Improved estimation of the degree of freedom parameter of multivariate t-distribution

Frédéric Pascal
Université Paris-Saclay, CNRS, CentraleSupélec
Laboratoire des signaux et systèmes
91190, Gif-sur-Yvette, France
frederic.pascal@centralesupelec.fr

Esa Ollila
Dept of Signal Processing and Acoustics
Aalto University
Helsinki, Finland
esa.ollila@aalto.fi

Daniel P. Palomar
The Hong Kong Univ.
of Science and Technology
Hong Kong
palomar@ust.hk

***Abstract*—The multivariate t (MVT)-distribution is a widely used statistical model in various application domains, mainly due to its adaptability to heavy-tailed data. However, estimating the degree of freedom (d.o.f) parameter, that controls the shape of the distribution, remains a challenging problem. In this work, we develop a novel methodology and design various algorithms for estimating the d.o.f parameter. More precisely, based on a key relationship between scatter and covariance matrices for the t-distribution, the estimator is derived from the expectation of a particular quadratic form and is proved to converge although the classical independence assumption is not fulfilled. finally, some preliminary simulations show the improvement of the proposed approach with respect to state-of-the-art methods.**

***Index Terms*—Multivariate t-distribution, M-estimators, Mahalanobis distance.**

I. Introduction

In many applications, the covariance matrix is a key parameter for data processing (*e.g.,* for dimension reduction, detection, clustering/classification). Since this parameter is unknown in practice, an estimator is required. On the other hand, the standard Gaussian assumption is not always adapted due to various phenomena: data heterogeneity, presence of outliers, etc. This is the case for instance in radar signal processing or in financial data [1]. In this work, we assume that the number n of observations is greater than each observation dimension p, *i.e.,* $n > p$ and, we consider an M-estimator of scatter matrix [2], adapted to various statistical models, possibly far from the Gaussian one.

We assume that the data set $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is an i.i.d. sample from a p-variate zero-mean *multivariate t(MVT)-distribution*, with $\nu > 0$ degrees of freedom (d.o.f.). The MVT distribution is a sub-class of Elliptically Symmetric (ES) distribution; see [3], [4]. The probability distribution function (pdf) of a zero-mean MVT distribution, denoted $\mathbf{x} \sim t_\nu(\mathbf{0}, \mathbf{\Sigma})$, is

$$f(\mathbf{x}) = C|\mathbf{\Sigma}|^{-1/2}\Big(1 + \frac{\mathbf{x}^\top \mathbf{\Sigma}^{-1}\mathbf{x}}{\nu}\Big)^{-(p+\nu)/2}, \qquad (1)$$

The work of F. Pascal has been partially supported by DGA under grant ANR-17-ASTR-0015. The work of D. P. Palomar was partially supported by the Hong Kong GRF 16207019 research grant.

where $\mathbf{\Sigma}$ is a positive definite symmetric matrix parameter, called the scatter matrix, and C is a normalizing constant ensuring that $f(\mathbf{x})$ integrates to 1. In this paper, both $\mathbf{\Sigma}$ and $\nu > 0$ are unknown parameters that need to be estimated. The d.o.f. parameter $\nu > 0$ determines the shape of the distribution. For $\nu \to \infty$, the pdf reduces to multivariate normal distribution, while $\nu = 1$ corresponds to multivariate Cauchy distribution.

For a given fixed value of ν, an estimate $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ can be found by maximizing the log-likelihood function of the data w.r.t. to $\mathbf{\Sigma}$. This leads to solving the following estimating equation

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \frac{p+\nu}{\nu + \mathbf{x}_i^\top \hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i}\mathbf{x}_i\mathbf{x}_i^\top. \qquad (2)$$

Note that $\hat{\mathbf{\Sigma}}$ depends on ν. When ν corresponds to the true value of the d.o.f. parameter of the data, then $\hat{\mathbf{\Sigma}}$ is the Maximum Likelihood estimator (MLE) of $\mathbf{\Sigma}$. When ν is misspecified, $\hat{\mathbf{\Sigma}}$ is an M-estimator of the scatter matrix $\mathbf{\Sigma}$. For a comprehensive study of M-estimators and the ML-estimation of t-distribution, see [5].

