Unified Framework to Regularized Covariance Estimation in Scaled Gaussian Models

Ami Wiesel, *Member, IEEE*

***Abstract*—We consider regularized covariance estimation in scaled Gaussian settings, e.g., elliptical distributions, compound-Gaussian processes and spherically invariant random vectors. Asymptotically in the number of samples, the classical maximum likelihood (ML) estimate is optimal under different criteria and can be efficiently computed even though the optimization is nonconvex. We propose a unified framework for regularizing this estimate in order to improve its finite sample performance. Our approach is based on the discovery of hidden convexity within the ML objective. We begin by restricting the attention to diagonal covariance matrices. Using a simple change of variables, we transform the problem into a convex optimization that can be efficiently solved. We then extend this idea to nondiagonal matrices using convexity on the manifold of positive definite matrices. We regularize the problem using appropriately convex penalties. These allow for shrinkage towards the identity matrix, shrinkage towards a diagonal matrix, shrinkage towards a given positive definite matrix, and regularization of the condition number. We demonstrate the advantages of these estimators using numerical simulations.**

***Index Terms*—Covariance estimation, hidden convexity, optimization on manifolds, regularization, robust statistics.**

I. Introduction

ESTIMATING a covariance matrix is a fundamental problem in statistical signal processing. Many techniques for detection and estimation, varying from array processing to functional genomics, rely on accurately estimated covariance matrices [1], [2]. The problem is well understood when the number of samples is much larger than the dimensions of the matrix, and when the underlying multivariate distribution is known to be Gaussian. In this case, the classical sample covariance coincides with the maximum likelihood (ML) estimator and is optimal under most criteria. In many modern applications neither of these assumptions holds. Dynamic large scale systems involve a large number of variables whose statistical properties remain stationary over a short period of samples. In many of these, it was empirically shown that the heavy tailed data does not fit a Gaussian distribution. Such settings require regularized and robust covariance estimation methods as considered in this paper.

The classical statistical approaches to high dimensional parameter estimation using a small number of samples are regularization, shrinkage and/or Bayesian priors. Most of these can be interpreted as adding constraints or penalties to the maximum likelihood solution in order to incorporate additional prior knowledge, and/or allow for a bias-variance tradeoff. Typical examples in the context of covariance estimation include shrinkage towards identity [3]–[6], shrinkage towards a diagonal structure [7], knowledge-aided estimation [8], [9], and constraints on the condition number of the estimate [10].

Robust statistics provide estimation methods which are not sensitive to small departures from the model assumption [11]. In the context of covariance estimation, a common approach is to replace the Gaussian assumption with a more general scaled Gaussian model. This model is related to the family of Elliptical distributions, spherically invariant random vectors (SIRV) and compound Gaussian processes. It has been empirically shown that these models are appropriate for describing different real worlds signals as speech, radar, wireless fading channels and more [12]–[20]. A well-studied robust covariance estimator in this setting is Tyler's ML estimator [21]–[23]. This ML optimization is nonconvex, yet it has been shown that its global solution can be found using a simple fixed point iteration. Recent contributions in this topic include the generalizations to the complex case [23] and to the case of incomplete data [24], and its asymptotic eigenstructure analysis [25], [26].

In an attempt to enjoy the best of both worlds, different authors proposed to regularize Tyler's estimator and obtain a robust high dimensional covariance estimator. A diagonal loading approach has shown promising performance using real data from high-frequency over-the-horizon-radar [27]. Minimum mean squared error and maximum *a posteriori* estimation methods using Bayesian priors were considered in the context of knowledge-aided space time adaptive processing [9]. Tyler's estimator with shrinkage towards the identity, a variant of diagonal loading, was recently proposed in [28]. This work addressed the existence, uniqueness and convergence properties of the estimator, provided a closed form data-dependent choice for the regularization parameter, and was successfully applied to anomaly detection in a real-world sensor network. These results demonstrated the advantages of regularized robust covariance estimation. However, they lacked many of the appealing properties associated with Tyler's original method. The proposed methods were not always proven to converge; were not scale invariant and required renormalization procedures after each iteration; and were not shown to be the global solution to any likelihood based optimization.

Manuscript received April 28, 2011; revised August 11, 2011; accepted September 12, 2011. Date of publication October 06, 2011; date of current version December 16, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Alfred Hanssen. This work was partially supported by Israel Science Foundation Grant No. 786/11.

Parts of this research have been presented in IEEE CAMSAP, San Juan, Puerto Rico, December 13–16, 2011.

The author is with the Rachel and Selim Benin School of Computer Science and Engineering, Hebrew University of Jerusalem, 91904 Jerusalem, Israel (e-mail: amiw@cs.huji.ac.il).

Digital Object Identifier 10.1109/TSP.2011.2170685

Our main contribution is a unified framework for robust covariance estimation based on regularized ML estimation. As explained, the main difficulty with the ML criterion is that the negative-log-likelihood is not convex in the classical sense. Recently, [29], [30] showed that the negative-log-likelihood is in fact convex on the geodesics of the quotient of positive definite matrices with determinant one. Exploiting this remarkable result, we propose to regularize the problem in a similarly convex manner. We begin with a low order diagonal model for the covariance. Through a simple change of variables, we convexify this constrained ML problem and propose a simple numerical method for finding its global solution. The resulting estimator is a simple yet highly applicable robust estimator of variances. In addition, its derivation leads the way to our unified regularization framework which is based on a more complicated notion of convexity. Here, we propose to regularize the negative-log-likelihood using penalty functions which are convex on the geodesics of the manifold of positive definite matrices. Our penalties are constructed to allow shrinkage towards the identity matrix, shrinkage towards an arbitrary positive definite matrix, shrinkage towards a diagonal matrix, and regularization of the condition number. We provide simple fixed point iterations for minimizing these optimization problems. Finally, we propose a novel cross-validation procedure for tuning their regularization parameters. This procedure takes into account the inherent scaling ambiguity in scaled Gaussian distributions. We demonstrate the accuracy advantages of our proposed methods using numerical experiments.

The paper is organized as follows. Section II provides a brief review of scaled Gaussian distributions, Tyler's original covariance estimation method and its generalizations. In Section III, we derive the diagonal version of Tyler's estimator and prepare the grounds to the unified regularization framework proposed in Section IV. Simulations results are presented in Section IV, and concluding remarks are offered in Section V.

The following notation is used. Boldface upper case letters denote matrices, boldface lower case letters denote column vectors, and standard lower case letters denote scalars. We use indexes in the subscript $[\mathbf{x}]_a$ or $[\mathbf{x}]_{a,b}$ to denote subvectors or submatrices, respectively. The superscripts $(\cdot)^T$ and $(\cdot)^{-1}$ denote the transpose and matrix inverse, respectively. The operator $||\cdot||$ denotes the Euclidean norm, and $\mathbf{X} \succ \mathbf{0}$ means that $\mathbf{X}$ is positive definite. The vectors $\mathbf{e}_i$ for $i = 1, \ldots, p$ denote the standard length p unit vectors, and $\mathrm{diag}\{a_j\}$ denotes a diagonal matrix with the elements a_j. For a vector $\mathbf{z}$, we use $e^{\mathbf{z}}$ and $\log\ \mathbf{z}$ to denote vectors with the elements e^{z_j} and $\log\ z_j$, respectively. The operators $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximal and minimal eigenvalues of a matrix, respectively. Finally, $a \setminus b$ for indexes sets a and b is the set difference operator.

II. Background

In this section, we review Tyler's covariance estimator [21]–[23] and its previous generalizations [9], [27], [28]. Tyler's estimator can be naturally derived as the ML estimator of the covariance in scaled Gaussian distributions. We define

$$\mathbf{x} = \nu \mathbf{u} \tag{1}$$

where ν is a random or deterministic scalar and $\mathbf{u}$ is an independent zero mean Gaussian vector of length p with covariance $\mathbf{\Sigma}^{\mathrm{true}} \succ \mathbf{0}$. Let $\mathbf{x}_i = \nu_i \mathbf{u}_i$ for $i = 1, \ldots, n$ be independent and identically distributed (i.i.d.) realizations of $\mathbf{x}$. Given these observations, our goal is to estimate the unknown covariance matrix. Without further assumptions on ν_i, there is a scaling ambiguity and we only expect to estimate the covariance up to an unknown scaling factor.

