Shrinking the eigenvalues of M-estimators of covariance matrix

Esa Ollila, Daniel Palomar, Frédéric Pascal

Shrinking the eigenvalues of M-estimators of covariance matrix

Esa Ollila, *Member, IEEE*, Daniel P. Palomar, *Fellow, IEEE* and Frédéric Pascal, *Senior Member, IEEE*

***Abstract*—A highly popular regularized (shrinkage) covariance matrix estimator is the shrinkage sample covariance matrix (SCM) which shares the same set of eigenvectors as the SCM but shrinks its eigenvalues toward the grand mean of the eigenvalues of the SCM. 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, Tyler'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 an unspecified heavy-tailed elliptically symmetric distribution. Also, real-world and synthetic stock market data validate the performance of the proposed method in practical applications.**

***Index Terms*—M-estimators, sample covariance matrix, shrinkage, regularization, elliptically symmetric distributions**

I. 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 mean vector equal to zero (i.e., $\mathbb{E}[\mathbf{x}] = \mathbf{0}$). One of the first task 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 heavier 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, defined as

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

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 of the SCM $\mathbf{S}$. 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 $\bar{d} = p^{-1}\sum_{j=1}^p d_j$. Optimal computation of β such that $\mathbf{S}_\beta$ has minimum mean squared error (MMSE) has been developed for example in [1], [2] or in [3] for certain structured target matrices. A Bayesian approach has been considered in [4].

However, the estimator in (1) remains sensitive to outliers and non-Gaussianity. M-estimators of scatter [5] 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 under mild conditions on the data (see [6]) and can thus 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 the optimal parameter β_o attaining the minimum MSE 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 weight 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 and in the presence of outliers.

Earlier works, e.g., in [7], [8], [9], [10], [11], [12], [13], [14], [15], 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 a conventional M-estimator and shrinks its eigenvalues to the grand mean of the eigenvalues. Our approach permits computation of the optimal shrinkage parameter for any M-estimation weight function. Preliminary study of the proposed estimators has appeared in a conference proceeding [16].

Finally, we note some related but different approaches to what is pursed in this paper. For example, covariance matrix estimation in the low sample size large dimensionality setting commonly arises in radar signal processing, where often some constrained, mismatched or structured estimation framework of the covariance matrix is exploited. See e.g., [17], [18], [19], [20], [21], [22]. On the other hand, there are also other approaches for parameter tuning of regularized covariance matrix estimators such as cross-validation or expected likelihood approach [23], [24], [25].

The paper is structured as follows. Section II introduces the proposed shrinkage M-estimator framework while Section III discusses automatic computation of the optimal shrinkage parameter under the assumption of sampling from unspecified elliptical distribution. Extension to complex case is discussed

E. Ollila is with the Department of Signal Processing and Acoustics, Aalto University, P.O. Box 15400, FI-00076 Aalto, Finland. D. Palomar is with the Hong Kong University of Science and Technology, Hong Kong. F. Pascal is with Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France.

Manuscript received November xx, 202X; revised xx, 202X.

in subsection III-A. Section IV addresses the most commonly used M-estimators, namely, the Gaussian weight function, Huber's weight function, Tyler's weight and t-distribution weight function. We provide simulation studies in Section V and experimental results in Section VI. In Section VI we validate the promising performance of the proposed approach both in the case of real-world and synthetic stock market data. Finally, Section VII concludes, while proofs of theorems and lemmas are kept in the Appendix.

II. Shrinkage M-estimators of scatter

In this paper, we assume that $n > p$ (except for Gaussian loss) and consider an M-estimator of scatter matrix [5] that solves an estimating equation

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

where $u : [0,\infty) \to [0,\infty)$ is a non-increasing *weight function*. An M-estimator is a sort of adaptively weighted SCM with weights determined by function $u(\cdot)$. To guarantee existence of the solution, it is required that the data verifies the condition stated in [6]. An M-estimator of scatter which shrinks the eigenvalues towards the grand mean of the eigenvalues is then defined as:

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

Thus $\hat{\boldsymbol{\Sigma}}_\beta$ is indexed by the shrinkage parameter $\beta \in [0,1]$. If $\beta = 1$, then $\hat{\boldsymbol{\Sigma}}_\beta$ coincides with the conventional M-estimator in (2) while if $\beta = 0$, then $\hat{\boldsymbol{\Sigma}}_\beta$ equals an identity matrix scaled by mean of the eigenvalues of $\hat{\boldsymbol{\Sigma}}$. Next we discuss the commonly used weight functions u.

The RSCM $\mathbf{S}_\beta$ in (1) is obtained when one uses the *Gaussian weight function* $u_{\mathrm{G}}(t) = 1\ \forall t$. Terminology 'Gaussian' stems from the fact that $\hat{\boldsymbol{\Sigma}} = \mathbf{S}$ is the maximum likelihood estimate (MLE) of the covariance matrix of MVN distribution. We return to this in Section III. *Huber's weight function* is defined as

$$u_{\mathrm{H}}(t;c) = \begin{cases} 1/b, & \text{for } t \leqslant c^2 \\ c^2/(tb), & \text{for } t > c^2 \end{cases} \tag{4}$$

where $c > 0$ is a user defined tuning constant that determines the robustness and efficiency of the estimator and b is a scaling factor; see subsection IV-B for more details. Another popular choice is *MVT-weight function* [6]:

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

in which case the corresponding M-estimator $\hat{\boldsymbol{\Sigma}}$ is also the MLE of the scatter matrix parameter of the *multivariate t (MVT) distribution* with $\nu > 0$ degrees of freedom (d.o.f.). We return to this estimator in subsection IV-D. Finally, another classic choice, with nice robustness properties, is Tyler's [26] M-estimator, in which case the weight function is

$$u_{\mathrm{Tyl}}(t) = \frac{p}{t}. \tag{6}$$

Both Huber's and MVT-weight function yield Tyler's weight function as special cases; namely, for $\nu = 0$, one notices that $u_{\mathrm{T}}(t;\nu=0) = u_{\mathrm{Tyl}}(t)$ and in the limit case as $c \to 0$ Huber's weight function tends to Tyler's weight function.

We would like to stress that $n > p$ is a necessary assumption for all but Gaussian weight functions for a solution $\hat{\boldsymbol{\Sigma}}$ to (2) to exist. We do not include the limit case $n = p$ since any affine equivariant M-estimator $\hat{\boldsymbol{\Sigma}}$ when $n = p$ and data is in general position is just a scalar multiple of the SCM $\mathbf{S}$, that is, $\hat{\boldsymbol{\Sigma}} = \gamma\mathbf{S}$ for some $\gamma > 0$ [27]. For example, M-estimator based on Huber's or t-weights are affine equivariant. Moreover, note that Theorem 2 in [28] ensures similar results in the large sample regime. Namely, the following convergence is proved

$$\|\hat{\boldsymbol{\Sigma}} - \hat{\mathbf{S}}_{np}\| \xrightarrow[n,p\to\infty]{a.s} 0$$

with $n/p \to c \in (0,1)$ and $\hat{\mathbf{S}}_{np}$ is an appropriate weighted SCM, and the norm denotes the spectral norm.

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

$$\boldsymbol{\Sigma}_0 = \mathbb{E}\big[u(\mathbf{x}^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{x})\mathbf{x}\mathbf{x}^\top\big]. \tag{7}$$

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

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{x}_i^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top, \tag{8}$$

we can compute

$$\mathbf{C}_\beta = \beta\mathbf{C} + (1-\beta)[\mathrm{tr}(\mathbf{C})/p]\mathbf{I}, \tag{9}$$

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

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

$$\beta_o = \arg\min_\beta \Big\{\mathrm{MSE}(\hat{\boldsymbol{\Sigma}}_\beta) = \mathbb{E}\Big[\|\hat{\boldsymbol{\Sigma}}_\beta - \boldsymbol{\Sigma}_0\|_{\mathrm{F}}^2\Big]\Big\}, \tag{10}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius matrix norm (i.e., $\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{A}^\top\mathbf{A})$ and $\|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{A}^{\mathsf{H}}\mathbf{A})$ for real-valued and complex-valued matrices, respectively, where $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose). Since solving (10) is not feasible due to the implicit form of M-estimators, we instead solve the following much simpler problem:

$$\beta_o^{\mathrm{app}} = \arg\min_\beta \Big\{\mathrm{MSE}(\mathbf{C}_\beta) = \mathbb{E}\Big[\big\|\mathbf{C}_\beta - \boldsymbol{\Sigma}_0\big\|_{\mathrm{F}}^2\Big]\Big\}. \tag{11}$$

Such approach was also used in [29] to derive an optimal parameter for the shrinkage Tyler's M-estimator of scatter proposed by the authors,

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

$$\gamma = \frac{p\,\mathrm{tr}(\boldsymbol{\Sigma}_0^2)}{\mathrm{tr}(\boldsymbol{\Sigma}_0)^2}. \tag{12}$$

Sphericity γ [30], [31] 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 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is an i.i.d. random sample from any p-variate distribution, and u is a weight function for which the expectation $\mathbb{E}[\mathrm{tr}(\mathbf{C}^2)]$ exists. The oracle parameter β_o^{app} in* (11) *is*

$$\beta_o^{\mathrm{app}} = \frac{\|\mathbf{\Sigma}_0 - \eta_o \mathbf{I}\|_{\mathrm{F}}^2}{\mathbb{E}\Big[\big\|\mathbf{C} - (\mathrm{tr}(\mathbf{C})/p)\mathbf{I}\big\|_{\mathrm{F}}^2\Big]} \tag{13}$$

$$= \frac{p(\gamma - 1)\eta_o^2}{\mathbb{E}[\mathrm{tr}(\mathbf{C}^2)] - p^{-1}\mathbb{E}[\mathrm{tr}(\mathbf{C})^2]} \tag{14}$$

where $\eta_o = \mathrm{tr}(\mathbf{\Sigma}_0)/p$ and γ is defined in (12)*. Note that $\beta_o^{\mathrm{app}} \in [0, 1)$ and the value of the MSE at the optimum is*

$$\mathrm{MSE}(\mathbf{C}_{\beta_o^{\mathrm{app}}}) = \frac{\mathbb{E}[\mathrm{tr}(\mathbf{C})^2] - \mathrm{tr}(\mathbf{\Sigma}_0)^2}{p} + (1 - \beta_o^{\mathrm{app}})\big\|\mathbf{\Sigma}_0 - \eta_o \mathbf{I}\big\|_{\mathrm{F}}^2. \tag{15}$$

Proof. The proof is postponed to Appendix A. □

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

III. Shrinkage parameter for elliptical samples

Maronna [5] developed M-estimators of scatter matrices originally within the framework of elliptically symmetric distributions [32], [33]. The probability density function (p.d.f.) of centered (zero mean) elliptically distributed random vector, denoted by $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$, is

$$f(\mathbf{x}) = C_{p,g} |\mathbf{\Sigma}|^{-1/2} g\big(\mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{x}\big), \tag{16}$$

where $\mathbf{\Sigma}$ is a 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 $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ is obtained when $g(t) = \exp(-t/2)$ and the t-distribution with ν d.o.f., denoted $\mathbf{x} \sim t_\nu(\mathbf{0}, \mathbf{\Sigma})$, is obtained when $g(t) = (1 + t/\nu)^{-(p+\nu)/2}$. Then the weight function for the MLE of scatter matrix corresponds to the case that the weight function is of the form $u(t) \propto -g'(t)/g(t)$. This yields (5) as the weight function for the MLE of scatter for t-distribution. 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 [33] for details.