Since the t-distribution is a compound Gaussian distribution (i.e., a scale mixture of Gaussian), it has a stochastic decomposition

$$\mathbf{x}_i =_d \tau_i\, \mathbf{z}_i, \qquad (3)$$

where for $i = 1, \ldots, n$, $\mathbf{z}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ and τ_i positive random variable, with $\tau_i^{-1} \sim Gam(\nu/2, 2/\nu)$, independent of $\mathbf{z}_i$. Notations $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ and $Gam(a, b)$ respectively stand for the multivariate Normal distribution with zero-mean and covariance matrix $\mathbf{\Sigma}$, and the Gamma distribution with shape parameter a and scale parameter b. As in [6], we call $\mathbf{z}_i$-s as Gaussian cores. The scatter matrix parameter $\mathbf{\Sigma}$ is a scaled copy of the covariance matrix $\mathbf{R} = \mathbb{E}[\mathbf{x}_i\mathbf{x}_i^\top]$, namely,

$$\mathbf{R} = \theta\mathbf{\Sigma} \qquad (4)$$

where

$$\theta = \frac{\mathbb{E}[r^2]}{p} = \frac{\nu}{\nu - 2} \tag{5}$$

and $r^2 = \|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|^2$ is the squared Mahalanobis distance of $\mathbf{x}$ w.r.t. $\mathbf{\Sigma}$. Note that (3) implies that $r^2 =_d \tau^2\|\mathbf{\Sigma}^{-1/2}\mathbf{z}_i\|^2$, where $\|\mathbf{\Sigma}^{-1/2}\mathbf{z}_i\|^2 \sim \chi^2_p$. Thus the expected value of r^2 is just the product of $\mathbb{E}[\tau^2] \cdot p$ since $\mathbb{E}[\chi^2_p] = p$. This yields then (5) by using that $\tau^{-1} \sim Gam(\nu/2, 2/\nu)$.

It is worthwhile to point out that if the interest is to estimate the covariance matrix $\mathbf{R}$, then one is forced to estimate both ν and $\mathbf{\Sigma}$ simultaneously. Moreover, in most situation, the d.o.f. parameter ν is unknown and needs to estimated. The quality of the estimate of $\mathbf{\Sigma}$ depends really on how accurately we are able to estimate ν. This is the main purpose of this work.

The paper is organized as follows: Section II presents the background on MVT distributions as well as existing methods for estimating ν. Then, Section IV provides the main contribution of this work by deriving an improved estimator of ν and the associated algorithm. Finally, Section V highlights the interest of the proposed approach on simulated data while Section VI draws some conclusions and perspectives.

II. Alternative estimators of ν

Recently, in [7], we proposed a method to estimate ν which we describe below. Denote $\eta = \mathrm{tr}(\mathbf{\Sigma})/p$ and let

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top.$$

denote the sample covariance matrix (SCM) based on the data. Then by (4) and (5), one has that $\mathrm{tr}(\mathbf{R})/p = (\nu/(\nu - 2))\,\mathrm{tr}(\mathbf{\Sigma})/p$. This means that

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

from which we obtain the relation

$$\nu = \frac{2\eta_{\text{ratio}}}{\eta_{\text{ratio}} - 1}. \tag{6}$$

Then given that one has an estimate $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ found by solving Eq. (2) with the current guess of d.o.f. parameter ν, one may compute an estimate $\hat{\eta}_{\text{ratio}} = \mathrm{tr}(\mathbf{S})/\,\mathrm{tr}(\hat{\mathbf{\Sigma}})$ which provides an estimate, denoted $\hat{\nu}^{(0)}$ via (6). This idea gives rise to an iterative algorithm to estimate ν detailed in [7]. The initial estimate is $\hat{\nu}_K = 2/\max(0, \hat{\kappa}) + 4$, where $\hat{\kappa}$ is the estimate of elliptical kurtosis (see [8]).

Other estimators for ν have been proposed in the literature. One can cite for instance, the ML estimation of $\hat{\nu}$ via the Expectation-Maximization (EM) approach [9]. Unfortunately, this estimator is rather unstable [10]. Another estimator based on the Hill estimator [11] has been recently proposed (see Eq. (26) of [12]).

