Regularized M-estimators of scatter matrix

Esa Ollila, *Member, IEEE,* and David E. Tyler

***Abstract*—In this paper, a general class of regularized M-estimators of scatter matrix are proposed which are suitable also for low or insufficient sample support (small n and large p) problems. The considered class constitutes a natural generalization of M-estimators of scatter matrix (Maronna, 1976) and are defined as a solution to a penalized M-estimation cost function. Using the concept of geodesic convexity, we prove the existence and uniqueness of the regularized M-estimators of scatter and the existence and uniqueness of the solution to the corresponding M-estimating equations under general conditions. Unlike the non-regularized M-estimators of scatter, the regularized estimators are shown to exist for any data configuration. An iterative algorithm with proven convergence to the solution of the regularized M-estimating equation is also given. Since the conditions for uniqueness do not include the regularized versions of Tyler's M-estimator, necessary and sufficient conditions for their uniqueness are established separately. For the regularized Tyler's M-estimators, we also derive a simple, closed form and data dependent solution for choosing the regularization parameter based on shape matrix matching in the mean squared sense. Finally, some simulations studies illustrate the improved accuracy of the proposed regularized M-estimators of scatter compared to their non-regularized counterparts in low sample support problems. An example of radar detection using normalized matched filter (NMF) illustrate that an adaptive NMF detector based on regularized M-estimators are able to maintain accurately the preset CFAR level.**

***Index Terms*—Geodesic convexity, Complex elliptically symmetric distributions, M-estimator of scatter, Regularization, Robustness, Normalized matched filter**

I. Introduction

MANY data mining and classic multivariate analysis techniques require an estimate of the covariance matrix or some nonlinear function of it, e.g., the inverse covariance matrix or its eigenvalues/eigenvectors. Given an i.i.d. sample $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{C}^p$ from a centered, i.e., $\mathbb{E}[\mathbf{z}] = \mathbf{0}$, (unspecified) p-variate distribution $\mathbf{z} \sim F$, the *sample covariance matrix* (SCM) $\hat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^{\mathrm{H}} \in \mathbb{C}^{p\times p}$ is the most commonly used estimator of the unknown covariance matrix $\mathbf{R} = \mathbb{E}[\mathbf{z}\mathbf{z}^{\mathrm{H}}]$. However, in high-dimensional (HD) problems, there are many cases that the SCM simply can not be computed, is completely corrupted, or is inaccurate. For example, low sample support (LSS) (i.e., p is of the same magnitude as n) is a commonly occurring problem in diverse HD data analysis problems such as chemometrics and medical imaging. In the case of *insufficient sample support (ISS)*, i.e., $p > n$, the inverse of the SCM can not be computed. Thus, for example, classic beamforming techniques such as MVDR beamforming or the adaptive normalized matched filter cannot be realized since they require an estimate of the inverse covariance matrix.

Robust estimation is also a key property in HD data analysis problems. Partly because outliers are more difficult to glean from HD data sets by conventional techniques, but also due to an increase of impulsive measurement environments and outliers in practical sensing systems. The SCM is well-known to be vulnerable to outliers and to be a highly inefficient estimator when the samples are drawn from a heavy-tailed non-Gaussian distribution. HD data poses additional problems and difficulties since most robust estimators such as M-estimators of scatter matrix [17] can not be computed in ISS scenarios, or are equivalent to the SCM [29].

In this paper, we address this issue and propose a general class of regularized M-estimators of scatter matrix. This class provides practical and actionable estimators of the covariance (scatter) matrix even in the problematic ISS case. The proposed class constitutes a natural generalization of M-estimators of scatter [17] and their complex-valued generalizations [18], [22], and are defined as a solution to a penalized M-estimation cost function that includes a fixed regularization parameter $\alpha > 0$. We prove the existence and uniqueness of the regularized M-estimators of scatter and the existence and uniqueness of the solution to the corresponding M-estimating equations under general conditions. Our derivations are based on the concept of geodesic convexity which has been previously utilized in [30], [33] in studying the uniqueness of the non-regularized M-estimators of scatter and in [31] which studied regularized Tyler's M-estimator of scatter matrix using a particular scale invariant geodesically convex penalty function. An iterative algorithm with proven convergence to the solution of the regularized M-estimating equation is also given. Our class include as a special case, when using a tuned cost function corresponding to p-variate complex normal samples, the commonly used *shrinkage* estimator of the sample covariance matrix

$$\hat{\mathbf{R}}_{\alpha,\beta} = \beta\hat{\mathbf{R}} + \alpha\mathbf{I}. \tag{1}$$

which, in finance literature, is commonly called the Ledoit-Wolf shrinkage estimator [16]. In a recent paper [7] in the field of adaptive beamforming it was termed the general linear combination (GLC) estimator, the term which we adopt in this paper. It should be noted however that in [7], [16], $\hat{\mathbf{R}}_{\alpha,\beta}$ was not proposed as a minimizer to any optimization problem.

Our general conditions on uniqueness do not apply to the regularized versions of Tyler's [27] M-estimator and hence this estimator is treated separately, with necessary and sufficient conditions being established to ensure the uniqueness of solution for the penalized Tyler's cost function. Special

E. Ollila is with the Department of Signal Processing and Acoustics, Aalto University, Espoo, P.O. Box 13000, FIN-00076 Aalto; e-mail: esa.ollila@aalto.fi (see http://users.spa.aalto.fi/esollila/). D.E. Tyler is with the Department of Statistics & Biostatistics, Rutgers – The State University of New Jersey, Piscataway NJ 08854, USA; e-mail: dtyler@rci.rutgers.edu.
Manuscript received May 11th, 2014

cases of regularized versions of Tyler's M-estimator have also been recently studied in [24] under more strict conditions on the sample, and also in [2], [4], but not in the context as a solution to a penalized M-estimation cost function. Estimation of the regularization parameter using the expected likelihood approach was proposed in [1], [3] for the regularized Tyler's M-estimator of [2], [4], whereas [6] based their analysis on random matrix theory (both n and p are large). For the regularized Tyler's M-estimators, we also derive a simple, closed form and data dependent solution to compute the regularization parameter α based on shape matrix matching in the mean squared sense. We illustrate the usefulness of the regularized M-estimators of scatter in radar detection application using normalized matched filter.

Finally, we note that although our derivations in the paper are for complex-valued case, they generalize in an straightforward manner to real-valued case as well.

The paper is organized as follows. Section II reviews complex elliptically symmetric (CES) distributions and the maximum likelihood (ML) and M-estimators of the scatter matrix parameters of the CES distributions [22]. Section III then introduces the penalized M-estimation cost function. The stationary points are shown to be solutions to shrinkage type M-estimation equations. Interpretation of regularization parameters are discussed and specific examples of regularized M-estimators are given. In Section IV, general conditions are presented to ensure the uniqueness of solution, with the proof of uniqueness being based on the concept of geodesic convexity. The regularized Tyler's M-estimator is then considered in Section V and numerical examples are given in Section VI. Some of the proofs are reserved for the Appendix.

Notations: Let $\mathcal{H}(p)$ denote the class positive definite Hermitian (PDH) $p \times p$ matrices, $|\mathbf{A}|$ the determinant of a square matrix $\mathbf{A}$. Furthermore, $\|\cdot\|$ (resp. $\|\cdot\|_1$) denotes the ℓ_2-norm (resp. ℓ_1-norm) defined as $\|\mathbf{A}\|^2 = \mathrm{Tr}(\mathbf{A}^{\mathrm{H}}\mathbf{A}) = \sum_i \sum_j |a_{ij}|^2$ (resp. $\|\mathbf{A}\|_1 = \sum_i \sum_j |a_{ij}|$) for any $m \times n$ matrix $\mathbf{A}$.

II. Preliminaries

A. Elliptical distributions

A continuous symmetric random vector (r.v.) $\mathbf{z} \in \mathbb{C}^p$ has a centered *complex elliptically symmetric (CES) distribution* [22] if its p.d.f. is of the form:

$$f(\mathbf{z}) = C_{p,g}|\mathbf{\Sigma}|^{-1} g\big(\mathbf{z}^{\mathrm{H}}\mathbf{\Sigma}^{-1}\mathbf{z}\big),$$

where $\mathbf{\Sigma} \in \mathcal{H}(p)$ is the unknown parameter, called the *scatter matrix*, $g : \mathbb{R}_0^+ \to \mathbb{R}^+$ is a fixed function called the *density generator* and $C_{p,g} > 0$ is a normalizing constant ensuring that $f(\mathbf{z})$ integrates to one. We denote this case by $\mathbf{z} \sim \mathrm{CE}_p(\mathbf{0}, \mathbf{\Sigma}, g)$. If the covariance matrix $\mathbf{R} = \mathbb{E}[\mathbf{z}\mathbf{z}^{\mathrm{H}}]$ of $\mathbf{z}$ exists, then

$$\mathbf{R} = c \cdot \mathbf{\Sigma} \quad \text{(for some } c > 0\text{)}.$$

For example, when $g(t) = \exp(-t)$, one obtains the p-variate complex normal (CN) distribution, denoted $\mathbf{z} \sim \mathbb{C}\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$; In this case, $\mathbf{R} = \mathbf{\Sigma}$. For a detailed account on properties of CES distributions, we refer the reader to [22]. Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ denote an i.i.d. random sample from an unspecified p-variate CES distribution as stated above.

The maximum likelihood estimator (MLE) of scatter matrix, denoted $\hat{\mathbf{\Sigma}}$, minimizes the negative log-likelihood function (divided by n)

$$\mathcal{L}(\mathbf{\Sigma}) = \frac{1}{n}\sum_{i=1}^{n} \rho(\mathbf{z}_i^{\mathrm{H}}\mathbf{\Sigma}^{-1}\mathbf{z}_i) - \ln|\mathbf{\Sigma}^{-1}| \tag{2}$$

where $\rho(t) = -\ln g(t)$. More appropriate notation would be $\mathcal{L}_n(\mathbf{\Sigma}|\rho)$ to emphasize the dependence on ρ and the sample. Critical points are then solutions to the estimating equation

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{z}_i^{\mathrm{H}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}_i)\mathbf{z}_i\mathbf{z}_i^{\mathrm{H}} \tag{3}$$

where $u = \rho' = -g'/g$.

B. M-estimators of scatter

M-estimators of scatter are generalizations of the ML-estimators of the scatter matrix of an elliptical distribution. They can be defined by allowing a general ρ functions in (2), not necessarily related to any elliptical density g, in which case we refer to (2) as a general cost function. The function ρ is usually chosen so that the corresponding weight function $u = \rho'$ is non-negative, continuous and non-increasing. Equation (3) is then referred to as an M-estimating equation. Some examples of M- and ML-estimators are given below.

SCM (the Gaussian MLE). In the Gaussian case, $\rho(t) = t$ and $u(t) = \rho'(t) \equiv 1$, so eq. (2) becomes

$$\mathcal{L}(\mathbf{\Sigma}) = \mathrm{Tr}(\hat{\mathbf{R}}\mathbf{\Sigma}^{-1}) - \ln|\mathbf{\Sigma}^{-1}|$$

where $\hat{\mathbf{R}}$ denotes the SCM. The (well-known) *unique* minimizer (assuming $n \geq p$) of this function is the sample covariance matrix, i.e., $\hat{\mathbf{\Sigma}} = \hat{\mathbf{R}}$.

