IEEE Open Journal of Signal Processing

Received 21 September 2025; revised 30 December 2025; accepted 7 January 2026. Date of publication 16 January 2026; date of current version 27 January 2026. The review of this article was arranged by Associate Editor D. Cai.

Digital Object Identifier 10.1109/OJSP.2026.3654886

Efficient Learning of Regularized Tyler's *M*-Estimator

KARIM T. ABOU–MOUSTAFA (Senior Member, IEEE)

Intel TDA Research and Development, Intel Corporation, Chandler, AZ 85226 USA

(e-mail: karim.abou-moustafa@intel.com)

ABSTRACT 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 number of variables p. Unfortunately, TME is not defined when $p > n$, and various research works have proposed regularized versions of TME using the spirit of Ledoit & Wolf estimator whose performance depends on a carefully chosen *shrinkage coefficient* $\alpha \in (0, 1)$. In this paper, we consider the problem of estimating an optimal shrinkage coefficient α for Regularized TME (RTME). In particular, we propose to estimate an optimal shrinkage coefficient by setting α as the solution to a suitably chosen objective function; namely the leave-one-out cross-validated (LOOCV) log-likelihood loss. LOOCV, however, is computationally prohibitive even for moderate values of n. To this end, we propose a computationally efficient approximation for the LOOCV log-likelihood loss that eliminates the need for invoking the RTME procedure n times for each sample left out during the LOOCV procedure. This approximation yields an $O(n)$ reduction in the running time complexity for the LOOCV procedure, which results in a significant speedup for computing the LOOCV estimate. We demonstrate the efficacy of the proposed approach on synthetic high-dimensional data sampled from heavy-tailed elliptical distributions, as well as on real high-dimensional datasets for object and face recognition. Our experiments demonstrate that the proposed method is efficient and consistently more accurate than other methods in the literature for shrinkage coefficient estimation.

INDEX TERMS Elliptical distributions, heavy-tailed distributions, scatter matrix estimation, robust covariance matrix estimation, Tyler's M-estimator, leave-one-out cross-validation.

I. INTRODUCTION

Covariance matrices, and scatter matrices (scaled covariance matrices), are ubiquitous in statistical models and procedures for machine learning, pattern recognition, signal processing, and various other fields of science and engineering. The performance of procedures such as principal component analysis (PCA) and its extensions [1], linear discriminant analysis (LDA) and its extensions [2], canonical correlation analysis (CCA) [3], portfolio optimization for investment diversification [4], and outlier detection using robust Mahalanobis distance [5], all depend on an accurate estimate of the covariance matrix. Unfortunately, the process of accurately estimating a covariance matrix is challenging since the number of unknown parameters grows quadratically with the number of variables (or features) p. The problem is well-understood when the number of samples n is much larger than p and the data's underlying distribution is a multivariate Gaussian. In this case, the sample covariance matrix (SCM) is an accurate estimate of the covariance matrix, and is optimal under most criteria [6].

In various modern applications, however, p may be comparable to, or greater than n, and the data's underlying distribution may be non-Gaussian and/or *heavy-tailed*. This challenge is further exacerbated if the data are also contaminated with outliers. In such settings, the SCM is known to be a poor estimate of the covariance matrix and one needs to consider estimators that are more accurate and robust than the SCM. In this work, we are interested in a particular estimator from the family of *robust* and *affine-invariantM*-estimators of scatter matrices proposed by Maronna [7] – namely Tyler's

M-estimator [8], [9] – in the setting where the data's distribution is *heavy-tailed* and the sample support is relatively low; i.e. p is large, and $p \geq n$.[1]

Various approaches were proposed for estimating high-dimensional covariance matrices when $p \geq n$; shrinkage-based approaches [12], [13], [14], [15]; specifying an appropriate prior distribution for the covariance matrix [16]; regularization-based approaches [17], [18], [19]; approaches that exploit sparsity assumptions (banding, tapering, thresholding) [20], [21], [22], [23]; and approaches developed in the robust statistics literature [24], [25]. Except for some approaches from the robust statistics literature, most of the other approaches assume that the data's underlying distribution is a multivariate Gaussian, which may not be a reasonable assumption for handling outliers, or samples from heavy-tailed distributions.

Elliptical distributions (introduced shortly) are the generalization of the multivariate Gaussian distribution and are suitable for modeling empirical distributions with heavy-tails, where such heavy-tails may be due to the existence of outliers in the data [26]. The elliptical family encompasses distributions such as Gaussian, Cauchy, Laplace, and Student-t. 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 $n \gg p$. In this setting, and under some mild assumptions on the data, TME has various attractive properties [8], [9]. TME is 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 [8]). Unfortunately, in the $p > n$ regime, TME is not defined. Various research works have proposed regularized versions of TME using the spirit of Ledoit–Wolf linear shrinkage estimator [15] whose performance depends on a *carefully chosen* regularization parameter, or *shrinkage coefficient* $\alpha \in (0, 1)$ [6], [27], [28], [29], [30], [31], [32], [33]. Our work here addresses the question of shrinkage coefficient estimation for *Regularized* TME, and proposes a computationally efficient algorithm for obtaining a near-optimal estimate for this parameter.[2]

Unfortunately, the recursive nature of TME's procedure makes estimating an optimal shrinkage coefficient for this estimator a non-trivial problem. Arguably, two broad approaches were considered to address this problem: (*i*) oracle and random matrix theory (RMT) based approaches [28], [31], [32], [36], [37]; (*ii*) data-dependent approaches based on *Cross Validation* (CV) techniques [6], [27], [30], [38]; and (*iii*) maximum likelihood (ML) based approaches [33]. However, none of these methods jointly address accuracy and computational efficiency.

Oracle-based approaches are computationally efficient due to their *closed-form* solutions but may come short in terms of accuracy due to their assumptions on the data distribution, as well as the assumptions on their asymptotic estimates [28], [31], [32], [36], [37]. Unfortunately, such assumptions may not hold in finite-sample high-dimensional settings even when the samples are from an elliptical distribution. For example, as shown in Fig. 2, closed-form solutions (e.g., CWH [28] and ZW [32]) tend to under-estimate the value of α as p is growing greater than n (see more discussion in Section V). CV 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, especially for high-dimensional data, which makes CV techniques not a favorable option for various applications.

In this paper, we propose a more general approach for estimating an optimal shrinkage coefficient α^* for RTME. Our proposed approach formulates the problem of finding an optimal shrinkage coefficient as an optimization problem with respect to parameter α. In particular, we define an optimal shrinkage coefficient α^* as the minimizer for the following loss function; the leave-one-out cross-validated (LOOCV) ative log-likelihood (NLL) for the estimated scatter matrix with respect to parameter α (12). Since LOOCV is computationally prohibitive even for moderate values of n, we propose a computationally efficient *approximation* for the LOOCV NLL loss that *eliminates* the need for computing the Regularized TME (RTME) n times for each sample left out during the LOOCV procedure. This approximation yields an $O(n)$ reduction in the running time complexity for the LOOCV procedure, which results in a significant speedup in computing the LOOCV NLL loss.

