Shrinkage Coefficient Estimation for Regularized Tyler's M–Estimator: A Leave One Out Approach

Karim Abou-Moustafa
Intel Corp. Chandler, AZ 85226
Email: Karim.Abou-Moustafa@intel.com

2023 IEEE Information Theory Workshop (ITW) | 979-8-3503-0149-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITW55543.2023.10161652

***Abstract*—We consider the problem of estimating a regularization parameter, or shrinkage coefficient $\alpha \in (0, 1)$ for regularized Tyler M-estimators (RTME). In particular, we propose a data-dependent approach for estimating an optimal α based on maximizing a suitably chosen leave-one-out cross-validated (LOOCV) likelihood function. Since the LOOCV approach scales linearly with the number of samples n and hence is computationally intensive, we propose a computationally efficient approximation for the LOOCV likelihood function that permits selecting a near-optimal choice for the shrinkage coefficient α. We demonstrate the efficiency and accuracy of our proposed approach on high-dimensional data sampled from heavy-tailed elliptical distributions, and show that it is consistently better than other methods in the literature for shrinkage coefficient estimation.**

I. Introduction

Tyler's M-estimator (TME) is an accurate and efficient robust estimator for the scatter matrix when the data are samples from an *elliptical distribution* with heavy-tails, and the number of samples n is larger than the data dimensionality p [1], [2]. Elliptical distributions (introduced shortly) are generalization of the multivariate Gaussian distribution and are suitable for modelling empirical distributions with heavy-tails, where such heavy-tails may be due to the existence of outliers in the data [3]. In this setting, and under some mild assumptions on the data, TME is shown to be strongly consistent, asymptotically normal, and is the *most robust* estimator for the scatter matrix for an elliptical distribution in a *minimax* sense; minimizing the maximum asymptotic variance (see Remark 3.1 in [1]). Unfortunately in the $p > n$ regime, TME is not defined. Various research works have proposed regularized versions of TME based on Ledoit & Wolf [4] linear shrinkage estimator whose performance depends on a *carefully chosen* regularization parameter, or *shrinkage coefficient* $\alpha \in (0, 1)$ [5]–[13].[1] Our work here addresses the question of shrinkage coefficient estimation for regularized TME (RTME), and proposes a computationally efficient algorithm for estimating a near-optimal value for this parameter.

Unfortunately, the recursive nature for the regularized TME procedure makes selecting an optimal shrinkage coefficient for this estimator a non-trivial problem. Arguably, three broad approaches were considered for addressing this problem: (*i*) oracle and random matrix theory (RMT) based approaches [6], [10], [11], [16]–[18]; (*ii*) data-dependent approaches based on cross validation techniques [5], [7], [9], [19]; and (*iii*) maximum likelihood based approaches [13]. Oracle based approaches are computationally efficient due their closed-form solutions but may come short in terms of accuracy due to their implicit assumptions on the data distribution, and the implicit assumptions in their asymptotic estimates. Cross validation techniques, on the other hand, are more accurate than oracle based methods since they are data-dependent approaches. This accuracy, however, comes at the cost of intensive computations which makes cross-validation techniques not a favorable option for various applications. The maximum likelihood approach was considered in [13] where the Authors develop an approach, namely the expected likelihood (EL) method, for selecting a near-optimal shrinkage coefficient for RTME when used for specific problems in wireless communications such as adaptive-filtering and estimating the signal's direction of arrival. While in such applications the noisy data samples may be reasonably assumed to have an elliptical distribution, the EL method cannot be considered a general approach for estimating the shrinkage coefficient for RTME due to the specific context and better controlled environments for such problems in wireless communications.

In this work we propose a more general approach for estimating an optimal shrinkage coefficient α for RTME. In particular, we propose a data-dependent approach for selecting an optimal α based on maximizing a suitably chosen leave-one-out cross-validated (LOOCV) likelihood function. Since the LOOCV approach scales linearly with the number of samples n and hence is computationally demanding, we propose a computationally efficient approximation for the LOOCV likelihood functions that eliminates the need for estimating the regularized TME n times for each sample left out during the cross-validation procedure. This efficient approximation results in a *significant speedup* in the computation for the LOOCV estimate, and permits selecting a near-optimal value for the shrinkage coefficient α. On experiments using high-dimensional data sampled from heavy-tailed elliptical distributions, we show that the proposed approach is efficient and consistently more accurate than other methods in the literature at the expense of some moderate computations.

A. Notation and Setup

Scalars and indices are denoted by lowercase letters: x, y and i, j, respectively. Vectors are denoted by lowercase bold letters: $\mathbf{x}, \mathbf{y}$, and matrices by uppercase bold letters: $\mathbf{X}, \mathbf{Y}$.

[1] Shrinkage coefficient estimation for the sample covariance matrix and *generalized* M-estimators for elliptically distributed data was considered in [14]–[16].

Sets are denoted by calligraphic letters: $\mathcal{X}, \mathcal{Y}$, and spaces are denoted by double-bold uppercase letters: $\mathbb{R}, \mathbb{S}$. The identity matrix is denoted by $\mathbf{I}$, and $\mathbf{0}$ is the vector with all zeros, both with suitable dimensions from the context. For $\mathbf{x} \in \mathbb{R}^p$, $\|x\|$ is the Euclidean norm. For a matrix $\mathbf{A} = (a_{ij})$, $\|\mathbf{A}\|_F$ is the Frobenius norm, $\mathrm{Tr}(\mathbf{A})$ is the matrix trace, and $\det(\mathbf{A})$ is the matrix determinant. The space of symmetric and positive definite (PD) matrices is denoted by $\mathbb{S}_+^p$. The unit sphere in $\mathbb{R}^p$ is denoted by $\mathcal{S}^p$, where $\mathcal{S}^p = \{\mathbf{x} \in \mathbb{R}^p \ s.t. \ \|\mathbf{x}\| = 1\}$.