Complex Tyler's [27] M-estimator is based on the functions

$$\rho(t) = p\ln t \quad \text{and} \quad u(t) = \rho'(t) = \frac{p}{t}.$$

Note that this ρ-function is *not* related to any elliptical density and the optimization problem (2) is now non-convex. Nevertheless, the estimator is actionable: a unique solution (up to a scale) exists under mild conditions and the global solution can be computed via simple fixed-point iterations; see [22], [23], [27]. It should be noted that for Tyler's M-estimator, the summations in both (2) and (3) are taken only over $\mathbf{z}_i \neq \mathbf{0}$. In the radar community, Tyler's M-estimator is often referred to as a fixed-point estimator, and it is known to admit numerous ML-interpretations as shown in [5], [9], [11], [21], [28] in the real and complex cases.

Complex Huber's M-estimator is based on a weight function of the form [19] $u(s) = bu_c(t)$, where

$$u_c(t) = \begin{cases} 1, & \text{for } t \leq c^2 \\ c^2/t, & \text{for } t > c^2 \end{cases}$$

where $c > 0$ is a tuning constant that controls robustness/efficiency of the method and $b > 0$ is a scaling constant usually chosen so that the resulting M-estimator is consistent to the covariance matrix for Gaussian data. As a consequence the value of the scaling constant b depends on c. See [10] for more details. Note that for $c \to \infty$, Huber's estimator approaches the SCM (i.e., constant weight function), and for $c \to 0$, the estimator approaches Tyler's M-estimator.

III. Regularized M-estimators of scatter matrix

To stabilize the optimization problem an additive penalty term $\alpha \cdot \mathcal{P}(\mathbf{\Sigma})$ can be introduced to the cost function (2), where $\alpha \geq 0$ denotes a fixed regularization parameter. A popular focus in the literature has been to enforce sparsity on the *precision matrix* $\mathbf{K} = \mathbf{\Sigma}^{-1}$ by using ℓ_1-penalty function

$$\mathcal{P}_{\ell_1}(\mathbf{\Sigma}) = \|\mathbf{\Sigma}^{-1}\|_1 \tag{4}$$

as is done in the real-valued case in [8], [32]. The use of the ℓ_1-penalty, though, to help enforce a sparse precision matrix is dependent on the cost function (2) being convex in $\mathbf{\Sigma}^{-1}$, which holds whenever $\rho(t)$ itself is convex. However, robust M-estimates of scatter typically have decreasing weight functions $u(t)$ and hence concave ρ-functions.

In this paper, we take a different approach and focus on a penalty function of the form

$$\mathcal{P}^*(\mathbf{\Sigma}) = \|\mathbf{\Sigma}^{-1/2}\|^2 = \mathrm{Tr}(\mathbf{\Sigma}^{-1}).$$

Notice that

$$\mathrm{Tr}(\mathbf{\Sigma}^{-1}) = \sum_{j=1}^{p} \frac{1}{\lambda_j(\mathbf{\Sigma})},$$

where $\lambda_j(\mathbf{\Sigma})$'s denote the ordered eigenvalues of $\mathbf{\Sigma}$. Thus the penalty term restricts $\frac{1}{\lambda_j(\mathbf{\Sigma})}$ from growing without bound; this is necessary in the ill-conditioned ISS case ($n < p$). In general, our *penalized cost function* is of the form

$$\mathcal{L}_\alpha(\mathbf{\Sigma}) = \frac{1}{n}\sum_{i=1}^{n} \rho(\mathbf{z}_i^{\mathrm{H}}\mathbf{\Sigma}^{-1}\mathbf{z}_i) - \ln|\mathbf{\Sigma}^{-1}| + \alpha\mathcal{P}(\mathbf{\Sigma}), \tag{5}$$

where $\alpha \geq 0$ is a (fixed) regularization parameter. For the case $\mathcal{P}(\mathbf{\Sigma}) = \mathcal{P}^*(\mathbf{\Sigma})$ this becomes

$$\mathcal{L}^*_\alpha(\mathbf{\Sigma}) = \frac{1}{n}\sum_{i=1}^{n} \rho(\mathbf{z}_i^{\mathrm{H}}\mathbf{\Sigma}^{-1}\mathbf{z}_i) - \ln|\mathbf{\Sigma}^{-1}| + \alpha\mathrm{Tr}(\mathbf{\Sigma}^{-1}) \tag{6}$$

As will be illustrated below the parameter α can be best described as *ridge (or sphering)* parameter.

Let $\hat{\mathbf{\Sigma}}$ denote the minimizer of $\mathcal{L}^*_\alpha(\mathbf{\Sigma})$. The solution $\hat{\mathbf{\Sigma}}$ naturally depends on α but this is not made explicit for notational convenience. It is easy to verify using matrix differential rules that a critical point of the penalized cost function (6) is a solution to

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{z}_i^{\mathrm{H}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}_i)\mathbf{z}_i\mathbf{z}_i^{\mathrm{H}} + \alpha\mathbf{I} \tag{7}$$

which is weighted and diagonally loaded form of the classic M-estimating equation obtained when $\alpha = 0$. Expressing the regularized M-estimating equation in the form

$$\mathbf{I} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{z}_i^{\mathrm{H}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}_i)\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}_i\mathbf{z}_i^{\mathrm{H}} + \alpha\hat{\mathbf{\Sigma}}^{-1},$$

and then taking the trace shows that the solution $\hat{\mathbf{\Sigma}}$ must satisfy

$$\alpha\mathrm{Tr}(\hat{\mathbf{\Sigma}}^{-1}) = p - \frac{1}{n}\sum_{i=1}^{n} \psi(\mathbf{z}_i^{\mathrm{H}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}_i) \tag{8}$$

where $\psi(t) = tu(t)$.

GLC estimators. A class of regularized SCM can be obtained by considering the cost functions of the form $\rho(t) = \beta t$, where $\beta > 0$ is a fixed scalar. In this case, the penalized cost function (6) simplifies to the form

$$\mathcal{L}^*_\alpha(\mathbf{\Sigma}) = \mathrm{Tr}\{(\beta\hat{\mathbf{R}} + \alpha\mathbf{I})\mathbf{\Sigma}^{-1}\} - \ln|\mathbf{\Sigma}^{-1}|$$

where $\hat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i^{\mathrm{H}}$ denotes the SCM. The unique minimizer $\hat{\mathbf{\Sigma}}$ of the function above is easily shown to widely used GLC estimator (1), i.e., $\hat{\mathbf{\Sigma}} = \hat{\mathbf{R}}_{\alpha,\beta}$. For $\beta = 1$, the solution is the diagonally loaded SCM, $\hat{\mathbf{R}}_\alpha = \hat{\mathbf{R}} + \alpha\mathbf{I}$. The interpretation of the GLC estimator as a solution to an optimization problem (6) differs from the motivation for the GLC estimator given in [16] or [7]. Note that the eigenvalues of $\hat{\mathbf{R}}_{\alpha,\beta}$ are $\hat{\lambda}_i = \beta\hat{\lambda}_{\hat{\mathbf{R}},i} + \alpha$, where $\hat{\lambda}_{\hat{\mathbf{R}},i}$, $i = 1,\ldots,p$ denote the eigenvalues of $\hat{\mathbf{R}}$. Thus α can be viewed as a ridge parameter as it provides a ridge down the diagonal and guarantees a non-singular solution. It can be also described as a *spherizing parameter* since the larger the value of α, the more "spherical" the solution (i.e., as α gets larger, the solution is shrunk towards the scaled identity matrix $\alpha\mathbf{I}$).

Regularized Tyler's M-estimators. Penalization of Tyler's M-estimator, i.e., choosing $\rho(t) = p\log t$ and hence $u(t) = p/t$, is not possible since for this case $\psi(t) = p$, and so the right hand side of (8) is zero. Alternatively, for some fixed $0 < \beta < 1$, consider the function $\rho(t) = p\beta\log t$, which gives the weight function $u(t) = p\beta/t$. The corresponding regularized M-estimating equations (7) are then given by

$$\hat{\mathbf{\Sigma}} = \frac{p\beta}{n_*}\sum_{i=1,\mathbf{z}_i\neq\mathbf{0}}^{n} \frac{\mathbf{z}_i\mathbf{z}_i^{\mathrm{H}}}{\mathbf{z}_i^{\mathrm{H}}\hat{\mathbf{\Sigma}}^{-1}\mathbf{z}_i} + \alpha\mathbf{I}, \tag{9}$$

where $n_* = \#\{\mathbf{z}_i \neq 0; i = 1,\ldots,n\}$. Hereafter, when using this estimator, we assume without loss of generality that $n_* = n$. Note that the solution $\hat{\mathbf{\Sigma}}$ in (9) depends on $\mathbf{z}_i$ only through $\mathbf{z}_i/\|\mathbf{z}_i\|$, and so $\hat{\mathbf{\Sigma}}$ has the same the distribution-free property over elliptical distributions as the unregularized Tyler scatter matrix. That is, when $\mathbf{z} \sim \mathrm{CE}_p(\mathbf{0},\mathbf{\Sigma},g)$, the distribution of $\mathbf{z}/\|\mathbf{z}\|$, and consequently the distribution of $\hat{\mathbf{\Sigma}}$ does not depend on the function g.

A curious property of the regularized Tyler's M-estimators are that their shapes do not depend on the penalization tuning parameter α. That is, for a given value of $0 < \beta < 1$, suppose we consider two different values of α, say α_1 and α_2, and let $\hat{\mathbf{\Sigma}}_1$ and $\hat{\mathbf{\Sigma}}_2$ represent the respective solutions to (9). It then easily follows that

$$\hat{\mathbf{\Sigma}}_1 = \frac{\alpha_1}{\alpha_2}\cdot\hat{\mathbf{\Sigma}}_2 \tag{10}$$

and so, for any fixed $0 < \beta < 1$, the regularized Tyler's M-estimators are proportional to one another as α varies. Consequently, when the main interest is on estimation of the covariance matrix or scatter matrix parameter up to a scale, as is the case in most applications, one can set without loss of generality $\alpha = 1-\beta$, or equivalently $\beta = 1-\alpha$, when using a regularized Tyler's M-estimator. For this choice, the constraint (8) becomes simply $\mathrm{Tr}(\hat{\mathbf{\Sigma}}^{-1}) = p$.

Remark 1. In general, for a given ρ-function, say $\rho_1(t)$, a class of ρ-functions can be generated by defining $\rho_\beta(t) = \beta\rho_1(t)$ for $\beta > 0$. The parameter β then represents an additional tuning constant which can be used to help obtain desirable properties of the estimator.

Remark 2. It readily follows from its definition, that the regularized M-estimators of scatter $\hat{\mathbf{\Sigma}}$ are *unitary equivariant*. That is, if $\hat{\mathbf{\Sigma}}$ denotes the solution to the penalized cost function $\mathcal{L}^*_\alpha(\mathbf{\Sigma})$ in (6) based on the data set $\mathbf{z}_i$, $i = 1,\dots,n$, then for any given unitary matrix $\mathbf{U}$, the estimator based on the transformed data set $\mathbf{z}_i^\star = \mathbf{U}\mathbf{z}_i$, $i = 1,\dots,n$ is given by

$$\hat{\mathbf{\Sigma}}^\star = \mathbf{U}\hat{\mathbf{\Sigma}}\mathbf{U}^{\mathrm{H}} \tag{11}$$

Note that non-regularized M estimators are affine equivariant, i.e., for this case (11) holds for any non-singular $\mathbf{U}$.

