Tyler's Estimator Performance Analysis

Ilya Soloveychik and Ami Wiesel,
Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel

***Abstract*—This paper analyzes the performance of Tyler's M-estimator of the scatter matrix in elliptical populations. We focus on non-asymptotic performance analysis of Tyler's estimator. Given n samples of dimension $p < n$, we show that the squared Frobenius norm of the error of the inverse estimator is proportional to $p^2/(1-c^2)^2 n$ with high probability, where c is the coherence coefficient of the properly scaled estimator. Under additional group symmetry conditions we improve the obtained bound, utilizing the inherent sparsity properties of group symmetry.**

***Index Terms*—Elliptical distribution shape matrix estimation, scatter matrix M-estimators, Tyler's scatter estimator, concentration bounds.**

I. Introduction

Estimation of large covariance matrices, particularly in the high-dimensional regime, when the data dimension p and the sample size n are of close magnitudes has recently attracted considerable attention. Estimators in this field can be classified based on the underlying distribution and the additional structure assumptions. Most of the research is traditionally devoted to the multivariate Gaussian setting which is currently well understood. When the number of samples is greater than the dimension, the Maximum Likelihood Estimator (MLE) of the covariance exists with probability one and coincides with the Sample Covariance Matrix (SCM). During the last decade there has emerged a great amount of literature on regularized versions of this estimator, their parameter tuning and performance analysis, such as shrinkage-estimator, see e.g. [1, 2]. Recently, similar challenges have appeared in the more ambitious setting of non-Gaussian and robust estimation. A prominent approach in this area is Tyler's scatter estimator [3]. The goal of this paper is to analyze its non-asymptotic behavior as defined below more rigorously.

In many applications the underlying multivariate distribution is non-Gaussian and robust covariance estimation methods are required. This occurs whenever the probability distribution of the measurements is heavy-tailed or a small proportion of the samples represents outlier behavior, [4, 5]. A common robust estimator of scatter is due to Tyler [3]. Given n independent, identically distributed (i.i.d.) proper measurements $\mathbf{x}_i \in \mathbb{C}^p, i = 1, \ldots, n$, Tyler's shape matrix estimator is defined as the solution to the fixed point equation

$$\widehat{\mathbf{\Theta}} = \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^H}{\mathbf{x}_i^H \widehat{\mathbf{\Theta}}^{-1} \mathbf{x}_i}. \tag{1}$$

When $\mathbf{x}_i$ are Generalized Elliptically (GE) distributed [6], their shape matrix $\mathbf{\Theta}_0$ is positive definite and $n > p$, Tyler's estimator exists with probability one and is a consistent estimator of $\mathbf{\Theta}_0$ up to a positive scaling factor. The GE family includes as particular cases generalized Gaussian distribution, compound Gaussian, elliptical and many others [6]. Therefore, it has been successfully used to replace the SCM in many applications such as anomaly detection in wireless sensor networks [7], antenna array processing [8] and radar detection [9–12].

Recently there is an increasing interest in non-asymptotic analysis of high-dimensional estimators, providing their high probability error bounds as functions of n and p. In the case of Gaussian populations, such error bounds for the SCM estimator were thoroughly studied [13–15] and tight sample complexity bounds were achieved. Regularized covariance estimation in the Gaussian case and its performance analysis have also been addressed in [2, 16, 17]. A common thread to all of these works is that the estimators are defined as solutions to convex optimization problems, and the analysis is directly related to the notion of strong convexity.