1) Elliptical Distributions: Let $\mathbf{z}$ be a p dimensional random vector (RV) generated by the following model [3], [20]:

$$\mathbf{z} = \boldsymbol{\mu} + u\mathbf{S}^{\frac{1}{2}}\mathbf{y} = \boldsymbol{\mu} + u\tilde{\mathbf{x}} \ , \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a location vector, $\mathbf{S} \in \mathbb{S}_+^p$ is a *scatter* or *shape* matrix, $\mathbf{y}$ is drawn uniformly from $\mathcal{S}^p$, and u is a nonnegative random variable (r.v.) stochastically independent of $\mathbf{y}$. The resulting RV $\mathbf{z}$ from the model in (1) is an *Elliptically Distributed* (ED) RV. Note that $\mathbf{S}$ in (1) is not unique since it can be arbitrarily scaled with $1/u$ absorbing the scaling factor u. The distribution function of u, known as the *generating distribution function*, constitutes the particular elliptical distribution family of the RV $\mathbf{z}$. If $\mathbf{z}$ is an ED RV, its probability density function (PDF) is defined as:

$$f(\mathbf{z}; \boldsymbol{\mu}, \mathbf{S}, g_u) = \det(\mathbf{S})^{-\frac{1}{2}} \, g_u\left(\bar{\mathbf{z}}^\top \mathbf{S}^{-1} \bar{\mathbf{z}}\right), \tag{2}$$

where $\bar{\mathbf{z}} = (\mathbf{z} - \boldsymbol{\mu})$, and $g_u : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is a nonnegative decreasing function known as the *density generator function* and is not dependent on $\boldsymbol{\mu}$ and $\mathbf{S}$, but dependent on the generating distribution function of u. The density generator function determines the shape of the PDF, as well as the *tail decay* of the distribution. For any elliptical distribution, if its population covariance matrix $\boldsymbol{\Sigma}$ exists, then $\boldsymbol{\Sigma} = c_g \mathbf{S}$ for some constant $c_g > 0$ that is dependent on g_u.

II. Regularized Tyler's M-Estimator

Let $\mathcal{Z}_n = (\mathbf{z}_i)_{i=1}^n$ be a sample of n *independent* and *identically distributed* (*i.i.d.*) realizations from the model in (1) with location vector $\boldsymbol{\mu} = \mathbf{0}$ and scatter matrix $\mathbf{S}$. Tyler's M-estimator (TME) can be derived as a maximum likelihood (ML) estimator of the shape matrix for the *Angular Central Gaussian* (ACG) distribution (introduced shortly) based on the sample $\mathcal{Z}_n$ [2]. Note that $\mathcal{Z}_n$ can be written as $(u_1\tilde{\mathbf{x}}_1, \ldots, u_n\tilde{\mathbf{x}}_n)$. However, since the scalars $u_1, \ldots, u_n$ are unknown, there is a scaling ambiguity and one can only expect to estimate $\mathbf{S}$ up to a scaling factor. TME overcomes this limitation by working with the normalized samples: $\mathbf{x}_i = \mathbf{z}_i / \|\mathbf{z}_i\| = \tilde{\mathbf{x}}_i / \|\tilde{\mathbf{x}}_i\|$, $1 \le i \le n$, where the scalars u_i cancels out. The PDF for the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is given by:

$$f(\mathbf{x}; \mathbf{S}) = (2\pi)^{-\frac{p}{2}} \Gamma(\tfrac{1}{2}) \det(\mathbf{S})^{-\frac{1}{2}} \left(\mathbf{x}^\top \mathbf{S}^{-1} \mathbf{x}\right)^{-\frac{p}{2}}, \tag{3}$$

where $\mathbf{x} \in \mathcal{S}^p$, $\Gamma(\cdot)$ is the Gamma function, and $\Gamma(p/2)/(2\pi)^{\frac{p}{2}}$ is the surface area of $\mathcal{S}^p$. The ACG density in (3) represents the *distribution of directions* for samples drawn from a multivariate Gaussian distribution with zero mean and covariance matrix $\mathbf{S}$ [2]. Thus, only the directions of outliers can affect TME's performance but not their magnitude. Given an *i.i.d.* random sample $\mathcal{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ from a distribution having the ACG density in (3), the likelihood of $\mathcal{X}_n$ with respect to $\mathbf{S}$ is *proportional* to:

$$L(\mathcal{X}_n; \mathbf{S}) = \det(\mathbf{S})^{-n/2} \prod_{i=1}^n \left(\mathbf{x}_i^\top \mathbf{S}^{-1} \mathbf{x}_i\right)^{-\frac{p}{2}} \ . \tag{4}$$

Taking negative $\log$ of $L(\mathcal{X}_n; \mathbf{S})$ yields the following loss function which will be relevant for our next discussions:

$$\mathcal{L}(\mathcal{X}_n; \mathbf{S}) = \frac{p}{2} \sum_{i=1}^n \log\left(\mathbf{x}_i^\top \mathbf{S}^{-1} \mathbf{x}_i\right) + \frac{n}{2} \log\det(\mathbf{S}) \, . \tag{5}$$