III. Preliminaries

Given ν is known, it has been shown in [6] that the MLE $\hat{\mathbf{\Sigma}}$ has the same properties than a *Gaussian-Core Wishart Equivalent*, defined as the SCM built from the Gaussian cores $\mathbf{z}_1, \ldots, \mathbf{z}_n$, i.e.,

$$\hat{\mathbf{\Sigma}}_{\text{GCWE}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i^\top \tag{7}$$

where the $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are the Gaussian-distributed vectors in the stochastic decomposition of Eq. (3). This estimator is referred to as the GCWE of $\hat{\mathbf{\Sigma}}$. It is important to notice that this matrix cannot be computed in practice (since $\mathbf{z}_i$-s are unobserved), but represents a theoretical equivalent.

Then, equipped with this new concept, one has that (see real case of Theorem III.1 of [6] for details)

$$\hat{\mathbf{\Sigma}} \underset{app}{\sim} \sigma\hat{\mathbf{\Sigma}}_{\text{GCWE}} \tag{8}$$

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

$$\mathbb{E}\left[\psi\left(\frac{r^2}{\sigma}\right)\right] = p, \tag{9}$$

where $\psi(t) = t(p+\nu)/(\nu+t)$ and $r^2 = \|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|^2$ as earlier, Note when ν is known, then $\sigma = 1$.

Then using the Shermon-Morrison formula[1], it holds that

$$\frac{\tilde{r}_i^2}{p} = \frac{\mathbf{z}_i^\top\hat{\mathbf{\Sigma}}_{\text{GCWE}}^{-1}\mathbf{z}_i}{p} = \frac{1}{p}\frac{\mathbf{z}_i^\top\hat{\mathbf{\Sigma}}_{\text{GCWE},(i)}^{-1}\mathbf{z}_i}{1 + \frac{1}{n}\mathbf{z}_i^\top\hat{\mathbf{\Sigma}}_{\text{GCWE},(i)}^{-1}\mathbf{z}_i},$$

where

$$\hat{\mathbf{\Sigma}}_{\text{GCWE},(i)} = \hat{\mathbf{\Sigma}}_{\text{GCWE}} - \frac{1}{n}\mathbf{z}_i\mathbf{z}_i^\top = \frac{1}{n}\sum_{\substack{k=1\\k\neq i}}^{n} \mathbf{z}_k\,\mathbf{z}_k^\top, \tag{10}$$

implying that $\hat{\mathbf{\Sigma}}_{\text{GCWE},(i)}$ is independent of $\mathbf{z}_i$.

The most important consequence of (8) is that the statistical properties of $\hat{\mathbf{\Sigma}}$ are very well approximated by the $\hat{\mathbf{\Sigma}}_{\text{GCWE}}$, that follows a Wishart distribution since $\mathbf{z}_1, \ldots, \mathbf{z}_n$ is an i.i.d. sample from $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. Consequently, one will study

$$\tilde{r}_i^2 = \mathbf{x}_i^\top\hat{\mathbf{\Sigma}}_{\text{GCWE}}^{-1}\mathbf{x}_i$$

instead of $\hat{r}_i^2 = \mathbf{x}_i^\top\hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_i$ in the following. See also [6] for equivalent Mahalanobis distances. We now prove a result that will be used in the sequel.

[1]For any invertible square matrix $\mathbf{A}$, p-vector $\mathbf{z}$, and positive scalar τ, it holds that $(\mathbf{A} + \tau\mathbf{z}\mathbf{z}^\top)^{-1}\mathbf{z} = \mathbf{A}^{-1}\mathbf{z}/(1 + \tau\mathbf{z}^\top\mathbf{A}^{-1}\mathbf{z})$.

Proposition 1. *Given* $\{\mathbf{z}_i\}_{i=1}^n \overset{iid}{\sim} \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$*, one has that*

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{z}_i^\top \hat{\mathbf{\Sigma}}_{\mathrm{GCWE},(i)}^{-1} \mathbf{z}_i}{p} \xrightarrow[n\to\infty]{a.s.} 1 \left(= \frac{\mathbb{E}[\mathbf{z}_i^\top \mathbf{\Sigma}^{-1} \mathbf{z}_i]}{p} \right).$$

where $\hat{\mathbf{\Sigma}}_{\mathrm{GCWE},(i)}$ *is defined in* (10)*.*

Proof. Proof is detailed in Appendix A. □

IV. An improved estimate of ν

In this section, we derive a new estimator, denoted $\hat{\nu}_{POP}$, of the d.o.f. parameter ν.