When $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \mathbf{\Sigma}, g)$, then the M-functional $\mathbf{\Sigma}_0$ in (7) is related to underlying scatter matrix parameter $\mathbf{\Sigma}$ via the relationship

$$\mathbf{\Sigma}_0 = \sigma \mathbf{\Sigma}, \tag{17}$$

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

$$\mathbb{E}\left[\psi\left(\frac{r^2}{\sigma}\right)\right] = p, \tag{18}$$

where $\psi(t) = u(t)t$ and $r = \|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|$. Often σ needs to be solved numerically from (18) 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 (14). These depend on a constant ψ_1 (which depends on weight function u via $\psi(t) = u(t)t$) as follows:

$$\psi_1 = \frac{1}{p(p+2)} \mathbb{E}\left[\psi\Big(\frac{r^2}{\sigma}\Big)^2\right], \tag{19}$$

where the statistical expectation is again computed w.r.t. distribution of the positive random variable $r = \|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|$.

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*

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

and

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

given that expectation (19) *exists.*

Proof. The proof is given in Appendix B. □

This then yields the following main result.

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*

$$\beta_o^{\mathrm{app}} = \frac{n(\gamma - 1)}{n(\gamma - 1)(1 - 1/n) + \psi_1(1 - 1/p)(2\gamma + p)} \tag{20}$$

where γ is defined in (12) *and ψ_1 in* (19)*.*

Proof. Follows from Theorem 1 after substituting the values for $\mathbb{E}\big[\,\mathrm{tr}\big(\mathbf{C}^2\big)\big]$ and $\mathbb{E}\big[\,\mathrm{tr}(\mathbf{C})^2\big]$ derived in Lemma 1 into the denominator of β_o^{app} in (14). □

A closely related result is derived in [34, Theorem 1]. Namely, [34] considers oracle estimator as in (9) but using a shrinkage target equal to identity matrix $\mathbf{I}$ instead of $[\mathrm{tr}(\mathbf{C})/p]\mathbf{I}$ as in this paper. This is due to the fact that [34] assumes that $\mathrm{tr}(\mathbf{\Sigma}) = p$. Another difference is that [34] assumes that $\mathbf{\Sigma}_0 = \mathbf{\Sigma}$ (so $\sigma = 1$), i.e., that the used M-estimator is consistent to the scatter matrix of the underlying elliptical population. This assumption implies knowledge of the underlying elliptical distribution in which case it is natural to use the ML-weight $u_{\mathrm{ML}}(t) = -2g(t)/g'(t)$ as was done in [34]. Thus Theorem 2 compared to [34, Theorem 1] holds in the more general case when the scale $\mathrm{tr}(\mathbf{\Sigma})$ is not known *apriori* and no assumption on the knowledge of the elliptical distribution is imposed. Furthermore, in the next subsesction, we extend the result to the complex-valued case which was not considered in [34].

Lemma 1 also allows to construct an unbiased estimate of $\vartheta = \mathrm{tr}(\mathbf{\Sigma}_0^2)/p$ as is shown next.

Theorem 3. *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is an i.i.d. random sample from $\mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$. Then an unbiased estimate of $\vartheta = \mathrm{tr}(\boldsymbol{\Sigma}_0^2)/p$ for any finite n and any p is*

$$\hat{\vartheta} = b_n \left(\frac{\mathrm{tr}(\mathbf{C}^2)}{p} - \psi_1 a_n \frac{p}{n} \left[\frac{\mathrm{tr}(\mathbf{C})}{p} \right]^2 \right), \tag{21}$$

where

$$a_n = \frac{n}{n + \psi_1 - 1} \quad \text{and} \quad b_n = \frac{n}{n-1} \left(\frac{n - 1 + \psi_1}{n - 1 + 3\psi_1} \right) \tag{22}$$

given that expectation (19)*, defining ψ_1, exists.*

Proof. First note that

$$\mathbb{E}[\hat{\vartheta}] = b_n(\mathbb{E}[\mathrm{tr}(\mathbf{C}^2)]/p - a_n(p/n)\mathbb{E}[\mathrm{tr}(\mathbf{C})^2]/p^2). \tag{23}$$

Then substituting the values of $\mathbb{E}[\mathrm{tr}(\mathbf{C})^2]$ and $\mathbb{E}\big[\mathrm{tr}\big(\mathbf{C}^2\big)\big]$ from Lemma 1 into (23) yields $\mathbb{E}[\hat{\vartheta}] = \mathrm{tr}(\boldsymbol{\Sigma}_0^2)/p$. □

It is instructive to consider in more detail the case of Gaussian loss. In this case, $\mathbf{C}$ equals the SCM, $\mathbf{C} = \mathbf{S}$, and the statistic $\hat{\vartheta}$ no longer depends on the unknown $\boldsymbol{\Sigma}_0$. Furthermore, if data is generated from a Gaussian distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, then $\psi_1 = 1$, $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}$ and $\hat{\vartheta}$ in (21) reduces to the estimator that is identical to one proposed by [31, Lemma 2.1] in the case that location is known ($\boldsymbol{\mu} = \mathbf{0}$); see also [3, Theorem 2]. Moreover, [2, Theorem 4] is obtained in the general elliptical case, again assuming known location parameter ($\boldsymbol{\mu} = \mathbf{0}$).

A. An extension to the complex-valued case

Consider the case that $\mathbf{x}_i$, $i = 1, \ldots, n$ are complex-valued (i.e., $\mathbf{x}_i \in \mathbb{C}^p$) and represent a random sample from a (circular) complex elliptically symmetric (CES) distribution [33]. The p.d.f. of a random vector $\mathbf{x} \in \mathbb{C}^p$ with centered (zero mean) CES distribution, denoted using same notation $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$, is

$$f(\mathbf{x}) = C_{p,g} |\boldsymbol{\Sigma}|^{-1} g\big(\mathbf{x}^{\mathsf{H}} \boldsymbol{\Sigma}^{-1} \mathbf{x}\big), \qquad \mathbf{x} \in \mathbb{C}^p,$$

where $\boldsymbol{\Sigma}$ is the positive definite hermitian (PDH) 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 $\boldsymbol{\Sigma}$, and $C_{p,g}$ is a normalizing constant ensuring that $f(\mathbf{x})$ integrates to 1.

An M-estimator of scatter matrix $\hat{\boldsymbol{\Sigma}}$ is a PDH matrix that solves an estimating equation

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n u(\mathbf{x}_i^{\mathsf{H}} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^{\mathsf{H}}, \tag{24}$$

where $u : [0, \infty) \to [0, \infty)$ is a non-increasing weight function. Again, $u(t) = 1$ gives the SCM $\mathbf{S} = \frac{1}{n} \sum_i \mathbf{x}_i \mathbf{x}_i^{\mathsf{H}}$, Huber's and Tyler's weight functions are as earlier, stated in (4) and (6), respectively, whereas $u_{\mathrm{T}}(t; \nu) = \frac{2p+\nu}{\nu+2t}$ corresponds to MLE of the scatter matrix parameter when sampling from a p-variate complex t-distribution with ν d.o.f. We refer to [33], [35] for more details. We may now define the shrinkage M-estimator as in (3). Definitions (7)-(9) hold also in the complex-valued case with obvious modifications (replacing transpose by the Hermitian transpose).

Theorem 1 did not require an assumption that random vectors are real-valued, i.e., it holds also when $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. complex-valued random vectors. This means that we only need to derive expectations in Lemma 1 in the case of random sampling from a CES distribution. First, we define the parameter ψ_1 in the complex-valued case as

$$\psi_1 = \frac{1}{p(p+1)} \mathbb{E}\Big[\psi\Big(\frac{r^2}{\sigma} \Big)^2 \Big], \tag{25}$$

where the expectation is w.r.t. $r = \|\boldsymbol{\Sigma}^{-1/2} \mathbf{x}\|$, where $\mathbf{x} \sim \mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$. The analog of Lemma 1 to complex case is given next.

Lemma 2. *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is an i.i.d. random sample from a (circular) complex elliptically symmetric distribution $\mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$. Then*

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

and

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

given that expectation (25) *exists.*

Proof. The proof is given in Appendix C. □

Plugging in the expectations above into β_o^{app} derived in Theorem 1 yields the following result.

Theorem 4. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ denote an i.i.d. random sample from a (circular) p-variate complex elliptically symmetric distribution $\mathcal{E}_p(\mathbf{0}, \boldsymbol{\Sigma}, g)$ and assume that expectation* (25) *exists. Then the oracle parameter β_o^{app} that minimizes the MSE in Theorem 1 is*

$$\beta_o^{\mathrm{app}} = \frac{n(\gamma - 1)}{n(\gamma - 1)(1 - 1/n) + \psi_1(1 - 1/p)(\gamma + p)}, \tag{26}$$

where γ is defined in (12) *and ψ_1 in* (25)*.*

Furthermore, $\hat{\vartheta}$ defined as in (21) *with*

$$a_n = \frac{n}{n + \psi_1 - 1} \quad \text{and} \quad b_n = \frac{n}{n-1} \left(\frac{n - 1 + \psi_1}{n - 1 + 2\psi_1} \right)$$

is an unbiased estimate of $\vartheta = \mathrm{tr}(\boldsymbol{\Sigma}_0^2)/p$ for any finite n and any p.

B. Computing the shrinkage parameter

In order to construct a data-adaptive estimate of β_o^{app} (either in real- or complex-valued cases), all we need to estimate is the sphericity γ and the constant ψ_1. An estimate $\hat{\psi}_1$ of ψ_1 is constructed separately for each weight function (Gaussian, Huber's, MVT- and Tyler's weight function) in Section IV. However, it is also possible to use an empirical (sample mean) version of (19). Next we discuss computation of the sphericity estimator $\hat{\gamma}$.

As an estimator $\hat{\gamma}$ we use the estimate derived in [14], which was named as Ell1-estimator in [2], and defined as

$$\hat{\gamma}^{\mathrm{Ell1*}} = \frac{n}{n-1} \left(p \ \mathrm{tr} \left(\frac{1}{n} \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\|\mathbf{x}_i\|^2} \right) - \frac{p}{n} \right) \tag{27}$$

which for complex-valued case is defined analogously (transpose replaced with the Hermitian transpose).

Next recall that $\hat{\vartheta}$ defined in (21) is an unbiased estimator of $\mathrm{tr}(\boldsymbol{\Sigma}_0^2)/p$. This statistics depends on $\mathbf{C}$ and ψ_1 which are unknown, but a plug-in estimate of $\hat{\vartheta}$ can be constructed by replacing $\mathbf{C}$ and ψ_1 with $\hat{\boldsymbol{\Sigma}}$ and $\hat{\psi}_1$, respectively. Dividing this plug-in statistic further by $(\mathrm{tr}(\hat{\boldsymbol{\Sigma}})/p)^2$, leads to another estimator of sphericity, named as Ell2-estimator, and defined as

$$\hat{\gamma}^{\mathrm{Ell2*}} = \hat{b}_n \left(\frac{p\,\mathrm{tr}(\hat{\boldsymbol{\Sigma}}^2)}{\mathrm{tr}(\hat{\boldsymbol{\Sigma}})^2} - \hat{\psi}_1 \hat{a}_n \frac{p}{n} \right), \tag{28}$$

where the constants $\hat{a}_n$ and $\hat{b}_n$ are as defined in Theorem 3 (and Theorem 4 in complex case) but with ψ_1 replaced by its estimate $\hat{\psi}_1$. When one uses Gaussian weight function, then $\hat{\boldsymbol{\Sigma}} = \mathbf{S}$ and $\hat{\psi}_1 = 1 + \hat{\kappa}$, where $\hat{\kappa}$ is an estimate of elliptical kurtosis (*cf.* subsection IV-A). In this case, $\hat{\gamma}^{\mathrm{Ell2*}}$ corresponds to Ell2-estimator of sphericity proposed in [2, Sect. IV.B].