Taking the derivative of $\mathcal{L}(\mathcal{X}_n; \mathbf{S})$ with respect to $\mathbf{S}$ and equating it to zero, the ML estimator for $\mathbf{S}$ is the solution to the following fixed point equation:

$$\mathbf{S}_n = \frac{p}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top / (\mathbf{x}_i^\top \mathbf{S}_n^{-1} \mathbf{x}_i) \ , \tag{6}$$

where it is assumed that for $i = 1, \ldots, n$, $\mathbf{x}_i \neq \mathbf{0}$ since samples lying at the origin provide no directional information on the scatter matrix. The solution to equation (6) can be found using the following fixed point iteration (FPI) algorithm [21]:

$$\widehat{\mathbf{S}}_{t+1} = \frac{p}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top / (\mathbf{x}_i^\top \widehat{\mathbf{S}}_t^{-1} \mathbf{x}_i) \ , \tag{7}$$

with $\widehat{\mathbf{S}}_0 = \mathbf{I}$, or any arbitrary initial $\widehat{\mathbf{S}}_0 \in \mathbb{S}_+^p$ [21]. Theorem 2.2 and Corollaries 2.3 & 2.3 in [1] show that under some mild assumptions on the data, the FPI algorithm in (7) *almost surely* converges to the solution of (6), and the limiting solution $\widehat{\mathbf{S}} = \widehat{\mathbf{S}}_T$ computed at the last iterate T is unique up to a positive multiplicative scalar. To avoid the scaling ambiguity in $\widehat{\mathbf{S}}$, [2], [21] proposed the following iterations:

$$\widehat{\mathbf{S}}_{t+1} = p \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\mathbf{x}_i^\top \widehat{\mathbf{S}}_t^{-1} \mathbf{x}_i} \Bigg/ \mathrm{Tr}\left(\sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\mathbf{x}_i^\top \widehat{\mathbf{S}}_t^{-1} \mathbf{x}_i}\right), \tag{8}$$

with $\widehat{\mathbf{S}}_0 = \mathbf{I}$. If $n > p(p-1)$, the FPI algorithm in (8) *almost surely* generates the solution to (6) and satisfies the constraint $\mathrm{Tr}(\widehat{\mathbf{S}}) = p$ (Corollary 2.2 in [1]).

Unfortunately, when $p > n$, TME is not defined; the LHS of (6) must be a full rank symmetric PD matrix, while the RHS is rank deficient.[2] Various researchers have proposed different flavours of a regularized TME (RTME) using the spirit of Ledoit & Wolf linear shrinkage estimator [4]. In particular, the works in [5]–[10] proposed slight variants from the following regularized FPI algorithm:

$$\widehat{\mathbf{S}}_{t+1}(\alpha) = (1-\alpha)\frac{p}{n} \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^\top}{(\mathbf{x}_i^\top \widehat{\mathbf{S}}_t^{-1}(\alpha) \mathbf{x}_i)} + \alpha \mathbf{I} \ , \tag{9}$$

where $\alpha > 0$ is a regularization parameter (or a *shrinkage coefficient*) that controls the amount of shrinkage applied to

[2]For TME, regularization may still be needed for $p \le n \le p(p-1)$ when the points are not in general position, and/or the samples are not drawn from an elliptical distribution.

scatter matrix $\mathbf{S}$ towards the identity matrix $\mathbf{I}$. In this work, we consider the RTME algorithm proposed by Chen, Wiesel & Hero (CWH) [6] for its better numerical properties over other RTME variants. CWH's algorithm requires that $\alpha \in (0,1)$. In addition, to avoid scaling ambiguity in $\widehat{\mathbf{S}}$ and attain a unique solution, CWH's algorithm requires a trace normalization step after each FPI in (9) to ensure that $\mathrm{Tr}(\widehat{\mathbf{S}}) = p$; i.e.

$$\tilde{\mathbf{S}}_{t+1}(\alpha) = (1-\alpha)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top(\widehat{\mathbf{S}}_t(\alpha))^{-1}\mathbf{x}_i} + \alpha\mathbf{I} \tag{10}$$

$$\widehat{\mathbf{S}}_{t+1}(\alpha) = \frac{p}{\mathrm{Tr}(\tilde{\mathbf{S}}_{t+1}(\alpha))}\tilde{\mathbf{S}}_{t+1}(\alpha)\ . \tag{11}$$

If $\alpha = 0$, one restores the original unbiased TME in (7), and if $\alpha = 1$ the estimator reduces to the uncorrelated scatter matrix defined by the scaled identity matrix $\alpha\mathbf{I}$. When $p < n$, and the samples are drawn from an elliptical distribution, α is expected to be zero (or close to zero) and results for existence and uniqueness of the estimator still hold [1], [8]. When $p \geq n$, α is expected to be large; however to ensure the *existence and uniqueness* of the estimator, α needs to be strictly greater than $1 - n/p$ [8], [9].