From (5) one obtains the following key equation

$$\nu = \frac{2\,\theta}{\theta - 1} \tag{11}$$

Consequently, estimating ν will rely on the capability to estimate the quantity $\theta = \mathbb{E}\big[r^2\big]/p$, that is the expectation of the Mahalanobis distance with $\boldsymbol{\mu} = \mathbf{0}$. However, $\mathbf{\Sigma}$ is unknown. We propose a strategy to estimate θ that is well-suited to the case of high-dimensional settings, namely $n, p \to \infty$ and $p/n \to c \in [0, 1)$, referred to as *RMT regime*.

Proposition 2. *Suppose* $\{\mathbf{x}_i\}_{i=1}^n \overset{iid}{\sim} t_\nu(\mathbf{0}, \mathbf{\Sigma})$*, where d.o.f. parameter* $\nu > 0$ *is known and* $\mathbf{\Sigma}$ *is unknown. Then* $\hat{\theta}$ *below is a consistent estimator of* θ*:*

$$\hat{\theta} = (1 - p/n)\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}_{(i)}^{-1} \mathbf{x}_i}{p}, \tag{12}$$

where $\hat{\mathbf{\Sigma}}_{(i)}$ *is the MLE defined in equation* (2) *computed with all observations except* $\mathbf{x}_i$*, i.e.,*

$$\hat{\mathbf{\Sigma}}_{(i)} = \frac{1}{n}\sum_{j\neq i} \frac{p+\nu}{\nu + \mathbf{x}_j^\top \hat{\mathbf{\Sigma}}_{(i)}^{-1} \mathbf{x}_j} \mathbf{x}_j \mathbf{x}_j^\top. \tag{13}$$

Proof. Let us consider the quantity $\tilde{r}_{(i)}^2 = \mathbf{x}_i^\top \hat{\mathbf{\Sigma}}_{\mathrm{GCWE}}^{-1} \mathbf{x}_i$. In a RMT regime, it is well-known that $\tilde{r}_{(i)}^2 \xrightarrow[\substack{n,p\to\infty \\ p/n\to c}]{a.s.} \frac{1}{1-c}$ (see *e.g.*, [13]).

Combining this result and the one of Proposition 1, a good candidate for estimating $\mathbb{E}[r_i^2]/p$ is

$$\hat{\theta} = (1 - p/n)\sum_{i=1}^{n} \hat{r}_{(i)}^2$$

where $\hat{r}_{(i)}^2$ is built with the t-MLE, computed from the samples $\{\mathbf{x}_j\}_{j\neq i}$ (without the i^{th} observation). Thus, one obtains:

$$\hat{\nu}_{POP} = \frac{2\,(1 - p/n)\sum_{i=1}^{n} \hat{r}_{(i)}^2}{(1 - p/n)\sum_{i=1}^{n} \hat{r}_{(i)}^2 - 1}$$

□

Thus, if ν is misspecified but close to the true d.o.f. parameter, then Proposition 2 and relation (11), allows us to obtain an improved estimate of ν as follows:

$$\hat{\nu} = \frac{2\,\hat{\theta}}{\hat{\theta} - 1}.$$

This property will be then used similarly as in our eatlier work [7] to iteratively estimate ν. Details of this estimator computation are contained in Algorithm 1. Then, once $\hat{\nu}_{POP}$ is computed, we use it to estimate the scatter matrix $\mathbf{\Sigma}$ by solving (2) with obtained estimate $\nu = \hat{\nu}_{POP}$.

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

Input : data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^p$

Initialize: Compute $\nu_0 = 2/\max(0, \hat{\kappa}) + 4$, where $\hat{\kappa}$ is an estimate of κ explained in [8].

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

Set $\hat{\mathbf{\Sigma}}_{(i),\mathtt{t}}$ as the MLE of $\mathbf{\Sigma}$ based on current estimate of d.o.f. parameter $\nu = \nu_\mathtt{t}$ computed when removing the i^{th} observation:

$$\hat{\mathbf{\Sigma}}_{(i),\mathtt{t}} = \frac{1}{n}\sum_{j\neq i} \frac{p+\nu_\mathtt{t}}{\nu_\mathtt{t} + \mathbf{x}_j^\top \hat{\mathbf{\Sigma}}_{(i),\mathtt{t}}^{-1} \mathbf{x}_j} \mathbf{x}_j \mathbf{x}_j^\top.$$

Update $\hat{\theta}_\mathtt{t} = (1 - p/n)\,\frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{x}_i^\top \hat{\mathbf{\Sigma}}_{(i),\mathtt{t}}^{-1} \mathbf{x}_i}{p}$.