In this article we focus on the non-asymptotic behavior of Tyler's estimator for moderate values of n and p based on the concentration of measure phenomenon. The estimator is not given in closed form and has to be iteratively computed using the fixed point iteration (1). In order to exploit the optimization based machinery mentioned above, we rely on an alternative derivation of Tyler's estimator. In particular, the estimator can also be obtained as an MLE of normalized GE distributed vectors defined as $\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$, [6]. The solution to this optimization problem exists with probability one and is scale invariant, which makes Tyler's estimator unidentifiable [18]. In order to restrict this invariance some constraints should be imposed diminish the number of degrees of freedom. Interestingly, introduction of such constraints makes the behavior of the estimator dependent on the coherence coefficient

$$c^2(\mathbf{\Omega}_0) = 1 - \left(\frac{\mathrm{Tr}\left(\mathbf{\Omega}_0\right)}{\|\mathbf{I}\|_F \, \|\mathbf{\Omega}_0\|_F} \right)^2, \tag{2}$$

where $\mathbf{\Omega}_0 = \mathbf{\Theta}_0^{-1}$. Our non-asymptotic performance bounds are therefore based on the analysis of the restricted Negative Log-Likelihood (NLL) function of the normalized GE distribution. In particular, we prove that as long as n is larger than p the Frobenius norm of the error in inverse matrices decays like $\frac{p}{(1-c^2(\mathbf{\Omega}_0))\sqrt{n}}$ with high probability. We also show that our performance bounds exhibit correct asymptotic behavior when n and p grow large together.

At the end of the paper we extend the proposed analysis to high-dimensional non-Gaussian settings with group symmetric constraints. Robust models with a priori known structure have recently attracted considerable attention due to significant engineering advances and necessity to deal with high-dimensional data. The a priori knowledge is usually introduced to diminish the number of degrees of freedom in a model and improve the estimators performance. Usually such knowledge is given as convex constraints, such as Toeplitz, sparse, banded, circulant, proper, group symmetric, etc, [16, 19–22]. It was recently discovered that both Tyler's target function [23] and group symmetric constraints [24] are g-convex, which allowed us to develop a new STyler estimator incorporating group symmetric structure into non-Gaussian models, [24]. Interestingly group symmetry is closely related to block diagonal structure, [25]. Using this phenomenon we derive high probability error bounds for the STyler estimator.

The paper is organized as following: first we introduce notations, state the problem, the main result and provide a discussion of it. Then we outline the proof. Finally we extend the obtained result to the group symmetric case and provide numerical simulations illustrating the derived bounds. The body of the article contains only sketches of proof due to lack of space, the full proofs can be found in [26].

Denote by $\mathcal{S}(p)$ the linear space of $p \times p$ hermitian matrices and by $\mathcal{P}(p) \subset \mathcal{S}(p)$ the open cone of positive definite matrices. $\mathbf{I}$ stands for the identity matrix of a proper dimension. We endow $\mathcal{S}(p)$ with

This work was partially supported by the Intel Collaboration Research Institute for Computational Intelligence, the Kaete Klausner Scholarship and ISF Grant 786/11.

the scalar product $(\mathbf{A},\mathbf{B}) = \mathrm{Tr}\,(\mathbf{AB})$, which induces the Frobenius norm on it. $\|\cdot\|$ will denote the Euclidean norm for vectors, $\|\cdot\|_F$ - the Frobenius norm and $\|\cdot\|_2$ - the spectral norm for matrices. Given a matrix $\mathbf{A}$ we denote by $\mathrm{vec}\,(\mathbf{A})$ a column vector obtained by stacking the columns of $\mathbf{A}$. For a matrix $\mathbf{A} \in \mathcal{P}(p)$ denote by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ its minimal and maximal eigenvalues, correspondingly, and by $\kappa(\mathbf{A}) = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$ its condition number. $|\mathbf{A}|$ stands for the determinant of $\mathbf{A}$. For n instances $a_1,\dots,a_n$ of scalars, vectors, matrices or functions we denote by $\widehat{a}$ their arithmetic average, when the index of summation is obvious from the context.

II. Tyler's estimator as a MLE and the Main Result

We define Tyler's estimator as an MLE of a covariance matrix parameter of a specific complex spherical p-dimensional distribution.