Computational Complexity: A preliminary analysis for Tyler's FPI algorithm shows that the running time for each iteration is $O(np^2 + p^3)$ where $O(np^2)$ is the time needed to compute the sum of rank-one matrices, and $O(p^3)$ is the time needed to compute the inverse matrix $\widehat{\mathbf{S}}_t^{-1}(\alpha)$. Since $\widehat{\mathbf{S}}_t(\alpha)$ is PD, an efficient computation for the inverse can be done using Cholesky factorization: $\widehat{\mathbf{S}}_t(\alpha) = \mathbf{L}\mathbf{L}^\top$, where $\mathbf{L}$ is a lower triangular matrix. Cholesky factorization requires $\frac{1}{3}p^3$ flops: $\frac{1}{6}p^3$ multiplications, and $\frac{1}{6}p^3$ additions. Finally inverting a triangular matrix will require p^2 flops. If T iterations are needed for the FPI algorithm to converge, its total running time complexity will be $O(T(np^2 + p^3))$.

III. Optimal Choice of Shrinkage Coefficient

Our objective is to find an appropriate α that is *optimal* under a suitable loss function. If the true scatter matrix $\mathbf{S}$ is known, one can choose a shrinkage coefficient that minimizes an appropriate distance metric between $\widehat{\mathbf{S}}$ and $\mathbf{S}$. Since $\mathbf{S}$ is unknown, our approach will depend on the likelihood function of $\mathcal{X}_n$ with respect to $\mathbf{S}$ in (4). In particular, for a *fixed* $\bar{\alpha} \in (0,1)$, suppose that $\widehat{\mathbf{S}}(\bar{\alpha})$ is an estimate of the true scatter matrix $\mathbf{S}$. Given the sample $\mathcal{X}_n$, one can assess the quality of $\widehat{\mathbf{S}}(\bar{\alpha})$ with respect to $\mathcal{X}_n$ using the likelihood function $L(\mathcal{X}_n; \mathbf{S})$ in (4) – or equivalently using the loss function $\mathcal{L}(\mathcal{X}_n; \mathbf{S})$ in (5) – by replacing $\mathbf{S}$ with $\widehat{\mathbf{S}}(\bar{\alpha})$. Using this approach, an optimal α with respect to $\mathcal{X}_n$, denoted α^*, will be the one that minimizes $\mathcal{L}(\mathcal{X}_n, \widehat{\mathbf{S}}(\alpha))$ over the range of $\alpha \in (0,1)$. That is,

$$\alpha^* = \underset{\alpha\in(0,1)}{\arg\min}\quad \mathcal{L}(\mathcal{X}_n, \widehat{\mathbf{S}}(\alpha))\ . \tag{12}$$

The problem with this direct approach is that $\widehat{\mathbf{S}}(\alpha)$ needs to be computed using the sample $\mathcal{X}_n$. That is, the sample $\mathcal{X}_n$ will be used twice; first time to compute $\widehat{\mathbf{S}}(\alpha)$, and a second time to assess the quality of $\widehat{\mathbf{S}}(\alpha)$ using $\mathcal{L}(\mathcal{X}_n, \widehat{\mathbf{S}}(\alpha))$ in (5). This is known as *double dipping* and inevitably it leads to an *overfit* estimate of the shrinkage coefficient α.

Cross validation (CV) techniques overcome this problem by splitting the data into two non-overlapping samples [22]; one sample for estimating $\mathbf{S}$ and the other sample for estimating the loss $\mathcal{L}$. Here, we propose to use the *Leave-One-Out* CV (LOOCV) method for estimating $\mathbf{S}$ and $\mathcal{L}$. In particular, for $1 \leq i \leq n$, LOOCV splits $\mathcal{X}_n$ into two sub-samples: the sample $\mathcal{X}_{n\setminus i} = (\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n)$, and the sample $(\mathbf{x}_i)$ which contains the single data point $\mathbf{x}_i$. The sample $\mathcal{X}_{n\setminus i}$ will be used to estimate $\mathbf{S}(\alpha)$ using CWH's algorithm in (10) & (11), while the single sample $(\mathbf{x}_i)$ will be used to estimate $\mathcal{L}(\mathbf{x}_i, \widehat{\mathbf{S}}(\alpha))$. This process is repeated n times and the LOOCV estimate will be the average of all $\mathcal{L}(\mathbf{x}_i, \widehat{\mathbf{S}}(\alpha))$, $1 \leq i \leq n$. Using LOOCV, an optimal α can be computed as follows:

$$\widehat{\alpha}^*_{CV} = \underset{\alpha\in(0,1)}{\arg\min}\quad \mathcal{L}_{CV}(\mathcal{X}_n, \alpha)\ , \tag{13}$$

where $\mathcal{L}_{CV}(\cdot)$ is the *Average* CV *Loss* (ACVL) defined as:

$$\mathcal{L}_{CV}(\mathcal{X}_n, \alpha) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathbf{x}_i, \widehat{\mathbf{S}}(\alpha; \mathcal{X}_{n\setminus i}))\ , \tag{14}$$

and $\widehat{\mathbf{S}}(\alpha; \mathcal{X}_{n\setminus i})$ is the regularized scatter matrix estimated from the samples in $\mathcal{X}_{n\setminus i}$ using CWH's algorithm in (10) & (11).

In practice, one possible approach to solve problem (13) can be using a simply *grid search*: (*i*) define a discrete range of increasing values of α: $(\alpha_1, \ldots, \alpha_j, \ldots, \alpha_m)$; (*ii*) evaluate $\mathcal{L}_{CV}(\mathcal{X}_n, \alpha_j)$ for each α_j using (14); and (*iii*) choose α_j with the minimum $\mathcal{L}_{CV}(\cdot)$.[3] For a *reasonably fine* discretization for the range of α's, using this direct estimation approach will yield an estimate for α that is *reasonably close* to its optimal value. With little abuse of terminology, and for reasons that will be discussed shortly, we refer to this method as the *Exact ACVL* method.

