operatorname{Tr}\hat{\mathbf{S}}_N^{\epsilon_n} - L_N^{\epsilon_n} \xrightarrow{\text{a.s.}} 0$.

C. Proof of the results in Section V

1) Theorem 4: As for the proof of Theorem 2, following the framework of standard interference functions [32], it can be proven that the system of equations (19) in Theorem 4 admits a unique solution $\{\gamma, \alpha\}$ for a given $\rho \in \mathcal{R}$, from which the existence and uniqueness of $\hat{\mathbf{S}}_N(\rho)$ unfolds. The convergence of the spectral norm of $\hat{\mathbf{C}}_N(\rho) - \hat{\mathbf{S}}_N(\rho)$ for $\rho \in \mathcal{R}$ is a direct extension of the proof of [19, Theorem 1], adapted to account for the introduction of the regularization parameter $\rho \in \mathcal{R}$.

2) Corollary 2: Define

$$\overline{\mathrm{MI}}(\rho, \epsilon_n) \triangleq \left\|\mathbb{E}\left[\frac{\hat{\mathbf{S}}_N^0(\rho)}{L^0(\rho)} - \frac{\hat{\mathbf{S}}_N^{\epsilon_n}(\rho)}{L^{\epsilon_n}(\rho)}\right]\right\|,$$

with $L_N^{\epsilon_n}(\rho) \triangleq v(\gamma^{\epsilon_n})(1-\rho)(1-\epsilon_n) + v(\alpha^{\epsilon_n})(1-\rho)\epsilon_n + \rho$. The convergence result is a direct consequence of Theorem 4 and the fact that, for $\rho \in \mathcal{R}$, $\frac{1}{N}\operatorname{Tr}\hat{\mathbf{C}}_N^{\epsilon_n}(\rho) - L_N^{\epsilon_n}(\rho) \xrightarrow{\text{a.s.}} 0$ and $\frac{1}{N}\operatorname{Tr}\hat{\mathbf{S}}_N^{\epsilon_n}(\rho) - L_N^{\epsilon_n}(\rho) \xrightarrow{\text{a.s.}} 0$. The derivation of $\overline{\mathrm{MI}}(\rho, \epsilon_n)$ is straightforward by expanding

$$\begin{aligned}
\overline{\mathbf{L}}(\rho, \epsilon_n) &\triangleq \mathbb{E}\left[\frac{\hat{\mathbf{S}}_N^0(\rho)}{L^0(\rho)} - \frac{\hat{\mathbf{S}}_N^{\epsilon_n}(\rho)}{L^{\epsilon_n}(\rho)}\right] \\
&= \frac{(1-\rho)(1-\epsilon_n)v(\gamma^{\epsilon_n})\mathbf{C}_N + (1-\rho)\epsilon_n v(\alpha^{\epsilon_n})\mathbf{D}_N + \rho\mathbf{I}_N}{(1-\rho)(1-\epsilon_n)v(\gamma^{\epsilon_n}) + (1-\rho)\epsilon_n v(\alpha^{\epsilon_n}) + \rho} \\
&\quad - \frac{(1-\rho)v(\gamma^0)\mathbf{C}_N + \rho\mathbf{I}_N}{(1-\rho)v(\gamma^0) + \rho} \\
&= \frac{U(\epsilon_n, \rho)}{V(\epsilon_n, \rho)},
\end{aligned}$$

with $U(\epsilon_n, \rho)$ and $V(\epsilon_n, \rho)$ given in the corollary.

3) Corollary 3: The result follows directly by taking the limit of $\frac{U(\epsilon_n,\rho)}{V(\epsilon_n,\rho)}$, as given in Corollary 2.

4) Computation of $\overline{\mathrm{IMI}}(\rho)$*:* In order to compute $\overline{\mathrm{IMI}}(\rho)$ (20) for arbitrary ρ, we need to compute $\frac{dv}{d\epsilon_n}\big|_{\epsilon_n=0} = \frac{dv}{d\gamma}\big|_{\gamma=\gamma^0} \times \frac{d\gamma}{d\epsilon_n}\big|_{\epsilon_n=0}$. For this, let us adopt the following notations:

$$\begin{aligned}
\mathbf{A}_N(\rho) &= \frac{(1-\rho)v(\gamma^0)}{1+(1-\rho)c\gamma^0 v(\gamma^0)}\mathbf{C}_N + \rho\mathbf{I}_N \\
\gamma^0 &\triangleq \lim_{n\to\infty}\gamma^{\epsilon_n} = \frac{1}{N}\operatorname{Tr}\mathbf{C}_N\mathbf{A}_N^{-1}(\rho) \\
\alpha^0 &\triangleq \lim_{n\to\infty}\alpha^{\epsilon_n} = \frac{1}{N}\operatorname{Tr}\mathbf{D}_N\mathbf{A}_N^{-1}(\rho).
\end{aligned}$$