At a high-level, the resulting procedure, namely the Approximate Cross-Validated Likelihood (ACVL) method, exploits mild computation and the given finite sample to select a (*data-dependent*) *optimal* shrinkage coefficient α for RTME. In addition, the ACVL method is amenable to parallel computation, and is directly applicable to sparse covariance matrix estimation by means of thresholding the RTME [39]. We demonstrate the efficiency and accuracy of the ACVL method on synthetic high-dimensional data sampled from heavy-tailed elliptical distributions, as well as on real high-dimensional datasets for face recognition (Yale B), object recognition (CIFAR10 and CIFAR 100), and handwritten digit recognition (USPS). Our experiments demonstrate that, with some additional mild computation, the ACVL method is efficient and more accurate than other methods in the literature.

An earlier version of this work with preliminary results appeared in the 2023 IEEE Information Theory Workshop [40]. The present journal article broadens both the scope and depth of the original contribution. Beyond refining the exposition, we streamline the derivation for RTME for any desired target matrix, and we introduce new theoretical insights that connect the proposed approximation to uniform stability and leave-one-out risk estimation. The experimental section has also been significantly expanded to include additional experiments on synthetic high-dimensional data, as well as real-world high-dimensional datasets for face and object recognition. In

[1] See [10], [11] for a recent overview and results on this family of estimators.

[2] Shrinkage coefficient estimation for SCM, and *generalized M*-estimators for elliptically distributed data, were considered in [34], [35].

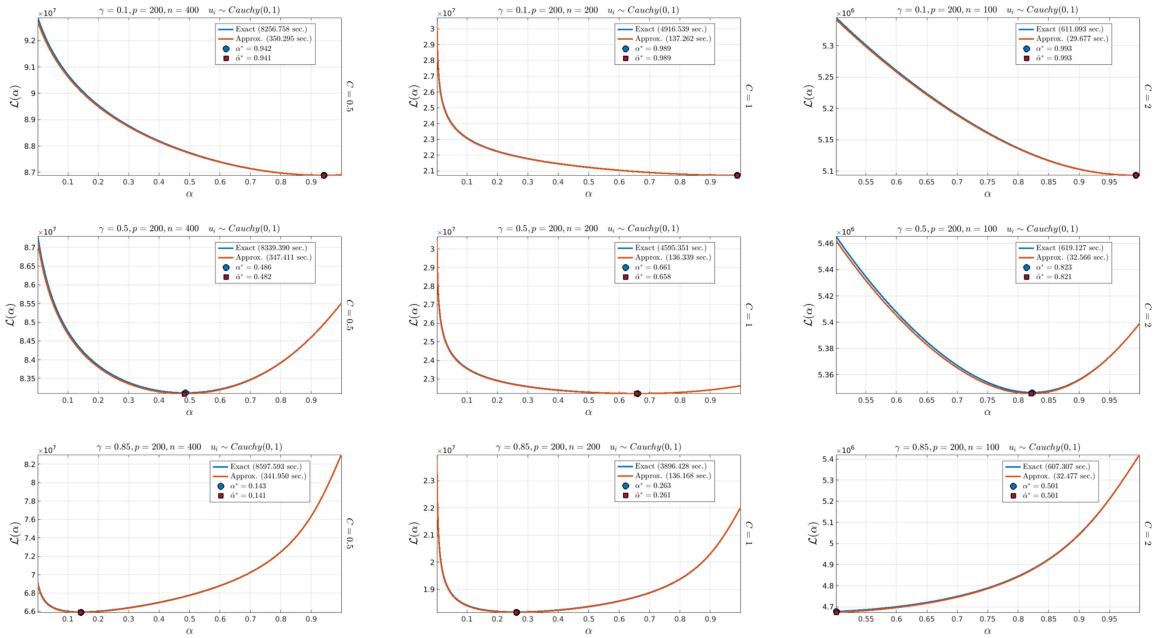

FIGURE 1. **Comparison between *Exact* and *Approximate* Cross-Validated Loss (CVL) for samples drawn from a multivariate Cauchy distribution with population scatter matrix $\mathbf{S}$ defined as the Toeplitz matrix $\mathbf{S} = (S_{i,j}) = \gamma^{|i-j|}$, where $\gamma = 0.1$ (first row), $\gamma = 0.5$ (second row), and $\gamma = 0.85$ (third row). See Section V for more details. For each population matrix configuration, we consider three different settings for p and n; $p < n$ (left column), $p = n$ (middle column), and $p > n$ (right column). The blue and red solid lines are for the Exact and Approximate CV loss for each value of α, respectively. The blue circle and red square indicate the optimal values for α obtained from the Exact and Approximate CVL methods, respectively. The running times (in seconds) for the Exact and Approximate CVL methods are shown in the legend. The speedup for the Approximate CVL method for each sub-figure is: (first row) 24.3x, 35.8x, 20.6x; (second row) 24.0x, 33.7x, 19.0x; (third row) 25.0x, 28.6x, 18.7x.**

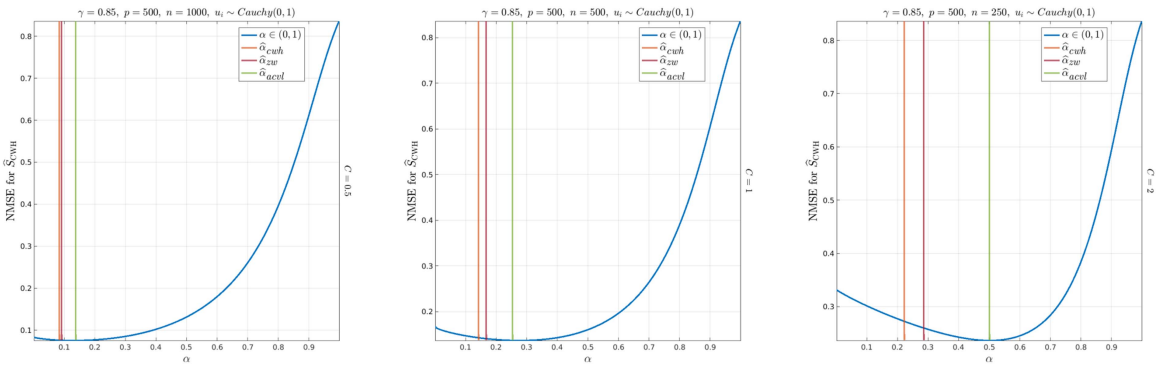

FIGURE 2. **Comparison between the estimated α from three different methods (CWH, ZW, and ACVL) on multivariate Cauchy distributed data with population matrix $\mathbf{S}$. The solid blue line shows the NMSE between the population matrix $\mathbf{S}$ and the scatter matrix $\widehat{\mathbf{S}}$ estimated using SBP's RFPI algorithm for $\alpha \in (0, 1)$ and $p = 500$, in three different settings: $p < n$ (left), $p = n$ (middle), and $p > n$ (right). The population scatter matrix $\mathbf{S}$ in this experiment is the Toeplitz matrix with $\gamma = 0.85$; i.e. $\mathbf{S}$ is a near-singular matrix. See Section V for more details. The orange, red, and green solid vertical lines indicate the values for optimal $\hat{\alpha}_{cwh}$, $\hat{\alpha}_{zw}$, and $\hat{\alpha}_{acvl}$ obtained using the methods in [28, Eq. (13)], [32, Eq. (12)], and the ACVL method, respectively.**

addition, we expand the discussions on the limitations of the proposed method, and potential directions for future research work. Collectively, these additions provide a more complete treatment of the proposed methodology than was possible in the workshop version.