IV. Geodesic convexity, uniqueness and algorithm

In this section, we show under general conditions that there exists a unique minimizer to the penalized likelihood or cost function given by (6). Hereafter, it is assumed that the function $\rho(t)$ satisfies the following condition.

Condition 1. *The function $\rho(t)$ is nondecreasing and continuous for $0 < x < \infty$. Also, the function $r(x) = \rho(e^x)$ is convex for $-\infty < x < \infty$*

If the function $\rho(t)$ in differentiable, then the above condition holds if and only if the weight function $u(t) \geq 0$ and the function $\psi(t) = tu(t)$ is nondecreasing. It readily follows that Huber's and Tyler's M-estimators as well as the Gaussian MLE satisfies Condition 1.

The concept of geodesic convexity for functions of PDH matrices plays a key role in our proof of uniqueness. This concept has been previously utilized in [30], [33] in studying the uniqueness of the non-regularized M-estimates of scatter and in [31] in the case of regularized Tyler's cost function. A review of geodesic convexity for positive definite matrices can be found in the aforementioned papers as well as in [26], wherein further references can be found. We briefly review here some important results.

Rather than treating the class $\mathcal{H}(p)$ as a convex cone in $\mathbb{C}^p$ and using notions from complex Euclidean geometry, one can treat $\mathcal{H}(p)$ as a differentiable Riemannian manifold with the geodesic path from $\mathbf{\Sigma}_0 \in \mathcal{H}(p)$ to $\mathbf{\Sigma}_1 \in \mathcal{H}(p)$ being

$$\mathbf{\Sigma}_t = \mathbf{\Sigma}_0^{1/2}\left(\mathbf{\Sigma}_0^{-1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_0^{-1/2}\right)^t\mathbf{\Sigma}_0^{1/2} \text{ for } t \in [0,1]. \tag{12}$$

Note that $\mathbf{\Sigma}_t \in \mathcal{H}(p)$ for $0 \leq t \leq 1$ and consequently $\mathcal{H}(p)$ is said to form a *geodesically convex set*. A function $h : \mathcal{H}(p) \to \mathbb{R}$ is then a *geodesically convex function* if

$$h(\mathbf{\Sigma}_t) \leq (1-t)\ h(\mathbf{\Sigma}_0) + t\ h(\mathbf{\Sigma}_1) \text{ for } t \in (0,1). \tag{13}$$

If the inequality is strict, then h is said to be geodesically strictly convex. In the $p = 1$ dimensional real setting, geodesic convexity/strict convexity is equivalent to the function $h(e^x)$ being convex/strictly convex in $x \in \mathbb{R}$. Thus, Condition 1 presumes $\rho(t)$ to be geodesically convex.

The concept of geodesic convexity enjoys properties similar to those of convexity in complex Euclidean space. In particular, if h is geodesically convex on $\mathcal{H}(p)$ than any local minimum is a global minimum. Furthermore, if a minimum is obtained in $\mathcal{H}(p)$ then the set of all minimums form a geodesically convex subset of $\mathcal{H}(p)$. If h is geodesically strictly convex and a minimum is obtained in $\mathcal{H}(p)$, then it is a unique minimum.

The following key result is given in [33] for real positive definite symmetric matrices, although it also holds for $\mathcal{H}(p)$. We omit the proof for the complex case since it is analogous to the proof for the real case given in [33].

Lemma 1. *If $\rho(t)$ satisfies Condition 1, then the cost function $\mathcal{L}(\mathbf{\Sigma})$ in (2) is geodesically convex in $\mathbf{\Sigma} \in \mathcal{H}(p)$. In addition, if $r(x)$ is strictly convex and $span\{\mathbf{z}_1,\dots,\mathbf{z}_n\} = \mathbb{C}^p$, then $\mathcal{L}(\mathbf{\Sigma})$ is geodesically strictly convex in $\mathbf{\Sigma} \in \mathcal{H}(p)$.*

Recall that when using the notion of convexity in complex Euclidean space the cost function $\mathcal{L}(\mathbf{\Sigma})$ is convex in $\mathbf{\Sigma}^{-1} \in \mathcal{H}(p)$, but not in $\mathbf{\Sigma} \in \mathcal{H}(p)$, whenever $\rho(t)$ is a convex function. This includes the well studied Gaussian case $\rho(t) = t$. As shown below, geodesic convexity has the interesting property that if $\mathcal{L}(\mathbf{\Sigma})$ is geodesically convex in $\mathbf{\Sigma} \in \mathcal{H}(p)$, then it is also geodesically convex in $\mathbf{\Sigma}^{-1} \in \mathcal{H}(p)$.

From Lemma 1, we readily obtain the following corollary, which follows since the sum of two geodesically convex functions is easily seen to be geodesically convex, and the sum of a geodesically convex function and a geodesically strictly convex function is geodesically strictly convex.

Corollary 1. *For $\rho(t)$ satisfying Condition 1, if $\mathcal{P}(\mathbf{\Sigma})$ is geodesically convex/strictly convex in $\mathbf{\Sigma} \in \mathcal{H}(p)$, then the penalized cost function $\mathcal{L}_\alpha(\mathbf{\Sigma})$ in (5) is geodesically convex/strictly convex in $\mathbf{\Sigma} \in \mathcal{H}(p)$ respectively.*

As Lemma 2 below shows, Corollary 1 applies to the penalty function of interest here, i.e., to $\mathcal{P}^*(\mathbf{\Sigma}) = \mathrm{Tr}(\mathbf{\Sigma}^{-1})$. Before proceeding, some further results and notations are reviewed. For Hermitian matrices $\mathbf{A}$ and $\mathbf{B}$ of the same order, the partial ordering $\mathbf{A} \leq \mathbf{B}$ or $\mathbf{A} < \mathbf{B}$ holds if and only if $\mathbf{B} - \mathbf{A}$ is positive semi-definite or positive definite, respectively. The matrix $\mathbf{\Sigma}_{1/2}$ can be viewed as the geometric mean of $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ [26], and as in the case of positive real numbers, it is known to be less than the arithmetic mean in the following sense,

$$\mathbf{\Sigma}_{1/2} \leq (\mathbf{\Sigma}_0 + \mathbf{\Sigma}_1)/2, \tag{14}$$

with equality holding if and only if $\mathbf{\Sigma}_0 = \mathbf{\Sigma}_1$. It readily follows from its definition (12) that for $\mathbf{K} = \mathbf{\Sigma}^{-1}$

$$\mathbf{K}_t = \mathbf{K}_0^{1/2}\left(\mathbf{K}_0^{-1/2}\mathbf{K}_1\mathbf{K}_0^{-1/2}\right)^t\mathbf{K}_0^{1/2} = \mathbf{\Sigma}_t^{-1}, \tag{15}$$

and consequently (14) also holds to $\mathbf{\Sigma}^{-1}$. Equation (15) together with the definition of geodesic convexity shows that geodesic convexity in $\mathbf{\Sigma}$ implies geodesic convexity in $\mathbf{\Sigma}^{-1}$.

Taking the trace on both side of (14) when applied to $\mathbf{\Sigma}^{-1}$ then gives

$$\mathrm{Tr}(\mathbf{\Sigma}_{1/2}^{-1}) < \left\{\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}) + \mathrm{Tr}(\mathbf{\Sigma}_1^{-1})\right\}/2,$$

for $\mathbf{\Sigma}_0 \neq \mathbf{\Sigma}_1$. That is, $\mathrm{Tr}(\mathbf{\Sigma}^{-1})$ is midpoint geodesically strictly convex in $\mathbf{\Sigma}$. As with convex functions, midpoint geodesic strict convexity along with $\mathrm{Tr}(\mathbf{\Sigma}^{-1})$ being continuous in $\mathbf{\Sigma} \in \mathcal{H}(p)$ is sufficient to imply geodesically strict convexity and hence we obtain our desired result.

Lemma 2. *The penalty term $\mathcal{P}^*(\mathbf{\Sigma}) = \mathrm{Tr}(\mathbf{\Sigma}^{-1})$ is geodesically strictly convex in $\mathbf{\Sigma} \in \mathcal{H}(p)$.*

We note that another interesting geodesically convex penalty function was proposed by Wiesel [31, Proposition 3]. Wiesel's penalty has a specific property of being scale invariant.

To this point, it has been shown that under the stated conditions on ρ, the regularized loss function (6) is geodesically strictly convex. To show that it has a unique minimum in $\mathcal{H}(p)$, and consequently to show the regularized M-estimating equation (7) admits a unique solution, it only needs to be shown that the minimum of (6) occurs in the interior of $\mathcal{H}(p)$. The following lemma shows that this holds and consequently implies the subsequent theorem.

Lemma 3. *If $\rho(t)$ is bounded below, then $\mathcal{L}_\alpha^*(\mathbf{\Sigma}) \to \infty$ as $\mathbf{\Sigma} \to \partial\mathcal{H}(p)$, i.e., the boundary of $\mathcal{H}(p)$.*

Proof: Since $\rho(t)$ is bounded below, it only needs to be shown that if $\mathbf{\Sigma} \to \partial\mathcal{H}(p)$ then

$$-\ln|\mathbf{\Sigma}^{-1}| + \alpha\mathrm{Tr}(\mathbf{\Sigma}^{-1}) = \sum_{j=1}^{p}\left(\frac{\alpha}{\lambda_j(\mathbf{\Sigma})} + \ln\lambda_j(\mathbf{\Sigma})\right) \to \infty.$$

However, $\mathbf{\Sigma} \to \partial\mathcal{H}(p)$ if and only if $\lambda_1(\mathbf{\Sigma}) \to \infty$ and/or $\lambda_p(\mathbf{\Sigma}) \to 0$. In either case, $\alpha/\lambda + \ln\lambda \to \infty$ and so the lemma is established. ■

Theorem 1. *If $\rho(t)$ is bounded below and satisfies Condition 1, then the penalized cost function $\mathcal{L}_\alpha^*(\mathbf{\Sigma})$ in* (6) *has a unique minimum in $\mathcal{H}(p)$. Furthermore, if $\rho(t)$ is also differentiable, then the minimum corresponds to the unique solution $\hat{\mathbf{\Sigma}} \in \mathcal{H}(p)$ to the regularized M-estimating equation* (7).

Remark 3. The existence and uniqueness of the regularized M-estimates *do not require any conditions on the sample* $\mathbf{z}_1, \ldots, \mathbf{z}_n$ for any $n \geq 1$. This is in contrast to the non-regularized M-estimates which requires a bound on the proportion of the data that can lie in any subspace [13]. Furthermore, non-regularized M-estimates exist and are unique for sparse samples, i.e. when $p < n$, whereas non-regularized M-estimators require $n \geq p$.

