M-Estimators of Scatter with Eigenvalue Shrinkage

Esa Ollila, Daniel P. Palomar, Frédéric Pascal

M-ESTIMATORS OF SCATTER WITH EIGENVALUE SHRINKAGE

Esa Ollila[⋆] *Daniel P. Palomar*[†] *Frédéric Pascal*[‡]

[⋆]Department of Signal Processing and Acoustics, Aalto University, Finland
[†] The Hong Kong University of Science and Technology, Hong Kong
[‡] L2S / CentraleSupelec, University Paris-Saclay, France

ABSTRACT

A popular regularized (shrinkage) covariance estimator is the shrinkage sample covariance matrix (SCM) which shares the same set of eigenvectors as the SCM but shrinks its eigenvalues toward its grand mean. In this paper, a more general approach is considered in which the SCM is replaced by an M-estimator of scatter matrix and a fully automatic data adaptive method to compute the optimal shrinkage parameter with minimum mean squared error is proposed. Our approach permits the use of any weight function such as Gaussian, Huber's, or t weight functions, all of which are commonly used in M-estimation framework. Our simulation examples illustrate that shrinkage M-estimators based on the proposed optimal tuning combined with robust weight function do not loose in performance to shrinkage SCM estimator when the data is Gaussian, but provide significantly improved performance when the data is sampled from a heavy-tailed distribution.

Index Terms— M-estimators, sample covariance matrix, shrinkage, regularization, elliptical distributions

1. INTRODUCTION

Consider a sample of p-dimensional vectors $\{\mathbf{x}_i\}_{i=1}^n$ sampled from a distribution of a random vector $\mathbf{x}$ with $\mathbb{E}[\mathbf{x}] = \mathbf{0}$. One of the first tasks in the analysis of high-dimensional data is to estimate the covariance matrix. The most commonly used estimator is the sample covariance matrix (SCM), $\mathbf{S} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$, but its main drawbacks are its loss of efficiency when sampling from distributions which have longer tails than the multivariate normal (MVN) distribution and its sensitivity to outliers. Although being unbiased estimator of the covariance matrix $\mathrm{cov}(\mathbf{x}) = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ for any sample length $n \geq 1$, it is well-known that the eigenvalues are poorly estimated when n is not orders of magnitude larger than p. In such cases, one commonly uses a regularized SCM (RSCM) with a linear shrinkage towards a scaled identity matrix,

$$\mathbf{S}_\beta = \beta\mathbf{S} + (1-\beta)\frac{\mathrm{tr}(\mathbf{S})}{p}\mathbf{I}, \tag{1}$$

where $\beta \in (0, 1]$ is the regularization parameter. The RSCM $\mathbf{S}_\beta$ shares the same set of eigenvectors as the SCM $\mathbf{S}$, but its eigenvalues are shrinked towards the grand mean of the eigenvalues. That is, if $d_1, \ldots, d_p$ denote the eigenvalues of $\mathbf{S}$, then $\beta d_j + (1-\beta)\bar{d}$ are the eigenvalues of $\mathbf{S}_\beta$, where $\bar{d} = p^{-1}\sum_j d_j$. Optimal computation of β such that $\mathbf{S}_\beta$ has minimum mean squared error (MMSE) has been developed for example in [1, 2].

The estimator in (1) remains sensitive to outliers and non-Gaussianity. M-estimators of scatter [3] are popular robust alternatives to SCM. We consider the situation where $n > p$ and hence a conventional M-estimator of scatter $\hat{\mathbf{\Sigma}}$ exists and can be used in place of the SCM $\mathbf{S}$ in (1). We then propose a fully automatic data adaptive method to compute the optimal shrinkage parameter β. First, we derive an approximation for parameter β that attains the minimum MMSE and then propose a data adaptive method for its computation. The benefit of our approach is that it can be easily applied to any M-estimator using any weight function $u(t)$. Our simulation examples illustrate that a shrinkage M-estimator using the proposed tuning and a robust loss function does not loose in performance to optimal shrinkage SCM estimator when the data is Gaussian, but is able to provide significantly improved performance in the case of heavy-tailed data.