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: $\mathbf{X}$, $\mathbf{Y}$. 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 $\mathcal{S}^p = \{\mathbf{x} \in \mathbb{R}^p \text{ s.t. } ||\mathbf{x}|| = 1\}$.

B. ELLIPTICAL DISTRIBUTIONS

We will use the stochastic model due to [41] and recently used in the literature to define *elliptical* random vectors (RV) [39], [42]. Let $\mathbf{z}$ be a p dimensional RV generated by the following model:

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

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 $\boldsymbol{\mathcal{S}}^p$, 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 defined up to scale, and hence 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 $\mathbf{z}$ (i.e. Gaussian, Cauchy, Laplace, etc.). 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 $\boldsymbol{\mathcal{Z}}_n = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ 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 $\mathbf{S}$. We are interested in computationally efficient and statistically accurate algorithms for estimating the population scatter matrix $\mathbf{S}$ using the samples in $\boldsymbol{\mathcal{Z}}_n$ in the setting where $p > n$. Here we do not make *a priori* sparsity assumptions on the scatter matrix $\mathbf{S}$. Without any *a priori* knowledge on c_g and g_u, it may seem less probable to obtain a good estimator for $\mathbf{S}$. In addition, for some elliptical distributions – such as the multivariate Cauchy distribution – they may have infinite second moments in which case the population covariance matrix $\boldsymbol{\Sigma}$ does not exist. Thus it may always be better to consider and estimate the *normalized* scatter matrix $\mathbf{S}$ which is always defined [29].

TME can be derived as a ML estimator of the shape matrix for the *Angular Central Gaussian* (ACG) distribution (defined in (3)) based on the sample $\boldsymbol{\mathcal{Z}}_n$ [9]. With $\boldsymbol{\mu} = \mathbf{0}$, the sample $\boldsymbol{\mathcal{Z}}_n$ can be written as $(u_1\tilde{\mathbf{x}}_1, \ldots, u_n\tilde{\mathbf{x}}_n)$. Since the scalars $u_1, \ldots, u_n$ are unknown, there is a scaling ambiguity and one can only expect to estimate matrix $\mathbf{S}$ 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:

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

where $\mathbf{x} \in \boldsymbol{\mathcal{S}}^p$, $\Gamma(\cdot)$ is the Gamma function, and $\Gamma(p/2)/(2\pi)^{\frac{p}{2}}$ is the surface area of $\boldsymbol{\mathcal{S}}^p$. The ACG density in (3) represents the *distribution of directions* for samples drawn from a multivariate Gaussian distribution with zero mean and covariance matrix $\mathbf{S}$ [9]. Thus, only the directions of outliers can affect TME's performance but not their magnitude. Given an *i.i.d.* random sample $\boldsymbol{\mathcal{X}}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ from a distribution having the ACG density in (3), the likelihood of $\boldsymbol{\mathcal{X}}_n$ with respect to $\mathbf{S}$ is *proportional* to:

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

Taking $-\log$ of $L(\boldsymbol{\mathcal{X}}_n; \mathbf{S})$ yields the loss function:

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

which will be needed for our next discussions. Taking the derivative of $\boldsymbol{\mathcal{L}}\,(\boldsymbol{\mathcal{X}}_n; \mathbf{S})$ 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 $\mathbf{x}_i \neq \mathbf{0}$, for $i = 1, \ldots, n$ since samples lying at the origin provide no directional information on $\mathbf{S}$. Note that (6) can be interpreted as a *reweighted* covariance estimate. If $n > p(p-1)$, Theorem 1 in [9] states that with probability one, the ML estimate of $\mathbf{S}$ *exists*, corresponds to the solution in (6), and is *unique* up to a positive multiplicative scalar. The solution to (6) can be found using the following fixed point iteration (FPI) algorithm:

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

with $\widehat{\mathbf{S}}_0 = \mathbf{I}$, or any arbitrary initial $\widehat{\mathbf{S}}_0 \in \mathbb{S}_+^p$ [43]. Theorem 2.2 and Corollaries 2.2 & 2.3 in [8] show that if $n > p + 1$ and assuming that every p samples out of $\boldsymbol{\mathcal{X}}_n$ are *linearly independent* with probability one, and that the maximum likelihood estimate of $\mathbf{S}$ exists, then 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.

TME has various attractive properties and is asymptotically optimal under different criteria. In particular, TME is 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 [8]). 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.[3] Various researchers have proposed different flavors of RTME using the spirit of [15] linear shrinkage estimator [6], [27], [28], [29], [31]. In particular, Sun et al. (SBP) [30] proposed the following penalized log-likelihood function to derive a regularized version of TME:

$$\mathcal{L}_{\mathcal{P}}(\mathcal{X}_n;\mathbf{S}) = \mathcal{L}(\mathcal{X}_n;\mathbf{S}) + \beta\mathcal{P}(\mathbf{S}), \tag{8}$$

where $\mathcal{L}(\mathcal{X}_n;\mathbf{S})$ is defined in (5), and $\mathcal{P}(\mathbf{S})$ is a penalty function defined as: $\mathcal{P}(\mathbf{S}) = \mathrm{Tr}(\mathbf{S}^{-1}\mathbf{T}) + \log\det(\mathbf{S})$, with $\beta > 0$ is the regularization parameter (or shrinkage coefficient). Matrix $\mathbf{T} \in \mathbb{S}_p^+$ is a given target matrix with some desirable structural properties (diagonal, banded, Toeplitz, etc.). Letting $\alpha = \beta/(1+\beta)$, the solution to (8) has to satisfy the fixed point equation:

$$\mathbf{S}_n = (1-\alpha)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^\top}{\mathbf{x}_i^\top\mathbf{S}_n^{-1}\mathbf{x}_i} + \alpha\mathbf{T}. \tag{9}$$

Note that $\alpha \in (0,1)$ for any $0 < \beta < \infty$. Starting from an arbitrary $\widehat{\mathbf{S}}_0 \in \mathbb{S}_p^+$, the final solution can be obtained using the *Regularized* FPI (RFPI) algorithm:

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

where $\alpha \in (0,1)$ is the *shrinkage coefficient* that controls the amount of shrinkage applied to scatter matrix $\mathbf{S}$ towards the target matrix $\mathbf{T}$. Theorem 11 and Proposition 13 in [30] establish the necessary and sufficient conditions for the *existence* and *uniqueness* of the solution to Equation (9), while Proposition 18 ensures that the RFPI in (10) converges to the unique solution of (9).