The regularized M-estimating equation (7) gives rise to the fixed point algorithm stated in Theorem below. The proof of convergence, given in the Appendix, is similar to the convergent proof for the non-regularized M-estimators used in [12]

Theorem 2. *Assume that $\rho(t)$ is continuously differentiable, satisfies Condition 1 and that $u(t) = \rho'(t)$ is non-increasing. If the M-estimating equation* (7) *has a unique solution $\hat{\mathbf{\Sigma}}$, then the iterations*

$$\hat{\mathbf{\Sigma}}_{k+1} = \frac{1}{n}\sum_{i=1}^{n} u(\mathbf{z}_i^{\mathrm{H}}\hat{\mathbf{\Sigma}}_k^{-1}\mathbf{z}_i)\mathbf{z}_i\mathbf{z}_i^{\mathrm{H}} + \alpha\mathbf{I}, \qquad (16)$$

for $k = 0, 1, \ldots$, converges to the solution of (7) *for any initial value $\hat{\mathbf{\Sigma}}_0 \in \mathcal{H}(p)$.*

Note that conditions for uniqueness of the regularized M-estimators are given in Theorem 1. For Tyler's M-estimator. the conditions for uniqueness are given in next Section, in Theorems 3 and 4

V. The Regularized Tyler's M-estimators

A. Existence and uniqueness

Important cases for which Lemma 3 and Theorem 1 do not hold are the regularized Tyler's M-estimators since for these cases $\rho(t) = p\beta\ln t$ is not bounded below. Hence these cases requires special treatment.

Theorem 3. *For $\rho(t) = p\beta\ln t$, with a fixed $0 \leq \beta < 1/p$, the penalized cost function $\mathcal{L}_\alpha^*(\mathbf{\Sigma})$ in* (6)*, for a given $\alpha > 0$, has a unique minimum in $\mathcal{H}(p)$, with the minimum being obtained at the unique solution $\hat{\mathbf{\Sigma}} \in \mathcal{H}(p)$ to* (9).

Proof: Since $\mathbf{z}_i^{\mathrm{H}}\mathbf{\Sigma}^{-1}\mathbf{z}_i \geq \mathbf{z}_i^{\mathrm{H}}\mathbf{z}_i/\lambda_1(\mathbf{\Sigma})$, it follows that

$$\mathcal{L}_\alpha^*(\mathbf{\Sigma}) \geq C - p\beta\ln\lambda_1(\mathbf{\Sigma}) + \sum_{j=1}^{p}\left(\frac{\alpha}{\lambda_j(\mathbf{\Sigma})} + \ln\lambda_j(\mathbf{\Sigma})\right),$$

where $C = \frac{p\beta}{n}\sum_{i=1}^{n}\ln(\mathbf{z}_i^{\mathrm{H}}\mathbf{z}_i)$ does not depend on $\mathbf{\Sigma}$. Again, the lemma follows since for any $c > 0$, $\alpha/\lambda + c\ln\lambda \to \infty$ as $\lambda \to 0$ or as $\lambda \to \infty$. ■

Theorem 3 does not require any condition on the sample. However, to extend this result to $1/p \leq \beta < 1$, the following Condition A is a sufficient condition and the following Condition B is a necessary conditions. These conditions holds for $n/p > \beta$ whenever the sample is in "general position", which occurs with probability one when sampling from a continuous complex multivariate distribution. Note that the sufficient Condition A and the necessary Condition B only differ when equality in the conditions is possible. Consequently, there is little room for improvement on Conditions A.

Condition A. For any subspace $\mathcal{V}$ of $\mathbb{C}^p$, $1 \leq dim(\mathcal{V}) < p$, the inequality $\frac{\#\{\mathbf{z}_i \in \mathcal{V}\}}{n} < \frac{dim(\mathcal{V})}{p\beta}$ holds.

Condition B. For any subspace $\mathcal{V}$ of $\mathbb{C}^p$, $1 \leq dim(\mathcal{V}) < p$, the inequality $\frac{\#\{\mathbf{z}_i \in \mathcal{V}\}}{n} \leq \frac{dim(\mathcal{V})}{p\beta}$ holds.

We then have the following general result, the proof of which can be found in the Appendix.

Theorem 4. *Suppose $\rho(t) = p\beta\ln t, \alpha > 0$ and $0 \leq \beta < 1$.*

a) *If condition A holds, then* (6) *has a unique minimum in $\mathcal{H}(p)$, with the minimum being obtained at the unique solution $\hat{\mathbf{\Sigma}} \in \mathcal{H}(p)$ to* (9).
b) *If condition B does not hold, then* (6) *does not have a minimum in $\mathcal{H}(p)$, and* (9) *has no solution in $\mathcal{H}(p)$.*

The existence and uniqueness of the regularized Tyler's M-estimator, for the case $\beta = 1 - \alpha$, has also been established in

[24], but only under the condition that the data are in general position and $n > p$. For such samples, Conditions A and B are automatically satisfied.

Remark 4. A related regularized M-Tyler's estimator was proposed by Abramovich and Spencer [2] as the limit of the algorithm

$$\mathbf{\Sigma}_{k+1} \leftarrow (1-\alpha)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{z}_i\mathbf{z}_i^{\mathrm{H}}}{\mathbf{z}_i^{\mathrm{H}}\mathbf{V}_k^{-1}\mathbf{z}_i} + \alpha\mathbf{I}$$
$$\mathbf{V}_{k+1} \leftarrow p\mathbf{\Sigma}_{k+1}/\mathrm{Tr}(\mathbf{\Sigma}_{k+1}),$$

where $\alpha \in (0,1)$ is a fixed regularization parameter. This algorithm represents a diagonally loaded (DL) version of the fixed-point algorithm given for Tyler's M-estimator. Later, it was shown by Chen, Wiesel and Hero [4], that the recursive algorithm above converges to a unique solution, referred to hereafter as the CWH estimator, regardless of the initialization. Here, convergence means convergence in $\mathbf{V}_k$ and not necessarily in $\mathbf{\Sigma}_k$. It is not clear whether this estimator can be derived as a solution to a penalized cost function.

B. Estimation of the regularization parameter

Let us define a scale measure of $\mathbf{\Sigma} \in \mathcal{H}_p$ as

$$\tau(\mathbf{\Sigma}) = p/\mathrm{Tr}(\mathbf{\Sigma}^{-1}), \tag{17}$$

with $\mathbf{V} = \mathbf{\Sigma}/\tau(\mathbf{\Sigma})$ being the respective shape matrix. Note that his implies the shape matrix is standardized so that $\mathrm{Tr}(\mathbf{V}^{-1}) = p$. Recall that the regularized Tyler's M-estimator $\hat{\mathbf{\Sigma}}$ using $\beta = 1-\alpha$, $\alpha \in (0,1)$, represents an estimator of the shape matrix $\mathbf{V}$ since it satisfies $\mathrm{Tr}(\hat{\mathbf{\Sigma}}^{-1}) = p$. We now focus on this particular estimator, i.e., the regularized Tyler's M-estimator with $\beta = 1-\alpha$, and derive an oracle estimator of the parameter α, or equivalently β, using a MSE criterion for similarity in shape.

Let $\mathbf{\Sigma}_\alpha$ denote a *clairvoyant estimator* of $\hat{\mathbf{\Sigma}}$ given $\mathbf{\Sigma}_0 = \mathbf{V}$,

$$\mathbf{\Sigma}_\alpha = (1-\alpha)\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{z}_i\mathbf{z}_i^{\mathrm{H}}}{\mathbf{z}_i^{\mathrm{H}}\mathbf{\Sigma}_0^{-1}\mathbf{z}_i} + \alpha\mathbf{I}. \tag{18}$$

This clairvoyant estimator corresponds to the first step of the algorithm (16), with $u(t) = (1-\alpha)p/t$, if we take $\mathbf{\Sigma}_0$ as the initial value. Since we are only interested in $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_\alpha$ up to a scale, our aim is to then choose α such that $\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha$ is as close as possible to being a scaled copy of an identity matrix. Thus, we define the oracle shrinkage parameter α_o as the minimizer of the following MSE criterion

$$\alpha_o = \arg\min_{\alpha} \mathbb{E}\big[\|\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha - \tfrac{1}{p}\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha)\mathbf{I}\|^2\big].$$

A similar approach has been used in [4] for deriving an oracle value for the shrinkage parameter α for the CWH estimator.

Theorem 5. *The oracle estimator α_0 is given by*

$$\alpha_o = \frac{p\mathrm{Tr}(\mathbf{\Sigma}_0) - 1}{p\mathrm{Tr}(\mathbf{\Sigma}_0) - 1 + n(p+1)\{p^{-1}\mathrm{Tr}(\mathbf{\Sigma}_0^{-2}) - 1\}}. \tag{19}$$

In the real case, the oracle estimator is

$$\alpha_{o,R} = \frac{p-2+p\mathrm{Tr}(\mathbf{\Sigma}_0)}{p-2+p\mathrm{Tr}(\mathbf{\Sigma}_0) + n(p+2)\{p^{-1}\mathrm{Tr}(\mathbf{\Sigma}_0^{-2}) - 1\}}$$

Since $\mathbf{\Sigma}_0$ is unknown, we estimate α_o in (19) by the simple plug-in estimator

$$\hat{\alpha}_o = \frac{p\mathrm{Tr}(\hat{\mathbf{\Sigma}}) - 1}{p\mathrm{Tr}(\hat{\mathbf{\Sigma}}) - 1 + n(p+1)\{p^{-1}\mathrm{Tr}(\hat{\mathbf{\Sigma}}^{-2}) - 1\}}, \tag{20}$$

with $\hat{\mathbf{\Sigma}}$ being Tyler's M-estimator normalized so that $\mathrm{Tr}(\hat{\mathbf{\Sigma}}^{-1}) = p$ whenever $n \geq p$. For $n < p$, one can choose $\hat{\mathbf{\Sigma}}$ to be a regularized Tyler's estimator with $\beta < n/p$ and $\alpha = 1-\beta$.