Upate the d.o.f. parameter $\nu_{\mathtt{t}+1} = \frac{2\theta_\mathtt{t}}{\theta_\mathtt{t} - 1}$.

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

break

Output : $\hat{\nu}_{POP} = \nu_{\mathtt{t}+1}^{(k)}$

V. Experiments

Figures 1 and 2 illustrate the interest of the proposed methods for estimating ν, outperforming the others approaches in all settings. The proposed estimator $\hat{\nu}_{POP}$ is compared to the Kurtosis-based estimator $\hat{\nu}_K$ [8] and to the recent $\hat{\nu}^{(0)}$ of [7]. The simulations settings are as follows: $\nu = 3, 10$, $p = 10, 30, 40$.

More precisely, in the very heavy-tailed scenario, $\nu = 3$, on can see in Fig. 1 that the standard method based on the Kurtosis is not appropriate. Moreover, the proposed method strongly outperforms the one proposed in [7] that was designed for heavy-tailed distributions. Note that the improvement is even bigger for small p, small n. This is expected since the proposed estimator has been designed under the RMT regime. Moreover, even when n increases, one can notice in the zoom ($n = 150$ to 300) that $\hat{\nu}_{POP}$ performance are still better than $\hat{\nu}^{(0)}$ ones. Another remark is that increasing the dimension p (from 10 to 30) improves the

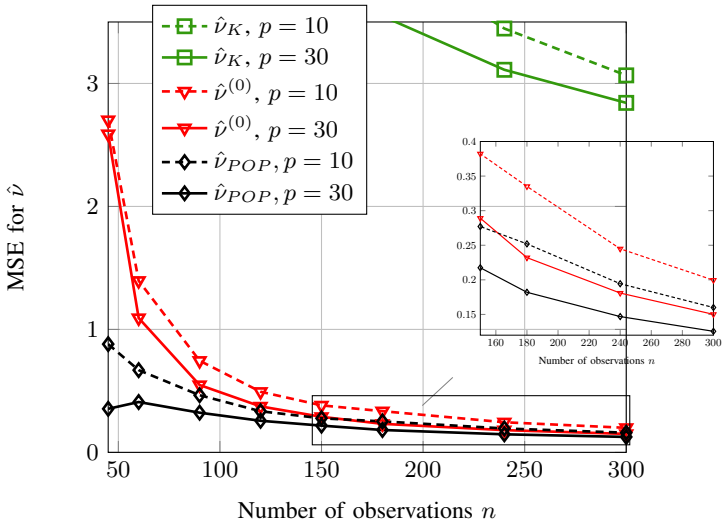

Fig. 1: Mean Square Errors of $\hat{\nu}$ versus the number n of observations by running the different algorithms until convergence (MAX 5 iterations), using initial value ν_0. The samples are generated from a p-variate real $t_3(\mathbf{0}, \mathbf{\Sigma})$-distribution, where $\mathbf{\Sigma}$ has an AR(1) covariance matrix structure with $\varrho = 0.6$ and $p = 10$ (dashed curves) and 30 (plain curves); 5000 Monte Carlo runs.

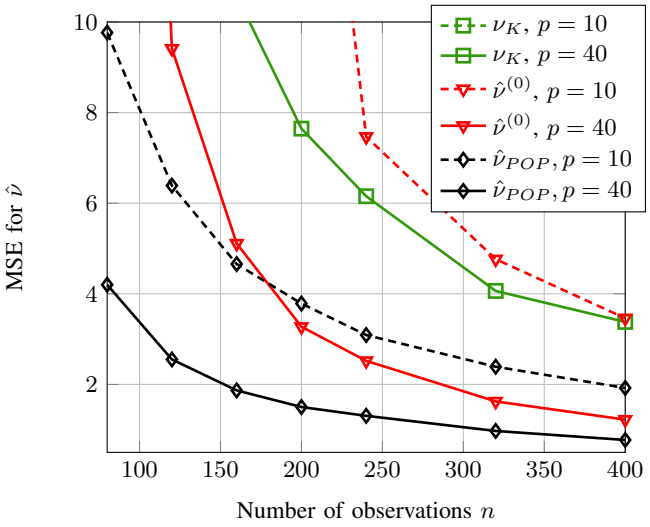

Fig. 2: Mean Square Errors of $\hat{\nu}$ versus the number n of observations by running the different algorithms until convergence (MAX 5 iterations), using initial value ν_0. The samples are generated from a p-variate real $t_{10}(\mathbf{0}, \mathbf{\Sigma})$-distribution, where $\mathbf{\Sigma}$ has an AR(1) covariance matrix structure with $\varrho = 0.6$; $p = 10$ (dashed curves) and 40 (plain curves); 5000 Monte Carlo runs.

estimation accuracy for all methods, which is expected since the methods rely on quadratic forms.