Without loss of generality, if $\mathbf{T} = \mathbf{I}$ and $\alpha = 0$, one restores the unbiased TME in (7), and if $\alpha = 1$ the estimator reduces to the uncorrelated scatter 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 the existence and uniqueness of the estimator still hold [29]. When $p \geq n$, α is expected to be large;[4] however to ensure the *existence and uniqueness* of the estimator, α needs to be strictly greater than $1 - n/p$ [29], [30].

A. RUNTIME ANALYSIS FOR THE RFPI ALGORITHM

The magnitude of α has an impact on the accuracy of the final estimate $\widehat{\mathbf{S}} = \widehat{\mathbf{S}}_T$, as well as on the convergence speed for the RFPI algorithm. In particular, Lemma 1 in [39] gives a result on the *uniform linear convergence* of the algorithm in (10) to a unique solution; for desired accuracy ε, convergence ratio r, and sufficiently large $\alpha > p/n - 1$, *at most* $\lceil \log_{1/r}(1/\varepsilon) \rceil$ iterations are needed for (10) to converge to the unique solution of (9). A preliminary analysis of the RFPI 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 [44]: $\widehat{\mathbf{S}}_t(\alpha) = \mathbf{L}\mathbf{L}^\top$, 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 RFPI algorithm to converge, its total running time complexity will be $O(T(np^2 + p^3))$.

III. OPTIMAL CHOICE OF SHRINKAGE COEFFICIENT α

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 loss function of $\mathcal{X}_n$ with respect to $\mathbf{S}$ in (5). 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 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 $\alpha \in (0,1)$; i.e.

$$\alpha^* = \underset{\alpha\in(0,1)}{\arg\min}\ \mathcal{L}(\mathcal{X}_n;\widehat{\mathbf{S}}(\alpha)). \tag{11}$$

The problem with this direct approach is that $\widehat{\mathbf{S}}(\alpha)$ needs to be computed using the sample $\mathcal{X}_n$. That is, $\mathcal{X}_n$ 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 α.

CV techniques address this problem by splitting the data into two non-overlapping samples; one sample for estimating $\mathbf{S}$ and the other sample for estimating the loss $\mathcal{L}$. Here, we propose to use *Leave-One-Out* CV (LOOCV) 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 the RFPI algorithm in (10), 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:

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

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

$$\mathcal{L}_{CV}(\mathcal{X}_n,\alpha) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathbf{x}_i;\widehat{\mathbf{S}}(\alpha;\mathcal{X}_{n\setminus i})), \tag{13}$$

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

[4] If $p < n$ and the samples are heavy-tailed and not from an elliptical distribution, α is expected to be large as well.

and $\widehat{\mathbf{S}}(\alpha; \mathcal{X}_{n\setminus i})$ is estimated from the points in $\mathcal{X}_{n\setminus i}$ using the RFPI algorithm (10). In practice, one possible approach to solve problem (12) can be using a simple *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 (13); and (*iii*) choose α_j with the minimum $\mathcal{L}_{CV}(\cdot)$.[5] For a *reasonably fine* discretization for the range of α's, this direct estimation approach will yield an estimate for α that is *reasonably close* to its optimal value. For simplicity, we refer to this method for estimating α^* as the *Exact CVL* method.

A. SOME PROPERTIES OF RTME, $\mathcal{L}(\mathcal{X}_n; \mathbf{S})$, AND LOOCV

TME is strongly consistent and asymptotically normal [8]. The sample applies to RTME for a properly chosen $\bar{\alpha} \in (\alpha_{\min} + \epsilon, 1 - \epsilon)$, where $\epsilon > 0$, and $\alpha_{\min} = 1 - n/p$ to ensure existence and uniqueness of $\widehat{\mathbf{S}}_T \in \mathbb{S}_+^p$. 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 [43]. For $\mathbf{S} \in \mathbb{S}_+^p$, the log likelihood function $\mathcal{L}(\mathcal{X}_n; \mathbf{S})$ in (5) is geodesically convex with respect to $\mathbb{S}_+^p$ [38], [45], and properties for this type of likelihood functions has been studied in [46]. In particular, $\mathcal{L}(\mathcal{X}_n; \mathbf{S})$ maintains the three main properties of maximum likelihood estimators: *consistency*, *efficiency*, and *functional invariance*. Thus, *asymptotically*, $\mathcal{L}(\mathcal{X}_n; \mathbf{S})$ will concentrate around its expectation.

The LOOCV estimate is *almost an unbiased* estimate in the following sense [47, Ch. 24]: for fixed p and $\bar{\alpha}$,

$$\mathbb{E}\mathcal{L}_{CV}(\mathcal{X}_n, \bar{\alpha}) = \mathbb{E}\mathcal{L}(\mathcal{X}'_{n-1}; \widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n-1})), \text{ where}$$

the expectations are w.r.t the random samples $\mathcal{X}_n$, $\mathcal{X}'_{n-1}$, $\mathcal{X}_{n-1}$, and $\mathcal{X}'_{n-1} \perp\!\!\!\perp \mathcal{X}_{n-1}$; i.e. $\mathcal{L}_{CV}(\mathcal{X}_n, \bar{\alpha})$ is an estimator for $\mathcal{L}^*(\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n-1}))$ rather than for $\mathcal{L}^*(\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n))$, where

$$\mathcal{L}^*(\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n-1})) = \mathbb{E}[\mathcal{L}(\mathcal{X}'_{n-1}; \widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n-1})) \,|\, \mathcal{X}_{n-1}],$$

$$\text{and } \mathcal{L}^*(\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n)) = \mathbb{E}[\mathcal{L}(\mathcal{X}'_n; \widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n)) \,|\, \mathcal{X}_n].$$

Thus, $\mathcal{L}_{CV}(\mathcal{X}_n, \bar{\alpha})$ converges to the expected risk $\mathcal{L}^*(\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n-1}))$. However, both random quantities $\mathcal{L}^*(\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n-1}))$ and $\mathcal{L}^*(\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n))$ converge with probability one, and asymptotically the difference between $\mathcal{L}^*(\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n))$ and $\mathcal{L}^*(\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n-1}))$ will be negligible. The asymptotic properties of RTME, $\mathcal{L}(\mathcal{X}_n; \mathbf{S})$, and LOOCV, encourage us to postulate the following proposition which will be useful for introducing our approximation approach discussed in Section IV.