In order to guarantee that the estimators remain in the valid interval, $1 \leq \gamma \leq p$, we use

$$\hat{\gamma}^{\mathrm{Ell1}} = \min(p, \max(1, \hat{\gamma}^{\mathrm{Ell1*}})) \tag{29}$$

as our final estimator (and similarly for Ell2-estimator). The related shrinkage parameter can then computed as

$$\beta = \beta_o^{\mathrm{app}}(\hat{\gamma}^{\mathrm{Ell1}}, \hat{\psi}_1),$$

and again similarly for Ell2-estimator.

IV. Important special cases

A. Regularized SCM (RSCM) estimator

If one uses Gaussian weight function $u_{\mathrm{G}}(t) \equiv 1$, then a necessary assumption is that the underlying elliptical distribution possesses finite 4th-order moments. As discussed earlier, one may then assume w.l.o.g. that the scatter matrix parameter equals the covariance matrix, i.e., $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x})$. When $u_{\mathrm{G}}(t) \equiv 1$, one has that $\hat{\boldsymbol{\Sigma}} = \mathbf{S}$ and $\mathbf{C}_\beta = \mathbf{S}_\beta$ and hence $\beta_o = \beta_o^{\mathrm{app}}$, meaning that the approximate MMSE solution is exact in this case. Finally, we may relate ψ_1 with an elliptical kurtosis [36] parameter κ:

$$\psi_1 = 1 + \kappa = \begin{cases} \dfrac{\mathbb{E}[r^4]}{p(p+2)}, & \text{real case} \\ \dfrac{\mathbb{E}[r^4]}{p(p+1)}, & \text{complex case} \end{cases} \tag{30}$$

where the expectation is over the distribution of the random variable $r = \|\boldsymbol{\Sigma}^{-1/2}\mathbf{x}\|$. The elliptical kurtosis parameter is defined as a generalization of the kurtosis parameter to the vector case, and as such it vanishes (so $\kappa = 0$) when $\mathbf{x}$ has MVN distribution (denoted $\mathbf{x} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$). Since $\mathbf{S}$ exists for any $n \geq 1$, we can drop the assumption that $n > p$ in this case.

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

$$\beta_o = \arg\min_\beta \mathbb{E}\big[\|\mathbf{S}_\beta - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2\big] = \frac{n(\gamma - 1)}{n(\gamma - 1) + p + a}, \tag{31}$$

where

$$a = \begin{cases} \kappa(2\gamma(1 - 1/p) + p - 1) + \gamma(1 - 2/p), & \textit{real case} \\ \kappa(\gamma(1 - 1/p) + p - 1) - \gamma/p, & \textit{complex case} \end{cases}$$

Proof. The result follows from Theorem 2 and Theorem 4 since $\mathbf{C}_\beta = \mathbf{S}_\beta$ and the M-functional for Gaussian loss is $\boldsymbol{\Sigma}_0 = \mathrm{cov}(\mathbf{x}) = \boldsymbol{\Sigma}$ and $\sigma = 1$. Since for Gaussian loss, $\psi(t) = t$, we notice from (19) and (25) hat

$$\psi_1 = 1 + \kappa. \tag{32}$$

Plugging $\psi_1 = 1 + \kappa$ into β_o^{app} in Theorem 2 and Theorem 4 yields the stated expressions, respectively. □

The elliptical kurtosis parameter κ can be easily estimated using the following relationship to kurtosis even in the cases when $p > n$. First, recall that kurtosis of a random variable x in the real and complex case is defined as

$$\mathsf{kurt}(x) = \frac{\mathbb{E}[x^4]}{(\mathbb{E}[x^2])^2} - 3 \quad \text{and} \quad \mathsf{kurt}(x) = \frac{\mathbb{E}[|x|^4]}{(\mathbb{E}[|x|^2])^2} - 2, \tag{33}$$

respectively. Kurtosis vanishes when the random variable has real or complex Gaussian distribution with variance $\mathbb{E}[|x|^2]$. The following result establishes the relationship of elliptical kurtosis parameter with marginal kurtosis.

Lemma 3. *Assume that $\mathbf{x}$ is a random vector from real or complex elliptically symmetric distribution with covariance matrix $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x})$ possessing finite 4th order moments. Then*

$$\kappa = \begin{cases} \frac{1}{3}\mathsf{kurt}(x_j), & \textit{real case} \\ \frac{1}{2}\mathsf{kurt}(x_j), & \textit{complex case} \end{cases} \tag{34}$$

where x_j is any jth component of $\mathbf{x}$ ($j \in \{1, \ldots, p\}$).

Proof. The proof is given in Appendix D. □

Since all marginal variables possess the same kurtosis, an estimate $\hat{\kappa}$ can be formed simply as the mean of marginal sample kurtosis statistics. This is the same estimate of the elliptical kurtosis proposed in [2]. Note that [2] only considered the real-valued case, and thus Corollary 1 allows us to extend the RSCM estimator in [2] to complex-valued case.

In the sequel, we use acronym **RSCM-Ell1** to refer to estimator $\mathbf{S}_\beta$ with β computed as $\beta = \beta_o(\hat{\kappa}, \hat{\gamma}^{\mathrm{Ell1}})$ with β_o given by (31) and $\hat{\gamma}^{\mathrm{Ell1}}$ being the estimate of sphericity defined in (29) and $\hat{\kappa}$ an estimate of elliptical kurtosis described above. An **RSCM-Ell2** estimator is defined similarly but now using Ell2-estimator of sphericity.

A natural competitor for RSCM-Ell1 or RSCM-Ell2 estimators (at least in the real-valued case) is the estimator proposed by Ledoit and Wolf [1], referred to as **LWE**. We note that LWE also uses RSCM $\mathbf{S}_\beta$, but the parameter β is computed in a different manner. An extra benefit of our approach is that an estimator of the optimal shrinkage parameter can be computed for real- or complex-valued observations while LWE assumes real-valued observations.

B. Regularized Huber's M-estimator (RHub)

Next consider the Huber's weight function $u_{\text{H}}(t;c)$ in (4). Note that $b>0$ is a scaling constant; if $\hat{\mathbf{\Sigma}}$ is Huber M-estimator of scatter when $b=1$, then the Huber M-estimator of scatter when $b=b_o$ is simply $b_o\hat{\mathbf{\Sigma}}$. The scaling constant b is usually chosen so that the resulting scatter estimator is Fisher consistent for the covariance matrix at MVN distribution, i.e, $\sigma=1$ when $\mathbf{x}\sim\mathcal{N}_p(\mathbf{0},\mathbf{\Sigma})$. In the real case, this holds when

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

where $F_{\chi^2_p}(\cdot)$ denotes the cumulative distribution function (c.d.f.) of chi-squared distribution with p d.o.f. Since $r^2=\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|^2$ has a χ^2_p-distribution when $\mathbf{x}\sim\mathcal{N}_p(\mathbf{0},\mathbf{\Sigma})$, the tuning constant c^2 is chosen as qth upper quantile of χ^2_p-distribution:

$$q=\Pr(r^2\leq c^2)\Leftrightarrow F^{-1}_{\chi^2_p}(q)=c^2 \tag{35}$$

for some $q\in(0,1]$. Tuning constant c and scaling factor b can be determined similarly in the complex-valued case; see [33], [35] for details.

Let us define a winsorized observation $\mathbf{w}$ based on $\mathbf{x}\sim\mathcal{E}_p(\mathbf{0},\mathbf{\Sigma},g)$ as

$$\mathbf{w}=\mathsf{wins}(\mathbf{x};\mathbf{\Sigma},c)=\frac{1}{\sqrt{b}}\begin{cases}\mathbf{x}, & \|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|^2\leqslant c^2\\ c\dfrac{\mathbf{x}}{\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|}, & \|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|^2>c^2\end{cases}$$

where c is the threshold c of Huber's weight function and b is the respective scaling factor. The winsorized r.v. $\mathbf{w}$ also has an elliptically symmetric distribution since the contours remain elliptical in shape (so the p.d.f. is still defined by (16) but for a truncated density generator g) and thus it shares the properties of elliptical random vectors.

If we take $\sigma=1$ (which holds at least when $\mathbf{x}$ has MVN distribution), then the constant ψ_1 can be written as

$$\begin{aligned}\psi_1&=\frac{\mathbb{E}[\psi^2_{\text{H}}(\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|^2;c)]}{p(p+2)}=\frac{\mathbb{E}\big[\|\mathbf{\Sigma}^{-1/2}\mathbf{w}\|^4\big]}{p(p+2)}\\&=1+\kappa_{\mathbf{w}}\end{aligned}$$

where $\kappa_{\mathbf{w}}$ is the elliptical kurtosis parameter (*cf.* Lemma 3) of an elliptical random vector $\mathbf{w}$. An estimate $\hat{\psi}_1$ of ψ_1 can be then calculated similarly as ψ_1 for RSCM-Ell1 or RSCM-Ell2 estimators defined earlier (recall relation (32)). The only difference is that κ is now computed for winsorized data $\{\mathbf{w}_i\}_{i=1}^n$, where $\mathbf{w}_i=\mathsf{wins}(\mathbf{x}_i;\hat{\mathbf{\Sigma}},c)$ and $\hat{\mathbf{\Sigma}}$ denotes the Huber's M-estimator.

In the sequel, we use acronym **RHub-Ell1** or **RHub-Ell2** to refer to shrinkage M-estimator $\hat{\mathbf{\Sigma}}_\beta$ that uses Huber's weight $u(\cdot)=u_{\text{H}}(\cdot;c)$ with threshold c^2 determined from (35) for user specified q and shrinkage parameter $\beta=\beta_o^{\text{app}}(\hat{\gamma}^{\text{Ell1}},\hat{\psi}_1)$. or $\beta=\beta_o^{\text{app}}(\hat{\gamma}^{\text{Ell2}},\hat{\psi}_1)$, respectively.

C. Regularized Tyler's M-estimator (RTyl)

Let $\mathbf{V}$ denote a *shape matrix* (normalized scatter matrix), defined as

$$\mathbf{V}=p\mathbf{\Sigma}/\operatorname{tr}(\mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ denotes the scatter matrix parameter of the ES distribution. Note that $\operatorname{tr}(\mathbf{V})=p$. If one uses Tyler's weight function in (6), then (17) holds with $\sigma=p/\operatorname{tr}(\mathbf{\Sigma})$, i.e., $\mathbf{\Sigma}_0=\mathbf{V}$, that is, Tyler's M-estimator is an estimate of the shape matrix. The following result hence follows at once from Theorem 2 and Theorem 4.

Corollary 2. *Let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ denote an i.i.d. random sample from a (real or complex) elliptical distribution $\mathcal{E}_p(\mathbf{0},\mathbf{\Sigma},g)$. When using Tyler's weight* (6)*, it holds that*

$$\begin{aligned}\beta_o^{\text{app}}&=\arg\min_\beta\ \mathbb{E}\big[\|\mathbf{C}_\beta-\mathbf{V}\|^2_{\text{F}}\big]\\&=\begin{cases}\dfrac{n(\gamma-1)}{n(\gamma-1)(1-\frac{1}{n})+\frac{p-1}{p+2}(2\gamma+p)}, & \textit{real case}\\[2ex] \dfrac{n(\gamma-1)}{n(\gamma-1)(1-\frac{1}{n})+\frac{p-1}{p+1}(\gamma+p)}, & \textit{complex case}\end{cases}\end{aligned}$$

Proof. The real-valued case follows by noting that ψ_1 in (19) is equal to $\psi_1=p/(p+2)$ for all random vectors $\mathbf{x}\sim\mathcal{E}_p(\mathbf{0},\mathbf{\Sigma},g)$ regardless of g (i.e., the functional form of the density generator) and that $\mathbf{\Sigma}_0=\mathbf{V}$. Plugging $\psi_1=p/(p+2)$ into (20) yields the stated expression. The complex-valued case follows similarly. □

Since Tyler's M-estimator verifies $\operatorname{tr}(\hat{\mathbf{\Sigma}})=p$, the shrinkage estimator in (3) simplifies to

$$\hat{\mathbf{\Sigma}}_\beta=\beta\hat{\mathbf{\Sigma}}+(1-\beta)\mathbf{I}, \tag{36}$$

where $\hat{\mathbf{\Sigma}}$ is Tyler's M-estimator, i.e., an M-estimator based on weight (6). By **RTyl-Ell1** we refer to (36), where the shrinkage parameter is computed as $\beta=\beta_o^{\text{app}}(\hat{\gamma}^{\text{Ell1}})$ with β_o^{app} given by Corollary 2 and $\hat{\gamma}^{\text{Ell1}}$ by (29).