Relations to prior work: Earlier work, [4, 5, 6, 7, 8, 9], proposed regularized M-estimators of scatter matrix either by adding a penalty function to M-estimation objective function or a diagonal loading term to the respective first-order solution (M-estimating equation). We consider a simpler approach that uses conventional M-estimator and shrinks its eigenvalues to grand mean of the eigenvalues. Our approach permits computation of the optimal shrinkage parameter for any M-estimation weight function.

The paper is structured as follows. Section 2 introduces the proposed shrinkage M-estimator framework. Section 3 discusses automatic computation of the optimal shrinkage parameter under the assumption of sampling from unspecified elliptical distribution. Section 4 contains simulation studies.

2. SHRINKAGE M-ESTIMATORS OF SCATTER

In this paper, we assume that $n > p$ and consider an M-estimator of scatter matrix [3] that solves an estimating equa-

tion

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top, \tag{2}$$

where $u : [0,\infty) \to [0,\infty)$ is a non-increasing weight function. An M-estimator is a sort of adaptively weighted sample covariance matrix with weights determined by function $u(\cdot)$. To guarantee existence of the solution, it is required that the data verifies the condition stated in [10]. An M-estimator of scatter which shrinks the eigenvalues towards the grand mean of the eigenvalues is then defined as:

$$\hat{\mathbf{\Sigma}}_\beta = \beta\hat{\mathbf{\Sigma}} + (1-\beta)\frac{\operatorname{tr}(\hat{\mathbf{\Sigma}})}{p}\mathbf{I}. \tag{3}$$

For example, the RSCM $\mathbf{S}_\beta$ is obtained when one uses the Gaussian weight function $u(t) = 1\ \forall t$ since then $\hat{\mathbf{\Sigma}} = \mathbf{S}$. Other popular choices are Huber's weight function

$$u_{\text{H}}(t;c) = \max(-c^2, \min(t, c^2))/b, \tag{4}$$

where $c > 0$ is a tuning constant, chosen by the user, and b is a scaling factor used to obtain Fisher consistency at the multivariate normal (MVN) distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$:

$$b = F_{\chi^2_{p+2}}(c^2) + c^2(1 - F_{\chi^2_p}(c^2))/p.$$

We choose c^2 as qth upper quantile of χ^2_p: $c^2 = F^{-1}_{\chi^2_p}(q)$. Another popular choice is t-MLE weight function

$$u_{\text{T}}(t;\nu) = \frac{p+\nu}{\nu+t} \tag{5}$$

in which case the corresponding M-estimator $\hat{\mathbf{\Sigma}}$ is also the maximum likelihood estimate (MLE) of the scatter matrix parameter of a p-variate t-distribution with $\nu > 0$ degrees of freedom.

An M-estimator is a consistent estimator of the M-functional of scatter matrix, defined as

$$\mathbf{\Sigma}_0 = \mathbb{E}\big[u(\mathbf{x}^\top\mathbf{\Sigma}_0^{-1}\mathbf{x})\mathbf{x}\mathbf{x}^\top\big]. \tag{6}$$

If the population M-functional $\mathbf{\Sigma}_0$ is known, then by defining a *1-step estimator*

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top\mathbf{\Sigma}_0^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top \tag{7}$$

we can compute

$$\mathbf{C}_\beta = \beta\mathbf{C} + (1-\beta)[\operatorname{tr}(\mathbf{C})/p]\mathbf{I} \tag{8}$$

as a proxy for $\hat{\mathbf{\Sigma}}_\beta$. Naturally, such an estimator is fictional, as the initial value $\mathbf{\Sigma}_0$ is unknown. The 1-step estimator $\mathbf{C}$ is an unbiased estimator of $\mathbf{\Sigma}_0$, i.e., $\mathbb{E}[\mathbf{C}] = \mathbf{\Sigma}_0$.

Ideally we would like to find the value of $\beta \in [0,1]$ for which the corresponding estimator $\hat{\mathbf{\Sigma}}_\beta$ attains the minimum MSE, that is,

$$\beta_o = \arg\min_\beta \Big\{\operatorname{MSE}(\hat{\mathbf{\Sigma}}_\beta) = \mathbb{E}\Big[\big\|\hat{\mathbf{\Sigma}}_\beta - \mathbf{\Sigma}_0\big\|_{\text{F}}^2\Big]\Big\}, \tag{9}$$

where $\|\cdot\|_{\text{F}}$ denotes the Frobenius matrix norm ($\|\mathbf{A}\|_{\text{F}}^2 = \operatorname{tr}(\mathbf{A}^\top\mathbf{A})$). Since solving (9) is not doable due to the implicit form of M-estimators, we look for an approximation:

$$\beta_o^{\text{app}} = \arg\min_\beta \Big\{\operatorname{MSE}(\mathbf{C}_\beta) = \mathbb{E}\Big[\big\|\mathbf{C}_\beta - \mathbf{\Sigma}_0\big\|_{\text{F}}^2\Big]\Big\}. \tag{10}$$

Such approach was also used in [11] in deriving an optimal parameter for shrinkage Tyler's M-estimator of scatter.