Proposition 1 (Stability of RTME): Let $\mathbf{x} = \frac{\mathbf{z}}{\|\mathbf{z}\|}$ be a random vector from the model in (1) s.t. $\mathbf{x}$ is independent of $\mathcal{X}_n$, and let p and $\bar{\alpha}$ be predefined fixed values, where $\bar{\alpha} \in (\alpha_{\min} + \epsilon, 1 - \epsilon)$, and $\epsilon > 0$. Then, under the following asymptotic conditions: (i) the samples in $\mathcal{X}_n$ are *i.i.d.*, (ii) the consistency of RTME, and (iii) the consistency and concentration of $\mathcal{L}(\mathcal{X}_n, \mathbf{S})$, we have that for large values of n, and for any i chosen (randomly) from $i = 1, \ldots, n$, the following holds:

1) $\|\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n) - \widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n\setminus i}))\|_2 = O(1/n)$, and
2) $|\mathcal{L}(\mathbf{x}; \widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n)) - \mathcal{L}(\mathbf{x}; \widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n\setminus i}))| = O(1/n)$,

where $\|\cdot\|_2$ is the matrix operator (spectral) norm.[6]

Proposition (1) asserts, based on an asymptotic argument, that for a fixed p and $\bar{\alpha}$, and as n is increasing, $\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n) \approx \widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n\setminus i})$, or more generally, $\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n) \approx \widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n-1})$; i.e. the estimator for $\mathbf{S}$ is *not too sensitive* to the removal of one sample from $\mathcal{X}_n$. Consequently, the difference between $\mathcal{L}(\mathbf{x}; \widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n))$ and $\mathcal{L}(\mathbf{x}; \widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n\setminus i}))$ will be small for any sample $\mathbf{x}_i$ randomly chosen from $\mathcal{X}_n$, where $i = 1, \ldots, n$. The notion of sensitivity of an estimator with respect to the removal (or replacement) of one sample from $\mathcal{X}_n$ is known as *Algorithmic Stability* and it has been extensively leveraged in learning theory to derive generalization bounds on the risk of various learning algorithms [48], [49]. Our approximation approach introduced in the following section will leverage the previous insight from Proposition (1) to approximate $\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n\setminus i})$ and speedup the computation for the LOOCV estimate in (13).

B. THE COMPUTATIONAL OVERHEAD OF LOOCV

Unfortunately, LOOCV is well known for its substantial computational overhead. For a fixed $\bar{\alpha}$ and for n samples in $\mathcal{X}_n$, LOOCV will make n calls for the RFPI algorithm in order to compute $\mathcal{L}(\mathbf{x}_i, \widehat{\mathbf{S}}(\alpha; \mathcal{X}_{n\setminus i}))$ in the RHS of (13). Thus, for m values of α_j from $(\alpha_1, \ldots, \alpha_m)$, the Exact CVL method in (12) will require mn calls for the RFPI algorithm, which is prohibitive even for moderate values of n. If the RFPI algorithm requires T iterations to converge, then the RFPI algorithm will consume $O(mn.T(np^2 + p^3))$ time from the Exact CVL method in (12), where $O(T(np^2 + p^3))$ is the running time for a single call for the RFPI algorithm.

The above analysis motivates an approximation that reuses a single RTME solution to accelerate the Exact ACVL method. Specifically, our objective in the following section is to reduce the computational complexity from $O(mn.T(np^2 + p^3))$ to $O(m.T(np^2 + p^3))$. The improvement in speed due to this approximation while preserving the accuracy of the estimated α is shown in Fig. 1 for the multivariate Cauchy distribution. In particular, Fig. 1 depicts the *Exact* CVL method vs. the *approximation* developed in the following section in terms of the average CV loss in (13), running time (in seconds), and the optimal α obtained from each method (details in Section V).

IV. APPROXIMATE LOOCV

The approximation approach proposed here is motivated by our observation from Proposition (1) that under the *i.i.d.* assumption on $\mathcal{X}_n$, and for large n, the estimator for $\mathbf{S}$ is not too sensitive to the removal of one sample from $\mathcal{X}_n$; i.e. $\widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_n) \approx \widehat{\mathbf{S}}(\bar{\alpha}; \mathcal{X}_{n-1})$. For a fixed $\bar{\alpha}$, the RFPI algorithm

[5] Note that when $p > n$, and for existence and uniqueness results to hold, α needs to be strictly greater than $1 - n/p$ [29], [30], and hence there is no need to evaluate $\mathcal{L}_{CV}(\cdot)$ for $\alpha \leq 1 - n/p$.

[6] The formal proof for Proposition (1) is omitted due to space limitations.

in (10) can be expressed in a more enlightening form as follows:

$$\widehat{\mathbf{S}}_{t+1}(\bar{\alpha}) = (1-\bar{\alpha})p\left(\frac{1}{n}\sum_{i=1}^{n} w_{t,i}^{-1}\mathbf{x}_i\mathbf{x}_i^\top\right) + \bar{\alpha}\mathbf{T}, \tag{14}$$

where $w_{t,i} = \mathbf{x}_i^\top \widehat{\mathbf{S}}_t^{-1}(\bar{\alpha})\mathbf{x}_i$, and $t = 1, \ldots, T$. The first term in the RHS of (14) involves a weighted sample covariance matrix using the weights $w_{t,i}$ and the RFPI algorithm iteratively estimates these weights until convergence. For initial matrix $\widehat{\mathbf{S}}_0 \in \mathbb{S}_+^p$, let $(\widehat{w}_1, \widehat{w}_2, \ldots, \widehat{w}_n)$ be the optimal weights estimated using $\boldsymbol{\mathcal{X}}_n$ and the RFPI in (14). Then, the *final* estimate for the scatter matrix can be written as:

$$\widehat{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_n) = (1-\bar{\alpha})\frac{p}{n}\sum_{i=1}^{n}\frac{1}{\widehat{w}_i}\mathbf{x}_i\mathbf{x}_i^\top + \bar{\alpha}\mathbf{T}. \tag{15}$$

Similar to (15), using $\bar{\alpha}$ and initial matrix $\widehat{\mathbf{S}}_0$, the *final* estimate $\widehat{\mathbf{S}}$ using $\boldsymbol{\mathcal{X}}_{n\setminus i}$ and the RFPI in (14) will be:

$$\widehat{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_{n\setminus i}) = (1-\bar{\alpha})\frac{p}{n-1}\sum_{j=1, j\neq i}^{n}\frac{1}{\widehat{v}_j}\mathbf{x}_j\mathbf{x}_j^\top + \bar{\alpha}\mathbf{T}, \tag{16}$$

where $(\widehat{v}_1, \ldots, \widehat{v}_{i-1}, \widehat{v}_{i+1}, \ldots, \widehat{v}_n)$ are the optimal weights estimated using $\boldsymbol{\mathcal{X}}_{n\setminus i}$. In terms of computations, and for a fixed $\bar{\alpha} \in (0, 1)$, computing the final estimate $\widehat{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_{n\setminus i})$ for each $i = 1, \ldots, n$ requires invoking the RFPI algorithm n times during the LOOCV procedure. This yields a total running time of $O(nT(np^2 + p^3))$ which is inefficient even for moderate values of n and p.