There are two approaches to this problem which both lead to same solution. First, we can use the normalized vectors which are invariant to ν_i:

$$\mathbf{s}_i = \frac{\mathbf{x}_i}{||\mathbf{x}_i||} = \frac{\mathbf{u}_i}{||\mathbf{u}_i||}, \quad i = 1, \ldots, n. \tag{2}$$

Computing the distribution of these normalized samples leads to the following ML estimation problem [31]:

$$\min_{\mathbf{\Sigma} \succ \mathbf{0}} l\left(\{\mathbf{s}_i\}_{i=1}^n ; \mathbf{\Sigma}\right) \tag{3}$$

where the negative-log-likelihood is given by

$$l\left(\{\mathbf{s}_i\}_{i=1}^n ; \mathbf{\Sigma}\right) = \frac{p}{n} \sum_{i=1}^n \log\left(\mathbf{s}_i^T \mathbf{\Sigma}^{-1} \mathbf{s}_i\right) + \log|\mathbf{\Sigma}|. \tag{4}$$

An alternative approach is to treat the scalings ν_i as deterministic unknown parameters, solve for them and use the concentrated likelihood [15], [22]. This leads to an identical optimization problem. As expected, the negative-log-likelihood objective is invariant to scaling of $\mathbf{\Sigma}$ by a positive constant.

Tyler and others proved that a global solution to (3)–(4) exists and can be efficiently found when $n > p$. These results are highly nontrivial as the optimization is nonconvex. Tyler's original solution to (3) was based on a simple fixed point iteration. We provide an alternative derivation of the same iteration using a majorization-minimization algorithm, e.g., [32]. This iteration will be easier to generalize in the sequel. Specifically, we propose to solve (3) using the iterations

$$\mathbf{\Sigma}_{k+1} = \arg\min_{\mathbf{\Sigma} \succ \mathbf{0}} Q\left(\mathbf{\Sigma}, \mathbf{\Sigma}_k\right) \tag{5}$$

where the majorization function satisfies

$$Q\left(\mathbf{\Sigma}, \mathbf{\Sigma}_k\right) \geq l\left(\{\mathbf{s}_i\}_{i=1}^n ; \mathbf{\Sigma}\right), \quad \forall \quad \mathbf{\Sigma}_k \tag{6}$$

$$Q\left(\mathbf{\Sigma}, \mathbf{\Sigma}\right) = l\left(\{\mathbf{s}_i\}_{i=1}^n ; \mathbf{\Sigma}\right). \tag{7}$$

Under suitable technical conditions, these properties ensure monotonicity of the algorithm and attainment of a local minimum. In particular, we use the following surrogate function

$$Q\left(\mathbf{\Sigma}, \mathbf{\Sigma}_k\right) = \frac{p}{n} \sum_{i=1}^n \log\left(\mathbf{s}_i^T \mathbf{\Sigma}_k^{-1} \mathbf{s}_i\right) + \frac{p}{n} \sum_{i=1}^n \frac{\mathbf{s}_i^T \mathbf{\Sigma}^{-1} \mathbf{s}_i}{\mathbf{s}_i^T \mathbf{\Sigma}_k^{-1} \mathbf{s}_i} - p + \log|\mathbf{\Sigma}| \tag{8}$$

which bounds the negative-log-likelihood due to the inequality

$$\log(x) \leq \log(a) + \left(\frac{x}{a} - 1\right) \quad x \geq 0, \tag{9}$$

and the concavity of the logarithm function. Ignoring constants, each iteration is then

$$\mathbf{\Sigma}_{k+1} = \arg\min_{\mathbf{\Sigma}\succ\mathbf{0}} \mathrm{Tr}\left\{\left[\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\mathbf{\Sigma}_k^{-1}\mathbf{s}_i}\right]\mathbf{\Sigma}^{-1}\right\} + \log|\mathbf{\Sigma}|\,. \tag{10}$$

When $n > p$ the matrix in the squared brackets is positive definite with probability one, and the iteration reduces to

$$\mathbf{\Sigma}_{k+1} = \frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\mathbf{\Sigma}_k^{-1}\mathbf{s}_i}, \quad k = 1,2,\ldots \tag{11}$$

beginning with any initial $\mathbf{\Sigma}_0 \succ \mathbf{0}$. Some intuition to this iteration can be obtained by rewriting it as an iteratively reweighted sample covariance

$$\mathbf{\Sigma}_{k+1} = \frac{1}{n}\sum_{i=1}^{n}[\beta_i]_k\,\mathbf{s}_i\mathbf{s}_i^T, \tag{12}$$

with the adaptive weights

$$[\beta_i]_k = \frac{p}{\mathbf{s}_i^T\mathbf{\Sigma}_k^{-1}\mathbf{s}_i}, \tag{13}$$

for $k = 1,2,\ldots$ and $i = 1,\ldots,n$.

Asymptotically, when $n \gg p$, Tyler's estimator is optimal under many criteria. Unfortunately, this is not true in many practical applications involving finite sample settings. The traditional approach for improving finite sample performance is known as regularization. To our knowledge, the first regularized version of Tyler's method was considered in [27]. Specifically, the authors proposed to regularize the fixed point iteration in (11) using diagonal loading[1]

$$\tilde{\mathbf{\Sigma}}_{k+1} = \frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\mathbf{\Sigma}_k^{-1}\mathbf{s}_i} + \alpha\mathbf{I} \tag{14}$$

$$\mathbf{\Sigma}_{k+1} = \frac{\tilde{\mathbf{\Sigma}}_{k+1}}{\mathrm{Tr}\left\{\tilde{\mathbf{\Sigma}}_{k+1}\right\}} \tag{15}$$

where α is a diagonal loading coefficient. This approach has shown significant performance advantages in numerical simulations. Following this work, a rigorous existence, uniqueness and convergence analysis based on concave Perron Frobenius theory was provided in [28]. An important difference between the original iteration in (11) and the regularized iteration in (14)-(15) is the rescaling in (15). Indeed, unlike the original iteration, its diagonally loaded version is not scale invariant and rescaling is required in order to ensure convergence. Moreover, unlike Tyler's method, it is unclear what optimization (if any) these iterations are trying to solve.

Recently, a different class of regularized Tyler's estimators was considered in [9]. Here, a knowledge-aided approach was considered which exploits prior knowledge in the form a prior Bayesian distribution on $\mathbf{\Sigma}$. Maximum *a posteriori* optimizations and their corresponding fixed point iterations were proposed, as well as promising numerical results. However, the optimization problems were not convex nor scale invariant. The iterations were also not scale invariant but were not rescaled, and it is not clear whether convergence was guaranteed.

[1]More precisely, the original iteration allowed an adaptive regularization coefficient α_k which depends on the iteration index k.

III. Diagonal Tyler's Estimator

In this section, we consider an extension to Tyler's estimator assuming a diagonal covariance matrix. This estimator is a nontrivial extension which may be useful in some applications. In addition, it is valuable as an introduction to our unified framework for regularized Tyler's estimator presented in the next section.

A classical approach to high dimensional parameter estimation in finite sample setting is to resort to simple low order models. In the context of covariance estimation an appealing model is the diagonal case. Clearly, estimating only the variances on the diagonal is much easier than estimating the full covariance matrix, and may provide a good bias-variance tradeoff. A typical application is high dimensional classification. Indeed, the successful diagonal linear discriminant analysis (DLDA) is based on plugging a diagonal Gaussian covariance estimate into the standard linear discriminant [33]. We now extend this estimate to the scaled Gaussian case.

Before we continue it is important to emphasize that this extension is not trivial. In the Gaussian case, a diagonal covariance means uncorrelated elements and the solution is decoupled and trivial:

$$\sigma_j = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{s}_i]_j^2, \quad j = 1,\ldots,p. \tag{16}$$

This is not true in the non-Gaussian case in (1), where the elements of $\mathbf{x}$ are stochastically correlated (or deterministically coupled) through their common scaling ν.