A. Properties of LOOCV and its computational overhead

The Riemannian manifold of symmetric PD matrices $\mathbb{S}_+^p$ is a subset of $\mathbb{R}^{p(p+1)/2}$ and is a compact space [21]. The log likelihood function $\mathcal{L}(\mathcal{X}_n, \mathbf{S})$ in (5) is geodesically convex with respect to $\mathbb{S}_+^p$ [23], [19], and properties for this type of likelihood functions has been studied in [24]. In particular, $\mathcal{L}(\mathcal{X}_n, \mathbf{S})$ maintains the three main properties of maximum likelihood estimators [25]: *consistency*, *efficiency*, and *functional invariance*. On the other hand, the LOOCV estimate is *almost an unbiased* estimate in the following sense: for a fixed $\bar{\alpha}$, $\mathbb{E}\mathcal{L}_{CV}(\mathcal{X}_n, \bar{\alpha}) = \mathbb{E}\mathcal{L}(\mathcal{X}_{n-1}, \widehat{\mathbf{S}}(\bar{\alpha}))$ [26, Ch. 24]; i.e. LOOCV is an estimator for $\mathcal{L}(\mathcal{X}_{n-1}, \widehat{\mathbf{S}}(\bar{\alpha}))$ rather than for $\mathcal{L}(\mathcal{X}_n, \widehat{\mathbf{S}}(\bar{\alpha}))$. Thus, from the consistency of $\mathcal{L}(\mathcal{X}_n, \mathbf{S})$, and for most interesting cases, we have that for large values of n, the difference between $\mathcal{L}(\mathcal{X}_n, \widehat{\mathbf{S}}(\bar{\alpha}))$ and $\mathcal{L}(\mathcal{X}_{n-1}, \widehat{\mathbf{S}}(\bar{\alpha}))$ will be negligible. In this sense we can state the following proposition which will be useful for our approximation approach introduced in the next section.

[3] Note that when $p > n$, and for existence and uniqueness results to hold, α needs to be strictly greater than $1 - n/p$ [8], [9], and hence there is no need to evaluate $\mathcal{L}_{CV}(\cdot)$ for $\alpha \leq 1 - n/p$.

Proposition 1. *Under the i.i.d assumption for the samples in $\mathcal{X}_n$ and from the consistency of $\mathcal{L}(\mathcal{X}_n, \mathbf{S})$, we have that for large values of n, with high probability, the difference between $\mathcal{L}_{CV}(\mathcal{X}_n, \bar{\alpha})$ and $\mathcal{L}_{CV}(\mathcal{X}_{n-1}, \bar{\alpha})$ will be arbitrarily small.*

LOOCV is also known for its high computational overhead. Indeed, for a fixed $\bar{\alpha}$ and for n samples in $\mathcal{X}_n$, LOOCV will make n calls for the FPI algorithm in order to compute $\mathcal{L}_{CV}(\mathcal{X}_n, \bar{\alpha})$ in (14). Thus, for m values of α_j, for $j = 1, \ldots, m$, the Exact ACVL method in (13) will require mn calls for the FPI algorithm, which is prohibitive even for moderate values of n. If the FPI algorithm requires T iterations to converge, then the FPI algorithm will consume $O(mn * T(np^2 + p^3))$ time from the Exact ACVL method in (13), where $O(T(np^2 + p^3))$ is the running time for a single call for the FPI algorithm. Our objective in the following section is to reduce the time consumed by the FPI algorithm in the Exact ACVL method to be $O(m * T(np^2 + p^3))$. In particular, we propose an efficient approximation for $\widehat{\mathbf{S}}(\alpha, \mathcal{X}_{n\setminus i})$ in (14) so that the FPI algorithm is invoked m times only instead of mn times to compute $\mathcal{L}_{CV}(\mathcal{X}_n, \alpha)$ in (13).

IV. Efficient Approximation of ACVL

The approximation approach proposed here is motivated by the consistency of $\mathcal{L}(\mathcal{X}_n, \widehat{\mathbf{S}}(\alpha))$, the unbiased property for the LOOCV estimate, and Proposition (1). For a fixed $\bar{\alpha}$, the regularized FPI algorithm in (9) can be expressed as follows:

$$\widehat{\mathbf{S}}_{t+1}(\bar{\alpha}) = (1-\bar{\alpha})p\left(\frac{1}{n}\sum_{i=1}^{n} w_{t,i}^{-1}\mathbf{x}_i\mathbf{x}_i^\top\right) + \bar{\alpha}\mathbf{I}, \text{ where} \quad (15)$$

$$w_{t,i} = \mathbf{x}_i^\top \widehat{\mathbf{S}}_t^{-1}(\bar{\alpha})\mathbf{x}_i \ , \quad (16)$$

and $1 \leq t \leq T$. That is, the first term for the regularized FPI algorithm involves a weighted sample covariance matrix using the weights $w_{t,i}$ and the FPI algorithm iteratively estimates these weights until convergence. Let $(\widehat{w}_1, \widehat{w}_2, \ldots, \widehat{w}_n)$ be the optimal weights estimated using the sample $\mathcal{X}_n$ and the FPI algorithm in (15). For initial matrix $\widehat{\mathbf{S}}_0 \in \mathbb{S}_+^p$, the final estimate for the scatter matrix can be written as:

$$\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n) = (1-\bar{\alpha})\frac{p}{n}\sum_{i=1}^{n}\frac{1}{\widehat{w}_i}\mathbf{x}_i\mathbf{x}_i^\top + \bar{\alpha}\mathbf{I} \ . \quad (17)$$