When the d.o.f. is bigger (ν=10), meaning the data distribution is closer to the Gaussian one, one can see in Fig. 2 that the Kurtosis-based estimator improves while the performance of $\hat{\nu}^{(0)}$ decreases. Again, the proposed approach strongly outperforms other methods, obtaining very good performance for any number n of data. Finally, one observes the same behavior when p increases from 10 to 40.

A general conclusion is that $\hat{\nu}_{POP}$ leads to the best estimation performance, with a significant improvement when both n and p are small (let's say of same order).

VI. Concluding remarks

In this paper, we have developed an improved estimator of the d.o.f. parameter for multivariate t-distributed observations, together with the associated recursive algorithm. This estimator is based on relationship between the d.o.f. parameter and the expectation of a Mahalanobis distance. The MVT distribution fits various heavy-tailed distributions thanks to this d.o.f. We have shown on simulations that the proposed estimator outperforms state-of-the-art estimators in all scenarios from very heavy-tailed distributions to the ones close to the Gaussian distribution. Future works will focus on statistical analysis of this estimator as well as experimentations on real data in finance.

Appendix A
Appendix: Proof of Proposition 1

Proof. For brevity, we use shorthands $\mathbf{S}_{(i)} = \hat{\mathbf{\Sigma}}_{\mathrm{GCWE},(i)}$ and $\mathbf{S} = \hat{\mathbf{\Sigma}}_{\mathrm{GCWE}}$. For simplicity, let us first define $d_i = \mathbf{z}_i^\top \mathbf{S}^{-1} \mathbf{z}_i$ and $d_{(i)} = \mathbf{z}_i^\top \mathbf{S}_{(i)}^{-1} \mathbf{z}_i$. To prove the result of Proposition 1, one cannot simply rely on the law of large numbers since the $d_{(i)}$'s are by definition not independent.

Let us consider $a_n = \left\| \frac{1}{np} \sum_{i=1}^n d_{(i)} - 1 \right\|$ which can be rewritten as

$$\left\| \frac{1}{np} \sum_{i=1}^n \left(d_{(i)} - d_i \right) + \frac{1}{n} \sum_{i=1}^n d_i/p - 1 \right\| \leq \frac{1}{np} \sum_{i=1}^n \underbrace{\| d_{(i)} - d_i \|}_{=\|b_i\|} + \underbrace{\left\| \frac{1}{n} \sum_{i=1}^n d_i/p - 1 \right\|}_{=\|c_i\|} \quad (14)$$

First, the following lemma leads to $c_i = 0$.

Lemma 1.

$$\frac{1}{n} \sum_{i=1}^n d_i = p.$$

Proof.

$$\frac{1}{n} \sum_{i=1}^n d_i = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i^T \mathbf{S}^{-1} \mathbf{z}_i = \frac{1}{n} \sum_{i=1}^n \mathrm{tr} \left(\mathbf{S}^{-1} \mathbf{z}_i \mathbf{z}_i^T \right) = \mathrm{tr} \left(\mathbf{S}^{-1} \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^T \right) = \mathrm{tr}(\mathbf{I}) = p.$$

□

Let us now focus on the first term b_i. Let us remind that from the Woodbury equality, on has:

$$d_i = \frac{d_{(i)}}{1 + \frac{1}{n}\, d_{(i)}} \Leftrightarrow d_{(i)} = \frac{1}{1 - \frac{1}{n}\, d_i}. \tag{15}$$

leading to

$$b_i = \frac{\frac{1}{n}\, d_{(i)}^2}{1 + \frac{1}{n}\, d_{(i)}} \leq \frac{1}{n} d_{(i)}^2,$$

where the denominator is always greater than 1 since $d_{(i)}$ is positive as a quadratic form.