We define the diagonal Tyler's estimator as the diagonal positive definite matrix

$$\mathbf{\Sigma} = \mathrm{diag}\{\sigma_j\} \tag{17}$$

which minimizes the negative-log-likelihood

$$\min_{\sigma_1,\ldots,\sigma_p>0} l\left(\{\mathbf{s}_i\}_{i=1}^n;\mathrm{diag}\{\sigma_j\}\right). \tag{18}$$

The estimator requires the solution of the following p dimensional optimization problem:

$$\min_{\sigma_1,\ldots,\sigma_p>0}\frac{p}{n}\sum_{i=1}^{n}\log\left(\mathbf{s}_i^T\mathrm{diag}\left\{\frac{1}{\sigma_j}\right\}\mathbf{s}_i\right) + \log\left|\mathrm{diag}\{\sigma_j\}\right|. \tag{19}$$

Like Tyler's original ML, this optimization is nonconvex and seems difficult. It is possible to follow Tyler's original derivations and show that its local minimas are also global and can be found efficiently, but this approach is cumbersome, nonconstructive and does not provide any insight. Instead, we now present an alternative approach based on hidden convexity.

Problem (19) can be transformed into a convex form using a simple change of variables. Defining

$$\sigma_j = e^{z_j}, \quad j = 1,\ldots,p, \tag{20}$$

the problem can be expressed as

$$\min_{z_1,\ldots,z_p} \frac{p}{n}\sum_{i=1}^{n}\log\left(\mathbf{s}_i^T \operatorname{diag}\left\{e^{-z_j}\right\}\mathbf{s}_i\right)+\log\left|\operatorname{diag}\left\{e^{z_j}\right\}\right|, \quad (21)$$

or more simply

$$\min_{z_1,\ldots,z_p} \frac{p}{n}\sum_{i=1}^{n}\log\left(\sum_{j=1}^{p}[\mathbf{s}_i]_j^2 e^{-z_j}\right)+\sum_{j=1}^{p} z_j. \quad (22)$$

This last optimization is convex in $z_1 \ldots, z_p$ since the first sum involves convex log-sum-exp expressions and the second term is linear. As such, it can be efficiently solved using standard methods with provable convergence to the global solution. Indeed, this problem is a special case of geometric programming (GP) which has recently made considerable impact in the signal processing community [34], [35].

Numerical solutions to (19) can be obtained via off-the-shelves GP toolboxes, or any unconstrained convex optimization technique. For simplicity, we follow our previous majorization-minimization approach in (5)–(8), and note that the upper bound holds for diagonal matrices. Its minimizer yields the following fixed point iteration:

$$[\sigma_j]_{k+1} = \frac{p}{n}\sum_{i=1}^{n}\frac{[\mathbf{s}_i]_j^2}{\mathbf{s}_i^T \operatorname{diag}\left\{\frac{1}{[\sigma_j]_k}\right\}\mathbf{s}_i}, \quad j=1,\ldots,p \quad (23)$$

where k is the iteration index, and the starting point is $[\sigma_j]_0 > 0$ for $j = 1,\ldots,p$. Similarly to Tyler's estimator, some intuition to this procedure can be obtained by expressing it as iteratively reweighted sample variances

$$[\sigma_j]_{k+1} = \frac{1}{n}\sum_{i=1}^{n}[\beta_i]_k [\mathbf{s}_i]_j^2, \quad (24)$$

with the adaptive weights

$$[\beta_i]_k = \frac{p}{\mathbf{s}_i^T \operatorname{diag}\left\{\frac{1}{[\sigma_j]_k}\right\}\mathbf{s}_i}, \quad (25)$$

for $j = 1,\ldots,p$, $i = 1,\ldots,n$ and $k = 1,\ldots$.

IV. Regularized Tyler's Estimator

We now present a unified framework for regularizing Tyler's estimator. As explained, the main difficulty in this extension is that the negative-log-likelihood in (4) is nonconvex in $\mathbf{\Sigma}$. In the diagonal case, we detected a hidden convexity which simplified the process. Similarly, the way to efficient regularization of Tyler's estimator is through a deeper understanding of its hidden convexity. In particular, [29] and [30] recently showed that the function is actually convex on the geodesics (shortest paths) of the quotient of positive definite matrices with determinant one. In a continuation of these works, we now relate this property to the diagonal case, and give a simple proof that the problem is convex on the geodesics of the Riemmanian manifold of positive definite matrices with a logarithmic metric which is more commonly used in signal processing [36]. Based on this understanding, we then provide a unified framework for regularizing Tyler's estimator.

A function $f(\mathbf{x})$ is convex in the standard definition if for any $\mathbf{x}_0$ and $\mathbf{x}_1$ in its convex domain, it satisfies

$$f(\mathbf{x}_t) \le t f(\mathbf{x}_1) + (1-t) f(\mathbf{x}_0) \quad (26)$$

where $\mathbf{x}_t$ belongs to the line segment

$$\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0 \quad (27)$$

and

$$t \in [0,1]. \quad (28)$$

Convexity on manifolds replaces this line definition with a geodesic[2] [37], [38]

$$\mathbf{x}_t = \operatorname{geodesic}(t;\mathbf{x}_0,\mathbf{x}_1), \quad t\in[0,1] \quad (29)$$

which is the shortest path using a specific metric between $\mathbf{x}_0$ and $\mathbf{x}_1$ parameterized by t. Thus, a function $f(\mathbf{x})$ is convex on the manifold if and only if $f(\mathbf{x}_t)$ is convex in t in the standard definition.

Let us illustrate these definitions in the diagonal case discussed in Section III. The function $f(\cdot)$ is now the negative-log-likelihood restricted to diagonal matrices

$$f(\boldsymbol{\sigma}) = l\left(\{\mathbf{s}_i\}_{i=1}^n; \operatorname{diag}\{\sigma_j\}\right). \quad (30)$$

It is not convex, but we used a change of variables in (20) to convexify it so that

$$f\left(e^{t\mathbf{z}_1+(1-t)\mathbf{z}_0}\right) \le t f(e^{\mathbf{z}_1}) + (1-t) f(e^{\mathbf{z}_0}), \quad (31)$$

for all $\mathbf{z}_0$, $\mathbf{z}_1$ and $t \in [0,1]$. In terms of $\boldsymbol{\sigma}_0 = e^{\mathbf{z}_0}$ and $\boldsymbol{\sigma}_1 = e^{\mathbf{z}_1}$, we have

$$e^{t\mathbf{z}_1+(1-t)\mathbf{z}_0} = e^{t\log\boldsymbol{\sigma}_1+(1-t)\log\boldsymbol{\sigma}_0} = \boldsymbol{\sigma}_1^t\boldsymbol{\sigma}_0^{1-t}, \quad (32)$$

and the inequality becomes

$$f\left(\boldsymbol{\sigma}_1^t\boldsymbol{\sigma}_0^{1-t}\right) \le t f(\boldsymbol{\sigma}_1) + (1-t) f(\boldsymbol{\sigma}_0). \quad (33)$$

In the nomenclature of manifolds, this means that the negative-log-likelihood is convex on the geodesics defined as

$$\boldsymbol{\sigma}_t = \boldsymbol{\sigma}_1^t\boldsymbol{\sigma}_0^{1-t}, \quad t\in[0,1]. \quad (34)$$

In the positive matrix case, we are not aware of any change of variables which transforms the problem into a convex one.[3] However, the geodesic in (34) has a well known generalization to the geodesic in the Riemmanian manifold of positive definite matrices with a logarithmic metric [36]:

$$\mathbf{\Sigma}_t = \mathbf{\Sigma}_0^{\frac{1}{2}}\left(\mathbf{\Sigma}_0^{-\frac{1}{2}}\mathbf{\Sigma}_1\mathbf{\Sigma}_0^{-\frac{1}{2}}\right)^t\mathbf{\Sigma}_0^{\frac{1}{2}}, \quad t\in[0,1], \quad (35)$$

[2]We assume that the geodesic lies within the domain of the function.

[3]The natural generalization of (20) is to parameterize $\mathbf{\Sigma} \succ \mathbf{0}$ via the matrix exponential operation $\exp(\mathbf{T})$ where $\mathbf{T}$ is a symmetric matrix. Unfortunately, $\log\left(\mathbf{s}^T\exp(\mathbf{T})\,\mathbf{s}\right)$ is not a convex function of $\mathbf{T}$.

for any $\mathbf{\Sigma}_0 \succ \mathbf{0}$ and $\mathbf{\Sigma}_1 \succ \mathbf{0}$. Indeed, in the diagonal case in which $\mathbf{\Sigma}_t = \mathrm{diag}\left\{[\boldsymbol{\sigma}_t]_j\right\}$, it is easy to see that (35) reduces to (34). The readers may also recognize the matrix $\mathbf{\Sigma}_{\frac{1}{2}}$ known as the geometric mean of the positive definite matrices $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$ which has recently attracted considerable attention in statistical signal processing [39]–[41]. Similarly to the diagonal case, our next result reveals the hidden convexity on the manifold.