Definition 1. *Assume $\mathbf{\Theta}_0 \in \mathcal{P}(p)$, the function*

$$p(\mathbf{x}) = \frac{(p-1)!}{\pi^p}\frac{1}{|\mathbf{\Theta}_0|(\mathbf{x}^H\mathbf{\Theta}_0^{-1}\mathbf{x})^p} \tag{3}$$

is a probability density function of a vector $\mathbf{x} \in \mathbb{C}^p$ lying on a unit sphere. This distribution is usually referred to as the Angular Central Gaussian (ACG) distribution on a sphere [27] and we denote it as $\mathbf{x} \sim \mathcal{U}(\mathbf{\Theta}_0)$. The matrix $\mathbf{\Theta}_0$ is referred to as a shape matrix of the distribution and is a multiple of the covariance matrix of $\mathbf{x}$.

The ACG distribution is closely related to the class of GE distributions, which includes Gaussian, compound Gaussian, elliptical, skew-elliptical, ACG and other distributions, [28]. An important property of the GE family is that the shape matrix of a population does not change when the vector is divided by its Euclidean norm [6, 28]. After normalization, any GE vector becomes ACG distributed. This allows us to treat all these distributions together using a single robust estimator.

Assuming $\mathbf{\Theta} \in \mathcal{P}(p)$ and given $n > p$ i.i.d. copies of a vector $\mathbf{x} \sim \mathcal{U}(\mathbf{\Theta})$: $\mathbf{x}_i, 1 = 1,\dots,n$ we derive the MLE estimator of the shape matrix. For this sake introduce the NLL function:

$$\widetilde{f}(\mathbf{\Theta};\mathbf{x}) = \log|\mathbf{\Theta}| + p\log(\mathbf{x}^H\mathbf{\Theta}^{-1}\mathbf{x}). \tag{4}$$

The function (4) is non-convex in $\mathbf{\Theta}$. Nevertheless, its critical points, given as the solution to (1), provide the global minima with probability one, [23, 29, 30].

The NLL (4) is invariant under multiplication of the shape matrix by a positive constant, thus we are only interested in the estimation of the shape matrix up to a positive scalar factor. In order to obtain a unique MLE we fix the scale of the estimator by assuming that $\mathrm{Tr}\left(\mathbf{\Theta}_0^{-1}\right)$ of the true covariance matrix is known (or arbitrarily fixed). Specifically, we define Tyler's estimator to be the solution to the program

$$\widehat{\mathbf{\Theta}} = \arg\begin{cases}\min\limits_{\mathbf{\Theta}} & \dfrac{1}{n}\sum\limits_{i=1}^{n}\widetilde{f}(\mathbf{\Theta};\mathbf{x}_i)\\ \text{subject to} & \mathrm{Tr}\left(\mathbf{\Theta}^{-1}\right) = \mathrm{Tr}\left(\mathbf{\Theta}_0^{-1}\right).\end{cases} \tag{5}$$

and state the following

Theorem 1. *Assume we are given $n > p$ i.i.d. copies of $\mathbf{x} \sim \mathcal{U}(\mathbf{\Theta}_0)$, then for $\theta \geq 0$ with probability at least*

$$1 - 2\exp\left(\frac{-\theta^2}{2(1+1.2\frac{\theta}{\sqrt{n}})}\right) - 2p^2\exp\left(-\frac{n(1-c^2(\mathbf{\Omega}_0))}{77\ln(7p)(1+\frac{1}{p})}\right)\left(1+\frac{8\cdot 10^3(1+\frac{1}{p})^4}{n^2(1-c^2(\mathbf{\Omega}_0))^4}\right)$$

Tyler's estimator (5) satisfies

$$\left\|\widehat{\mathbf{\Theta}}^{-1} - \mathbf{\Theta}_0^{-1}\right\|_F \leq \frac{15\theta}{\lambda_{\min}(\mathbf{\Theta}_0)(1-c^2(\mathbf{\Omega}_0))}\frac{p+1}{\sqrt{n}}. \tag{6}$$