Let $\mathcal{X}_{n\setminus i} = (\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n)$. Similar to (17), using $\bar{\alpha}$ and initial matrix $\widehat{\mathbf{S}}_0$, the final estimate for the scatter matrix based on $\mathcal{X}_{n\setminus i}$ will be:

$$\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n\setminus i}) = (1-\bar{\alpha})\frac{p}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{1}{\widehat{v}_j}\mathbf{x}_j\mathbf{x}_j^\top + \bar{\alpha}\mathbf{I} \ , \quad (18)$$

where $(\widehat{v}_1, \ldots, \widehat{v}_{i-1}, \widehat{v}_{i+1}, \ldots, \widehat{v}_n)$ are the optimal weights estimated using $\mathcal{X}_{n\setminus i}$. From Proposition (1), and using initial matrix $\widehat{\mathbf{S}}_0$ to obtain the estimates in (17) and (18), it is expected that for large values of n the difference between $\mathcal{L}_{CV}(\mathcal{X}_n, \bar{\alpha})$ and $\mathcal{L}_{CV}(\mathcal{X}_{n\setminus i}, \bar{\alpha})$ will be arbitrarily small. Note also that in this setting the difference between $\widehat{v}_j$ and $\widehat{w}_j$ will be arbitrarily small as well; i.e. $\widehat{v}_j \approx \widehat{w}_j$, for $j \neq i$, and $j = 1, \ldots, n$.

To introduce our proposed approximation, suppose that the true scatter matrix $\mathbf{S}^* \in \mathbb{S}_p^+$ is known and that $(\mathbf{S}^*)^{-1}$ has been computed. It follows that the final estimate $\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n)$ in (17) can be directly computed as follows:

$$\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n) = \frac{(1-\bar{\alpha})p}{n}\sum_{i=1}^{n}\frac{1}{\widehat{w}_i^*}\mathbf{x}_i\mathbf{x}_i^\top + \bar{\alpha}\mathbf{I}, \text{ where} \quad (19)$$

$$\widehat{w}_i^* = \mathbf{x}_i^\top(\mathbf{S}^*)^{-1}\mathbf{x}_i \ . \quad (20)$$

Similarly, using $(\mathbf{S}^*)^{-1}$, the final estimate $\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n\setminus i})$ in equation (18) can be directly computed as follows:

$$\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n\setminus i}) = \frac{(1-\bar{\alpha})p}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{1}{\widehat{v}_j^*}\mathbf{x}_j\mathbf{x}_j^\top + \bar{\alpha}\mathbf{I}, \text{ where} \quad (21)$$

$$\widehat{v}_j^* = \mathbf{x}_j^\top(\mathbf{S}^*)^{-1}\mathbf{x}_j \ . \quad (22)$$

Note that both $\widehat{w}_i^*$ in (20) and $\widehat{v}_j^*$ in (22) are dependent on the true but unknown scatter matrix $\mathbf{S}^*$ and in this case, $\widehat{v}_j^* = \widehat{w}_j^*$ for $j \neq i$, and $j = 1, \ldots, n$. Also, from Proposition (1), it is expected that for large values of n, the difference between $\mathcal{L}_{CV}(\mathcal{X}_n, \bar{\alpha})$ using $\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n)$ in (19) and $\mathcal{L}_{CV}(\mathcal{X}_{n\setminus i}, \bar{\alpha})$ using $\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n\setminus i})$ in (21) will be arbitrarily small. However since $\mathbf{S}^*$ is unknown, we propose to approximate $\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n\setminus i})$ in (21) using the following estimate:

$$\widetilde{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n\setminus i}) = \frac{(1-\bar{\alpha})p}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{1}{\widetilde{v}_j}\mathbf{x}_j\mathbf{x}_j^\top + \bar{\alpha}\mathbf{I}, \text{ where} \quad (23)$$

$$\widetilde{v}_j = \mathbf{x}_j^\top\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n)^{-1}\mathbf{x}_j \ . \quad (24)$$

That is, we plugin the regularized TME $\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n) \in \mathbb{S}_p^+$ from (17) into equation (22) to obtain the new weights $\widetilde{v}_j$, for $j \neq i$, $j = 1, \ldots, n$, then use the new weights $\widetilde{v}_j$ to obtain the new estimate $\widetilde{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n\setminus i})$ in (23). Using the previous approximation, the optimal shrinkage coefficient α^* can now be computed as follows:

$$\widehat{\alpha}_{CV}^* = \underset{\alpha\in(0,1)}{\arg\min} \quad \widetilde{\mathcal{L}}_{CV}(\mathcal{X}_n, \alpha) \ , \text{ where} \quad (25)$$

$$\widetilde{\mathcal{L}}_{CV}(\mathcal{X}_n, \alpha) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathbf{x}_i, \widetilde{\mathbf{S}}(\alpha; \mathcal{X}_{n\setminus i})) \ , \quad (26)$$

and $\widetilde{\mathcal{L}}_{CV}(\mathcal{X}_n, \alpha)$ is the *approximate* ACVL. Due to this approximation, we refer to the method in (25) for estimating the optimal shrinkage coefficient α^* as the *Approximate ACVL* method. For m values of α in $(\alpha_1, \ldots, \alpha_m)$, the FPI algorithm will now consume $O(m * T(np^2 + p^3))$ running time from the Approximate ACVL method since the FPI algorithm is not needed to compute $\widetilde{\mathbf{S}}(\alpha; \mathcal{X}_{n\setminus i})$ for every $i = 1, \ldots, n$.