Suppose that the true scatter matrix $\mathbf{S}^* \in \mathbb{S}_p^+$ is known and $(\mathbf{S}^*)^{-1}$ has been computed. Then, the final estimate $\widehat{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_n)$ in (15) can be *directly* computed without invoking the RFPI algorithm in (14):

$$\widehat{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_n) = \frac{(1-\bar{\alpha})p}{n}\sum_{i=1}^{n}\frac{1}{\widehat{w}_i^*}\mathbf{x}_i\mathbf{x}_i^\top + \bar{\alpha}\mathbf{T}, \tag{17}$$

where $\widehat{w}_i^* = \mathbf{x}_i^\top(\mathbf{S}^*)^{-1}\mathbf{x}_i$. Similarly, using $(\mathbf{S}^*)^{-1}$, the final estimate $\widehat{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_{n\setminus i})$ in (16) can be *directly* computed without invoking the RFPI algorithm in (14):

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

where $\widehat{v}_j^* = \mathbf{x}_j^\top(\mathbf{S}^*)^{-1}\mathbf{x}_j$. Note that both $\widehat{w}_i^*$ in (17) and $\widehat{v}_j^*$ in (18) 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 $j = 1, \ldots, n$. Since $\mathbf{S}^*$ is unknown, we propose to approximate $\widehat{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_{n\setminus i})$ in (18) using the following estimate:

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

$$\widetilde{v}_j = \mathbf{x}_j^\top\widehat{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_n)^{-1}\mathbf{x}_j. \tag{19}$$

That is, we plug in the *Regularized* TME $\widehat{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_n) \in \mathbb{S}_p^+$ from (15) into equation (18) 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 $\widetilde{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_{n\setminus i})$ in (19). Using this plug-in approximation, and for a fixed $\bar{\alpha} \in (0, 1)$, computing $\widetilde{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_{n\setminus i})$ does not require invoking the RFPI algorithm for each $i = 1, \ldots, n$. Instead, the RFPI algorithm will be invoked once to compute $\widehat{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_n)$ in (15), while $\widetilde{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_{n\setminus i})$ in (19) can be directly computed for each $i = 1, \ldots, n$. Using the plug-in approximation in (19), an optimal α can now be computed as follows:

$$\widehat{\alpha}_{CV}^* = \underset{\alpha\in(0,1)}{\arg\min}\ \widetilde{\mathcal{L}}_{CV}(\boldsymbol{\mathcal{X}}_n, \alpha), \text{ where} \tag{20}$$

$$\widetilde{\mathcal{L}}_{CV}(\boldsymbol{\mathcal{X}}_n, \alpha) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(\mathbf{x}_i, \widetilde{\mathbf{S}}(\alpha; \boldsymbol{\mathcal{X}}_{n\setminus i})\right), \tag{21}$$

and $\widetilde{\mathcal{L}}_{CV}(\boldsymbol{\mathcal{X}}_n, \alpha)$ is the *approximate cross-validated loss* (ACVL); and we refer to the problem in (20) as the ACVL method. For m values of α in $(\alpha_1, \ldots, \alpha_m)$, the RFPI algorithm will now consume $O(m * T(np^2 + p^3))$ running time from the ACVL method.

Discussion: The proposed approximation $\widetilde{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_{n\setminus i})$ is motivated by the properties of RTME, $\mathcal{L}(\boldsymbol{\mathcal{X}}_n; \mathbf{S})$, the LOOCV loss, and Proposition (1) in Section III-A. This approximation is expected to be accurate whenever the RTME exhibits *algorithmic stability* with respect to single perturbations, which is empirically observed for: (*i*) moderate to large sample sizes (n); (*ii*) α is away from its boundary values; i.e. $\widehat{\alpha}^* \in (1 - (n/p) + \epsilon, 1 - \epsilon)$, $\epsilon > 0$; and (*iii*) the data are elliptically distributed but with bounded influence points (i.e. the data distribution may be heavy-tailed but without *adversarial* points).

On the other hand, approximation accuracy may degrade in one or more of the following settings: (*i*) *ultra-small* sample regimes; (*ii*) α is very close to, or equal to, its boundary values which puts RTME at its existence boundary ($\alpha \approx 1 - n/p$); or (*iii*) in the presence of extreme leverage (or adversarial) observations. Finally, as with any likelihood-based approach, the approximation may be affected by *model misspecification*; i.e. if the data deviates substantially from elliptical symmetry as in the case of strongly multi-modal or mixture model settings.

V. EMPIRICAL EVALUATION

In this section, we evaluate the performance of the ACVL method on synthetic and real high-dimensional datasets, and compare it with other shrinkage coefficient estimation methods in the literature; in particular the methods proposed in [28], [32], and [15]. For synthetic data, and similar to other works in the literature on RTME [6], [28], [29], [30], [31], [39], we consider the Toeplitz matrix used in [20] to be the population scatter matrix $\mathbf{S}$ for the elliptical RV in (1); i.e. $\mathbf{S} = (s_{i,j}) = \gamma^{|i-j|}$, where $\gamma = \{0.1, 0.5, 0.85\}$. Note that $\mathbf{S}$ approaches a singular matrix when $\gamma \to 1$, and $\mathbf{S}$ approaches the identity matrix when $\gamma \to 0$.

The random quantities u and $\mathbf{y}$ in (1) are stochastically independent. We let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ 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. For this reason, we show here only the plots for the multivariate 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 RFPI algorithms is $\|\widehat{\mathbf{S}} - \mathbf{S}\|_F^2 < \epsilon$, where $\epsilon = 1.0e-9$ is the desired solution accuracy. For Fig. 2, p is set to 500, while n is set to three different values $\{1000, 500, 250\}$ 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 all figures is for the ratio p/n.

Fig. 1 compares the *Exact* CV loss to the *Approximate* CV loss developed in the previous section, for the multivariate Cauchy distribution (which has undefined moments). It can be seen that the Exact CV loss in (13) (solid blue line) and the Approximate CV loss in (21) (solid red line) are *almost identical* in all settings: $p < n$, $p = n$, $p > n$, and for all values of $\gamma = \{0.1, 0.5, 0.85\}$. Also note that the optimal α estimated using the ACVL method (red square) overlaps the optimal α estimated using the Exact CVL (blue circle) in all nine settings. In terms of speedup, the ACVL method is at least $20\times$ faster than the Exact CVL method in all the different settings (see exact runtimes in the legends of Fig. 1).