Now, one has to study the positive quantity $\frac{1}{n}\sum_{i=1}^{n} \frac{d_{(i)}^2}{n}$, since p is assumed to be fixed. Thus, let us consider $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{d_{(i)}^2}{n} > \delta\right)$, for any $\delta > 0$. One has,

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{d_{(i)}^2}{n} > \delta\right) &= \mathbb{P}\left(\sum_{i=1}^{n} \frac{d_{(i)}^2}{n} > n\,\delta\right) \\
&\leq \mathbb{P}\left(\bigcup_{i=1}^{n} \left\{\frac{d_{(i)}^2}{n} > \delta\right\}\right) \\
&\leq n\,\mathbb{P}\left(\frac{d_{(i)}^2}{n} > \delta\right) \\
&\leq n\,\mathbb{P}\left(d_{(i)} > \sqrt{n}\,\delta\right).
\end{aligned}$$

The first inequality arises from the fact that at least one event $\left\{\frac{d_{(i)}^2}{n} > \delta\right\}$ must occur for ensuring $\sum_{i=1}^{n} \frac{d_{(i)}^2}{n} > n\,\delta$. Now, remind that

$$d_{(i)} \sim \frac{n\,p}{(n-p)} F(p, n-p).$$

This implies that $\mathbb{E}\left[d_{(i)}^3\right] = \left(\frac{n\,p}{n-p}\right)^3 \beta_{n,p}$, where $\beta_{n,p}$ can be computed from the skewness of a Fisher distribution as follows

$$\beta_{n,p} = s_3\,\sigma^3 + 3\,\mu\,\sigma^2 + \mu^3,$$

where μ is the mean, σ^2 the variance and s_3 the skewness of a Fisher distribution with p and $n-p$ degrees of freedom. One has

$$\begin{aligned}
\mu &= \frac{n-p}{n-p-2}, \\
\sigma^2 &= \frac{2(n-p)^2(n-2)}{p(n-p-2)^2(n-p-4)}, \\
s_3 &= \frac{(n+p-2)\sqrt{8(n-p-4)}}{(n-p-6)\sqrt{p(n-2)}}.
\end{aligned}$$

After simplification, one obtains

$$\mathbb{E}\left[d_{(i)}^3\right] = \left(\frac{n\,p}{n-p-2}\right)^3 \left(\frac{8(n-2)(n-p-2)}{p^2(n-p-6)(n-p-4)} + \frac{3(n-2)}{p(n-p-4)} + 1\right).$$

Now using Chebyshev's inequality, one has

$$\mathbb{P}\left(d_{(i)} > \sqrt{n}\,\delta\right) \leq \frac{\mathbb{E}\left[d_{(i)}^3\right]}{n^{3/2}\,\delta^3},$$

leading to

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{d_{(i)}^2}{n} > \delta\right) \leq \frac{\mathbb{E}\left[d_{(i)}^3\right]}{n^{1/2}\,\delta^3},\ \forall \delta > 0.$$

Since the right-hand term tends to 0 when n tends to infinity, this concludes the proof.

□

REFERENCES

[1] Y. Feng and D. P. Palomar, *A Signal Processing Perspective on Financial Engineering*. Foundations and Trends in Signal Processing, Now Publishers, 2016.

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

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

[4] 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.

[5] 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.

[6] G. Draškovic and F. Pascal, "New insights into the statistical properties of M-estimators," *Signal Processing, IEEE Transactions on*, vol. 66, no. 16, pp. 4253–4263, August 2018.

[7] E. Ollila, D. P. Palomar, and F. Pascal, "Shrinking the eigenvalues of M-estimators of covariance matrix," *IEEE Transactions on Signal Processing*, vol. 69, pp. 256–269, 2021.

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

[9] C. Liu and D. B. Rubin, "ML estimation of the t-distribution using EM and its extensions, ECM and ECME," *Statistica Sinica*, vol. 5, no. 1, pp. 19–39, 1995.

[10] C. Fernandez and M. F. J. Steel, "Multivariate student-t regression models: Pitfalls and inference," *Biometrika*, vol. 86, no. 1, pp. 153–167, 1999.

[11] B. M. Hill, "A simple general approach to inference about the tail of a distribution," *The annals of statistics*, pp. 1163–1174, 1975.

[12] 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.

[13] R. Couillet and M. Debbah, *Random matrix methods for wireless communications*. Cambridge University Press, 2011.