V. Empirical Validation

Similar to other works in the literature on RTME [6]–[10], [27], we consider the Toeplitz matrix used in the work of Bickel & Levina [28] to be the population scatter (or shape)

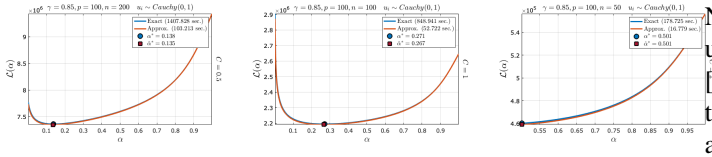

Fig. 1. Comparison between the Exact ACVL method (solid blue line) and Approximate ACVL method (solid red line) in three different settings: $p < n$ (left), $p = n$ (middle), and $p > n$ (right). The blue circle and red square indicate the optimal values for α obtained from the Exact and Approximate ACVL methods, respectively. The speedup in each setting is: 13×, 16×, and 11×, respectively.

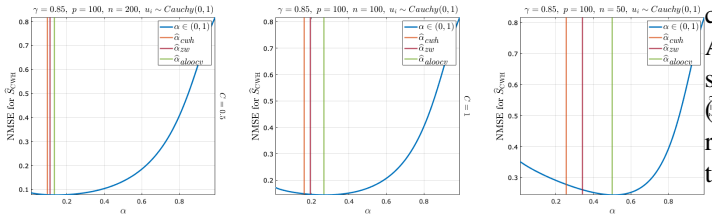

Fig. 2. The solid blue line shows the NMSE between the population matrix $\mathbf{S}$ and the scatter matrix $\widehat{\mathbf{S}}_{CWH}$ estimated using CWH's FPI algorithm for values of $\alpha \in (0, 1)$ in three different settings: $p < n$ (left), $p = n$ (middle), and $p > n$ (right). The orange, red, and green solid vertical lines indicate the shrinkage coefficients $\widehat{\alpha}_{cwh}$, $\widehat{\alpha}_{zw}$, and $\widehat{\alpha}_{aloocv}$, obtained using the method in [6, Eq. 13], the method in [11, Eq. 12], and the Approximate ACVL method, respectively.

matrix $\mathbf{S}$ for the elliptical RV in (1); that is $\mathbf{S} = (s_{i,j}) = \gamma^{|i-j|}$, where $\gamma = 0.85$. Note that as $\gamma \to 1$, $\mathbf{S}$ approaches being a singular matrix, while as $\gamma \to 0$, $\mathbf{S}$ approaches the identity matrix. The random quantities u and $\mathbf{y}$ in (1) are stochastically independent. We let $\mathbf{y}_1, \ldots, \mathbf{y}_n$'s be samples from a p-variate standard Gaussian distribution $N(\mathbf{0}, \mathbf{I})$. For r.v. u, we consider four different choices for heavy-tailed distributions: (*i*) $u_i = 1$, which makes $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ are *i.i.d.* samples from $N(\mathbf{0}, \mathbf{S})$; (*ii*) $u_i = \sqrt{d/\chi^2_d}$, a Student-T distribution with degrees of freedom $d = 3$; (*iii*) $u_i = \text{Laplace}(0, 1)$, a heavy-tailed distribution with finite moments; and (*iv*) $u_i = \text{Cauchy}(0, 1)$, a heavy-tailed distribution with undefined moments. Note that since TME and RTME operate on the normalized samples $\mathbf{x}_i$, the scalars u_i's cancel out, and the resulting plots become identical regardless of the distribution of u_i. Due to space limitations, we only show the plots for the Cauchy distribution. The accuracy of an estimator $\widehat{\mathbf{S}}$ is measured using the normalized mean squared error (NMSE) $\|\widehat{\mathbf{S}} - \mathbf{S}\|_F^2 / \|\mathbf{S}\|_F^2$. The convergence criterion for all RTME algorithms is $\|p\widehat{\mathbf{S}}/\text{Tr}(\widehat{\mathbf{S}}) - p\mathbf{S}/\text{Tr}(\mathbf{S})\|_F^2 < \epsilon$, where $\epsilon = 1.0e-9$ is the desired solution accuracy. The value of p was set to 100, while n was set to three different values $\{200, 100, 50\}$ to consider three different scenarios: $p < n$, $p = n$, and $p > n$, respectively. The value of C that appears on the right y-axis in Figures 1 and 2 is for the ratio p/n.

Figure (1) compares the *Exact* ACVL method and the *Approximate* ACVL method developed in the previous section in terms of (*i*) average CV loss for each method (solid blue line vs. solid red line), (*ii*) running time (in seconds), and (*iii*) the optimal shrinkage coefficient α obtained by each method. Note that in the $p > n$ setting, and for the existence and uniqueness results to hold for the estimated scatter matrix [8], [9], the search range for the shrinkage coefficient α was set to the interval $(\alpha_0, 1)$, where $\alpha_0 = 1 - n/p$. In terms of average CV loss, it can be seen that the Exact CV loss in (14) (solid blue line) and the Approximate CV loss in (26) (solid red line) are almost identical in the three settings: $p < n$, $p = n$, and $p > n$. This confirms that the Approximate CV loss proposed in the previous section is valid and sufficiently close to the Exact CV loss, and hence it can be used to obtain a near-optimal value for the shrinkage coefficient α. Indeed, it can be seen that the optimal α obtained by the Approximate ACVL method (red square) is reasonably close to the optimal shrinkage coefficient obtained by the Exact ACVL method (blue cricle) in the three settings. As expected, in terms of running time, the Approximate ACVL method is at least 10 times faster than the Exact ACVL method in the three settings.