Fig. 2 compares the shrinkage coefficient estimated using the ACVL method in (20), denoted by $\widehat{\alpha}_{acvl}$, with the shrinkage coefficients estimated from the closed-form expressions in [28, Eq. (13)] (CWH), denoted by $\widehat{\alpha}_{cwh}$, and [32, Eq. (12)] (ZW), denoted by $\widehat{\alpha}_{zw}$. While the methods in [28] and [32] are faster than the ACVL method due to their closed-form expressions, it can be seen that the ACVL method provides more accurate estimates for α especially when $p \geq n$. Also, it can be noticed that the estimates from [28] and [32] tend to *diverge* from the optimal value as p is growing greater than n. The tendency for methods based on asymptotic analysis and RMT results to over/under estimate the value for α is understandable since such methods make explicit assumptions about the data's underlying distribution. This over/under estimation of α leads to larger values of the NMSE as shown in Fig. 2, as well as larger values for the LOOCV NLL loss as demonstrated in the following experiments on real-world high-dimensional datasets.

A. EMPIRICAL VALIDATION ON REAL DATASETS

Tables 1–4 compare the LOOCV NLL loss for the scatter matrices estimated using Ledoit–Wolf (LW) estimator [15], and the RTME with shrinkage coefficients from CWH [28], ZW [32], and the ACVL method in (20).

TABLE 1. Comparison Results for the First 6 (Out of 38) Classes From the Extended Yale B Dataset; $n = 64$, $p = 1024$

Label	LW	CWH	ZW	ACVL
1	5677	5371	5643	**3705**
2	5613	5440	5598	**3706**
3	5768	5470	5749	**3826**
4	5403	5080	5362	**3455**
5	5824	5435	5786	**3716**
6	5797	5460	5761	**3790**

TABLE 2. Comparison Results for the First 6 (Out of 20) Classes From the CIFAR100 *test Set*; $n = 500$, $p = 1024$

Label	LW	CWH	ZW	ACVL
apple	849	866	**846**	**846**
aq.fish	807	810	799	**767**
baby	968	984	967	**932**
bear	810	794	803	**769**
beaver	869	846	859	**812**
bed	1051	1047	1043	**1008**

TABLE 3. Comparison Results for the First 6 (Out of 10) Classes From CIFAR10 *test Set*; $n = 1000$, $p = 1024$

Label	LW	CWH	ZW	ACVL
airp	631	593	612	**590**
auto	913	900	906	**894**
bird	727	705	718	**694**
cat	773	757	772	**755**
deer	769	753	761	**739**
dog	721	702	719	**699**

TABLE 4. Comparison Results for the First 6 (Out of 10) Classes From the USPS Dataset; $p = 256$. Note That n Varies for Each Digit's Class

Label	n	LW	CWH	ZW	ACVL
0	1585	268	259	261	**239**
1	1330	-269	-374	-309	**-475**
2	952	370	366	369	**342**
3	807	336	327	330	**298**
4	795	310	293	301	**249**
5	659	360	357	359	**337**

The comparison between the different estimators was carried out using four real-world high-dimensional datasets: (*i*) the Extended Yale B dataset for face recognition [50]; (*ii*) the test set for the CIFAR100 dataset for object recognition; (*iii*) the test set for the CIFAR10 dataset for object recognition; and (*iv*) the United States Postal Service (USPS) dataset for handwritten digits [51]. Note that the number of samples n varies for each digit's class from the USPS dataset.

The Extended Yale B dataset consists of 2414 frontal-face grayscale (intensity) images for 38 subjects – approx. 64 images per subject – where each image size (height × width) is 192 × 168 pixels. The images were captured under different poses, lighting conditions, and facial expressions. The exact face images are cropped and scaled to 32 × 32 pixels (i.e. $p = 1024$). The CIFAR10 and CIFAR100 datasets consist of colored (RGB) images for ten (10) and one hundred (100) objects, respectively, from different visual categories (trucks, ships, dogs, mountains, frogs, apples, roads, etc.) Each colored image has a size of (height × width × channels) 32 × 32 × 3 which is then converted to a grayscale (intensity) image with a final size of 32 × 32 pixels (i.e. $p = 1024$). [7] The USPS dataset consists of 9298 grayscale images each with a size of 16 × 16pixels (i.e. $p = 256$). The images are obtained from scanning handwritten numerals from envelopes by the U.S. Postal Service and they reflect a wide range of handwriting styles.

Tables 1 to 4 show that scatter matrices estimated using RTME result in lower LOOCV NLL loss compared to those estimated with the LW estimator. The better performance of RTME indicates that multivariate elliptical distributions may be more effective than Gaussian distributions for modeling high–dimensional data with unknown empirical distributions. Regarding shrinkage coefficients for RTME, the ACVL method consistently achieves lower LOOCV NLL loss than the methods in [28] and [32] across all cases. This supports the earlier observation that inaccurate estimation of the coefficient α increases LOOCV NLL loss and may compromise the performance of downstream inferential tasks.

VI. CONCLUDING REMARKS

Robust estimation of high-dimensional covariance matrices from empirical data remains a significant challenge, particularly when $p \geq n$. This work focuses on TME, which provides accurate and efficient robust estimation of the scatter matrix for samples drawn from elliptical distributions with heavy tails, under the condition that $n \gg p$. Because TME is undefined when $p \geq n$, several studies have introduced regularized versions of TME, where the performance depends critically on the choice of the shrinkage coefficient $\alpha \in (0, 1)$. Our work here builds on these efforts by proposing an alternative method for estimating the optimal α for RTME. Unlike existing methods, the proposed approach utilizes efficient computation and the available finite sample to estimate a near-optimal α for RTME. The key computational component is the Approximate LOOCV NLL loss for the estimated scatter matrix as a function of α (20). The resulting procedure, named the ACVL method, demonstrated competitive performance in experiments with both high–dimensional synthetic and real-world datasets.

The proposed approximation relies on asymptotic algorithmic stability for RTME; specifically, that the estimator for scatter matrix $\mathbf{S}$ is not too sensitive to the removal of one sample from $\boldsymbol{\mathcal{X}}_n$; $\widehat{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_n) \approx \widehat{\mathbf{S}}(\bar{\alpha}; \boldsymbol{\mathcal{X}}_{n-1})$. k-folds cross-validation (KFCV) is another popular technique for model selection that is computationally more efficient than LOOCV, but also *less accurate* than LOOCV. KFCV may seem a potential candidate to replace the LOOCV in our proposed learning framework. Unfortunately, from an algorithmic stability standpoint, KFCV will require a more stringent stability assumption for the estimator of scatter matrix $\mathbf{S}$ [49]; specifically, that the estimator for $\mathbf{S}$ is not too sensitive to the removal of $m = n/k$ samples from $\boldsymbol{\mathcal{X}}_n$, where $k > 1$ is the number of folds used for KFCV. A deeper theoretical understanding of RTME's stability and its interaction with KFCV remains an open and promising avenue for future research work, with potential impact beyond robust covariance matrix estimation.