Before stating the expression for β_o^{app} we introduce a *sphericity* measure of scatter:

$$\gamma = \frac{p\operatorname{tr}(\mathbf{\Sigma}_0^2)}{\operatorname{tr}(\mathbf{\Sigma}_0)^2}. \tag{11}$$

Sphericity γ measures how close $\mathbf{\Sigma}_0$ is to a scaled identity matrix: $\gamma \in [1,p]$ where $\gamma = 1$ if and only if $\mathbf{\Sigma}_0 \propto \mathbf{I}$ and $\gamma = p$ if $\mathbf{\Sigma}_0$ has rank equal to 1.

Theorem 1. *Suppose $\mathbf{x}_1,\dots,\mathbf{x}_n$ is an i.i.d. random sample from any p-variate distribution (not necessarily elliptical distribution), and u is a weight function for which the expectation $\mathbb{E}[\operatorname{tr}(\mathbf{C}^2)]$ exists. The oracle parameter β_o^{app} in* (10) *is*

$$\beta_o^{\text{app}} = \frac{\|\mathbf{\Sigma}_0 - \eta_o\mathbf{I}\|_{\text{F}}^2}{\mathbb{E}\big[\big\|\mathbf{C} - (\operatorname{tr}(\mathbf{C})/p)\mathbf{I}\big\|_{\text{F}}^2\big]} \tag{12}$$

$$= \frac{p(\gamma-1)\eta_o^2}{\mathbb{E}[\operatorname{tr}(\mathbf{C}^2)] - p^{-1}\mathbb{E}[\operatorname{tr}(\mathbf{C})^2]} \tag{13}$$

where $\eta_o = \operatorname{tr}(\mathbf{\Sigma}_0)/p$ and γ is defined in (11)*. Note that $\beta_o^{\text{app}} \in [0,1)$ and the value of the MSE at the optimum is*

$$\operatorname{MSE}(\mathbf{C}_{\beta_o^{\text{app}}}) = \frac{\mathbb{E}[\operatorname{tr}(\mathbf{C})^2] - \operatorname{tr}(\mathbf{\Sigma}_0)^2}{p} + (1-\beta_o^{\text{app}})\big\|\mathbf{\Sigma}_0 - \eta_o\big\|_{\text{F}}^2. \tag{14}$$

Proof. Write $L(\beta) = \operatorname{MSE}(\mathbf{C}_\beta) = \mathbb{E}[\|\mathbf{C}_\beta - \mathbf{\Sigma}_0\|_{\text{F}}^2]$. Then note that

$$\begin{aligned} L(\beta) &= \mathbb{E}\big[\big\|\beta\mathbf{C} + (1-\beta)p^{-1}\operatorname{tr}(\mathbf{C})\mathbf{I} - \mathbf{\Sigma}_0\big\|_{\text{F}}^2\big] \\ &= \mathbb{E}\Big[\big\|\beta(\mathbf{C} - \mathbf{\Sigma}_0) + (1-\beta)\big(p^{-1}\operatorname{tr}(\mathbf{C})\mathbf{I} - \mathbf{\Sigma}_0\big)\big\|_{\text{F}}^2\Big] \\ &= \beta^2 a_1 + (1-\beta)^2 a_2 + 2\beta(1-\beta)a_3, \end{aligned}$$

where $a_1 = \mathbb{E}\big[\big\|\mathbf{C} - \mathbf{\Sigma}_0\big\|_{\text{F}}^2\big] = \mathbb{E}\big[\operatorname{tr}(\mathbf{C}^2)\big] - \operatorname{tr}(\mathbf{\Sigma}_0^2)$, and