Figure (2) compares the shrinkage coefficient obtained using the Approximate ACVL method in (25), denoted by $\widehat{\alpha}_{aloocv}$, with the shrinkage coefficients obtained from the closed-form expressions derived in the works of Chen, Wiesel & Hero [6, Equation 13], denoted by $\widehat{\alpha}_{cwh}$, and Zhang & Wiesel [11, Equation 12], denoted by $\widehat{\alpha}_{zw}$. It can be seen that the Approximate ACVL method is consistent in providing better estimates for the shrinkage coefficient α than the methods in [6] and [11], especially for $p \geq n$ settings. Although the methods in [6] and [11] are faster than the Approximate ACVL method due to their closed-form expressions, it can be noticed that these methods tend to underestimate the optimal value for α, and their estimates tend to diverge from the optimal value as p is growing greater than n. This is unlike the data-dependent method proposed here which tend to obtain an estimate for the shrinkage coefficient that is reasonably close to its optimal value at the cost of some moderate computations.

VI. Concluding Remarks

This work proposes a new shrinkage coefficient estimator for regularized Tyler's M-estimators. The new estimator is data-dependent and is based on minimizing the leave-one-out cross-validated negative log-likelihood function for the estimated scatter matrix with respect to the shrinkage coefficient α. Since the LOOCV approach is computationally demanding and scales linearly with the number of samples n, we proposed an efficient approximation for the LOOCV loss that led to a significant speedup (by one order of magnitude) in computing a near-optimal estimate for the shrinkage coefficient α at the cost of some moderate computations. On experiments using high-dimensional data sampled from heavy-tailed elliptical distributions, our proposed approach showed to be efficient and consistently more accurate than other methods in the literature.

REFERENCES

[1] D. E. Tyler, "A Distribution-Free M-Estimator of Multivariate Scatter," *The Annals of Statistics*, vol. 15, no. 1, pp. 234–251, 1987.

[2] ——, "Statistical analysis for the angular central gaussian distribution on the sphere," *Biometrika*, vol. 74, no. 3, pp. 579–589, 09 1987.

[3] S. Cambanis, S. Huang, and G. Simons, "On the theory of elliptically contoured distributions," *Journal of Multivariate Analysis*, vol. 11, no. 3, pp. 368–385, 1981.

[4] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365 – 411, 2004.

[5] Y. I. Abramovich and N. K. Spencer, "Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering," in *2007 IEEE Int. Conf. on Acoustics, Speech and Signal Processing - ICASSP*, vol. 3, 2007, pp. 1105–1108.

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

[7] A. Wiesel, "Unified framework to regularized covariance estimation in scaled Gaussian models," *IEEE Trans. on Signal Processing*, vol. 60, no. 1, pp. 29–38, 2012.

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

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

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

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

[12] Y. I. 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 Transactions on Signal Processing*, vol. 61, no. 23, pp. 5807–5818, 2013.

[13] O. Besson and Y. I. Abramovich, "Regularized covariance matrix estimation in complex elliptically symmetric distributions using the expected likelihood approach—part 2: The under-sampled case," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5819–5829, 2013.

[14] 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, 2019.

[15] E. Ollila, D. P. Palomar, and F. Pascal, "Shrinking the eigenvalues of M-estimators of covariance matrix," *IEEE Transactions on Signal Processing*, vol. 69, no. 12, pp. 256–269, 2021.

[16] K. Ashurbekova, A. Usseglio-Carleve, F. Forbes, and S. Achard, "Optimal shrinkage for robust covariance matrix estimators in a small sample size setting," March 2021, working paper or preprint. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02378034

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

[18] Q. Hoarau, A. Breloy, G. Ginolhac, A. Atto, and J. Nicolas, "A subspace approach for shrinkage parameter selection in undersampled configuration for regularised Tyler estimators," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 3291–3295.

[19] L. Duembgen and D. E. Tyler, "Geodesic convexity and regularized scatter estimators," 2016. [Online]. Available: https://arxiv.org/abs/1607.05455

[20] G. Frahm, "Generalized elliptical distributions: Theory and applications," Ph.D. dissertation, Universität zu Köln, 2004.

[21] J. T. Kent and D. E. Tyler, "Maximum likelihood estimation for the wrapped Cauchy distribution," *Journal of Applied Statistics*, vol. 15, no. 2, pp. 247–254, 1988.

[22] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, 1974.

[23] A. Wiesel, "Geodesic convexity and covariance estimation," *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6182–6189, 2012.

[24] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry*, ser. AMS Translations of Mathematical Monographs, Vol. 191. Oxford University Press, 2000.

[25] G. Casella and R. Berger, *Statistical Inference, Second Edition*. Duxbury Resource Center, June 2002.

[26] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[27] J. Goes, G. Lerman, and B. Nadler, "Robust sparse covariance estimation by thresholding Tyler's M-estimator," *The Annals of Statistics*, vol. 48, no. 1, pp. 86–110, 2020.

[28] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, 2008.