A natural direction for future work is to establish a deeper theoretical foundation for the stability properties underlying the proposed approximation. In particular, it is of interest to:

i) Extend the current asymptotic results to derive asymptotic equivalence between the minimizers of the Approximate and Exact LOOCV objectives, thereby guaranteeing consistency of the proposed ACVL method.
ii) Derive explicit *finite-sample bounds* quantifying RTME's sensitivity to single-sample perturbations.

Beyond the specific setting considered here, the approximation principle introduced in this work suggests a broader applicability. In particular, it will be interesting to extend the stability-based argument to other fixed-point covariance estimators such as structured, sparse, or constrained robust M-estimators, and more broadly to the selection of regularization parameters and hyperparameters for various classes of learning algorithms. This could yield general and computationally efficient methodologies for hyperparameter selection in high-dimensional robust estimation problems.

[7] See Matlab's `rgb2 Gy()` function for more details.

REFERENCES

[1] I. Jolliffe, *Principal Component Analysis*, New York, NY, USA: Springer, 2002.
[2] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition (Probability and Statistics Series)*. Hoboken, NJ, USA: Wiley, 2004.
[3] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.
[4] H. Markowitz, "Portfolio selection," *J. Finance*, vol. 7, no. 1, pp. 77–91, 1952.
[5] E. Cabana, R. E. Lillo, and H. Laniado, "Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators," *Stat. Papers*, vol. 62, no. 4, pp. 1583–1609, Nov. 2019.
[6] A. Wiesel, "Unified framework to regularized covariance estimation in scaled Gaussian models," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 29–38, Jan. 2012.
[7] R. A. Maronna, "Robust *M*-estimators of multivariate location and scatter," *Ann. Statist.*, vol. 4, no. 1, pp. 51–67, 1976.
[8] D. E. Tyler, "A distribution-free *M*-estimator of multivariate scatter," *Ann. Statist.*, vol. 15, no. 1, pp. 234–251, 1987.
[9] D. E. Tyler, "Statistical analysis for the angular central Gaussian distribution on the sphere," *Biometrika*, vol. 74, no. 3, pp. 579–589, Sep. 1987.
[10] R. A. Maronna and V. J. Yohai, "Robust and efficient estimation of multivariate scatter and location," *Comput. Statist. Data Anal.*, vol. 109, pp. 64–75, 2017.
[11] A. Wiesel and T. Zhang, "Structured robust covariance estimation," *Foundations Trends Signal Process.*, vol. 8, no. 3, pp. 127–216, 2015.

[12] W. James and C. Stein, "Estimation with quadratic loss," in *Proc. Fourth Berkeley Symp. Math. Statist. Probability*, 1961, pp. 361–379.

[13] K. Dipak Dey and C. Srinivasan, "Estimation of a covariance matrix under stein's loss," *Ann. Statist.*, vol. 13, no. 4, pp. 1581–1591, 1985.

[14] J. Michael Daniels and R. E. Kass, "Shrinkage estimators for covariance matrices," *Biometrics*, vol. 57, no. 4, pp. 1173–1184, 2001.

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

[16] L. R. Haff, "Empirical Bayes estimation of the multivariate normal covariance matrix," *Ann. Statist.*, vol. 8, no. 3, pp. 586–597, 1980.

[17] A. d'Aspremont, O. Banerjee, and L. El Ghaoui, "First-order methods for sparse covariance selection," *SIAM J. Matrix Anal. Appl.*, vol. 30, pp. 56–66, 2008.

[18] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing L_1-penalized log-determinant divergence," *Electron. J. Statist.*, vol. 5, pp. 935–980, 2011.

[19] M. Pourahmadi, *High-Dimensional Covariance Estimation (Probability and Statistics Series)*. Hoboken, NJ, USA: Wiley, 2013.

[20] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *Ann. Statist.*, vol. 36, no. 6, pp. 2577–2604, 2008.

[21] N. El Karoui, "Spectrum estimation for large dimensional covariance matrices using random matrix theory," *Ann. Statist.*, vol. 36, no. 6, pp. 2757–2790, 2008.

[22] T. Cai and H. H. Zhou, "MiniMax estimation of large covariance matrices under $/ell_1$-norm," *Statistica Sinica*, vol. 20, no. 4, pp. 1305–1337, 2012.

[23] F. Han and H. Liu, "Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution," *Bernoulli*, vol. 23, no. 1, pp. 23–57, 2017.

[24] P. Huber, Ed., *Robust Statistics (Probability and Mathematical Statistics Series)*. Hoboken, NJ, USA: Wiley, 1981.

[25] R. F. Elvezio, M. H. Peter, J. R. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions (Probability and Statistics Series)*. Hoboken, NJ, USA: Wiley, 2011.

[26] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*. Hoboken, NJ, USA: Wiley-Interscience, 1982.

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

[28] 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, Sep. 2011.

[29] 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, Nov. 2014.

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

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

[32] T. Zhang and A. Wiesel, "Automatic diagonal loading for Tyler's robust covariance estimator," in *Proc. 2016 IEEE Stat. Signal Process. Workshop*, 2016, pp. 1–5.

[33] 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 Trans. Signal Process.*, vol. 61, no. 23, pp. 5819–5829, Dec. 2013.

[34] 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.

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

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

[37] Q. Hoarau, A. Breloy, G. Ginolhac, A. M. Atto, and J. M. Nicolas, "A subspace approach for shrinkage parameter selection in undersampled configuration for regularised Tyler estimators," in *Proc IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 3291–3295.

[38] L. Dümbgen and D. E. Tyler, "Geodesic convexity and regularized scatter estimators," 2016 *arXiv:1607.05455*.

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

[40] K. Abou-Moustafa, "Shrinkage coefficient estimation for regularized tyler's *M*-estimator. a leave one out approach," in *Proc. IEEE Inf. Theory Workshop*, 2023, pp. 335–340.

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

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

[43] J. T. Kent and D. E. Tyler, "Maximum likelihood estimation for the wrapped cauchy distribution," *J. Appl. Statist.*, vol. 15, no. 2, pp. 247–254, 1988.

[44] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: The Johns Hopkins University Press, 1996.

[45] A. Wiesel, "Geodesic convexity and covariance estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6182–6189, Dec. 2012.

[46] S. Amari and H. Nagaoka, *Methods of Information Geometry, AMS Translations of Mathematical Monographs*, vol. 191, New York, NY, USA: Oxford University Press, 2000.

[47] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Berlin, Germany: Springer, 1996.

[48] M. Kearns and D. Ron, "Algorithmic stability and sanity-check bounds for leave-one-out cross-validation," *Neural Computation*, vol. 11, no. 6, pp. 1427–1453, Aug. 1999.

[49] K. Abou-Moustafa and C. Szepesvári, "An exponential tail bound for L_q stable learning rules," in *Proc. 30th Int. Conf. Algorithmic Learn. Theory, in Mach. Learn. Res.*, 2019, vol. 98, pp. 31–63.

[50] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[51] D. Newman, S. Hettich, C. Blake, and C. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: www.ics.uci.edu/mlearn/MLRepository.html