Proposition 1: The negative-log-likelihood $l\left(\{\mathbf{s}_i\}_{i=1}^n; \mathbf{\Sigma}_t\right)$ in (4) is convex in $t \in [0,1]$ on the geodesic in (35).

Proof: Consider the eigenvalue decomposition

$$\mathbf{\Sigma}_0^{-\frac{1}{2}} \mathbf{\Sigma}_1 \mathbf{\Sigma}_0^{-\frac{1}{2}} = \mathbf{U} \mathrm{diag}\left\{d_j\right\} \mathbf{U}^T \tag{36}$$

where $d_j > 0$ are the eigenvalues and $\mathbf{U}$ is the matrix of eigenvectors. When we substitute $\mathbf{\Sigma}_t$ for $\mathbf{\Sigma}$, the first terms in (4) become

$$\begin{aligned} \log\ \mathbf{s}_i^T \mathbf{\Sigma}_t^{-1} \mathbf{s}_i &= \log\ \mathbf{s}_i^T \mathbf{\Sigma}_0^{-\frac{1}{2}} \left(\mathbf{\Sigma}_0^{-\frac{1}{2}} \mathbf{\Sigma}_1 \mathbf{\Sigma}_0^{-\frac{1}{2}}\right)^{-t} \mathbf{\Sigma}_0^{-\frac{1}{2}} \mathbf{s}_i \\ &= \log\ \mathbf{s}_i^T \mathbf{\Sigma}_0^{-\frac{1}{2}} \mathbf{U} \mathrm{diag}\left\{d_j^{-t}\right\} \mathbf{U}^T \mathbf{\Sigma}_0^{-\frac{1}{2}} \mathbf{s}_i \\ &= \log \sum_j \left[\mathbf{U}^T \mathbf{\Sigma}_0^{-\frac{1}{2}} \mathbf{s}_i\right]_j^2 d_j^{-t} \\ &= \log \sum_j \left[\mathbf{U}^T \mathbf{\Sigma}_0^{-\frac{1}{2}} \mathbf{s}_i\right]_j^2 e^{-t \log\ d_j} \end{aligned} \tag{37}$$

which are convex log-sum-exp expressions in t. The second term is linearized since

$$\begin{aligned} \log\ |\mathbf{\Sigma}_t| &= \log\ \left|\mathbf{\Sigma}_0^{\frac{1}{2}} \left(\mathbf{\Sigma}_0^{-\frac{1}{2}} \mathbf{\Sigma}_1 \mathbf{\Sigma}_0^{-\frac{1}{2}}\right)^t \mathbf{\Sigma}_0^{\frac{1}{2}}\right| \\ &= \log\ |\mathbf{\Sigma}_0| + t \log\ \left|\mathbf{\Sigma}_0^{-\frac{1}{2}} \mathbf{\Sigma}_1 \mathbf{\Sigma}_0^{-\frac{1}{2}}\right|. \end{aligned} \tag{38}$$

■

Geodesic convexity is useful as it gives a local characterization to global optimality [37], [38], [42]:

Proposition 2: Let $f(\mathbf{\Sigma}_t)$ be convex in $t \in [0,1]$ on the geodesic in (35). Then, any local minimum of $f(\mathbf{\Sigma})$ over the set of positive definite matrices is a global minimum.

Proof: Assume $\mathbf{\Sigma}_0 \succ \mathbf{0}$ and $\mathbf{\Sigma}_1 \succ \mathbf{0}$ are local minimums of $f(\mathbf{\Sigma})$. Assume in contradiction that only $\mathbf{\Sigma}_1$ is a global minimum. Let $\mathbf{\Sigma}_t$ be the geodesic between these two points as defined in (35). Then,

$$f(\mathbf{\Sigma}_t) \leq t f(\mathbf{\Sigma}_1) + (1-t) f(\mathbf{\Sigma}_0) \tag{39}$$

$$< f(\mathbf{\Sigma}_0), \quad \text{for all} \quad t \in (0,1] \tag{40}$$

where the first inequality is due to geodesic convexity and the second due to $f(\mathbf{\Sigma}_1) < f(\mathbf{\Sigma}_0)$. For sufficiently small t, this is a contradiction to local optimality of $\mathbf{\Sigma}_0$. See [37], [38], and [42] for more details.

These results shed more light on the convergence of Tyler's original estimator to the global solution. The seemingly non-convex ML problem is convex in a more generalized definition, and simple descent algorithms can attain its global solution. This observation is the starting point to our unified framework for regularizing Tyler's estimator. We define our regularized estimator as the solution to

$$\min_{\mathbf{\Sigma} \succ \mathbf{0}} l\left(\{\mathbf{s}_i\}_{i=1}^n; \mathbf{\Sigma}\right) + \alpha h(\mathbf{\Sigma}) \tag{41}$$

where α is a regularization parameter, and $h(\mathbf{\Sigma})$ is a penalty function which satisfies

- $h(\mathbf{\Sigma})$ is scale invariant;
- $h(\mathbf{\Sigma})$ is convex on the geodesics.

These ensure that the overall objective function, i.e., the negative-log-likelihood plus its regularization, is also scale invariant and convex on the geodesics. In the next subsections, we provide a few promising penalties that satisfy these requirements.

A. Shrinkage to an Identity Matrix

The most common approach to covariance regularization is shrinkage towards the identity matrix, also known as diagonal loading or ridge regularization [3]–[6]. It results in a well conditioned matrix and has an appealing bias-variance tradeoff. In our framework, such a regularization can be achieved using the following result.

Proposition 3: Consider the penalty function

$$h^{\mathrm{identity}}(\mathbf{\Sigma}) = p \log\left(\mathrm{Tr}\left\{\mathbf{\Sigma}^{-1}\right\}\right) + \log|\mathbf{\Sigma}|. \tag{42}$$

It is scale invariant and convex on the geodesics in (35). The solution to

$$\min_{\mathbf{\Sigma} \succ \mathbf{0}} h^{\mathrm{identity}}(\mathbf{\Sigma}) \tag{43}$$

is the set of positively scaled identity matrices.

Proof: The scale invariance is due to the difference of logarithms. The function is convex on the geodesics in (35) since the first term is a convex log-sum-exp function and the second is linear. Finally, taking the gradient with respect to $\mathbf{\Sigma}^{-1}$ and equating to zero yields the condition

$$\frac{\partial h^{\mathrm{identity}}(\mathbf{\Sigma})}{\partial \mathbf{\Sigma}^{-1}} = \frac{p}{\mathrm{Tr}\left\{\mathbf{\Sigma}^{-1}\right\}} \mathbf{I} - \mathbf{\Sigma} = \mathbf{0} \tag{44}$$

which is satisfied if and only if $\mathbf{\Sigma} = c\mathbf{I}$ for some c. ■

Thus, our shrinkage towards the identity estimator is defined as the solution to

$$\min_{\mathbf{\Sigma} \succ \mathbf{0}} l\left(\{\mathbf{s}_i\}_{i=1}^n; \mathbf{\Sigma}\right) + \alpha h^{\mathrm{identity}}(\mathbf{\Sigma}) \tag{45}$$

and penalizes estimates that are far from the identity. The problem can be addressed using standard descent methods, or advanced optimization on manifold techniques [38], [43]. The convexity properties guarantee that any local minimum found is globally optimal. For simplicity, we propose to extend our previous majorization-minimization approach. We bound the negative-log-likelihood by $Q(\mathbf{\Sigma}, \mathbf{\Sigma}_k)$ in (8) and the penalty function by

$$h^{\mathrm{identity}}(\mathbf{\Sigma}) \leq p \log\left(\mathrm{Tr}\left\{\mathbf{\Sigma}_k^{-1}\right\}\right) + \frac{p \mathrm{Tr}\left\{\mathbf{\Sigma}^{-1}\right\}}{\mathrm{Tr}\left\{\mathbf{\Sigma}_k^{-1}\right\}} - p + \log|\mathbf{\Sigma}|. \tag{46}$$

Ignoring constants, each iteration is then defined as

$$\mathbf{\Sigma}_{k+1} = \arg\min_{\mathbf{\Sigma} \succ \mathbf{0}} \mathrm{Tr}\left\{\left[\frac{p}{n} \sum_{i=1}^n \frac{\mathbf{s}_i \mathbf{s}_i^T}{\mathbf{s}_i^T \mathbf{\Sigma}_k^{-1} \mathbf{s}_i} + \frac{\alpha p \mathbf{I}}{\mathrm{Tr}\left\{\mathbf{\Sigma}_k^{-1}\right\}}\right] \mathbf{\Sigma}^{-1}\right\} + (1+\alpha) \log|\mathbf{\Sigma}|, \tag{47}$$

and results in

$$\mathbf{\Sigma}_{k+1} = \frac{1}{1+\alpha}\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\mathbf{\Sigma}_k^{-1}\mathbf{s}_i} + \frac{\alpha}{1+\alpha}\frac{p}{\mathrm{Tr}\left\{\mathbf{\Sigma}_k^{-1}\right\}}\mathbf{I}, \quad (48)$$

starting at any initial point $\mathbf{\Sigma}_0 \succ \mathbf{0}$.