A related regularized Tyler's estimator was proposed by [7] as the limit of the algorithm

$$\begin{aligned}&\mathbf{\Sigma}_{k+1}\leftarrow\beta\frac{p}{n}\sum_{i=1}^n\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top\mathbf{V}_k^{-1}\mathbf{x}_i}+(1-\beta)\mathbf{I}\\&\mathbf{V}_{k+1}\leftarrow p\mathbf{\Sigma}_{k+1}/\operatorname{tr}(\mathbf{\Sigma}_{k+1}),\end{aligned} \tag{37}$$

where $\beta\in(0,1)$ is a fixed shrinkage parameter. This algorithm represents a diagonally loaded version of the fixed-point algorithm given for Tyler's M-estimator. Uniqueness and convergence of the recursive algorithm has been later derived in [29], [12]. By **CWH** estimator we now refer to estimator obtained by iterating (37) using same value $\beta=\beta_o^{\text{app}}(\hat{\gamma}^{\text{Ell1}})$ as for RTyl-Ell1. An interesting question then is how different is RTyl-Ell1 in its performance from CWH. We explore this by simulation studies later. This is interesting as the former is simply shrinking the eigenvalues of Tyler's M-estimator towards its grand mean where as the latter does not have an explicit connection to Tyler's M-estimator $\hat{\mathbf{\Sigma}}$ for any $\beta\in(0,1)$.

D. Regularized M-estimator for MVT distribution (RMVT)

We assume that the data is arising from a MVT distribution $t_\nu(\mathbf{0},\mathbf{\Sigma})$ but the d.o.f. parameter ν is unknown and is adaptively estimated from the data using Algorithm 1 explained below. Once $\hat{\nu}$ is found, we use function $u(\cdot)=u_{\text{T}}(\cdot;\hat{\nu})$

to compute the underlying M-estimator $\hat{\mathbf{\Sigma}}$ for the postulated MVT distribution.

Yet, we need to address the question of how the constant ψ_1 is computed. Due to data adaptive estimation of ν, we can assume that $\sigma \approx 1$ since the scaling factor σ equals unity for an MLE of the scatter matrix parameter. We use the fact that for MVT distribution (i.e., when $\mathbf{x} \sim t_\nu(\mathbf{0}, \mathbf{\Sigma})$), the ψ_1 parameter is[1]

$$\psi_1 = \begin{cases} \dfrac{p+\nu}{2+p+\nu}, & \text{real case} \\ \dfrac{2p+\nu}{2+2p+\nu}, & \text{complex case} \end{cases}.$$

Hence, a natural estimate is $\hat{\psi}_1 = (p+\hat{\nu})/(2+p+\hat{\nu})$ in the real case. An estimate of $\hat{\psi}_1$ is constructed similarly in the complex case. We use acronym **RMVT-Ell1** to refer to shrinkage M-estimator $\hat{\mathbf{\Sigma}}_\beta$ that uses $u(\cdot;\hat{\nu})$ with shrinkage parameter calculated by $\beta = \beta_o^{\text{app}}(\hat{\gamma}^{\text{Ell1}}, \hat{\psi}_1)$. **RMVT-Ell2** is constructed similarly, but now Ell2-estimator of sphericity $\hat{\gamma}^{\text{Ell2}}$ is used.

Next we discuss our approach for estimating ν from the data. Assume $\mathbf{x} \sim t_\nu(\mathbf{0}, \mathbf{\Sigma}, g)$ and denote $\eta = \operatorname{tr}(\mathbf{\Sigma})/p$. Then,

$$\mathbf{R} = \operatorname{cov}(\mathbf{x}) = (\nu/(\nu-2))\mathbf{\Sigma}$$

and hence $\operatorname{tr}(\mathbf{R})/p = (\nu/(\nu-2))\operatorname{tr}(\mathbf{\Sigma})/p$. This means that

$$\frac{\nu}{\nu-2} = \frac{\operatorname{tr}(\mathbf{R})}{\operatorname{tr}(\mathbf{\Sigma})} = \eta_{\text{ratio}}$$

from which we obtain the relation

$$\nu = \frac{2\eta_{\text{ratio}}}{\eta_{\text{ratio}} - 1}. \tag{38}$$

The above relation holds true in both real and complex cases. Then given an estimate $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$, we may compute an estimate $\hat{\eta}_{\text{ratio}} = \operatorname{tr}(\mathbf{S})/\operatorname{tr}(\hat{\mathbf{\Sigma}})$ which in turn provides an estimate $\hat{\nu}$ via (38). This gives rise to an iterative algorithm to estimate ν detailed in Algorithm 1. In the simulations, the algorithm converged, but already 2 iterations are sufficient to yield accuracy to first decimal; see Figure 1 for an illustration. The initial estimate is $\nu_o = 2/(\max(0,\hat{\kappa})+\delta)+4$, where $\hat{\kappa}$ is an estimate of marginal kurtosis explained in subsection IV-A (see also [2] for more details) and $\delta > 0$ is a small number. The initial start ν_o is based on the following relationship with elliptical kurtosis parameter, $\kappa = 2/(\nu-4)$, i.e., $\nu = 2/\kappa + 4$ which holds true both in real and complex cases. Again the estimate $\hat{\nu}$ in the complex-valued case is constructed similarly. Note that also other estimators of ν are proposed in the literature, for example in [34].

V. Simulation Studies

In the simulation study, we generate samples from real ES distributions with a scatter matrix $\mathbf{\Sigma}$ following an AR(1) structure, $(\mathbf{\Sigma})_{ij} = \tau\varrho^{|i-j|}$, where $\varrho \in (0,1)$ and scale parameter $\tau = \operatorname{tr}(\mathbf{\Sigma})/p = 10$. When $\varrho \downarrow 0$, then $\mathbf{\Sigma}$ is close to an identity matrix scaled by τ, and when $\varrho \uparrow 1$, $\mathbf{\Sigma}$ tends to a singular matrix of rank 1. The results are reported for the

[1]Note that the t-weight in the complex case is [33]: $u_{\text{T}}(t;\nu) = \frac{2p+\nu}{\nu+2t}$.

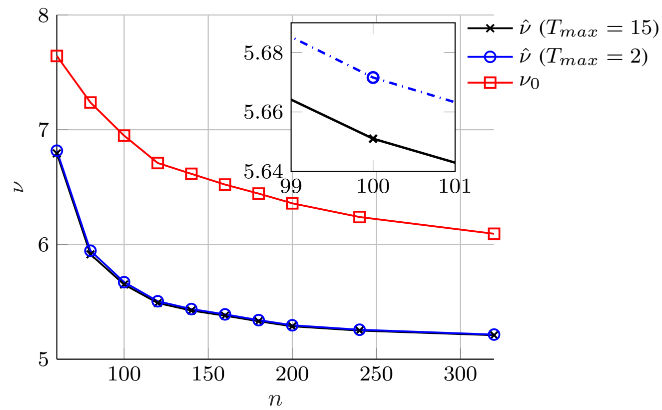


Fig. 1. Average $\hat{\nu}$ by running the Algorithm 1 with different choices of T_{max}. Also shown is the initial estimate ν_0. The samples are generated from a p-variate t-distribution with $\nu = 5$ d.o.f., where $\mathbf{\Sigma}$ follows the same AR(1) covariance matrix structure explained in the simulation set-up of Section V; $\varrho = 0.6$ and $p = 40$, As can be noted, $\hat{\nu}$ converge to $\nu = 5$ as n increases, albeit the convergence is a bit slow.

Algorithm 1: Automatic data-adaptive computation of the d.o.f. parameter ν

Input : data matrix $\mathbf{X}$ of size $n \times p$, maximum number of iterations T_{max}

Initialize: Compute $\nu_0 = 2/\max(0,\hat{\kappa}) + 4$, where $\hat{\kappa}$ is an estimate of κ explained in the text.

for $\mathrm{t} = 0, 1, \ldots, T_{max}$ **do**

Set $\hat{\mathbf{\Sigma}}_{\mathrm{t}} = \hat{\mathbf{\Sigma}}$, where $\hat{\mathbf{\Sigma}}$ denotes the t-MLE based on current estimate of d.o.f. parameter $\nu = \nu_{\mathrm{t}}$, i.e., solving the M-estimating equation

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

with $u(\cdot) = u_{\text{T}}(\cdot;\nu_{\mathrm{t}})$ is the MVT-weight function.

Update the ratio $\hat{\eta}_{\mathrm{t}} = \dfrac{\operatorname{tr}(\mathbf{S})}{\operatorname{tr}(\hat{\mathbf{\Sigma}}_{\mathrm{t}})}$.

Upate the d.o.f. parameter $\nu_{\mathrm{t}+1} = \dfrac{2\eta_{\mathrm{t}}}{\eta_{\mathrm{t}} - 1}$.

if $|\nu_{\mathrm{t}+1} - \nu_{\mathrm{t}}|/\nu_{\mathrm{t}} < 0.01$ **then**

break

Output : $\hat{\nu} = \nu_{\mathrm{t}+1}$

proposed shrinkage M-estimators using shrinkage parameter estimates based on Ell1-estimator of sphericity. However, for notational convenience, we drop the suffix -Ell1 from the proposed estimators. Thus the proposed estimators, described in Section IV are referred to as RSCM, RMVT, RHub, and RTyl. Furthermore, acronyms LW, CWH and RBLW are used to refer to estimators proposed in [1] (see also subsection IV-A), [29] (see also subsection IV-C and (37)) and [8], respectively. **RBLW** is the Rao-Blackwellized LW estimator, but unlike LW estimator, it assumes that the data distribution is Gaussian.

We also compare to RSCM estimator $\mathbf{S}_\beta$ in (1), but now the shrinkage parameter β is chosen via k-fold cross-validation (CV), where as cross-validation fit we use $\|\mathbf{S}_{\beta,\text{tr}} - \mathbf{S}_{\text{val}}\|_{\text{F}}$, where $\mathbf{S}_{\text{val}}$ is the SCM based on the validation set (data fold

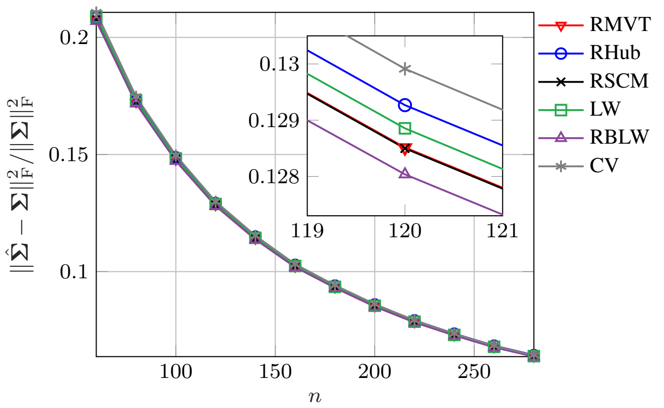


Fig. 2. NMSE as a function of n when samples are drawn from a MVN distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ with an AR(1) structure; $\varrho = 0.6$ and $p = 40$.