VI. Numerical examples

A. Simulations study

In our first simulation set-up, the covariance matrix $\mathbf{\Sigma}$ is a real-valued correlation matrix (i.e., components z_i have unit variances, real and imaginary parts are uncorrelated) of Toeplitz form

$$[\mathbf{\Sigma}]_{ij} = \rho^{|i-j|}, \quad \rho \in (0,1).$$

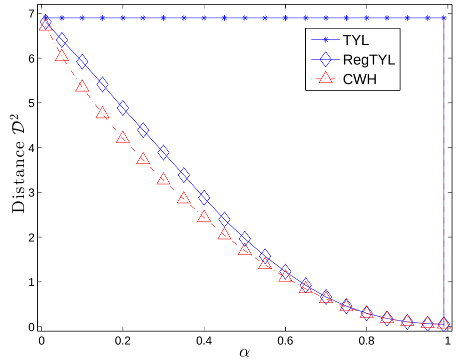
Note that when ρ is close to 0, then $\mathbf{\Sigma}$ is close to an identity matrix and when ρ tends to 1, $\mathbf{\Sigma}$ tends a singular matrix of rank 1. To assess the performance of the estimators, we use the distance measure

$$\mathcal{D}^2 \equiv \mathcal{D}^2(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}) = \|\{p/\mathrm{Tr}(\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}})\}\,\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} - \mathbf{I}\|^2$$

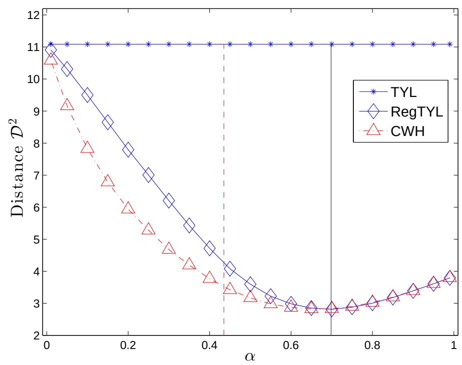
which measures the ability of the estimator $\hat{\mathbf{\Sigma}}$ to estimate the scatter matrix $\mathbf{\Sigma}$ up to its scale, since $\mathcal{D}^2(c_1\mathbf{\Sigma}, c_2\hat{\mathbf{\Sigma}}) = \mathcal{D}^2(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}})$ for any $c_1, c_2 > 0$ and $\mathcal{D}^2 = 0$ if $\mathbf{\Sigma} \propto \hat{\mathbf{\Sigma}}$. In the simulation we consider the regularized Tyler's M-estimators, taking without loss of generality $\beta = 1-\alpha$, and the CWH estimators discussed in Remark 4. We also compare the results to the (non-regularized) Tyler's M-estimator. The samples $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are generated from $\mathbb{C}\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, where the dimension of the data is $p = 12$. Recall that the simulation results would be the same if we sampled from any centered CES distribution, including compound Gaussian distributions, since the distribution of $\mathbf{z}_i/\|\mathbf{z}_i\|$ is the same for all such distributions.

Figure 1 depicts the graphs of $\mathcal{D}^2$ averaged over 1000 MC-trials as a function of α for the CWH estimators and the regularized Tyler's M-estimators (RegTYL) for the cases $\rho = 0.01, 0.5, 0.8$ and sample size is $n = 24$. Also included in Figure 1 is the non-regularized Tyler's M-estimator of scatter (TYL), which corresponds to RegTYL when $\alpha = 0$. Figure 2 gives the corresponding results for a sample size of $n = 48$. In both figures, the solid vertical line depicts the value of the oracle estimator α_o for the regularized Tyler's M-estimator given by Theorem 5 and the dotted vertical line depicts the value of the oracle estimator α_o^{CWH} of CWH estimator given by [4, Theorem 3].

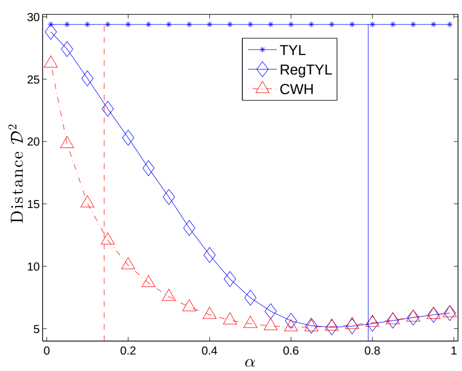
The simulation results show the following. First, although the performance of the regularized Tyler's M-estimator (RegTYL) tends to the performance of Tyler's M-estimator as $\alpha \to 0$, an observation also illustrated in [24], the performance of the CWH estimator can still be quite different from that of Tyler's M-estimator even for $\alpha \approx 0$. Second, the shape distance curves are very different for RegTYL and CWH



(a) $\rho = 0.05$

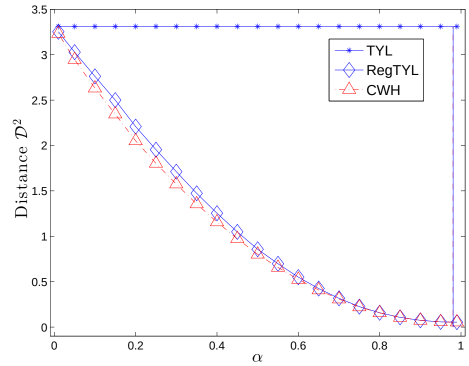


(b) $\rho = 0.5$

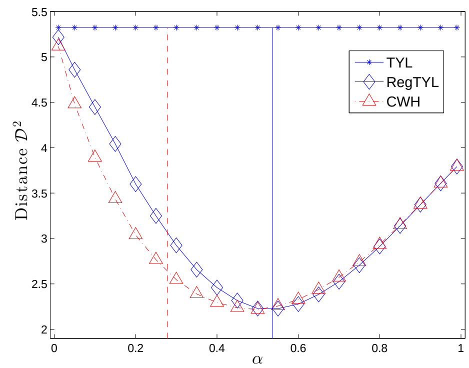


(c) $\rho = 0.8$

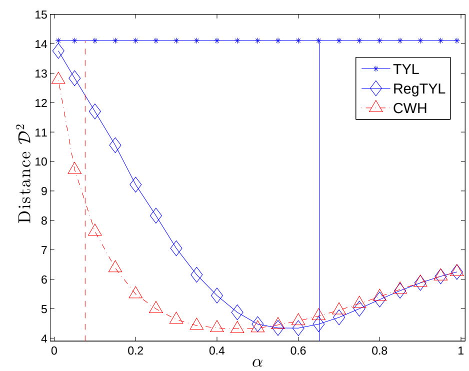
Fig. 1. Distance $\mathcal{D}^2$ of Tyler's M-estimator (TYL), regularized Tyler's M-estimator (RegTYL) and CWH estimator as a function of the shrinkage parameter α. Results for different correlation matrix $\mathbf{\Sigma}$ given by $\rho = 0.05, 0.5, 0.8$ are given from top to bottom. The dimension was $p = 12$, sample length was $n = 24$ and the results are averages of 1000 MC trials. The solid (resp. dotted) vertical line gives the oracle estimator α_0 of RegTYL estimator in Theorem 5 (resp. of CWH estimator in [4, Theorem 3]).



(a) $\rho = 0.05$



(b) $\rho = 0.5$



(c) $\rho = 0.8$

Fig. 2. Distance $\mathcal{D}^2$ for shrinkage estimators RegTYL and CWH as a function of the shrinkage parameter α. Set-up is as in Figure 1, but the sample size is twice larger $n = 48$.

estimators for the cases $\rho = 0.5$ and $\rho = 0.8$. Only for the case $\rho = 0.05$ are they similar. In general, though, the value of α play a different role in RegTYL and CWH, and so comparing the two estimators for the same α is not particularly meaningful. Third, of primary interest is the performance of the oracle estimators for RegTYL, obtained at α_o, and the performance of the CWH oracle estimator, obtained at say α_0^{CWH}. The figures illustrate that these two shrinkage generalizations of Tyler's scatter matrix provide fairly different estimators of the scatter matrix, and that RegTYL oracle estimator outperforms the CWH oracle estimator when $\mathcal{D}^2$ is used as a criterion. In all cases, for $\mathbf{\Sigma}$ having a Toeplitz form, the shrinkage estimators (RegTYL and CHW) outperform the (non-regularized) Tyler's M-estimator (TYL). For the case $\rho = 0.05$, which corresponds to $\mathbf{\Sigma}$ being close to an identity matrix, both α_0 and α_0^{CWH} are approximately one, as expected, i.e., both estimators are being heavily shrunk towards a scaled identity matrix.

B. Adaptive normalized matched filter example

We address the problem of detecting a known complex *signal vector* $\mathbf{p}$ in *received data* $\mathbf{z} = \gamma\mathbf{p} + \mathbf{c}$, where $\mathbf{c}$ represents the unobserved complex *noise* (clutter) r.v. and $\gamma \in \mathbb{C}$ is a signal parameter. The signal-absent vs. signal-present problem can then be expressed as $H_0 : |\gamma| = 0$ vs. $H_1 : |\gamma| > 0$. We assume that $\mathbf{c}$ follows a centered CES distribution with a positive definite hermitian (PDH) scatter matrix parameter $\mathbf{\Sigma}$. For this problem, we consider the widely used *normalized matched filter (NMF) detector* [9], [25]

$$\Lambda \equiv \Lambda(\mathbf{z}; \mathbf{p}, \mathbf{\Sigma}) = \frac{|\mathbf{p}^{\mathrm{H}}\mathbf{\Sigma}^{-1}\mathbf{z}|^2}{(\mathbf{z}^{\mathrm{H}}\mathbf{\Sigma}^{-1}\mathbf{z})(\mathbf{p}^{\mathrm{H}}\mathbf{\Sigma}^{-1}\mathbf{p})} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda \tag{21}$$

It is well known that the distribution of $\mathbf{\Lambda}$ under H_0 is $\mathrm{Beta}(1, p-1)$, i.e., it is distribution-free under the class of CES distributions [15], [21]. Hence the detector is CFAR under various commonly used clutter models (including the K-distribution, t-distribution, inverse Gaussian distribution). Thus, to obtain a probability of false alarm (PFA) equal to a desired level P_{FA} (e.g., $P_{\mathrm{FA}} = 0.01$), the rejection threshold λ can be set as

$$P_{\mathrm{FA}} = \Pr(\Lambda > \lambda | H_0) = (1-\lambda)^{p-1} \tag{22}$$

or $\lambda = 1 - {P_{\mathrm{FA}}}^{1/(p-1)}$. See e.g. [21].

In practice $\mathbf{\Sigma}$ is unknown and an *adaptive NMF detector* $\hat{\mathbf{\Lambda}}$ is obtained by replacing $\mathbf{\Sigma}$ by an estimate $\hat{\mathbf{\Sigma}}$ as in [5], [9], [14], [15]. Note that the detector requires an estimate of $\mathbf{\Sigma}$ only up to a scale since $\mathbf{\Lambda} = \mathbf{\Lambda}(\mathbf{z}; \mathbf{p}, c\mathbf{\Sigma})$ for all $c > 0$. Since the sample size n is rarely large compared to the dimension p (LSS/ISS cases) in many applications, the adaptive NMF detector $\hat{\mathbf{\Lambda}}$ based on the sample covariance matrix or any M-estimator of scatter does not retain the CFAR property since an M-estimator $\hat{\mathbf{\Sigma}}$ (although consistent) can be a highly inaccurate estimator in LSS/ISS cases.