$$\begin{aligned} a_2 &= \mathbb{E}\big[\big\|p^{-1}\operatorname{tr}(\mathbf{C})\mathbf{I} - \mathbf{\Sigma}_0\big\|_{\text{F}}^2\big] \\ &= a_3 + \operatorname{tr}(\mathbf{\Sigma}_0^2) - p\eta_o^2 = a_3 + p(\gamma-1)\eta_o^2 \\ a_3 &= p^{-1}\mathbb{E}\big[\operatorname{tr}(\mathbf{C})\operatorname{tr}(\mathbf{C} - \mathbf{\Sigma}_0)\big] = p^{-1}\mathbb{E}\big[\operatorname{tr}(\mathbf{C})^2\big] - \eta_o^2 p. \end{aligned}$$

Note that $L(\beta)$ is a convex quadratic function in β with a unique minimum given by

$$\beta_o^{\text{app}} = \frac{a_2 - a_3}{(a_1 - a_3) + (a_2 - a_3)}.$$

Substituting the expressions for constants a_1, a_2 and a_3 into β_o^{app} yields the stated result. □

Next we derive a more explicit form of β_o^{app} by assuming that the data is generated from unspecified elliptically symmetric distribution.

3. SHRINKAGE PARAMETER COMPUTATION

Maronna [3] developed M-estimators of scatter matrices originally within the framework of elliptically symmetric distributions [12, 13]. The probability density function (p.d.f.) of centered (zero mean) elliptically distributed random vector $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$ is

$$f(\mathbf{x}) = C_{p,g} |\mathbf{\Sigma}|^{-1/2} g\big(\mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{x}\big),$$

where $\mathbf{\Sigma}$ is the positive definite symmetric matrix parameter, called the scatter matrix, $g : [0, \infty) \to [0, \infty)$ is the *density generator*, which is a fixed function that is independent of $\mathbf{x}$ and $\mathbf{\Sigma}$, and $C_{p,g}$ is a normalizing constant ensuring that $f(\mathbf{x})$ integrates to 1. The density generator g determines the elliptical distribution. For example, the MVN distribution is obtained when $g(t) = \exp(-t/2)$ and the t-distribution with ν d.o.f., denoted $\mathbf{x} \sim t_\nu(\mathbf{0}, \mathbf{\Sigma}, g)$, is obtained when $g(t) = (1 + t/\nu)^{-(p+\nu)/2}$. Then the weight function for the MLE of scatter corresponds to the case that $u(t) \propto -g'(t)/g(t)$. This yields (5) as the weight function for the MLE of scatter when $\mathbf{x} \sim t_\nu(\mathbf{0}, \mathbf{\Sigma}, g)$. If the second moments of $\mathbf{x}$ exists, then g can be defined so that $\mathbf{\Sigma}$ represents the covariance matrix of $\mathbf{x}$, i.e., $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{x})$; see [13] for details.

When $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$, then the M-functional $\mathbf{\Sigma}_0$ in (6) is related to underlying scatter matrix parameter $\mathbf{\Sigma}$ via the relationship

$$\mathbf{\Sigma}_0 = \sigma \mathbf{\Sigma},$$

where $\sigma > 0$ is a solution to an equation

$$\mathbb{E}\bigg[\psi\bigg(\frac{\mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{x}}{\sigma}\bigg)\bigg] = p, \tag{15}$$

where $\psi(t) = u(t)t$. Often σ needs to be solved numerically from (15) but in some cases an analytic expression can be derived. If $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$ and the used weight function matches with the data distribution, so $u(t) \propto -g'(t)/g(t)$, then $\sigma = 1$.

Next we derive expressions for $\mathbb{E}[\mathrm{tr}(\mathbf{C})^2]$ and $\mathbb{E}[\mathrm{tr}(\mathbf{C}^2)]$ appearing in the denominator of β_o^{app} in (13). They depend on a constant ψ_1 (which depend on weight function u via $\psi(t) = u(t)t$) as follows:

$$\psi_1 = \frac{1}{p(p+2)} \mathbb{E}\Big[\psi\bigg(\frac{\mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{x}}{\sigma}\bigg)^2\Big], \tag{16}$$

where the expectation is w.r.t. $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$.