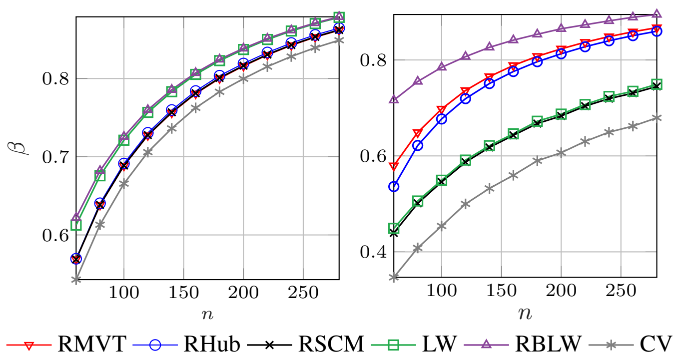


Fig. 3. Shrinkage parameter β as a function of n when samples are drawn from a MVN distribution (left panel) and a t-distribution with $\nu = 5$ d.o.f. (right panel), where $\mathbf{\Sigma}$ has an AR(1) structure; $\varrho = 0.6$ and $p = 40$.

that was left out) and $\mathbf{S}_{\beta,\text{tr}}$ is the RSCM computed on the training set (data based on remaining folds) for a given β. As a grid for β for the CV method we use a uniform grid in $[0, 1]$ with 0.05 increments and 5-fold cross-validation. We call this method as **RSCM-CV** or simply **CV**. All simulation results in this section are averaged over 2000 Monte-Carlo trials. Since $n > p$ is assumed for all estimators expect for RSCM, we do not consider the case low sample regime, $n \leq p$, in our simulation studies. Furthermore, we adopt the MSE (squared Frobenius norm) as our performance metric as it is used in deriving the optimal shrinkage parameters in this paper. It is important to keep in mind, however, that in the low sample regime and for different performance metrics, the performance differences between estimators can often be noticeable and even quite different than in the $n > p$ regime that is considered here; See e.g., [8], [4], [2] for numerical illustrations and [37], [38] for different distances between covariances that could be used instead of the MSE metric.

A. Gaussian data

The data is generated from MVN distribution $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ has an AR(1) covariance matrix structure with $\varrho = 0.6$. The dimension is $p = 40$ and n varies from 60 to 280. Value $q = 0.7$ determining the threshold c is used in (35) for Huber's weight. Since Huber's M-estimator is scaled to be consistent to the covariance matrix for Gaussian samples, the underlying population parameter $\mathbf{\Sigma}_0$ coincides with the covariance matrix $\mathbf{\Sigma}$ in this case. We also scaled the MVT-weight $u_{\text{T}}(t;\nu)$ so that it is consistent to $\mathbf{\Sigma}$ for Gaussian data. Figure 2 compares the normalized MSE (NMSE) $\|\hat{\mathbf{\Sigma}}_\beta - \mathbf{\Sigma}\|_{\text{F}}^2 / \|\mathbf{\Sigma}\|_{\text{F}}^2$ of different estimators w.r.t. increasing sample length n. It can be noted that all estimators provide essentially equally good estimator of the covariance matrix $\mathbf{\Sigma}$ for Gaussian data; RSCM and RMVT are performing equally well, largely due to the effect of data-adaptive estimation of d.o.f. parameter ν. It should be noted that their performance difference to LW or RHub estimators are still marginal and differences can be spotted only by zooming in as in the sub-plot of Figure 2. As expected, RBLW estimator has a slight advantage over the other estimators in this case. The left panel of Figure 3 shows the (average) shrinkage parameter β as a function of n. As can be noted, the average shrinkage parameter of the proposed RSCM estimator can be seen to be roughly an average of CV and LW shrinkage parameters.

B. Heavy-tailed data

Next we computed the NMSE curves when the data is generated from a heavy-tailed t-distribution with $\nu = 5$ and $\nu = 3$ d.o.f. Note that NMSE of each estimator is now compared against the underlying population parameter $\mathbf{\Sigma}_0$ of each M-estimator. Figure 4 displays the results. RBLW had a very poor performance which is due to its strict assumption of Gaussianity. It can be noted that CV method performs similarly, but slightly worse, than RSCM or LW. This can be partially attested to poor robustness properties of cross-validation. In the case of $\nu = 3$ d.o.f., also the non-robust RSCM and LW provided large NMSE and thus all non-robust estimators are not visible in the right panel of Figure 4. This was expected since t-distribution with $\nu = 3$ d.o.f. is very heavy-tailed with non-finite kurtosis. As can be noted, the proposed robust RHub and RMVT estimators provide significantly improved performance. We can also notice that RMVT estimator that adaptively estimates the d.o.f. ν from the data is able to outperform the regularized Huber's estimator (RHub).

The right panel of Figure 3 depicts the (average) shrinkage parameter β as a function of n in the case that samples are drawn from a t-distribution with $\nu = 5$ d.o.f. As can be noted the robust shrinkage estimators (RHub and RMVT) use larger shrinkage parameter value β than the non-robust RSCM and LW estimators. Compared to RSCM and LW, the RBLW (resp. CV) is seen to overestimate (resp. underestimate) the shrinkage parameter as it obtains much larger (resp. smaller) values.

Next we investigate how the estimators are able to estimate the shape matrix, i.e., the covariance matrix up to a scale. Figure 5 displays the NMSE, $\|\hat{\mathbf{V}} - \mathbf{V}\|_{\text{F}}^2 / \|\mathbf{V}\|_{\text{F}}^2$, of different shrinkage shape matrix estimators, defined as $\hat{\mathbf{V}} = p\hat{\mathbf{\Sigma}}_\beta / \operatorname{tr}(\hat{\mathbf{\Sigma}}_\beta)$ when samples are generated from a t-distribution with $\nu = 5$ d.o.f. Note that such normalization is not necessary for CWH or RTyl since they verify $\operatorname{tr}(\hat{\mathbf{\Sigma}}) = p$ in the first place. Figure 5 illustrate both the case when correlation parameter ϱ of the AR(1) scatter matrix parameter

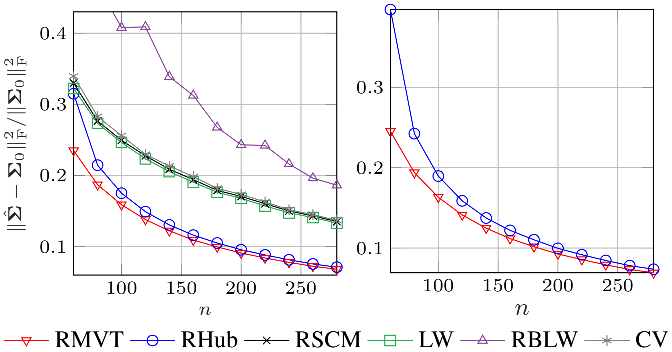


Fig. 4. NMSE as a function of n when samples are drawn from a p-variate t-distribution with $\nu = 5$ (left panel) and $\nu = 3$ (right panel) d.o.f. The scatter matrix follows an AR(1) structure; $\varrho = 0.6$ and $p = 40$.

Σ is fixed while n varies and the case that n is fixed while ϱ varies. As can be seen from the top panel of Figure 5, all robust shape estimators are performing well and very similarly. In fact, performance of RMVT and CWH is essentially the same. We can also observe that the two different approaches for shrinking Tyler's M-estimator, so CWH and the proposed RTyl are very similar. We can note from the bottom panels of Figure 5 that when $\varrho \approx 0$ (so Σ is close to a scaled identity matrix) all estimators perform similarly. This is because all estimators are shrunk heavily towards the scaled identity matrix (namely, $\beta \approx 0$ for all estimators). Similarly, when $\varrho \approx 1$ (so Σ is close to a singular matrix of rank 1), all estimators have a rather similar performance. This is because the true scatter matrix Σ is poorly conditioned ($\mathrm{cond}(\Sigma) \approx 7000$) and all estimators share similar difficulties of capturing the subspace structure due to limited training data and no *a priori* information about such structure. Indeed biggest differences between estimators are observed when Σ has no particular structure, i.e., ϱ in the range $[0.4, 0.7]$.

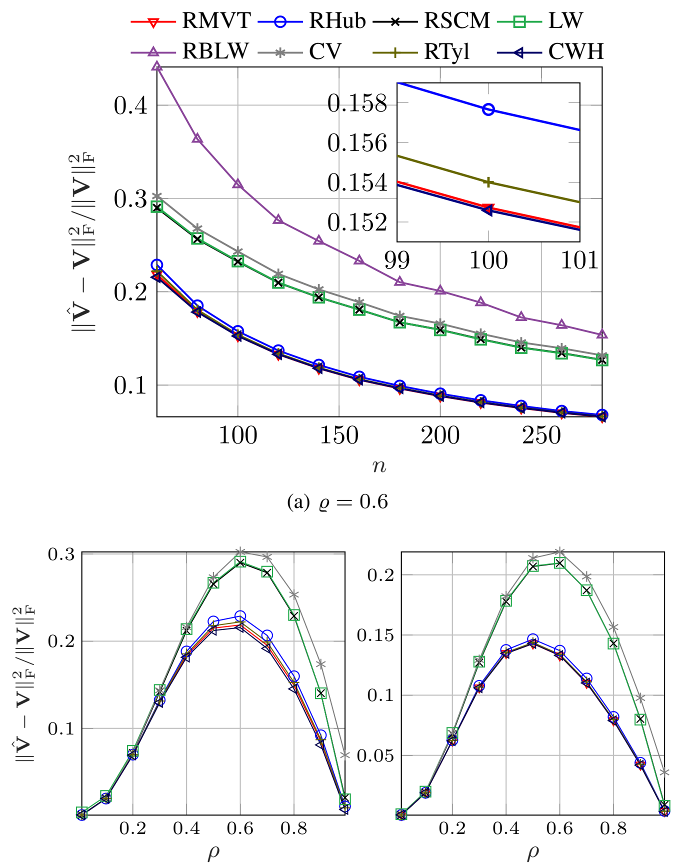
C. Complex-valued data

Finally, we note that an important property of our shrinkage method is that it can be used for complex-valued data as well. Some other methods in the previous study, such as RBLW or LW assume real-valued data. In the supplementary material, we provide the results of a simulation study in the same set-up, but now the data being generated from circular complex Gaussian and heavy-tailed t-distribution, respectively. These are distributions in the class of CES distributions. In our study we also include empirical Bayes diagonal loading estimator (EBDL) [4] which was developed for complex circular Gaussian data. Results obtained for complex-valued data attest the validity of the findings in the real-valued case.

VI. Application to Financial Data and Portfolio Design

A. Financial data

We use the S&P 500 stock market index (see Figure 6), which measures the stock performance of 500 large companies listed on stock exchanges in the United States, and its



(a) $\varrho = 0.6$

(b) $n = 60$ (c) $n = 120$

Fig. 5. NMSE of different shrinkage estimators of shape matrix $\mathbf{V}$ when samples are drawn a $p = 40$ variate t-distribution with $\nu = 5$ d.o.f. having an AR(1) structure: (a) $\varrho = 0.6$ and sample length n varies; (b) and (c) illustrate the case when ϱ varies while the sample length n is fixed.

Fig. 6. Log-prices of the S&P 500 index.

constituent stocks during the period 2016-01-01 to 2020-01-31.

We can easily observe from the returns shown in Figure 7 the effect of volatility clustering over time that is responsible for heavy tails. More concretely, if we assume that the data follows an MVT distribution, then we can compute the degrees of freedom ν on a rolling-window basis and verify that indeed the data has heavy tails with $\nu \approx 5$ (from mid-2017, ν varies between 4.5 and 6) as shown in Figure 8.