We illustrate that an adaptive NMF detector based on the regularized Tyler's M-estimator (RegTYL) using $\beta = 1 - \hat{\alpha}_o$ and $\alpha = \hat{\alpha}_o$, $\hat{\alpha}_o$ given by (20), is able to accurately retain the same CFAR property as the theoretical NMF based on the true scatter matrix $\mathbf{\Sigma}$ whereas the non-regularized Tyler's M-estimator (TYL) performs poorly due to the small sample size. In addition, the following shrinkage estimators are included in the study: the GLC, which again refers to $\hat{\mathbf{R}}_{\alpha,\beta}$ in (1), with α and β being estimated as proposed in [7, cf. Eq.'s (32) and (33)], and the CWH estimator of Remark 4 using the plug-in oracle estimator $\hat{\alpha}_o^{\text{CWH}}$ as proposed in [4, cf. Eq.'s (13) and (14)]. For each MC trial, the simulated data consists of the received data $\mathbf{z}$ (used as input to NMF detector) and the secondary data $\mathbf{z}_1, \ldots, \mathbf{z}_n$ (used as input to estimate $\hat{\mathbf{\Sigma}}$). The data sets are generated as i.i.d. random samples from a $p = 8$ variate K-distribution $\mathbb{C}K_{p,\nu}(\mathbf{0}, \mathbf{\Sigma})$ with $\nu = 4.5$. Since the chosen K-distribution is not heavy-tailed, the GLC estimator is also expected to produce reliable estimates. This would not be the case for choices of ν closer to 0. Using $10,000$ trials, the empirical P_{FA} (the proportion of incorrect rejections) was calculated for a fixed threshold λ. The true scatter matrix $\mathbf{\Sigma}$ differed from trial to trial and was generated randomly for each trial data set as follows: We first generated a random complex unitary $p \times p$ matrix $\mathbf{P}$ and a diagonal matrix $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_p)$, with the d_i's arising from independent $Unif(0,1)$ distributions. Using the EVD, the scatter matrix $\mathbf{\Sigma}$ is then taken to be $\mathbf{\Sigma} = \mathbf{PDP}^{\mathrm{H}}$. Since the null distribution of Λ is invariant to the signal vector $\mathbf{p}$, the signal vector can be be chosen arbitrarily. In our simulations, $\mathbf{p}$ is fixed at $\mathbf{p} = (1, \exp(\jmath\pi), \ldots, \exp(\jmath(p-1)\pi))$.

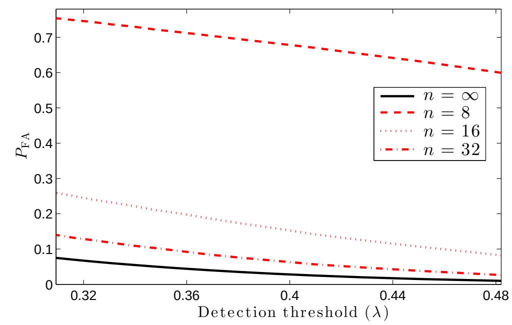
Given that the detector is invariant to the scale of $\mathbf{\Sigma}$, using a $Unif(0,1)$ distribution for the eigenvalues d_i is equivalent to using a $Unif(0,b)$ distribution for some fixed $b > 0$. Also, it is worth noting that due to the unitary equivariance (cf. Remark 2) of $\hat{\mathbf{\Sigma}}$, the simulation results do not depend on how the orthogonal matrices $\mathbf{P}$ are generated, and they would have been the same even if $\mathbf{P}$ had been set to $\mathbf{I}$ in every trial.

Figure 3 depicts empirical the PFA curves of the adaptive detectors. The solid curves ($n = \infty$) depict the theoretical PFA curve (22) for NMF Λ with known $\mathbf{\Sigma}$. As can be seen in Figure 3(a), when the detector is based on Tyler's M-estimator and the sample length is small $n = 8, 16, 32$, there exists a remarkably large gap between the observed PFA and the desired (theoretical) PFA. All of the shrinkage estimators are performing very well illustrating their usefulness in practical applications. RegTYL estimator has slightly better performance than others in the sense that it is able to maintain the empirical PFA very close to the theoretical (desired) PFA for sample lengths $n = 8, 16, 32$ considered. Note that same graphs would be obtained (on the average) for the TYL, RegTYL and CWH estimators if the simulation samples are drawn from any other CES distribution due to distribution-free property of these estimators. This is not true, though, for the GLC estimator whose performance depends on the underlying CES distribution. Furthermore, if the shape parameter ν of the K-distribution is close to zero, then the performance of GLC estimator degrades severely whereas the performance of RegTYL and CWH estimators remain unaffected.

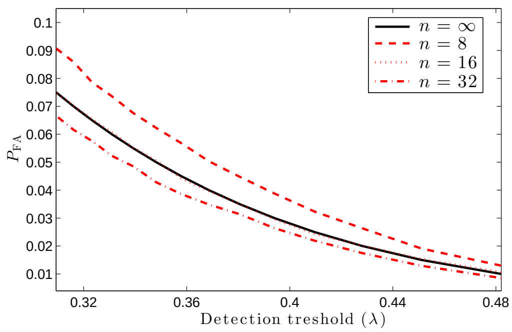
VII. Conclusions

A general class of regularized M-estimators was proposed that are suitable also for small n and large p problems. The considered class was defined as a solution to a penalized M-estimation cost function that depends on a regularization parameter $\alpha \geq 0$ which determines the shrinkage intensity to $\mathbf{I}$. General conditions for uniqueness of the solution were established using the concept of geodesic convexity. Remarkably, the regularized M-estimators do not require any conditions to be placed on the sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$ for any $n \geq 1$. Furthermore, the estimators are actionable: an iterative algorithm with proven convergence to the solution of the regularized M-estimating equation was provided. For the regularized Tyler's M-estimator, necessary and sufficient conditions for existence and uniqueness of the penalized Tyler's cost function were established separately and a closed form (data dependent) choice for the regularization parameter was derived. We also showed that in the special case of using a tuned Gaussian cost function, the unique solution to the penalized likelihood function is given by the widely used Ledoit-Wolf [16] (also called GLC [7]) shrinkage estimator of the sample covariance matrix.

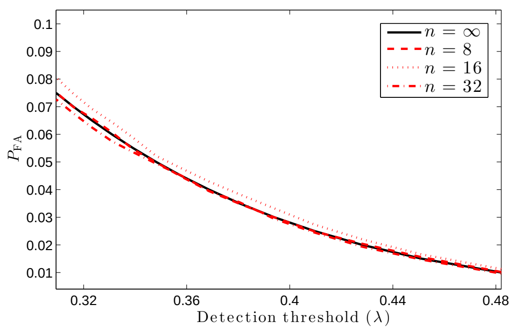
Finally, numerical examples illustrated the usefulness of the proposed estimators. In the signal detection problem using the adaptive normalized matched filter, the regularized covariance matrix estimators were accurately maintaining the



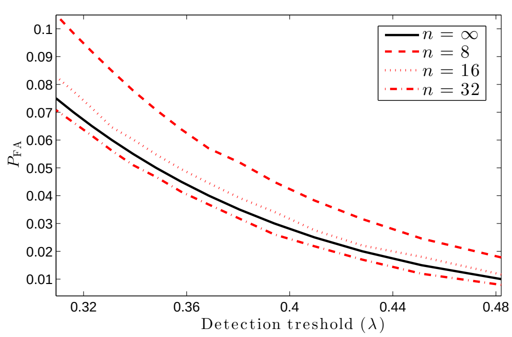
(a) TYL estimator



(b) GLC estimator using $\hat{\alpha}$ and $\hat{\beta}$



(c) Reg-TYL estimator using $\hat{\alpha}_o$



(d) CWH estimator using $\hat{\alpha}_o^{\text{CWH}}$

Fig. 3. Empirical P_{FA} for adaptive detector employing different scatter matrix estimators under K-distributed clutter with $\nu = 4.5$ and different sample lengths n of the secondary data. The dimension $m = 8$ and the clutter covariance matrix $\boldsymbol{\Sigma}$ was generated randomly for each 10000 trials.

preset CFAR. All of the considered regularized estimators outperformed the commonly used non-regularized estimator. This is in line with previous works [1]–[4], [24] which have nicely outlined the benefits of shrinkage type covariance matrix estimators in different engineering applications. It should also be noted that further benefits can be achieved when tuning parameters are chosen so that they optimize an application specific metric for the problem at hand. This will be a subject of future work.

APPENDIX
PROOF OF THEOREM 2

Proof: Let $\widehat{\boldsymbol{\Sigma}}$ be the unique solution to (7), and define $\mathbf{V}_k = \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_k \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$. Algorithm (16) can then be re-expressed as

$$\mathbf{V}_{k+1} = G(\mathbf{V}_k) \equiv \frac{1}{n} \sum_{i=1}^{n} u(\mathbf{y}_i^{\text{H}} \mathbf{V}_k^{-1} \mathbf{y}_i) \mathbf{y}_i \mathbf{y}_i^{\text{H}} + \alpha \widehat{\boldsymbol{\Sigma}}^{-1},$$

where $\mathbf{y}_i = \widehat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} \mathbf{z}_i$ for $i = 1, \ldots, n$. From (7), it follows that $G(\mathbf{I}_p) = \mathbf{I}_p$. Note that $\mathbf{V}_k \in \mathcal{H}(p)$, and so let $\lambda_{1,k} \geq \cdots \geq \lambda_{p,k} > 0$ denote the eigenvalues of $\mathbf{V}_k$. The objective is to then show that $\mathbf{V}_k \to \mathbf{I}_p$ as $k \to \infty$. We first give two lemmas.

Lemma 4.
a) $\lambda_{1,k} > 1 \Rightarrow \lambda_{1,k+1} < \lambda_{1,k}$.
b) $\lambda_{1,k} \leq 1 \Rightarrow \lambda_{1,k+1} \leq 1$.
c) $\lambda_{p,k} < 1 \Rightarrow \lambda_{p,k+1} > \lambda_{p,k}$.
d) $\lambda_{p,k} \geq 1 \Rightarrow \lambda_{1,k+1} \geq 1$.

Proof: (a) Since $u(t)$ in non-increasing, and $\psi(t) = tu(t)$ is non-decreasing, it follows that $u(\mathbf{y}^{\text{H}} \mathbf{V}_k^{-1} \mathbf{y}) \leq u(\mathbf{y}^{\text{H}} \mathbf{y} / \lambda_{1,k}) = \lambda_{1,k} \psi(\mathbf{y}^{\text{H}} \mathbf{y} / \lambda_{1,k}) / \mathbf{y}^{\text{H}} \mathbf{y} \leq \lambda_{1,k} u(\mathbf{y}^{\text{H}} \mathbf{y})$, and so

$$\begin{aligned} \mathbf{V}_{k+1} &\leq \lambda_{1,k} \frac{1}{n} \sum_{i=1}^{n} u(\mathbf{y}_i^{\text{H}} \mathbf{y}_i) \mathbf{y}_i \mathbf{y}_i^{\text{H}} + \alpha \widehat{\boldsymbol{\Sigma}}^{-1} \\ &= \lambda_{1,k} G(\mathbf{I}_p) + (1 - \lambda_{1,k}) \alpha \widehat{\boldsymbol{\Sigma}}^{-1}. \end{aligned}$$

Thus, $\mathbf{V}_{k+1} < \lambda_{1,k} G(\mathbf{I}_p) = \lambda_{1,k} \mathbf{I}_p$, and so part (a) follows.

(b) Since $u(t)$ is non-increasing, $u(\mathbf{y}^{\text{H}} \mathbf{V}_k^{-1} \mathbf{y}) \leq u(\mathbf{y}^{\text{H}} \mathbf{y} / \lambda_{1,k}) \leq u(\mathbf{y}^{\text{H}} \mathbf{y})$. Consequently, $\mathbf{V}_{k+1} \leq G(\mathbf{I}_p) = \mathbf{I}_p$, and so part (b) follows.

The proofs for parts (c) and (d) are analogous. ■

Lemma 5.
a) $\limsup \lambda_{1,k} \leq 1$.
b) $\liminf \lambda_{p,k} \geq 1$.