Lemma 1. *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is an i.i.d. random sample from $\mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$. Then*

$$\mathbb{E}\big[\,\mathrm{tr}\big(\mathbf{C}^2\big)\big] = \bigg(1 + \frac{2\psi_1 - 1}{n}\bigg)\,\mathrm{tr}(\mathbf{\Sigma}_0^2) + \frac{\psi_1}{n}\,\mathrm{tr}(\mathbf{\Sigma}_0)^2,$$

$$\mathbb{E}[\mathrm{tr}(\mathbf{C})^2] = \frac{2\psi_1}{n}\,\mathrm{tr}(\mathbf{\Sigma}_0^2) + \bigg(1 + \frac{\psi_1 - 1}{n}\bigg)\,\mathrm{tr}(\mathbf{\Sigma}_0)^2,$$

given that expectation (16) *exists.*

Proof. Omitted due to lack of space. □

Theorem 2. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ denote an i.i.d. random sample from an elliptical distribution $\mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$. Then the oracle parameter β_o^{app} that minimizes the MSE in Theorem 1 is*

$$\beta_o^{\mathrm{app}} = \frac{\gamma - 1}{(\gamma - 1)(1 - 1/n) + \psi_1(1 - 1/p)(2\gamma + p)/n} \tag{17}$$

where γ is the sphericity measure defined in (11)*.*

Proof. Follows from Theorem 1 after substituting the values for $\mathbb{E}\big[\,\mathrm{tr}\big(\mathbf{C}^2\big)\big]$ and $\mathbb{E}[\mathrm{tr}(\mathbf{C})^2]$ derived in Lemma 1 in the denominator of β_o^{app} in (13). □

If one uses Gaussian loss function $u(t) \equiv 1$, then one needs to assume that the 4th-order moments exists and one may assume w.l.o.g. that the scatter matrix parameter equals the covariance matrix [13], i.e., $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{x})$, so $\mathbf{\Sigma}_0 = \mathbf{\Sigma}$ and $\sigma = 1$. Furthermore, it holds that $\hat{\mathbf{\Sigma}} = \mathbf{S}$ and $\mathbf{C}_\beta = \mathbf{S}_\beta$ and hence $\beta_o = \beta_o^{\mathrm{app}}$. Finally, we may relate ψ_1 with an elliptical kurtosis [14] parameter κ, defined as

$$\kappa = \frac{\mathbb{E}[\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|^4]}{p(p+2)} - 1. \tag{18}$$

Elliptical kurtosis vanishes, i.e., $\kappa = 0$, when $\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$.

Corollary 1. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ denote an i.i.d. random sample from an elliptical distribution $\mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$ with finite 4th order moments and covariance matrix $\mathbf{\Sigma} = \mathrm{cov}(\mathbf{x})$. Then the optimal tuning parameter of the shrinkage SCM estimator $\mathbf{S}_\beta$ in* (1) *is*

$$\beta_o = \arg\min_\beta \mathbb{E}\big[\|\mathbf{S}_\beta - \mathbf{\Sigma}\|_{\mathrm{F}}^2\big] = \frac{\gamma - 1}{\gamma - 1 + a}, \tag{19}$$

where

$$a = \frac{\kappa(2\gamma(1 - 1/p) + p - 1)}{n} + \frac{\gamma(1 - 2/p) + p}{n}.$$

Proof. The result follows from Theorem 2 since $\mathbf{C}_\beta = \mathbf{S}_\beta$ and the M-functional for Gaussian loss is $\mathbf{\Sigma}_0 = \mathrm{cov}(\mathbf{x}) = \mathbf{\Sigma}$. Since for Gaussian loss, $\psi(t) = t$, we notice from (16) that $\psi_1 = 1 + \kappa$. Plugging $\psi_1 = 1 + \kappa$ into (17) yields the stated expression. □