Factor model. Let $\mathbf{x}_i$ denote the returns of the p stocks at time i. It turns out that the stock returns are largely driven by

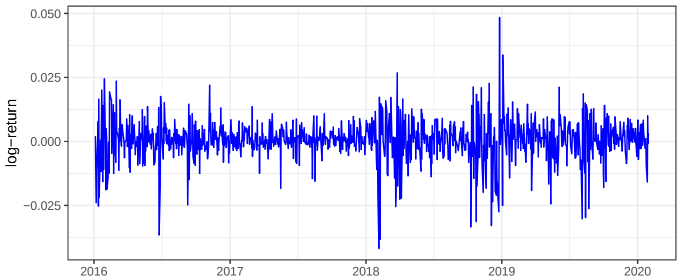


Fig. 7. Log-returns of the S&P 500 index.

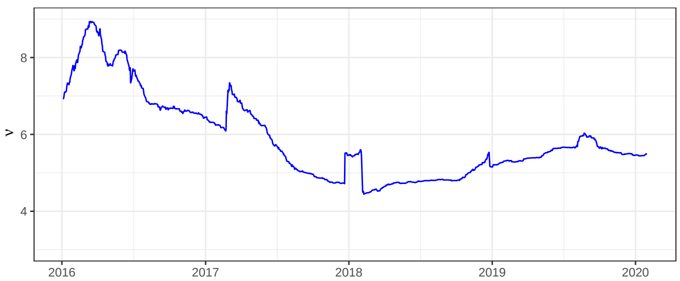


Fig. 8. Degrees of freedom ν of the S&P 500 index.

very few $k \ll p$ financial factors $\mathbf{f}_i$ as

$$\mathbf{x}_i = \mathbf{B}\mathbf{f}_i + \boldsymbol{\epsilon}_i, \tag{39}$$

where $\mathbf{B} \in \mathbb{R}^{p\times k}$ is the factor loading matrix (which is very tall since $k \ll p$) and $\boldsymbol{\epsilon}_i$ is the residual idiosyncratic component. As a consequence, the covariance matrix of these data has the form (assuming normalized factors):

$$\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^\top + \boldsymbol{\Psi} \tag{40}$$

where $\boldsymbol{\Psi}$ is the (diagonal) covariance matrix of the residuals. Typically, the term $\mathbf{B}\mathbf{B}^\top$ is much stronger than $\boldsymbol{\Psi}$ and this leads to the commonly used spike model in RMT (Random Matrix Theory) [39] which contains a few large eigenvalues and the rest small eigenvalues form the so-called bulk. Figure 9 shows the histogram of the empirical eigenvalues of the covariance matrix estimated from $p = 50$ stocks from the S&P 500 market data, where a very strong eigenvalue can be observed corresponding to the market factor.

B. Results in terms of MSE

Since our shrinkage estimators are derived to minimize the MSE in the estimation of the covariance matrix, we start by showing the obtained MSE in the context of financial data. We consider seven methods in our comparison:

- LWE is the Ledoit-Wolf estimator [1];
- RMVT-Ell1 described in subsection IV-D;
- MVT: equals RMVT-Ell1 with no shrinkage ($\beta = 1$);
- RSCM-Ell1 estimator described in subsection IV-A;
- SCM: equals RSCM-Ell1 with no shrinkage ($\beta = 1$);
- RHub-Ell1 described in subsection IV-B; and
- RSCM-Ell2, RMVT-Ell2, and RHub-Ell2 are as RSCM-Ell1, RMVT-Ell1, and RHub-Ell1, respectively, but using $\hat{\gamma}^{\mathrm{Ell2}}$ estimator of sphericity γ (*cf.* subsection III-B).

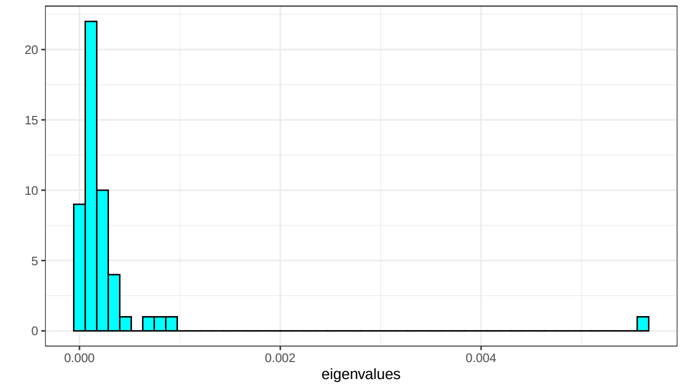


Fig. 9. Histogram of empirical eigenvalues obtained from the market data ($p = 50$) showing a strong market factor.

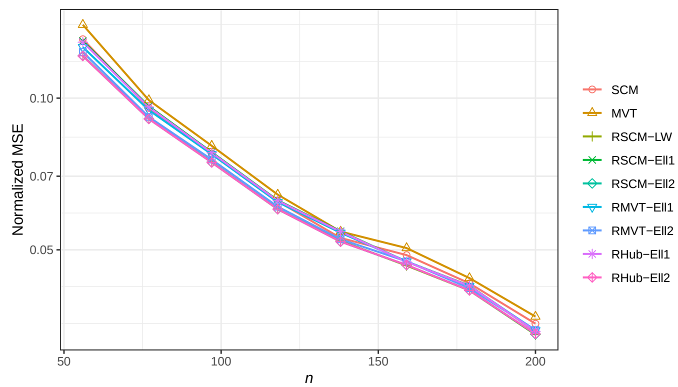


Fig. 10. Normalized MSE of covariance matrix vs number of observations for Gaussian data (with $p = 50$).

To make sure that the robust estimators do not underperform the benchmarks when the data is not heavy-tailed, we start by generating synthetic MVN data following the empirical covariance matrix previously obtained from market data (see Figure 9). Figure 10 displays the normalized MSE of the covariance matrix $\mathbb{E}\big[\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2\big]$ via 200 Monte-Carlo simulations. We do not observe any significant difference among the methods. Figure 11 shows the normalized MSE of the precision matrix $\mathbb{E}\big[\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|_{\mathrm{F}}^2\big]$ via 200 Monte-Carlo simulations. The main observation is that the two methods without shrinkage significantly underperform.

We now generate heavy-tailed synthetic data from a t-distribution following the empirical covariance matrix previously obtained from market data (see Figure 9) and with d.o.f. $\nu = 5$. Figure 12 shows the normalized MSE of the covariance matrix. We can clearly observe a striking difference showing the superiority of robust estimators (i.e., MVT, RMVT, RHub) over non-robust estimators. Among the robust estimators, the proposed shrinkage methods clearly outperform the non-shrinkage MVT in the low-sample regime, as expected. Figure 13 displays the normalized MSE of the precision matrix, with similar observations.

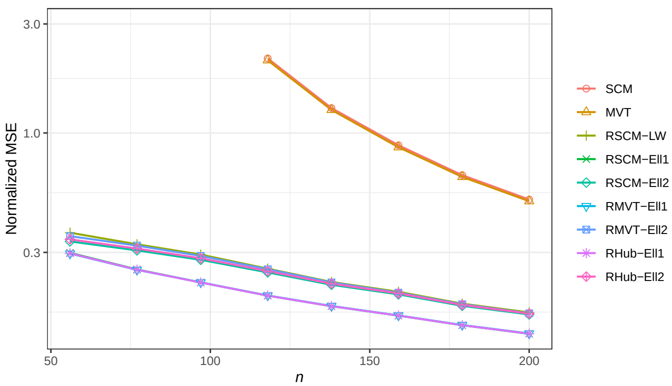


Fig. 11. Normalized MSE of precision matrix vs number of observations for Gaussian data (with $p = 50$).

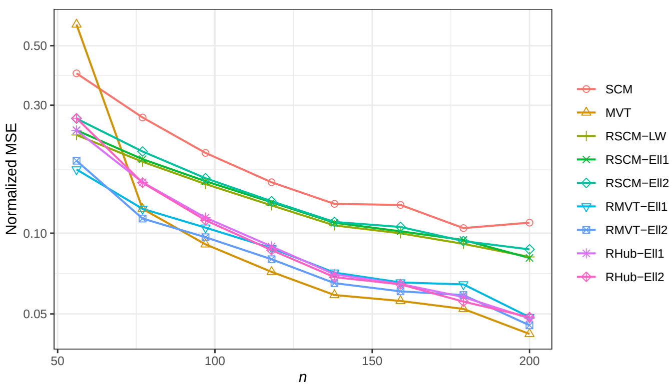


Fig. 12. Normalized MSE of covariance matrix vs number of observations for t-distributed data (with $p = 50$, $\nu = 5$).

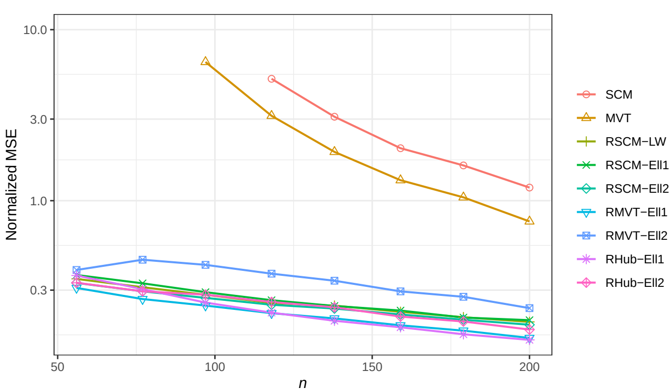


Fig. 13. Normalized MSE of precision matrix vs number of observations for t-distributed data (with $p = 50$, $\nu = 5$).

C. Results in terms of portfolio Sharpe ratio

After having shown the superiority of our proposed estimators, RMVT and RHub in terms of MSE in the estimation of the covariance matrix under heavy-tailed data, we now turn to assess the effects in terms of portfolio design. Note that an improvement of MSE in the covariance matrix may or may not translate into a significant improvement in terms of portfolio design; this depends on exactly what portfolio design is used and how it employs the estimated covariance matrix.

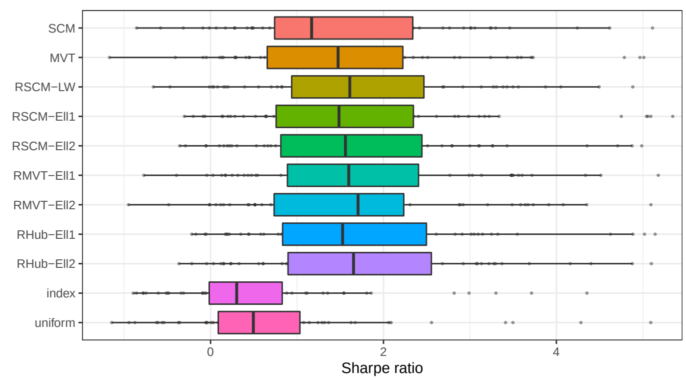


Fig. 14. Boxplot of Sharpe ratio obtained by the mean-variance portfolio according to different estimators for the covariance matrix.

For simplicity, we consider the most basic Markowitz portfolio design [40]:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} \\ \text{subject to} & \mathbf{w}^\top \boldsymbol{\mu} \geq \alpha \\ & \mathbf{1}^\top \mathbf{w} = 1, \end{array}$$

where $\boldsymbol{\mu}$ is the expected return of the returns and α is the minimum return desired for the portfolio.

We perform our backtest during the market period 2016-01-01 to 2020-01-31 on a rolling-window basis with a window length of 378 days (1.5 years). To make sure that our results are realistic, rather than performing a single backtest, we use the R package portfolioBacktest [41] to randomly select a large number of 200 datasets from the market data in the following way: each dataset chooses randomly $p = 200$ stocks from the universe of 500, as well as a random period of 504 days (2 years) among the available period from 2016-01-01 to 2020-01-31.