Proof: The proof is by contradiction. To show part (a), presume $\lambda_1 \equiv \limsup \lambda_{1,k} > 1$. By Lemma 4.b, this then implies that $\lambda_{1,k} > 1$ for all k. So, by Lemma 4.a, it follows that $\lambda_{1,k}$ is a strictly decreasing sequence and hence $\lambda_{1,k} \downarrow \lambda_1 > 1$.

Next, note that Lemma 4 also implies that the sequences $\lambda_{1,k}$ and $\lambda_{p,k}$ are both bounded away from 0 and ∞. Hence, there exists a convergent subsequence $\mathbf{V}_{k(j)} \to \mathbf{V} \in \mathcal{H}(p)$, with $\lambda_1(\mathbf{V}) = \lambda_1 > 1$. Here, $\lambda_1(\mathbf{V}) \geq \cdots \geq \lambda_p(\mathbf{V}) > 0$ denote the eigenvalues of $\mathbf{V}$. Furthermore, by continuity, $\mathbf{V}_{k(j)+1} \to G(\mathbf{V})$ with $\lambda_1\{G(\mathbf{V})\} = \lambda_1$. However, Lemma 4.a implies $\lambda_1 = \lambda_1\{G(\mathbf{V})\} < \lambda_1(\mathbf{V}) = \lambda_1$, a contradiction. Hence part (a) holds. The proof to part (b) is analogous. ■

So, by Lemma 5 we have $1 \leq \liminf \lambda_{p,k} \leq \limsup \lambda_{1,k} \leq 1$, which implies $\lim \lambda_{p,k} = \lim \lambda_{1,k} = 1$. Thus, $\mathbf{V}_k \to \mathbf{I}_p$ and hence Theorem 2 holds. ■

PROOF OF THEOREM 4

Proof: (a) Express $\Gamma = \mathbf{\Sigma}^{-1} = \gamma\mathbf{M}$ with $\mathrm{Tr}(\mathbf{M}) = 1$, and so $\mathcal{L}^*_{\alpha,\beta}(\mathbf{\Sigma}) = L_1(\gamma) + L_2(\mathbf{M})$, where

$$L_1(\gamma) = p(\beta-1)\ln(\gamma) + \alpha\gamma$$
$$L_2(\mathbf{M}) = \frac{p\beta}{n}\left\{\sum_{i=1}^{n}\ln(\mathbf{z}_i^{\mathrm{H}}\mathbf{M}\mathbf{z}_i)\right\} - \ln|\mathbf{M}|.$$

Now if $\mathbf{\Sigma} \to \partial\mathcal{H}(p)$ then either $\gamma \to 0$, $\gamma \to \infty$, or $\mathbf{M} \to \partial\mathcal{H}(p)$. If γ goes to zero or infinity, it readily follows that $L_1(\gamma) \to \infty$ since for any $c > 0$, $\alpha\gamma - c\ln\gamma \to \infty$ as $\gamma \to 0$ or as $\gamma \to \infty$.

So, we only need to consider what happens to $L_2(\mathbf{M})$ as $\mathbf{M} \to \partial\mathcal{H}(p)$. Since the set of positive semi-definite Hermitian matrices with trace one is compact, it is sufficient to consider a sequence $\mathbf{M}_k \to \mathcal{M}$, where $\mathcal{M}$ is a singular positive semi-definite Hermitian matrix with trace one. Hence $1 < \mathrm{rank}(\mathcal{M}) < p$. Let $\lambda_1(\mathbf{M}) \ge \cdots \ge \lambda_p(\mathbf{M})$ denote the eigenvalue of $\mathbf{M}$. Since eigenvalues are continuous functions, $\lambda_j(\mathbf{M}_k) \to \lambda_j(\mathcal{M})$. The spectral value decomposition gives $\mathbf{M}_k = \sum_{j=1}^{p}\lambda_j(\mathbf{M}_k)\boldsymbol{\theta}_{k,j}\boldsymbol{\theta}_{k,j}^{\mathrm{H}}$, where $\mathbf{M}_k\boldsymbol{\theta}_{k,j} = \lambda_j(\mathbf{M}_k)\boldsymbol{\theta}_{k,j}$ with $\boldsymbol{\theta}_{k,j}^{\mathrm{H}}\boldsymbol{\theta}_{k,m} = \delta_{j,m}$. By compactness, it can be assumed without loss of generality that $\boldsymbol{\theta}_{k,j} \to \boldsymbol{\theta}_j$, $j = 1,\ldots,p$, with $\boldsymbol{\theta}_j^{\mathrm{H}}\boldsymbol{\theta}_m = \delta_{j,m}$. For $j = 1,\ldots,p$, let S_j denote the subspace of $\mathbb{C}^p$ spanned by $\{\boldsymbol{\theta}_j,\ldots,\boldsymbol{\theta}_p\}$, $S_{p+1} = \{\mathbf{0}\}$ and $D_j = S_j\backslash S_{j+1} = \{\mathbf{z} \in \mathbb{C}^p \mid \mathbf{z} \in S_j, \mathbf{z} \notin S_{j+1}\}$. Also, let $n_j = \#\{\mathbf{z}_i \in D_j\}$ and $N_j = \#\{\mathbf{z}_i \in S_j\}$.

For $n_j \ge 1$ and $\mathbf{z}_i \in D_j$, $\mathbf{z}_i^{\mathrm{H}}\mathbf{M}_k\mathbf{z}_i \ge \lambda_j(\mathbf{M}_k)|\boldsymbol{\theta}_{k,j}^{\mathrm{H}}\mathbf{z}_i|^2 \ge \lambda_j(\mathbf{M}_k)c_{k,j}$, where

$$\begin{aligned}c_{k,j} &= \min\{|\boldsymbol{\theta}_{k,j}^{\mathrm{H}}\mathbf{z}_i|^2; \mathbf{z}_i \in D_j\}\\ &\to c_j = \min\{|\boldsymbol{\theta}_j^{\mathrm{H}}\mathbf{z}_i|^2; \mathbf{z}_i \in D_j\} > 0.\end{aligned}$$

For $n_j = 0$, let $c_{k,j} = c_j = 1$. Hence,

$$\begin{aligned}L_2(\mathbf{M}_k) \ge{}& \frac{p\beta}{n}\sum_{j=1}^{p} n_j\ln(c_{k,j})\\ &+ \sum_{j=1}^{p}\left(\frac{p\beta n_j}{n} - 1\right)\ln\{\lambda_j(\mathbf{M}_k)\}.\end{aligned}$$

The first term on the right converges to $\frac{p\beta}{n}\sum_{j=1}^{p} n_j\ln(c_j) > -\infty$ and for $j \le r = \mathrm{rank}(\mathcal{M}), 0 < \lambda_j(\mathcal{M}) < 1$. So, to complete the proof of part (a), it only needs to be shown that

$$L_3(\mathbf{M}_k) = \sum_{j=r+1}^{p}\left(\frac{p\beta n_j}{n} - 1\right)\ln\{\lambda_j(\mathbf{M}_k)\} \to \infty.$$

Condition A implies $\frac{p\beta N_j}{n} < p - j + 1$ for $j = 2,\ldots p$. Also, since $n_j = N_j - N_{j+1}$ with $N_{p+1} = 0$, it follows that $\left(\frac{p\beta n_j}{n} - 1\right) < a_j$, where $a_j = \left(p - j - \frac{p\beta N_{j+1}}{n}\right)$ for $j = 2,\ldots,p$. Condition A also insures that $a_j \le 0$ and so $\left(\frac{p\beta n_j}{n} - 1\right)$ is strictly negative. Finally, for $j = r+1,\ldots,p$, $\ln\{\lambda_j(\mathbf{M}_k)\} \to -\infty$. Thus, each term in $L_3(\mathbf{M}_k)$ must go to ∞.

b) If condition B does not hold, then there exists a subspace $\mathcal{V}_o$ such that $\frac{n_o}{n} > \frac{d_o}{p\beta}$, where $n_o = \#\{\mathbf{z}_i \in \mathcal{V}_o\}$ and $d_o = \dim(\mathcal{V}_o)$, with $1 \le d_o < p$. Construct a sequence $\Gamma_k = \mathbf{\Sigma}_k^{-1} \in \mathcal{H}(p)$ having eigenvalues 1 and $\gamma_{k,o}$ with multiplicities $p - d_o$ and d_o respectively, with $\gamma_{k,o} \to 0$. Also, for every k, let the eigenspace associated with $\gamma_{k,o}$ be $\mathcal{V}_o$. Part (b) will then follow by showing $\mathcal{L}^*_{\alpha}(\mathbf{\Sigma}_k) \to -\infty$.

To show this, note that $\mathcal{L}^*_{\alpha}(\mathbf{\Sigma}_k) = L_{a,k} + L_{o,k}$, where

$$L_{o,k} = \left(\frac{p\beta n_o}{n} - d_o\right)\ln(\gamma_{k,o}) \quad \text{and}$$

$$L_{a,k} = \frac{p\beta}{n}\left\{\sum_{\mathbf{z}_i\in\mathcal{V}_o}\ln(\mathbf{z}_i^{\mathrm{H}}\mathbf{z}_i) + \sum_{\mathbf{z}_i\notin\mathcal{V}_o}\ln(\mathbf{z}_i^{\mathrm{H}}\Gamma_k\mathbf{z}_i)\right\} - \alpha\mathrm{Tr}(\Gamma_k).$$

It readily follows that $L_{a,k} \to L_a < \infty$. Also, $L_{o,k} \to -\infty$ since $\log(\gamma_{o,k}) \to -\infty$ and $\frac{p\beta n_o}{n} > d_o$. ■

PROOF OF THEOREM 5

Proof: Denote

$$\mathbf{C} = \frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{z}_i\mathbf{z}_i^{\mathrm{H}}}{\mathbf{z}_i^{\mathrm{H}}\mathbf{\Sigma}_0^{-1}\mathbf{z}_i} = \mathbf{\Sigma}_0^{1/2}\Big(\frac{p}{n}\sum_{i=1}^{n}\mathbf{u}_i\mathbf{u}_i^{\mathrm{H}}\Big)\mathbf{\Sigma}_0^{1/2} \tag{23}$$

where $\mathbf{u}_i = \mathbf{\Sigma}_0^{-1/2}\mathbf{z}_i/\|\mathbf{\Sigma}_0^{-1/2}\mathbf{z}_i\|$ for $i = 1,\ldots,n$. Hence the clairvoyant estimator is $\mathbf{\Sigma}_\alpha = (1-\alpha)\mathbf{C} + \alpha\mathbf{I}$. First, note that the MSE criterion is

$$\begin{aligned}\Delta(\alpha) &= \mathbb{E}\big[\|\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha - \tfrac{1}{p}\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha)\mathbf{I}\|^2\big]\\ &= \mathrm{Tr}\Big(\mathbf{\Sigma}_0^{-2}\mathbb{E}\big[\mathbf{\Sigma}_\alpha^2\big]\Big) - \frac{1}{p}\mathbb{E}\big[\mathrm{Tr}^2(\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha)\big],\end{aligned}$$

and observe that

$$\begin{aligned}\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}\mathbf{\Sigma}_\alpha) &= \mathrm{Tr}\Big((1-\alpha)\Big(\frac{p}{n}\sum_{i=1}^{n}\mathbf{u}_i\mathbf{u}_i^{\mathrm{H}}\Big) + \alpha\mathbf{\Sigma}_0^{-1}\Big)\\ &= p(1-\alpha) + \alpha\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}) = p\end{aligned}$$

where the 3rd identity follows from the fact that $\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}) = p$. This result then implies that finding the minimum of $\Delta(\alpha)$ is equivalent to finding the minimum of $\Delta^*(\alpha) = \mathrm{Tr}\Big(\mathbf{\Sigma}_0^{-2}\mathbb{E}\big[\mathbf{\Sigma}_\alpha^2\big]\Big)$.