4. SIMULATION STUDIES

We compute different shrinkage M-estimators $\hat{\mathbf{\Sigma}}_\beta$ detailed below. We use acronym **Huber** to refer to the shrinkage M-estimator $\hat{\mathbf{\Sigma}}_\beta$ that uses Huber's weight $u(\cdot) = u_{\mathrm{H}}(\cdot; c)$ with threshold c^2 corresponding to $q = 0.7$ quantile. Shrinkage parameter is computed as $\beta = \beta_o^{\mathrm{app}}(\hat{\gamma}, \hat{\psi}_1)$. As an estimator $\hat{\gamma}$ of γ we use the same estimate as in [9, 2] and $\hat{\psi}_1$ is an estimate of ψ_1, computed as

$$\hat{\psi}_1 = \frac{1}{n} \sum_{i=1}^n \frac{[t_i u(t_i)]^2}{p(p+2)}, \tag{20}$$

where $t_i = \mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1} \mathbf{x}_i$ and $\hat{\mathbf{\Sigma}}$ is the corresponding Huber's M-estimator. Huber's weight function is standardized to be Fisher consistent for Gaussian samples, meaning that (15) holds with $\sigma = 1$ when $\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. Since (20) ignores estimation of σ, some loss in accuracy of this estimate of ψ_1 is expected for non-Gaussian data.

Acronym **t-MLE** refers to the shrinkage M-estimator of scatter using weight function $u(\cdot) = u_{\text{T}}(\cdot; \nu)$, where d.o.f. parameter ν is estimated from the data. This means that $\sigma = 1$ can be assumed since the scaling factor σ equals one for an MLE of the scatter matrix parameter. The shrinkage parameter is computed as $\beta = \beta_o^{\text{app}}(\hat{\gamma}, \hat{\psi}_1)$, where $\hat{\gamma}$ is as earlier and $\hat{\psi}_1$ is computed as in (20) but using $u(\cdot) = u_{\text{T}}(\cdot; \nu)$ and $\hat{\mathbf{\Sigma}}$ being the corresponding M-estimator.

Acronym **Gauss** refers to the shrinkage M-estimator of scatter using Gaussian weight function $u(t) = 1$, i.e., $\hat{\mathbf{\Sigma}}_\beta = \mathbf{S}_\beta$. The shrinkage parameter is computed as $\beta = \beta_o(\hat{\kappa}, \hat{\gamma})$ with β_o given by (19) and $\hat{\kappa}$ is an estimate of elliptical kurtosis κ proposed in [2]. Finally, acronym **LW** refers to estimator proposed by Ledoit and Wold [1]. LW estimator also uses RSCM $\mathbf{S}_\beta$, where parameter β is computed in a different manner than for Gauss estimator.

We generated the data from an elliptical distribution $\mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$, where the scatter matrix $\mathbf{\Sigma}$ has an AR(1) structure, $(\mathbf{\Sigma})_{ij} = \eta \varrho^{|i-j|}$, where $\varrho \in (0, 1)$ and scale parameter $\eta = \text{tr}(\mathbf{\Sigma})/p = 10$. When $\varrho \downarrow 0$, then $\mathbf{\Sigma}$ is close to an identity matrix scaled by η, and when $\varrho \uparrow 1$, $\mathbf{\Sigma}$ tends to a singular matrix of rank 1. Parameter ϱ is set to $\varrho = 0.6$. The dimension is $p = 40$ and n varies from 60 to 280.

In our first study, samples are drawn from a MVN distribution and the normalized MSE $\|\hat{\mathbf{\Sigma}}_\beta - \mathbf{\Sigma}_0\|_{\text{F}}^2 / \|\mathbf{\Sigma}_0\|_{\text{F}}^2$ as a function of sample length n is depicted in Figure 1. Results are averages over 2000 Monte-Carlo trials. All estimators are performing well; Gauss and t-MLE are performing slightly better than LW or Huber but differences are marginal.

Figure 2 shows the NMSE figures in the case that samples are from t_5- and t_3-distribution, respectively. In the latter case, the non-robust Gauss and LW estimator provided large NMSE and are therefore not shown in the plot. This was expected as t_3-distribution is heavy-tailed with non-finite kurtosis. As can be seen, the robust Huber and t-MLE shrinkage estimators provide significantly improved performance when the data is sampled from a heavy-tailed t_5 or t_3-distribution. We can also notice that t-MLE estimator that adaptively estimates the d.o.f. ν from the data is able to outperform the Huber's M-estimator due to the data adaptivity.

Figure 3 depicts the (average) shrinkage parameter β as a function of n in the case that samples are from a p-variate t_5 distribution. As can be seen the robust shrinkage estimators (Huber and t-MLE) use larger shrinkage parameter value β than Gauss and LW.