Figure 14 shows a boxplot with the Sharpe ratio[2] obtained mean-variance portfolio according to different estimators for the covariance matrix (along with two benchmarks: the index and the $1/N$ or uniform portfolio). We can observe that the two methods without shrinkage underperform the shrinkage methods (in particular the SCM). Among the shrinkage methods, we can see that our robust estimators slightly outperform the others, although the improvement is not extremely significative.

D. Supplementary studies

Supplementary material to this paper also contain comparison of the proposed methods in the set-up described in [2], where the global mean variance portfolio (GMVP) is used as portfolio optimization strategy and the net returns correspond to $p = 50$ stocks that are currently included in the Hang Seng Index (HSI). Compared methods include GMVP weight vector based on LW estimator and an estimator proposed in [42] that uses robust regularized Tyler's M-estimator with a

[2]The Sharpe ratio is defined as the expected return normalized with the volatility or standard deviation: $\mathsf{SR} = \frac{\mathbf{w}^\top \boldsymbol{\mu} - r_f}{\sqrt{\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}}}$, where r_f is the return of the risk-free asset.

tuning parameter selection optimized for GMVP strategy. We observed that GMVP based on the proposed RHub and RSCM are the best performing methods in terms of realized risk. We note that method of [42] was excluded in the study in subsection VI-C due to its high computational cost in the high-dimensional setting.

VII. 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 optimal one when the data is Gaussian while it outperforms shrinkage Gaussian-based methods when the data distribution turns out to be non-Gaussian. One of the benefits of the proposed adaptive shrinkage parameter selection is that it permits using real-valued or complex-valued data. Furthermore, a MATLAB toolbox called ShrinkM is freely available at http://users.spa.aalto.fi/esollila/shrinkM/ that includes functions to compute all of the proposed estimators (RHub, RTyl, RSCM, RMVT, and CV) as well as a script of one the simulation studies presented in the paper to reproduce the results. Furthermore, this paper opens several ways, notably considering the challenging cases where $p > n$ which is left for future work. Supplementary materials also provide additional examples illustrating the benefits of the proposed estimators.

Appendix

A. Proof of Theorem 1

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

$$\begin{aligned} L(\beta) &= \mathbb{E}\big[\big\|\beta\mathbf{C} + (1-\beta)p^{-1}\operatorname{tr}(\mathbf{C})\mathbf{I} - \boldsymbol{\Sigma}_0\big\|_\mathrm{F}^2\big] \\ &= \mathbb{E}\Big[\big\|\beta(\mathbf{C}-\boldsymbol{\Sigma}_0) + (1-\beta)\big(p^{-1}\operatorname{tr}(\mathbf{C})\mathbf{I} - \boldsymbol{\Sigma}_0\big)\big\|_\mathrm{F}^2\Big] \\ &= \beta^2 a_1 + (1-\beta)^2 a_2 + 2\beta(1-\beta)a_3 \end{aligned} \tag{41}$$

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

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

and $\eta_o = \operatorname{tr}(\boldsymbol{\Sigma}_0)/p$. Note that $L(\beta)$ is a convex quadratic function in β with a unique minimum given by

$$\beta_o^{\mathrm{app}} = \frac{a_2 - a_3}{(a_1 - a_3) + (a_2 - a_3)}. \tag{42}$$

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

The expression for MSE of $\mathbf{C}_{\beta_o^{\mathrm{app}}}$ then follows by substituting β_o^{app} into expression for $L(\beta)$ in (41) and using the relation, $(1-\beta_o^{\mathrm{app}})(a_2 - a_3) = \beta_o^{\mathrm{app}}(a_1 - a_3)$, which follows from (42). This yields the MSE expression

$$L(\beta_o^{\mathrm{app}}) = a_2 - \beta_o^{\mathrm{app}}(a_2 - a_3) = a_3 + (1-\beta_o^{\mathrm{app}})(a_2 - a_3).$$

This gives the stated MSE expression after noting that $a_2 - a_3 = \|\boldsymbol{\Sigma}_0 - \eta_0\mathbf{I}\|_\mathrm{F}^2$.

B. Proof of Lemma 1

First recall that $\boldsymbol{\Sigma}_0 = \sigma\boldsymbol{\Sigma}$, and hence

$$\begin{aligned} \mathbf{C} &= \frac{1}{n}\sum_{i=1}^n u(\mathbf{x}_i^\top \boldsymbol{\Sigma}_0^{-1}\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top \\ &= \boldsymbol{\Sigma}_0^{1/2}\bigg\{\frac{1}{n}\sum_{i=1}^n u\Big(\frac{r_i^2}{\sigma}\Big)\frac{r_i^2}{\sigma}\mathbf{v}_i\mathbf{v}_i^\top\bigg\}\boldsymbol{\Sigma}_0^{1/2}, \end{aligned} \tag{43}$$

where $\mathbf{v}_i = \boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i/\|\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i\|$ and $r_i^2 = \|\boldsymbol{\Sigma}^{-1/2}\mathbf{x}_i\|^2$. Recall that stochastic representation theorem of elliptical random vectors states that r_i is independent of $\mathbf{v}_i$ and $\mathbf{v}_i$-s are i.i.d. on a uniform distribution on the unit sphere $\mathcal{S}^{p-1} = \{\mathbf{v} : \mathbf{v}^\top\mathbf{v} = 1\}$. Then note that

$$\begin{aligned} &\mathbb{E}\big[\operatorname{tr}\big(\mathbf{C}^2\big)\big] \\ &= \frac{1}{n^2}\mathbb{E}\bigg[\operatorname{tr}\bigg\{\boldsymbol{\Sigma}_0^{1/2}\sum_{i=1}^n u\Big(\frac{r_i^2}{\sigma}\Big)\frac{r_i^2}{\sigma}\mathbf{v}_i\mathbf{v}_i^\top\boldsymbol{\Sigma}_0^{1/2} \\ &\qquad \cdot \boldsymbol{\Sigma}_0^{1/2}\sum_{j=1}^n u\Big(\frac{r_j^2}{\sigma}\Big)\frac{r_j^2}{\sigma}\mathbf{v}_j\mathbf{v}_j^\top\boldsymbol{\Sigma}_0^{1/2}\bigg\}\bigg] \\ &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\bigg[u\Big(\frac{r_i^2}{\sigma}\Big)\frac{r_i^2}{\sigma}u\Big(\frac{r_j^2}{\sigma}\Big)\frac{r_j^2}{\sigma}\bigg] \\ &\qquad \times \mathbb{E}\bigg[\underbrace{\operatorname{tr}\Big\{\mathbf{v}_i\mathbf{v}_i^\top\boldsymbol{\Sigma}_0\mathbf{v}_j\mathbf{v}_j^\top\boldsymbol{\Sigma}_0\Big\}}_{=(\mathbf{v}_i^\top\boldsymbol{\Sigma}_0\mathbf{v}_j)^2}\bigg] \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\bigg[u\Big(\frac{r_i^2}{\sigma}\Big)^2\frac{r_i^4}{\sigma^2}\bigg]\mathbb{E}\big[(\mathbf{v}_i^\top\boldsymbol{\Sigma}_0\mathbf{v}_i)^2\big] \\ &\qquad + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\bigg[u\Big(\frac{r_i^2}{\sigma}\Big)\frac{r_i^2}{\sigma}\bigg]\mathbb{E}\bigg[u\Big(\frac{r_j^2}{\sigma}\Big)\frac{r_j^2}{\sigma}\bigg]\mathbb{E}\big[(\mathbf{v}_i^\top\boldsymbol{\Sigma}_0\mathbf{v}_j)^2\big] \\ &= \frac{1}{n}\mathbb{E}\bigg[u\Big(\frac{r_1^2}{\sigma}\Big)^2\frac{r_1^4}{\sigma^2}\bigg]\mathbb{E}\big[(\mathbf{v}_1^\top\boldsymbol{\Sigma}_0\mathbf{v}_1)^2\big] \\ &\qquad + \Big(1-\frac{1}{n}\Big)\bigg(\mathbb{E}\bigg[u\Big(\frac{r_1^2}{\sigma}\Big)\frac{r_1^2}{\sigma}\bigg]\bigg)^2\mathbb{E}\big[(\mathbf{v}_1^\top\boldsymbol{\Sigma}_0\mathbf{v}_2)^2\big]. \end{aligned} \tag{44}$$

In the second identity, we used the fact that r_i is independent of $\mathbf{v}_i$ and $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$. In the 3rd identity we used that r_i is independent of r_j and in the 4th identity, we used that $\mathbf{v}_i$-s and r_i-s are i.i.d.

Note that $\mathbb{E}\big[u(r_1^2/\sigma)(r_1^2/\sigma)\big] = p$ due to (18) and $\mathbb{E}\big[u(r_1^2/\sigma)^2(r_1^4/\sigma^2)\big] = \psi_1 p(p+2)$. Using these facts and the following results from [8]:

$$\mathbb{E}\big[(\mathbf{v}_1^\top\boldsymbol{\Sigma}_0\mathbf{v}_1)^2\big] = \frac{2\operatorname{tr}(\boldsymbol{\Sigma}_0^2) + \operatorname{tr}(\boldsymbol{\Sigma}_0)^2}{p(p+2)} \tag{45}$$

$$\mathbb{E}\big[(\mathbf{v}_1^\top\boldsymbol{\Sigma}_0\mathbf{v}_2)^2\big] = \frac{\operatorname{tr}(\boldsymbol{\Sigma}_0^2)}{p^2}, \tag{46}$$

we get

$$\begin{aligned} &\mathbb{E}\big[\operatorname{tr}\big(\mathbf{C}^2\big)\big] \\ &= \frac{\psi_1}{n}\big(2\operatorname{tr}(\boldsymbol{\Sigma}_0^2) + \operatorname{tr}(\boldsymbol{\Sigma}_0)^2\ \big) + \Big(1-\frac{1}{n}\Big)\operatorname{tr}(\boldsymbol{\Sigma}_0^2) \\ &= \Big(1 + \frac{2\psi_1 - 1}{n}\Big)\operatorname{tr}(\boldsymbol{\Sigma}_0^2) + \frac{\psi_1}{n}\operatorname{tr}(\boldsymbol{\Sigma}_0)^2. \end{aligned}$$

Next note that

$$\operatorname{tr}(\mathbf{C})^2 = \left\{ \frac{1}{n} \sum_{i=1}^{n} u\Big(\frac{r_i^2}{\sigma}\Big) \frac{r_i^2}{\sigma} \mathbf{v}_i^\top \boldsymbol{\Sigma}_0 \mathbf{v}_i \right\}^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} u\Big(\frac{r_i^2}{\sigma}\Big) \frac{r_i^2}{\sigma} u\Big(\frac{r_j^2}{\sigma}\Big) \frac{r_j^2}{\sigma} \mathbf{v}_i^\top \boldsymbol{\Sigma}_0 \mathbf{v}_i \mathbf{v}_j^\top \boldsymbol{\Sigma}_0 \mathbf{v}_j$$

and hence

$$\begin{aligned}\mathbb{E}[\operatorname{tr}(\mathbf{C})^2] &= \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}\left[u\Big(\frac{r_i^2}{\sigma}\Big)^2 \frac{r_i^4}{\sigma^2} \right] \mathbb{E}\big[(\mathbf{v}_i^\top \boldsymbol{\Sigma}_0 \mathbf{v}_i)^2\big] \\ &\quad + \frac{1}{n^2} \sum_{i \neq j} \mathbb{E}\left[u\Big(\frac{r_i^2}{\sigma}\Big) \frac{r_i^2}{\sigma} \right] \mathbb{E}\left[u\Big(\frac{r_j^2}{\sigma}\Big) \frac{r_j^2}{\sigma} \right] \frac{\operatorname{tr}(\boldsymbol{\Sigma}_0)^2}{p^2} \\ &= \frac{\psi_1 p(p+2)}{n} \mathbb{E}\big[(\mathbf{v}_i^\top \boldsymbol{\Sigma}_0 \mathbf{v}_i)^2\big] + \Big(1 - \frac{1}{n}\Big) \operatorname{tr}(\boldsymbol{\Sigma}_0)^2.\end{aligned}$$

In the first identity we used that

$$\mathbb{E}[\mathbf{v}_i^\top \boldsymbol{\Sigma}_0 \mathbf{v}_i] = \operatorname{tr}\{\mathbb{E}[\mathbf{v}_i \mathbf{v}_i^\top] \boldsymbol{\Sigma}_0\} = \operatorname{tr}(\boldsymbol{\Sigma}_0)/p$$

as $\mathbb{E}[\mathbf{v}_i \mathbf{v}_i^\top] = (1/p)\mathbf{I}$ and the fact that r_i is independent of $\mathbf{v}_i$. In the second identity we used that samples are i.i.d. and $\mathbb{E}\big[u(r_i^2/\sigma)(r_i^2/\sigma)\big] = p$ due to (18) and $\mathbb{E}\big[u(r_i^2/\sigma)^2(r_i^4/\sigma^2)\big] = \psi_1 p(p+2)$. The result then follows by substituting (45) into the last equation.