It is interesting to compare (48) with (14)-(15) proposed in [27] and [28]. The iterations are very similar. Other than the definition of the regularization parameters, the main difference is that the diagonal loading in (48) is scaled by $\frac{p}{\mathrm{Tr}\left\{\mathbf{\Sigma}_k^{-1}\right\}}$ whereas (14)-(15) uses rescaling after each iteration. Both achieve scale invariance, but previous work does it indirectly and normalizes $\mathrm{Tr}\left\{\mathbf{\Sigma}\right\}$ whereas our normalization is done automatically through $\mathrm{Tr}\left\{\mathbf{\Sigma}^{-1}\right\}$. A drawback to our estimator, is that it has been shown that (14)-(15) always converges to a fixed point, whereas (48) may diverge when the objective is unbounded. It is not clear what is the meaning of the solution to (14)-(15) in the unbounded case, and our view is that this behavior should be avoided by a better choice of regularization parameter. This important issue will be addressed in future work.

B. Shrinkage to a Known Positive Definite Matrix

In some applications, it is reasonable to assume additional prior knowledge in the form of a known matrix $\mathbf{T} \succ \mathbf{0}$ which is close to $\mathbf{\Sigma}$ in some sense. For example, in knowledge-aided adaptive radar processing such priors can be obtained from secondary data from adjacent cells, digital elevation and terrain data, synthetic aperture radar imagery and other resources [9]. A straightforward extension to (42) promotes the solution to any positive definite shrinkage target $\mathbf{T}$ (up to a scaling factor).

Proposition 4: Consider the penalty function

$$h^{\mathrm{target}}(\mathbf{\Sigma}) = p\log\left(\mathrm{Tr}\left\{\mathbf{\Sigma}^{-1}\mathbf{T}\right\}\right) + \log|\mathbf{\Sigma}|. \quad (49)$$

It is scale invariant and convex on the geodesics in (35). The solution to

$$\min_{\mathbf{\Sigma}\succ\mathbf{0}} h^{\mathrm{target}}(\mathbf{\Sigma}) \quad (50)$$

is the set of matrices $c\mathbf{T}$ for any $c > 0$.

The proof is similar to that of Prop. 3 and therefore omitted. Similarly to (48), the regularized solution can be computed using the following fixed point iteration:

$$\mathbf{\Sigma}_{k+1} = \frac{1}{1+\alpha}\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\mathbf{\Sigma}_k^{-1}\mathbf{s}_i} + \frac{\alpha}{1+\alpha}\frac{p}{\mathrm{Tr}\left\{\mathbf{\Sigma}_k^{-1}\mathbf{T}\right\}}\mathbf{T}, \quad (51)$$

starting at any initial point $\mathbf{\Sigma}_0 \succ \mathbf{0}$.

C. Shrinkage to a Diagonal Positive Matrix

In some settings, the scaled identity is over simplistic and its reliance on equal variances is too restrictive. A more flexible model presumes a diagonal positive matrix which requires uncorrelated elements but allows for different scalings of the variables and more degrees of freedom. The shrinkage target is exactly the positive definite diagonal solution discussed in Section III, but now we use it as a regularization penalty. More details on the advantages of this approach in the Gaussian case are discussed in [7].

Proposition 5: Consider the penalty function

$$h^{\mathrm{diagonal}}(\mathbf{\Sigma}) = \log\prod_{i=1}^{p}\left[\mathbf{\Sigma}^{-1}\right]_{ii} + \log|\mathbf{\Sigma}|. \quad (52)$$

It is scale invariant and convex on the geodesics in (35). The solution to

$$\min_{\mathbf{\Sigma}\succ\mathbf{0}} h^{\mathrm{diagonal}}(\mathbf{\Sigma}) \quad (53)$$

is the set of positive diagonal matrices.

Proof: This function is a special case of the negative-log-likelihood function since

$$h^{\mathrm{diagonal}}(\mathbf{\Sigma}) = l\left(\{\mathbf{e}_i\}_{i=1}^{p};\mathbf{\Sigma}\right) \quad (54)$$

where $\mathbf{e}_i$ for $i = 1,\ldots,p$ are the length p unit vectors. Therefore, it is scale invariant and convex on the geodesics. It is bounded from below by zero due to Hadamard's inequality, and this bound is tight only for diagonal matrices. ■

Both the negative-log-likelihood and the penalty can be independently upper bounded using $Q(\mathbf{\Sigma},\mathbf{\Sigma}_k)$ in (8). A majorization-minimization algorithm can therefore be applied:

$$\mathbf{\Sigma}_{k+1} = \frac{1}{1+\alpha}\frac{p}{n}\sum_{i=1}^{n}\frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\mathbf{\Sigma}_k^{-1}\mathbf{s}_i} + \frac{\alpha}{1+\alpha}\sum_{i=1}^{p}\frac{\mathbf{e}_i\mathbf{e}_i^T}{\mathbf{e}_i^T\mathbf{\Sigma}_k^{-1}\mathbf{e}_i}, \quad (55)$$

starting at any initial point $\mathbf{\Sigma}_0 \succ \mathbf{0}$.

D. Regularization of the Condition Number

Recent work on covariance estimation in the Gaussian case has shown promising success to regularization of the condition number of the estimate [10]. This method has a similar effect to shrinkage towards the identity in the sense that it keeps the structure of the eigenvectors of the sample covariance while concentrating the eigenvalues towards their mean. The difference is that shrinkage towards the identity does this in a linear manner whereas regularizing the condition number uses hard thresholding of the extreme eigenvalues which may be advantageous. We now extend these results to the non-Gaussian case. The following proposition characterizes the convexity properties of the condition number.

Proposition 6: Consider the penalty function

$$h^{\mathrm{cond}}(\mathbf{\Sigma}) = \frac{\lambda_{\max}(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})}. \quad (56)$$

This function is not convex. It is quasi-convex, i.e., its sublevel sets $\left\{\mathbf{\Sigma} : h^{\mathrm{cond}}(\mathbf{\Sigma}) \le \alpha\right\}$ are convex sets for all α. Finally, it is convex on the geodesics in (35) and scale invariant.

Proof: A simple counter example with the matrices

$$\mathbf{A} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \quad \mathbf{B} = \begin{bmatrix}2 & 0\\ 0 & 3\end{bmatrix} \quad (57)$$

shows that

$$\frac{5}{4} = \frac{h^{\mathrm{cond}}(\mathbf{A})}{2} + \frac{h^{\mathrm{cond}}(\mathbf{B})}{2} < h^{\mathrm{cond}}\left(\frac{\mathbf{A}}{2}+\frac{\mathbf{B}}{2}\right) = \frac{4}{3} \quad (58)$$

and therefore the function is not convex. The function is quasi-convex since $\lambda_{\max}(\boldsymbol{\Sigma})$ is convex in $\boldsymbol{\Sigma} \succ \mathbf{0}$, $\lambda_{\min}(\boldsymbol{\Sigma})$ is concave in $\boldsymbol{\Sigma} \succ \mathbf{0}$, and a convex over concave function is quasi-convex [44, Example 3.38].