Let us first compute $\frac{d\gamma}{d\epsilon_n}\big|_{\epsilon_n=0}$. For this, we need to differentiate (19) with respect to ϵ_n. We can do so by using the fact that $\frac{d\mathbf{M}^{-1}}{d\zeta}(\zeta) = -\mathbf{M}^{-1}(\zeta)\frac{d\mathbf{M}}{d\zeta}(\zeta)\mathbf{M}^{-1}(\zeta)$. Taking the limit when $\epsilon_n \to 0$ in the resulting equation, we get (30) (see top of next page). It remains to compute $\frac{dv}{d\gamma}\big|_{\gamma=\gamma^0}$. It is challenging to find a general expression for $\frac{dv}{d\gamma}\big|_{\gamma=\gamma^0}$ for an arbitrary u function (since it requires computing $v(x) = u(g^{-1}(x))$, which does not necessarily take a tractable form). However, we can do so for the u functions $u_{\mathrm{M-Tyler}} = \frac{1}{c_N}\frac{1+t}{t+x}$ and $u_{\mathrm{M-Huber}} = \frac{1}{c_N}\min\{1, \frac{1+t}{t+x}\}$. Assume $\rho > 0$ (for $\rho = 0$ (when possible), we fall back into the non-regularized case). Then, for these two functions, the associated v functions can be approximated by

$$v_{\mathrm{M-Tyler}}(x) \simeq \frac{1}{c_N}\frac{1+t}{t+\rho x}$$

and

$$v_{\mathrm{M-Huber}}(x) \simeq \begin{cases} \frac{1}{c_N} & \text{if} \quad x \leq \frac{1}{\rho} \\ \frac{1}{c_N}\frac{1+t}{t+\rho x} & \text{if} \quad x \geq \frac{1}{\rho} \end{cases},$$

for t small, from which we can deduce $\frac{dv}{dx}\big|_{x=\gamma^0}$.[4] We can then substitute $\frac{dv}{d\epsilon}\big|_{\epsilon=0} = \frac{dv}{d\gamma}\big|_{\gamma=\gamma^0} \times \frac{d\gamma}{d\epsilon}\big|_{\epsilon=0}$ in (20). We can proceed similarly for $u_{\mathrm{M-Tyler}} = \frac{1+t}{t+x}$ and $u_{\mathrm{M-Huber}} = \min\{1, \frac{1+t}{t+x}\}$.

[4]Note however that $v_{\mathrm{M-Huber}}$ is only piece-wise differentiable. In particular, additional care is needed if $\gamma^0 = 1/\rho$.

$$\left.\frac{d\gamma}{d\epsilon_n}\right|_{\epsilon_n=0} = (1-\rho)\frac{\frac{1}{N}\operatorname{Tr}\left[\mathbf{A}_N^{-1}(\rho)\mathbf{C}_N\mathbf{A}_N^{-1}(\rho)\left(\frac{v(\gamma^0)}{1+(1-\rho)c\gamma^0 v(\gamma^0)}\mathbf{C}_N - \frac{v(\alpha^0)}{1+(1-\rho)c\alpha^0 v(\alpha^0)}\mathbf{D}_N\right)\right]}{1+\frac{(1-\rho)\left(\left.\frac{dv}{d\gamma}\right|_{\epsilon=0}-(1-\rho)cv(\gamma^0)^2\right)}{(1+(1-\rho)c\gamma^0 v(\gamma^0))^2}\frac{1}{N}\operatorname{Tr}\mathbf{A}_N^{-1}(\rho)\mathbf{C}_N\mathbf{A}_N^{-1}(\rho)\mathbf{C}_N} \tag{30}$$

REFERENCES

[1] Y. Abramovich and N. K. Spencer, "Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering," in *IEEE Int. Conf. Acoust. Signal Process.*, vol. 3, pp. III–1105, 2007.

[2] F. Pascal, Y. Chitour, and Y. Quek, "Generalized robust shrinkage estimator and its application to STAP detection problem," *IEEE Trans. Signal Process.*, vol. 62, pp. 5640–5651, Sept. 2014.