A closed-form expression for $\Delta^*(\alpha)$ can be obtained by using the following identities:

$$\mathbb{E}[\mathbf{C}] = \mathbf{\Sigma}_0 \tag{24}$$
$$\mathbb{E}[\mathbf{C}^2] = \frac{p\{\mathbf{\Sigma}_0^2 + \mathrm{Tr}(\mathbf{\Sigma}_0)\mathbf{\Sigma}_0\}}{n(p+1)} + \Big(\frac{n-1}{n}\Big)\mathbf{\Sigma}_0^2. \tag{25}$$

The proofs rely on the representation of $\mathbf{C}$ in (23) in terms of i.i.d. r.v.'s $\mathbf{u}_i$, which possess a uniform distribution on complex p-sphere, and properties of their moments as stated in [20, Lemma 4]. The derivations are similar to the Proof of Theorem 2 in [4] and are therefore omitted.

To conclude the proof of Theorem 5, note that

$$\begin{aligned}\mathbb{E}[\mathbf{\Sigma}_\alpha^2] &= \mathbb{E}[((1-\alpha)\mathbf{C} + \alpha\mathbf{I})^2]\\ &= 2\alpha(1-\alpha)\mathbb{E}[\mathbf{C}] + \alpha^2\mathbf{I} + (1-\alpha)^2\mathbb{E}[\mathbf{C}^2]\end{aligned}$$

and hence, using (24), (25) and the property $\mathrm{Tr}(\mathbf{\Sigma}_0^{-1}) = p$, we obtain

$$\begin{aligned}\Delta^*(\alpha) =& 2\alpha(1-\alpha)p + \alpha^2 \mathrm{Tr}(\mathbf{\Sigma}_0^{-2}) \\ &+ (1-\alpha)^2 \Big\{ \frac{p(p + p\mathrm{Tr}(\mathbf{\Sigma}_0))}{n(p+1)} + \Big(\frac{n-1}{n}\Big)p \Big\} \\ =& \alpha^2(\mathrm{Tr}(\mathbf{\Sigma}_0^{-2}) - p) + (1-\alpha)^2 \frac{p(p\mathrm{Tr}(\mathbf{\Sigma}_0) - 1)}{n(p+1)} + C\end{aligned}$$

where the constant C does not depend on α. The minimizer α_o of $\Delta^*(\alpha)$ (and hence of $\Delta(\alpha)$) is thus $\alpha_o = a/(a+b)$, where a (resp. b) denotes the multiplier term of $(1-\alpha)^2$ (resp. α^2) in the expression of $\Delta^*(\alpha)$ above. This then gives the stated result for the complex-valued case.

The proof for the real-case follows similarly, the only difference being that the identity in Eq. (25) in the real case becomes

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

■

REFERENCES

[1] Y. Abramovich and O. Besson, "Regularized covariance matrix estimation in complex elliptically symmetric distributions using the expected likelihood approach-part 1: The over-sampled case," *IEEE Trans. Signal Processing*, vol. 61, no. 23, pp. 5807–5818, 2013.

[2] Y. I. Abramovich and N. K. Spencer, "Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering," in *Proc. Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP'07)*, Apr. 15–20 2007, pp. 1105–1108.

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

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

[5] E. Conte, A. De Maio, and G. Ricci, "Recursive estimation of the covariance matrix of a compound-Gaussian process and its application to adaptive CFAR detection," *IEEE Trans. Signal Processing*, vol. 50, no. 8, pp. 1908 – 1915, 2002.

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

[7] L. Du, J. Li, and P. Stoica, "Fully automatic computation of diagonal loading levels for robust adaptive beamforming," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 1, pp. 449–458, 2010.

[8] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[9] F. Gini and M. Greco, "Covariance matrix estimation for CFAR detection in correlated heavy-tailed clutter," *Signal Processing*, vol. 82, no. 12, pp. 1847–1859, 2002.

[10] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[11] J. T. Kent, "Data analysis for shapes and images," *J. Statist. Plann. Inference*, vol. 57, no. 2, pp. 181–193, 1997.

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

[13] ——, "Constrained M-estimation for multivariate location and scatter," *Ann. Statist.*, vol. 24, no. 3, pp. 1346–1370, 1996.

[14] S. Kraut, L. L. Scharf, and R. W. Butler, "The adaptive coherence estimator: A uniformly most-powerful-invariant adaptive detection statistic," *IEEE Trans. Signal Processing*, vol. 53, no. 2, pp. 427–438, 2005.

[15] S. Kraut, L. L. Scharf, and L. T. McWhorter, "Adaptive subspace detectors," *IEEE Trans. Signal Processing*, vol. 49, no. 1, pp. 1–16, 2001.

[16] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Mult. Anal.*, vol. 88, pp. 365–411, 2004.

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

[18] E. Ollila and V. Koivunen, "Robust antenna array processing using M-estimators of pseudo-covariance," in *Proc. 14th IEEE Int. Symp. on Personal, Indoor and Mobile Radio Comm. (PIMRC'03)*, Beijing, China, Sep. 7–10, 2003, pp. 2659–2663.

[19] ——, "Robust space-time scatter matrix estimator for broadband antenna arrays," in *Proc. 58th IEEE Vehicular Technology Conference (VTC'03-Fall)*, vol. 1, Orlando, FL, USA, Oct. 6–9, 2003, pp. 55 – 59.

[20] ——, "Influence function and asymptotic efficiency of scatter matrix based array processors: Case MVDR beamformer," *IEEE Trans. Signal Processing*, vol. 57, no. 1, pp. 247 – 259, 2009.

[21] E. Ollila and D. E. Tyler, "Distribution-free detection under complex elliptically symmetric clutter distribution," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM'12)*, Hoboken, NJ, USA, Jun. 17–20, 2012.

[22] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor, "Complex elliptically symmetric distributions: survey, new results and applications," *IEEE Trans. Signal Processing*, vol. 60, no. 11, pp. 5597 – 5625, 2012.

[23] F. Pascal, Y. Chitour, J.-P. Ovarlez, P. Forster, and P. Larzabal, "Covariance structure maximum-likelihood estimates in compound Gaussian noise: existence and algorithm analysis," *IEEE Trans. Signal Processing*, vol. 56, no. 1, pp. 34 – 48, 2008.

[24] F. Pascal, Y. Chitour, and Y. Quek, "Generalized robust shrinkage estimator and its application to stap detection problem," http://arxiv.org/pdf/1311.6567arXiv:1311.6567, 2013, [online] ArXiv.

[25] L. L. Scharf and B. Friedlander, "Matched subspace detectors," *IEEE Trans. Signal Processing*, vol. 42, no. 8, pp. 2146–2157, 1994.

[26] S. Sra and R. Hosseini, "Geometric optimisation on positive definite matrices for elliptically contoured distributions," *Advances in Neural Information Processing Systems*, vol. 26, pp. 2562 – 2570, 2013, expanded version: "Conic geometric optimisation on the manifold of positive definite matrices," ArXiv: [1312.1039; math.FA], Dec 2013.

[27] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *Ann. Statist.*, vol. 15, no. 1, pp. 234–251, 1987.

[28] ——, "Statistical analysis for the angular central Gaussian distribution on the sphere," *Biometrika*, vol. 74, pp. 579–589, 1987.

[29] ——, "A note on affine equivariant location and scatter statistics for sparse data," *Statist. and Prob. Letters*, vol. 80, pp. 1409–1413, 2010.

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

[31] ——, "Unified framework to regularized covariance estimation in scaled gaussian models," *IEEE Trans. Signal Processing*, vol. 60, no. 1, pp. 29–38, 2012.

[32] D. M. Witten and R. Tibshirani, "Covariance-regularized regression and classification for high-dimensional problems," *J. Royal Stat. Soc., Ser. B*, vol. 71, no. 3, pp. 615–636, 2009.

[33] T. Zhang, A. Wiesel, and M. S. Greco, "Multivariate generalized Gaussian distribution: Convexity and graphical models," *IEEE Trans. Signal Processing*, vol. 61, no. 16, pp. 4141–4148, 2013.

Esa Ollila (M'03) received the M.Sc. degree in mathematics from the University of Oulu, in 1998, Ph.D. degree in statistics with honors from the University of Jyväskylä, in 2002, and the D.Sc.(Tech) degree with honors in signal processing from Aalto University, in 2010. From 2004 to 2007 he was a post-doctoral fellow of the Academy of Finland. He has also been a Senior Researcher and a Senior Lecturer at Aalto University and University of Oulu, respectively. Currently he is appointed as an Academy Research Fellow of the Academy of Finland at the Department of Signal Processing and Acoustics, Aalto University, Finland. He is also adjunct Professor (statistics) of University of Oulu. The Fall-term 2001 he was a Visiting Researcher with the Department of Statistics, Pennsylvania State University, State College, PA while the academic year 2010-2011 he spent as a Visiting Post-doctoral Research Associate with the Department of Electrical Engineering, Princeton University, Princeton, NJ. His research interests focus on theory and methods of statistical signal processing, independent component analysis and blind source separation, complex-valued signal processing, array and radar signal processing and robust statistical methods.

David E. Tyler received the M.Sc. and Ph.D. degree from the Department of Statistics, Princeton University, in 1976 and 1979, respectively, and M.Sc. degree from the Department of Mathematics and Statistics, University of Massachusetts. He is currently a Distinguished Professor in the Department of Statistics & Biostatistics at Rutgers – The State University of New Jersey, where he has been on the faculty since 1983, serving as chair of the department from 1993-1996. From 1978-1979, he was an assistant professor in the Department of Statistics, University of Florida, and from 1979-1983, assistant professor in the Department of Applied Mathematics, Old Dominion University. He has held numerous visiting positions in the USA and abroad, including the University of Pittsburgh, Columbia University, the University of Toronto, the University of Leeds, the Swiss Institute of Technology (ETH/EPFL), the University of Buenos Aires, the University of Jyväskylä, the University of Tampere, and Princeton University. Dr. Tyler is a fellow of *IMS*, the *Institute of Mathematical Statistics*.

He has served as an Associate Editor for various statistical journals, including the *Annals of Statistics* and the *Journal of the Royal Statistical Society*, as guest editor for special issues of the *Journal of Statistical Planning and Inference*, the *Canadian Journal of Statistics* and *Computer Vision and Image Understanding*, and is a member of the steering committee of *ICORS*, the *International Conferences on Robust Statistics*. His current research interests in statistics include robust statistics, multivariate statistical analysis and the spectral analysis of time series, while in the past he has worked is such diverse areas as psychometrics and computer vision.

David Tyler's research was supported in part by National Science Foundation Grant DMS-1407751.