In order to prove that it is convex on the geodesic, we use the variational characterization of extreme eigenvalues:

$$\lambda_{\max}(\boldsymbol{\Sigma}) = \max_{\mathbf{u}:\|\mathbf{u}\|=1} \mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u} \tag{59}$$

$$\frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})} = \lambda_{\max}\left(\boldsymbol{\Sigma}^{-1}\right) = \max_{\mathbf{v}:\|\mathbf{v}\|=1} \mathbf{v}^T \boldsymbol{\Sigma}^{-1} \mathbf{v}. \tag{60}$$

Due to the monotonicity of the logarithm, we obtain

$$h^{\mathrm{cond}}(\boldsymbol{\Sigma}) = e^{\max\limits_{\mathbf{u}:\|\mathbf{u}\|=1} \log\left(\mathbf{u}^T \boldsymbol{\Sigma} \mathbf{u}\right) + \max\limits_{\mathbf{v}:\|\mathbf{v}\|=1} \log\left(\mathbf{v}^T \boldsymbol{\Sigma}^{-1} \mathbf{v}\right)}. \tag{61}$$

Plugging in the geodesic in (35) yields convex log-sum-exp functions in the maximizations objective, the point-wise maximum of a set of convex functions is convex, and the exponent of a convex function is also convex. Additional recent contributions on the convexity properties of the condition number can be found in [45], [46].

Finally, the condition number is clearly invariant to scaling the minimal and maximal eigenvalues by a positive constant. ■

Motivated by these properties, we propose to use the regularization

$$\min_{\boldsymbol{\Sigma} \succ \mathbf{0}} l\left(\{\mathbf{s}_i\}_{i=1}^n; \boldsymbol{\Sigma}\right) \quad \text{s.t.} \quad h^{\mathrm{cond}}(\boldsymbol{\Sigma}) \leq \alpha \tag{62}$$

where $\alpha > 1$ is a fixed upper bound on the condition number. Note that, unlike the previous regularizations, we follow [10] and apply $h^{\mathrm{cond}}(\boldsymbol{\Sigma})$ as a constraint rather than a penalty. Using the majorization-minimization approach, we bound the negative-log-likelihood by $Q(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_k)$ and obtain the following iteration:

$$\boldsymbol{\Sigma}_{k+1} = \arg \begin{cases} \min_{\boldsymbol{\Sigma} \succ \mathbf{0}} & Q(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_k) \\ \text{s.t.} & h^{\mathrm{cond}}(\boldsymbol{\Sigma}) \leq \alpha. \end{cases} \tag{63}$$

Due to the quasi-convexity, each of the subproblems in (63) is a convex optimization problem which can be solved using standard numerical methods. Furthermore, it was recently shown that each of these has a simple solution based on efficient sorting and thresholding [10].

E. Parameter Tuning

We now briefly discuss the issue of parameter tuning. In order to enjoy the advantages of regularization it is vital to choose the parameter α in an efficient and accurate manner. Recently, [28] proposed a closed-form data-dependent choice for the regularization parameter in the case of shrinkage towards the identity. Other than that, there are numerous classical tuning methods and different error criteria. Choosing the best method for a practical application depends on the specific high level goal and is outside the scope of this paper. For completeness, we now present a simple approach which will be used in the following numerical analysis section.

We propose to choose the regularization parameter based on a K-fold cross validation procedure. Specifically, we modify the standard procedure to ensure scale invariance. We divide the indexes set $S = \{1, \ldots, n\}$ into $K = 10$ nonoverlapping groups S_k such that $S = \cup_{k=1}^K S_k$. We define a grid of parameters α_r for $r = 1, \ldots, R$ and let $\boldsymbol{\Sigma}_r[k]$ be the solution to (41) with α_r and the samples $\{\mathbf{x}_i\}_{i \in S \setminus S_k}$. Our criterion for choosing r is then mean squared Frobenius error compensated for the scaling ambiguity:

$$\begin{aligned} r &= \arg \min_{r \in [1,\ldots,R]} \sum_{k=1}^K \left[\min_{v>0} \left\| v \boldsymbol{\Sigma}_r[k] - \sum_{i \in S_k} \mathbf{x}_i \mathbf{x}_i^T \right\|^2 \right] \\ &= \arg \min_{r \in [1,\ldots,R]} \sum_{k=1}^K \left\| \sum_{i \in S_k} \left[\frac{\mathbf{x}_i^T \boldsymbol{\Sigma}_r[k] \mathbf{x}_i}{\mathrm{Tr}\left\{\boldsymbol{\Sigma}_r[k]^2\right\}} \boldsymbol{\Sigma}_r[k] - \mathbf{x}_i \mathbf{x}_i^T \right] \right\|^2. \end{aligned} \tag{64}$$

This criterion is easily motivated since $\boldsymbol{\Sigma}_r[k]$ and $\mathbf{x}_i$ are independent, and if $\boldsymbol{\Sigma}_r[k] = \boldsymbol{\Sigma}^{\mathrm{true}}$ then the expected value of the argument in the norm is scale invariant and equal to zero.

V. Numerical Results

We now provide numerical examples of our unified framework to both robust and regularized covariance estimation. The purpose of these examples is to demonstrate the different estimators, rather than a detailed practical application which is beyond the scope of this paper and will be pursued elsewhere.

In each simulation, we define a deterministic $p \times p$ true covariance $\boldsymbol{\Sigma}^{\mathrm{true}} \succ \mathbf{0}$ which we keep fixed throughout 200 statistically independent experiments. In each experiment, we generate n i.i.d. realizations of a zero mean multivariate normal of covariance $\boldsymbol{\Sigma}^{\mathrm{true}}$ with or without i.i.d. scaling factors generated according to a Chi-squared distribution with 3 degrees of freedom. These realizations are used to estimate the unknown covariance using the various estimators. In our implementation, the starting points for the fixed point iterations are properly scaled identity matrices and we use 10 iterations. When needed, we tune the regularization parameter using the cross validation method described in Section IV-E with the parameters $\alpha_i = \frac{\rho_i}{1-\rho_i}$ where ρ_i for $i = 1, \ldots, 10$ is a uniform grid over $[0, 1]$. Due to the scaling ambiguity, we normalize the true covariance and its estimates to have unit trace before computing the errors. We quantify the performance using the normalized mean-square errors (NMSEs) defined as

$$\mathrm{NMSE} = \frac{\mathrm{E}\left\{ \left\| \hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^{\mathrm{true}} \right\|^2 \right\}}{\left\| \boldsymbol{\Sigma}^{\mathrm{true}} \right\|^2} \tag{65}$$

where $\hat{\boldsymbol{\Sigma}}$ are the various estimators. The expectation in the NMSE is approximated by averaging over the independent experiments.

In the first simulation, we compare the classical sample variances in (16) with the diagonal generalization of Tyler's estimator in (24)–(25). We let the true covariance to be a $p = 20$ diagonal matrix with $\left[\boldsymbol{\Sigma}^{\mathrm{true}}\right]_{j,j} = j$ for $j = 1, \ldots, p$. The NMSEs in the Gaussian case (no scaling factors) and the Elliptical case (random scaling factors) are presented in Fig. 1. In the Gaussian case presented in the top graph, both estimators

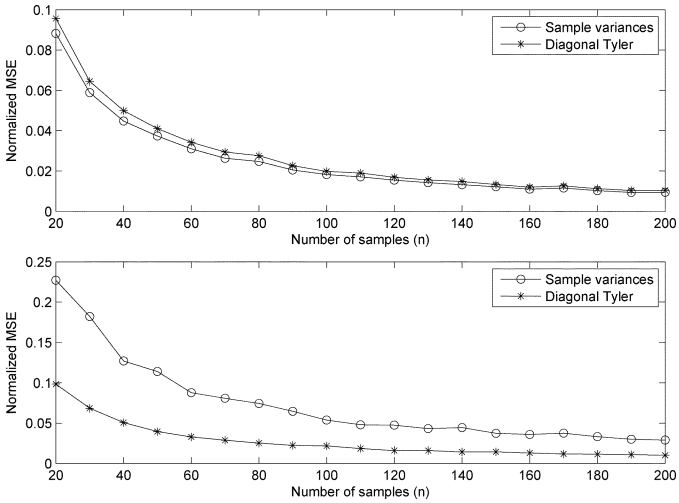

Fig. 1. Diagonal Tyler's estimator. Top: Gaussian distribution. Bottom: Elliptical distribution.