[3] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 1, pp. 1–182, 2004.

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

[5] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Stat. Applicat. Genetics Molecular Biology*, vol. 4, no. 1, 2005.

[6] K. D. Ward, "Compound representation of high resolution sea clutter," *Electron. Lett.*, vol. 17, no. 16, pp. 561–563, 1981.

[7] J. B. Billingsley, A. Farina, F. Gini, M. V. Greco, and L. Verrazzani, "Statistical analyses of measured radar ground clutter data," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 35, pp. 579–593, Apr. 1999.

[8] D. Kelker, "Distribution theory of spherical distributions and a location-scale parameter generalization," *Sankhyā: Indian J. Stat., Series A*, pp. 419–430, 1970.

[9] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor, "Complex elliptically symmetric distributions: Survey, new results and applications," *IEEE Trans. Signal Process.*, vol. 60, pp. 5597–5625, Aug. 2012.

[10] X. Mestre and M. Á. Lagunas, "Modified subspace algorithms for DoA estimation with large arrays," *IEEE Trans. Signal Process.*, vol. 56, pp. 598–614, Jan. 2008.

[11] B. Nadler, "Nonparametric detection of signals by information theoretic criteria: performance analysis and an improved estimator," *IEEE Trans. Signal Process.*, vol. 58, pp. 2746–2756, Feb. 2010.

[12] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Stat.*, vol. 35, no. 1, pp. 73–101, 1964.

[13] R. A. Maronna, "Robust M-estimators of multivariate location and scatter," *Ann. Stat.*, pp. 51–67, 1976.

[14] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *Ann. Stat.*, pp. 234–251, 1987.

[15] Y. Chen, A. Wiesel, and A. O. Hero, "Robust shrinkage estimation of high-dimensional covariance matrices," *IEEE Trans. Signal Process.*, vol. 59, pp. 4097–4107, Apr. 2011.

[16] R. Couillet, F. Pascal, and J. W. Silverstein, "Robust estimates of covariance matrices in the large dimensional regime," *IEEE Trans. Inf. Theory*, vol. 60, pp. 7269–7278, Sept. 2014.

[17] R. Couillet and M. R. McKay, "Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators," *J. Multivar. Anal.*, vol. 131, pp. 99–120, Oct. 2014.

[18] T. Zhang, X. Cheng, and A. Singer, "Marchenko-Pastur law for Tyler's and Maronna's M-estimators," *arXiv preprint arXiv:1401.3424*, 2014.

[19] D. Morales-Jimenez, R. Couillet, and M. R. McKay, "Large dimensional analysis of robust M-estimators of covariance with outliers," *IEEE Trans. Signal Process.*, vol. 63, pp. 5784–5797, Jul. 2015.

[20] Y. Abramovich, "A controlled method for adaptive optimization of filters using the criterion of maximum signal-to-noise ratio," *Radio Eng. Elect. Phys*, vol. 26, no. 3, pp. 87–95, 1981.

[21] B. D. Carlson, "Covariance matrix estimation errors and diagonal loading in adaptive arrays," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 24, pp. 397–401, Jul. 1988.

[22] E. Ollila and D. E. Tyler, "Regularized M-estimators of scatter matrix," *IEEE Trans. Signal Process.*, vol. 62, pp. 6059–6070, Sept. 2014.

[23] R. Couillet, F. Pascal, and J. W. Silverstein, "The random matrix regime of Maronna's M-estimator with elliptically distributed samples," *J. Multivar. Anal.*, vol. 139, pp. 56–78, Jul. 2015.

[24] J. T. Kent and D. E. Tyler, "Redescending M-estimates of multivariate location and scatter," *Ann. Stat.*, pp. 2102–2119, 1991.

[25] Y. Sun, P. Babu, and D. P. Palomar, "Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms," *IEEE Trans. Signal Process.*, vol. 62, pp. 5143–5156, Aug. 2014.

[26] P. J. Huber, *Robust Statistics*. Springer, 2011.

[27] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. Anal.*, vol. 88, pp. 365–411, Jul. 2004.

[28] L. Du, J. Li, and P. Stoica, "Fully automatic computation of diagonal loading levels for robust adaptive beamforming," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, Feb. 2010.

[29] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[30] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, vol. 114. John Wiley & Sons, 2011.

[31] Z. D. Bai and J. W. Silverstein, "No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices," *Ann. Prob.*, pp. 316–345, 1998.

[32] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, pp. 1341–1347, Sept. 1995.