C. Proof of Lemma 2

Using $\boldsymbol{\Sigma}_0 = \sigma \boldsymbol{\Sigma}$ and (43) one obtains as in Lemma 1 the following expression

$$\begin{aligned}&\mathbb{E}\big[\operatorname{tr}\big(\mathbf{C}^2\big)\big] \\ &= \frac{1}{n} \mathbb{E}\left[u\Big(\frac{r_1^2}{\sigma}\Big)^2 \frac{r_1^4}{\sigma^2} \right] \mathbb{E}\big[(\mathbf{v}_1^{\mathsf{H}} \boldsymbol{\Sigma}_0 \mathbf{v}_1)^2\big] \\ &\quad + \Big(1 - \frac{1}{n}\Big) \left(\mathbb{E}\left[u\Big(\frac{r_1^2}{\sigma}\Big) \frac{r_1^2}{\sigma} \right] \right)^2 \mathbb{E}\big[|\mathbf{v}_1^{\mathsf{H}} \boldsymbol{\Sigma}_0 \mathbf{v}_2|^2\big]. \quad (47)\end{aligned}$$

Using the facts that $\mathbb{E}\big[u(r_1^2/\sigma)(r^2/\sigma)\big] = p$ due to (18) and $\mathbb{E}\big[u(r^2/\sigma)^2(r^4/\sigma^2)\big] = \psi_1 p(p+1)$ along with the facts that [29, cf. eq. (66), (67)]

$$\mathbb{E}\big[(\mathbf{v}^{\mathsf{H}} \boldsymbol{\Sigma} \mathbf{v})^2\big] = [p(p+1)]^{-1} (\operatorname{tr}(\boldsymbol{\Sigma}^2) + \operatorname{tr}(\boldsymbol{\Sigma})^2), \quad (48)$$

$$\mathbb{E}\big[(\mathbf{v}_1^{\mathsf{H}} \boldsymbol{\Sigma} \mathbf{v}_2)^2\big] = p^{-2} \operatorname{tr}(\boldsymbol{\Sigma}^2), \quad (49)$$

we get

$$\begin{aligned}&\mathbb{E}\big[\operatorname{tr}\big(\mathbf{C}^2\big)\big] \\ &= \frac{\psi_1}{n} \big(\operatorname{tr}(\boldsymbol{\Sigma}_0^2) + \operatorname{tr}(\boldsymbol{\Sigma}_0)^2 \big) + \Big(1 - \frac{1}{n}\Big) \operatorname{tr}(\boldsymbol{\Sigma}_0^2) \\ &= \Big(1 + \frac{\psi_1 - 1}{n}\Big) \operatorname{tr}(\boldsymbol{\Sigma}_0^2) + \frac{\psi_1}{n} \operatorname{tr}(\boldsymbol{\Sigma}_0)^2.\end{aligned}$$

Using similar proof as in proof of Lemma 1, we obtain

$$\mathbb{E}[\operatorname{tr}(\mathbf{C})^2] = \frac{\psi_1 p(p+1)}{n} \mathbb{E}\big[(\mathbf{v}_i^{\mathsf{H}} \boldsymbol{\Sigma}_0 \mathbf{v}_i)^2\big] + \Big(1 - \frac{1}{n}\Big) \operatorname{tr}(\boldsymbol{\Sigma}_0)^2. \quad (50)$$

The result then follows by substituting the expression from (49) into the last equation.

D. Proof of Lemma 3

We show the result in the real case only as the result follows similarly for complex-valued case. Write $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}\mathbf{x}$ and note that $\mathbf{z} \sim \mathcal{E}_p(\mathbf{0}, \mathbf{I}, g)$. The result

$$\kappa = \frac{\mathbb{E}[\|\mathbf{z}\|^4]}{p(p+2)} - 1 = \frac{1}{3}\big(\mathbb{E}[z_1^4] - 3\big) \quad (51)$$

follows by recalling the stochastic decomposition. Namely, $\mathbf{z} =_d r\mathbf{v}$, where $r =_d \|\mathbf{z}\|$ is independent of $\mathbf{v}$, and $\mathbf{v}$ possesses a uniform distribution on the unit sphere $\mathcal{S}^{p-1}$. Thus we have that

$$\mathbb{E}[z_i^4] = \mathbb{E}[\|\mathbf{z}\|^4]\mathbb{E}[v_i^4] = 3\psi_1$$

where we used that $\mathbb{E}[v_i^4] = 3(p(p+2))^{-1}$ (see e.g. [43, Lemma A.1.] and that $\psi_1 = \mathbb{E}[\|\mathbf{z}\|^4]/p(p+2)$. Furthermore, since $\mathbb{E}[z_i^2] = 1$, (51) states that $\kappa = (1/3)\mathsf{kurt}(z_i)$. Then since $\mathsf{kurt}(z_i) = \mathsf{kurt}(x_i)$, we have the stated result that $\kappa = (1/3)\mathsf{kurt}(x_i)$.

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, May 2019.

[3] J. Li, J. Zhou, and B. Zhang, "Estimation of large covariance matrices by shrinking to structured target in normal and non-normal distributions," *IEEE Access*, vol. 6, pp. 2158–2169, 2017.

[4] A. Coluccia, "Regularized covariance matrix estimation via empirical bayes," *IEEE Signal Process. Lett.*, vol. 22, no. 11, pp. 2127–2131, 2015.

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

[6] 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.

[7] Y. I. Abramovich and N. K. Spencer, "Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering," in *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 15–20 2007, pp. 1105–1108.

[8] 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.

[9] 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, no. 1, pp. 449–458, 2010.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] M. Yi and D. E. Tyler, "Shrinking the covariance matrix using convex penalties on the matrix-log transformation," *Journal of Computational and Graphical Statistics*, no. just-accepted, pp. 1–10, 2020.

[16] E. Ollila, D. P. Palomar, and F. Pascal, "M-estimators of scatter with eigenvalue shrinkage," in *IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 4–8 2020, pp. 5305–5309.

[17] A. Aubry, A. De Maio, L. Pallotta, and A. Farina, "Radar detection of distributed targets in homogeneous interference whose inverse covariance structure is defined via unitary invariant functions," *IEEE Trans. Signal Process.*, vol. 61, no. 20, pp. 4949–4961, 2013.

[18] Y. Sun, A. Breloy, P. Babu, D. P. Palomar, F. Pascal, and G. Ginolhac, "Low-complexity algorithms for low rank clutter parameters estimation in radar systems," *IEEE Trans. Signal Process.*, vol. 64, no. 8, pp. 1986–1998, 2016.
[19] A. Aubry, A. De Maio, and L. Pallotta, "A geometric approach to covariance matrix estimation and its applications to radar problems," *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 907–922, 2018.
[20] A. De Maio, L. Pallotta, J. Li, and P. Stoica, "Loading factor estimation under affine constraints on the covariance eigenvalues with application to radar target detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 3, pp. 1269–1283, 2019.
[21] M. Tang, Y. Rong, X. R. Li, and J. Zhou, "Invariance theory for adaptive detection in non-gaussian clutter," *IEEE Trans. Signal Process.*, vol. 68, pp. 2045–2060, 2020.
[22] O. Besson, "Maximum likelihood covariance matrix estimation from two possibly mismatched data sets," *Signal Processing*, vol. 167, p. 107285, 2020.
[23] Y. Abramovich and O. Besson, "Regularized covariance matrix estimation in complex elliptically symmetric distributions using the expected likelihood approach-part 1: The over-sampled case," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5807–5818, 2013.
[24] O. Besson and Y. Abramovich, "Regularized covariance matrix estimation in complex elliptically symmetric distributions using the expected likelihood approach-part 2: The under-sampled case," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5819–5829, 2013.
[25] Y. I. Abramovich and O. Besson, "On the expected likelihood approach for assessment of regularization covariance matrix," *IEEE Signal Process. Lett.*, vol. 22, no. 6, pp. 777–781, 2015.
[26] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *Ann. Stat.*, vol. 15, no. 1, pp. 234–251, 1987.
[27] ——, "A note on affine equivariant location and scatter statistics for sparse data," *Statist. and Prob. Letters*, vol. 80, pp. 1409–1413, 2010.
[28] R. Couillet, F. Pascal, and J. W. Silverstein, "The Random Matrix Regime of Maronna's M-estimator with elliptically distributed samples," *J. Mult. Anal.*, vol. 139, pp. 56–78, July 2015. arXiv:1311.7034.
[29] 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.
[30] O. Ledoit, M. Wolf *et al.*, "Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size," *Ann. Stat.*, vol. 30, no. 4, pp. 1081–1102, 2002.
[31] M. S. Srivastava, "Some Tests Concerning the Covariance Matrix in High Dimensional Data," *Journal of the Japan Statistical Society*, vol. 35, no. 2, pp. 251–272, 2005.
[32] K.-T. Fang, S. Kotz, and K.-W. Ng, *Symmetric Multivariate and Related Distributions*. London: Chapman and hall, 1990.
[33] 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.
[34] K. Ashurbekova, A. Usseglio-Carleve, F. Forbes, and S. Achard, "Optimal shrinkage for robust covariance matrix estimators in a small sample size setting," *hal-02378034v2*, 2020.
[35] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robust Statistics for Signal Processing*. Cambridge, UK: Cambride University Press, Nov. 2018.
[36] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. New York: Wiley, 1982, 704 pages.
[37] S. T. Smith, "Covariance, subspace, and intrinsic Cramér-Rao bounds," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1610–1630, 2005.
[38] A. Breloy, G. Ginolhac, A. Renaux, and F. Bouchard, "Intrinsic Cramér-Rao bounds for scatter and shape matrices estimation in CES distributions," *IEEE Signal Process. Lett.*, vol. 26, no. 2, pp. 262–266, 2019.
[39] J. Bun, J.-P. Bouchaud, and M. Potters, "Cleaning correlation matrices," *Risk Management*, 2016.
[40] H. Markowitz, "Portfolio selection," *J. Financ.*, vol. 7, no. 1, pp. 77–91, 1952.
[41] D. P. Palomar and R. Zhou, *portfolioBacktest: Automated Backtesting of Portfolios over Multiple Datasets*, 2019, r package version 0.2.1. [Online]. Available: https://CRAN.R-project.org/package=portfolioBacktest
[42] L. Yang, R. Couillet, and M. R. McKay, "A robust statistics approach to minimum variance portfolio optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 24, pp. 6684–6697, 2015.
[43] E. Ollila, H. Oja, and C. Croux, "The affine equivariant sign covariance matrix: asymptotic behavior and efficiencies," *J. Mult. Anal.*, vol. 87, no. 2, pp. 328–355, 2003.