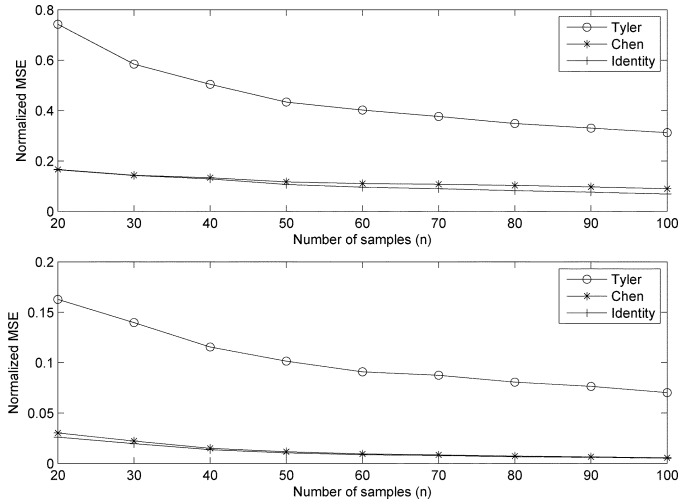

Fig. 2. Shrinkage to the identity. Top: $\beta = 0.4$. Bottom: $\beta = 0.95$.

perform similarly with negligible degradation in performance for our mismatched non-Gaussian estimator. On the other hand, this diagonal estimator significantly outperforms the Gaussian estimator in the Elliptical case presented in the bottom graph.

In the second simulation, we compare the two shrinkage to identity methods, namely (14)-(15) of [27] and [28] and the new iteration in (48). This is a difficult comparison as the results depend on the choice of regularization parameter and the true unknown covariance. In order to eliminate the effect of parameter tuning and focus on the iterations themselves, we did not use cross validation but chose the regularization parameter that yielded the smallest error for each estimator in each experiment. Clearly, this is not possible in practice as the true parameter is unknown, but we believe this procedure makes a fair comparison for our purposes. We let the true covariance be a Toeplitz matrix with

$$\left[\mathbf{\Sigma}^{\text{true}}\right]_{i,j} = \beta^{|i-j|} \tag{66}$$

and $p = 10$. The NMSEs are reported in Fig. 2 for $\beta = 0.4$ and $\beta = 0.95$. It is evident that the two methods perform roughly the same. Our experience with other simulations (not shown) suggest similar behavior. In practice, it is much easier to choose the regularization parameter in (14)-(15) via the closed form proposal in [28], and we find it preferable.

In the third simulation, we demonstrate the advantage of shrinkage towards a specific target matrix, e.g., in knowledge-aided systems. We choose the target covariance $\mathbf{T}$ as a Toeplitz matrix with $\beta = 0.7$, and let the true covariance be a Toeplitz matrix with $\beta = 0.8$. We then estimate it using Tyler's estimator, our shrinkage to identity estimator in (48), and the shrinkage to a target estimator in (51) which assumes prior knowledge of $\mathbf{T}$. The NMSEs are reported in Fig. 3. It is easy to see the performance improvement gained through the efficient utilization of this additional prior information.

In the fourth simulation, we illustrate the advantage of shrinkage towards a diagonal matrix via (55) rather than towards an identity via (48). We choose the true covariance as a Toeplitz matrix in (67) with $\beta = 0.4$ but scale its first row and

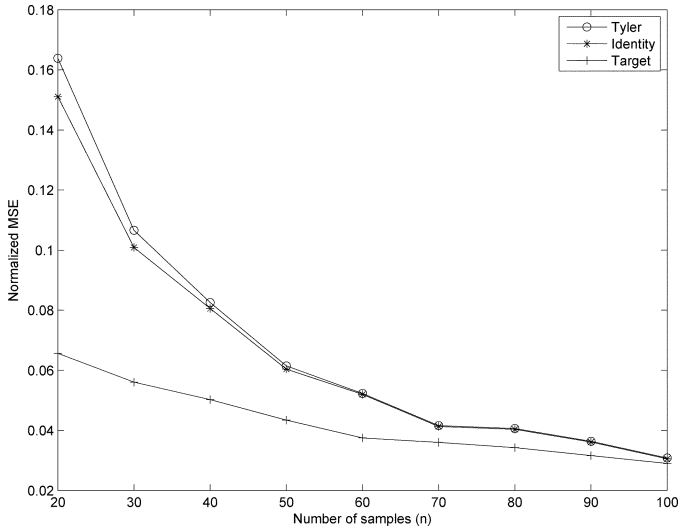

Fig. 3. Shrinkage to a target matrix.

column by a factor of two. This scaling serves to model a system in which some of the variables are of different magnitude. The NMSEs of Tyler's estimator, the shrinkage to identity estimator and the shrinkage to diagonal estimator are reported in Fig. 4. As expected, the diagonal shrinkage approach outperforms its competitors when the number of samples is small.

Finally, in the fifth simulation, we focus on regularizing the condition number of the estimates. We let the true covariance be a $p = 10$ Gaussian shaped covariance matrix with

$$\left[\mathbf{\Sigma}^{\text{true}}\right]_{i,j} = e^{-\frac{|i-j|^2}{2}}. \tag{67}$$

We compare four estimators: the naive sample covariance, Tyler's estimator, a regularized condition number version of the Gaussian ML estimator derived in [10], and our novel estimator which regularized the condition number of Tyler's estimate via (63). We choose the regularization parameter using the cross validation method described in Section IV-F, and a grid of 10 parameters $\alpha \in [1, 10]$ (the unknown condition number of the true covariance is approximately 4.98). The NMSEs are

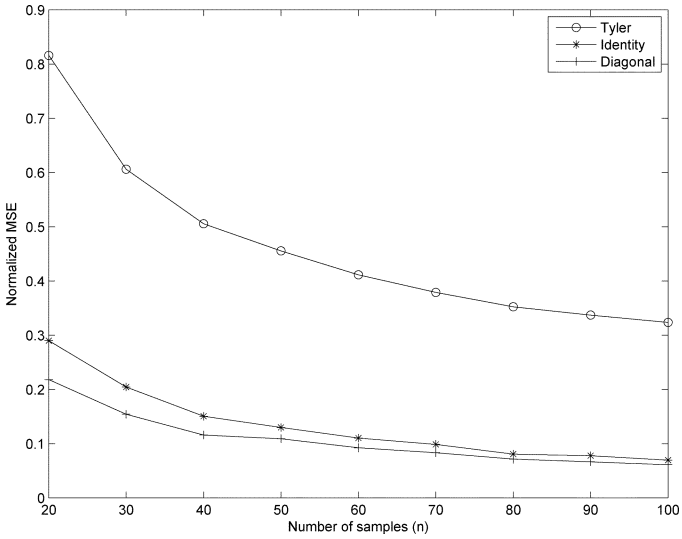

Fig. 4. Shrinkage to a diagonal matrix.

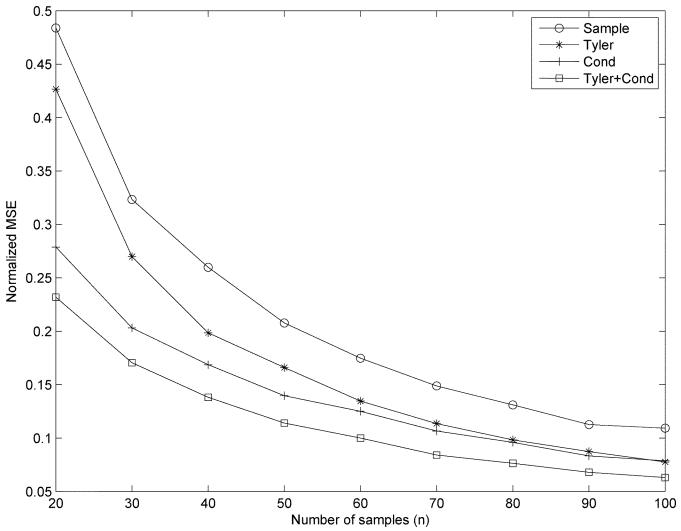

Fig. 5. Regularization of the condition number.

reported in Fig. 5 and illustrate the performance gain due to the combination of Tyler's method and the regularization.

VI. Discussion

In this work, we proposed a unified framework for both regularized and robust covariance estimation. The core ingredient of this framework is hidden convexity on manifolds. Using properly convex penalty functions, our framework allows regularization towards various shrinkage targets as required by different applications.

This research is based on the hypothesis that good estimators are based on a likelihood based optimization. Unlike previous works, we regularize the optimization problem rather than the solution. This allows us to enjoy the recent advances in convex optimization theory and methods. Our current choice of numerical algorithms is based on their simplicity. Future work should address more efficient optimization on manifolds algorithms with provable performance guarantees. Future work should also focus on a rigorous analysis of the boundedness of the problems and the uniqueness of their solutions. Such an analysis would provide a deeper understanding and will assist in tuning the shrinkage parameters so that the problems will be uniquely solvable.