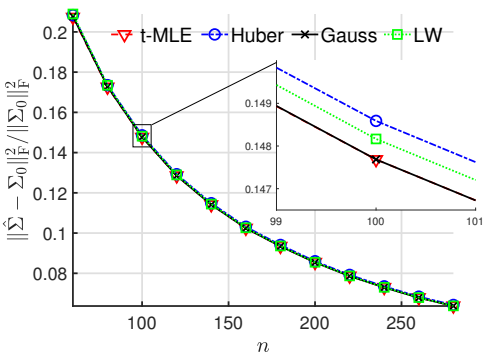

Fig. 1. NMSE of the estimators a function of n when samples are drawn from MVN distribution with an AR(1) covariance structure withs $\varrho = 0.6$ and $p = 40$.

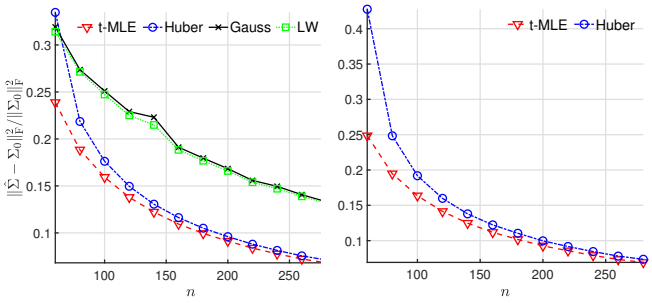

Fig. 2. NMSE of the estimators as a function of n when samples are draw from a p-variate t_5 (left panel) and t_3 (right panel) distribution with an AR(1) covariance structure; $\varrho = 0.6$ and $p = 40$.

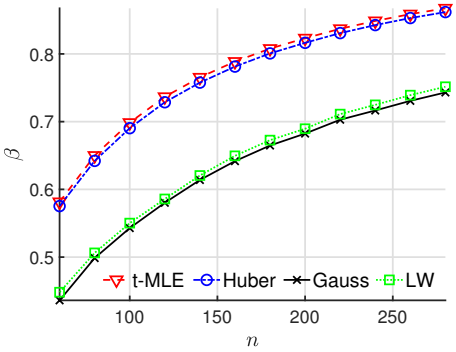

Fig. 3. Shrinkage parameter β as a function of n when samples are drawn from a p-variate t_5-distribution with an AR(1) covariance structure; $\varrho = 0.6$ and $p = 40$.

5. CONCLUSIONS AND PERSPECTIVES

This work proposed an original and fully automatic approach to compute an optimal shrinkage parameter in the context of heavy-tailed distributions and/or in presence of outliers. It has been shown that the performance of the method is similar to the optimal one when the data is Gaussian while it outperforms shrinkage Gaussian-based methods when the data distribution turns out to be non-Gaussian. This paper opens several ways, notably considering the case when $p > n$.

6. REFERENCES

[1] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Mult. Anal.*, vol. 88, no. 2, pp. 365–411, 2004.

[2] E. Ollila and E. Raninen, "Optimal shrinkage covariance matrix estimation under random sampling from elliptical distributions," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2707–2719, 2019.

[3] R. A. Maronna, "Robust M-estimators of multivariate location and scatter," *Ann. Stat.*, vol. 5, no. 1, pp. 51–67, 1976.

[4] E. Ollila and D. E. Tyler, "Regularized M-estimators of scatter matrix," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 6059–6070, 2014.

[5] F. Pascal, Y. Chitour, and Y. Quek, "Generalized robust shrinkage estimator and its application to STAP detection problem," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5640–5651, 2014.

[6] Y. Sun, P. Babu, and D. P. Palomar, "Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5143–5156, 2014.

[7] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for MMSE covariance estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5016–5029, 2010.

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

[9] T. Zhang and A. Wiesel, "Automatic diagonal loading for Tyler's robust covariance estimator," in *IEEE Statistical Signal Processing Workshop (SSP'16)*, 2016, pp. 1–5.

[10] J. T. Kent and D. E. Tyler, "Redescending M-estimates of multivariate location and scatter," *Ann. Stat.*, vol. 19, no. 4, pp. 2102–2119, 1991.

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

[12] K.-T. Fang, S. Kotz, and K.-W. Ng, *Symmetric Multivariate and Related Distributions*. London: Chapman and hall, 1990.

[13] 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, no. 11, pp. 5597–5625, 2012.

[14] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. New York: Wiley, 1982, 704 pages.