Acknowledgment

The author would like to sincerely thank Y. Chen and A. O. Hero, III, for numerous discussions on the topic which led to this work. In particular, Y. Chen provided the majorization-minimization interpretation of Tyler's fixed point iteration.

References

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[2] E. R. Dougherty, A. Datta, and C. Sima, "Research issues in genomic signal processing," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 46–68, Nov. 2005.

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

[4] C. Stein, "Estimation of a covariance matrix," in *Rietz Lecture, 39th Annu. Meeting IMS*, Atlanta, GA, 1975.

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

[6] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for MMSE covariance estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5016–5029, 2010.

[7] J. Schafer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statist. Appl. Genetics Mol. Biol.*, vol. 4, no. 1, pp. 1175–??, 2005.

[8] P. Stoica, L. Jian, Z. Xumin, and J. R. Guerci, "On using *a priori* knowledge in space-time adaptive processing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2598–2602, 2008.

[9] F. Bandiera, O. Besson, and G. Ricci, "Knowledge-aided covariance matrix estimation and adaptive detection in compound-Gaussian noise," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5391–5396, Oct. 2010.

[10] J. H. Won, J. Lim, S. J. Kim, and B. Rajaratnam, "Maximum likelihood covariance estimation with a condition number constraint," Statistics Dept., Stanford Univ., , Stanford, CA, Tech. Rep. no. 10, 2009.

[11] P. J. Huber and E. M. Ronchetti, *Robust Statistics*. New York: Wiley, 2009.

[12] J. B. Billingsley, "Ground clutter measurements for surface-sited radar," Mass. Inst. of Technol., Cambridge, MA, Tech. Rep. 780, Feb. 1993.

[13] M. Rangaswamy, "Statistical analysis of the nonhomogeneity detector for non-Gaussian interference backgrounds," *IEEE Trans. Signal Process.*, vol. 53, no. 6, pp. 2101–2111, 2005.

[14] J. Wang, A. Dogandzic, and A. Nehorai, "Maximum likelihood estimation of compound-Gaussian clutter and target parameters," *IEEE Trans. Signal Process.*, vol. 54, no. 10, pp. 3884–3898, 2006.

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

[16] F. Gini and A. Farina, "Vector subspace detection in compound-Gaussian clutter. Part I: Survey and new results," *IEEE Trans. Aerosp. Electron. Syst.,* , vol. 38, no. 4, pp. 1295–1311, 2002.

[17] F. Pascal, P. Forster, J. P. Ovarlez, and P. Larzabal, "Performance analysis of covariance matrix estimates in impulsive noise," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2206–2217, 2008.

[18] G. Vasile, J. P. Ovarlez, F. Pascal, and C. Tison, "Coherency matrix estimation of heterogeneous clutter in high-resolution polarimetric SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 4, pp. 1809–1826, 2010.

[19] F. Gini, "Sub-optimum coherent radar detection in a mixture of K-distributed and Gaussian clutter," in *Proc. Inst. Elect. Eng.—Radar, Sonar, Navig.—IET*, 2002, vol. 144, pp. 39–48.

[20] E. Conte, M. Lops, and G. Ricci, "Asymptotically optimum radar detection in compound-Gaussian clutter," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 31, no. 2, pp. 617–625, 2002.

[21] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *Ann. Statist.*, vol. 15, no. 1, pp. 234–251, 1987.
[22] 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 Process.*, vol. 50, no. 8, pp. 1908–1915, 2002.
[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 Process.*, vol. 56, no. 1, pp. 34–48, Jan. 2008.
[24] G. Frahm and U. Jaekel, "A generalization of Tyler's M-estimators to the case of incomplete data," *Comput. Statist. Data Anal.*, vol. 54, no. 2, pp. 374–393, 2010.
[25] G. Frahm and K. Glombek, "Semicircle law for Tyler's M-estimator," 2010 [Online]. Available: http://arxiv.org/pdf/1004.3938, to be published
[26] L. Dumbgen, "On Tyler's M-functional of scatter in high dimension," *Ann. Inst. Statist. Math.*, vol. 50, no. 3, pp. 471–491, 1998.
[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. (ICASSP)*, Apr. 15–20, 2007, vol. 3, pp. III-1105–III-1108, doi: 10.1109/ICASSP.2007.366877.
[28] Y. Chen, A. Wiesel, and A. O. Hero, III, "Robust Shrinkage estimation of high-dimensional covariance matrices," *IEEE Trans. Signal Process.*, 2010, to be published.
[29] C. Auderset, C. Mazza, and E. A. Ruh, "Angular Gaussian and Cauchy estimation," *J. Multivar. Anal.*, vol. 93, no. 1, pp. 180–197, 2005.
[30] C. Auderset, C. Mazza, and E. Ruh, Grassmannian Estimation Arxiv, 2008 [Online]. Available: arXiv:0809.3697, to be published
[31] G. Frahm, "Generalized elliptical distributions: theory and applications," Ph.D. dissertation, Univ. of Cologne, Cologne, Germany, 2004, unpublished.
[32] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization-Minimization algorithms for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2980–2991, 2007.
[33] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *J. Amer. Statist. Assoc.*, vol. 97, no. 457, pp. 77–87, 2002.
[34] S. Boyd, S. J. Kim, L. Vandenberghe, and A. Hassibi, "A tutorial on geometric programming," *Optim. Eng.*, vol. 8, no. 1, pp. 67–127, 2007.
[35] M. Chiang, C. W. Tan, D. P. Palomar, D. O'Neill, and D. Julian, "Power control by geometric programming," *IEEE Trans. Wireless Commun.*, vol. 6, no. 7, pp. 2640–2651, 2007.
[36] S. T. Smith, "Covariance, subspace, and intrinsic Cramér-Rao bounds," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1610–1630, 2005.
[37] T. Rapcsak, "Geodesic convexity in nonlinear optimization," *J. Optim. Theory Appl.*, vol. 69, no. 1, pp. 169–183, 1991.
[38] C. Udriste, *Convex Functions and Optimization Methods on Riemannian Manifolds*. New York: Springer, 1994.
[39] M. Moakher, "A differential geometric approach to the geometric mean of symmetric positive-definite matrices," *SIAM J. Matrix Anal. Appl.*, vol. 26, no. 3, pp. 735–747, 2005.
[40] F. Barbaresco, "Innovative tools for radar signal processing based on Cartan's geometry of SPD matrices & information geometry," in *Proc. IEEE Radar Conf. (RADAR)*, 2008, pp. 1–6.
[41] Sacristá, D. Murga, and A. Pascual-Iserte, "Differential feedback of MIMO channel Gram matrices based on geodesic curves," *IEEE Trans. Wireless Commun.*, vol. 9, no. 12, pp. 3714–3727, Dec. 2010.
[42] L. Liberti, "On a class of nonconvex problems where all local minima are global," *Publications de l'Institut Mathematique*, vol. 76, no. 90, pp. 101–109, 2004.
[43] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ: Princeton Univ. Press, 2008.
[44] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[45] C. Beltrán, J. P. Dedieu, G. Malajovich, and M. Shub, "Convexity properties of the condition number," 2008 [Online]. Available: http://arxiv.org/abs/0806.0395
[46] C. Beltrán, J. P. Dedieu, G. Malajovich, and M. Shub, "Convexity properties of the condition number II," 2009 [Online]. Available: http://arxiv.org/abs/0910.5936

Ami Wiesel (M'07) received the B.Sc. and M.Sc. degrees in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 2000 and 2002, respectively, and the Ph.D. degree in electrical engineering from the Technion—Israel Institute of Technology, Haifa, in 2007.

He was a Postdoctoral Fellow with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, from 2007 to 2009. Since January 2010, he has been a Faculty Member at the Rachel and Selim Benin School of Computer Science and Engineering at the Hebrew University of Jerusalem, Israel.

Dr. Wiesel was a recipient of the Young Author Best Paper Award for a 2006 paper in the IEEE TRANS. SIGNAL PROCESS. and a Student Paper Award for a 2005 Workshop on Signal Processing Advances in Wireless Communications (SPAWC) paper. He was awarded the Weinstein Study Prize in 2002, the Intel Award in 2005, the Viterbi Fellowship in 2005 and 2007, and the Marie Curie Fellowship in 2008.