Note that for the Gaussian populations the same technique provides a similar up to a constant factor bound for the SCM estimator, suggesting that both estimators are bounded by a multiple of $\frac{p}{\sqrt{n}}$ with high probability. Actually, in the Gaussian case stronger bounds can be obtained with the spectral norm. In fact it can be shown that the rate of convergence of the SCM matrix to the true covariance is of order $\sqrt{\frac{p}{n}}$ in the spectral norm, see [13] and references therein. In our case the problem of obtaining spectral norm non-asymptotic bounds is much more involved and remains an open question.

The presence of the coherence $c(\mathbf{\Omega}_0)$ in the denominator appears to be quite natural due to the way the bound is obtained. Indeed, the derivation below is based on the convexity properties of the Fisher Information Matrix (FIM) reflected by $1-c^2(\mathbf{\Omega}_0)$. It easily follows from (2) that

$$1 - c^2(\mathbf{\Omega}_0) \geq \frac{1}{\kappa^2(\mathbf{\Theta}_0)},$$

thus the coherence is close to 0 if the condition number $\kappa(\mathbf{\Theta}_0)$ is close to 1 and gets closer to 1 if the matrix $\mathbf{\Theta}_0$ has many small eigenvalues and few large ones. As we know the estimation of the matrix becomes less stable in the latter case, as the theorem suggests.

III. The Proof Outline

The proof of Theorem 1 follows a technique similar to that of [2, 17, 31] and is based on the second-order Taylor expansion of the sample average NLL. More exactly, our aim is to show that the average NLL is positive on a boundary of some ball around the true parameter and is strongly convex inside this ball, with high probability. It can be shown by means of g-convexity that our target possesses a unique minimum [23]. The method we propose ensures that this minimum belongs to the interior of the ball. The analysis becomes easier if the NLL is parametrized by the inverse shape matrix. We denote $\mathbf{\Omega} = \mathbf{\Theta}^{-1}$ and write

$$f(\mathbf{\Omega};\mathbf{x}) = \widetilde{f}(\mathbf{\Omega}^{-1};\mathbf{x}) = -\log|\mathbf{\Omega}| + p\log(\mathbf{x}^H\mathbf{\Omega}\mathbf{x}).$$

Denote

$$\widehat{f}_{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{\Omega};\mathbf{x}_i), \tag{7}$$

and note that $\widehat{f}_{\mathbf{\Omega}}$ is invariant under the linear map $'$ defined as

$$\mathcal{L}' = \mathbf{\Theta}_0^{1/2}\mathcal{L}\mathbf{\Theta}_0^{1/2},$$

applied to matrices $\mathbf{\Omega}_0, \Delta\mathbf{\Omega}$ followed by a change of random vectors from $\mathbf{x}_i \sim \mathcal{U}(\mathbf{\Theta}_0)$ to $\mathbf{x}_i' \sim \mathcal{U}(\mathbf{I})$ for $i = 1,\dots,n$. This linear transformation relocates the domain under consideration from the vicinity of $\mathbf{\Omega}_0$ into a vicinity of $\mathbf{I}$ and the Taylor polynomial formula with remainder in the Lagrange form near $\mathbf{I}$ reads as

$$\widehat{f}_{\mathbf{\Omega}'} - \widehat{f}_{\mathbf{I}} = \nabla\widehat{f}_{\mathbf{I}}(\Delta\mathbf{\Omega}') + \frac{1}{2}\nabla^2\widehat{f}_{\overline{\mathbf{\Omega}'}}(\Delta\mathbf{\Omega}'), \tag{8}$$

where $\mathbf{\Omega}' = \mathbf{I} + \Delta\mathbf{\Omega}', \overline{\mathbf{\Omega}'} = \mathbf{I} + \alpha\Delta\mathbf{\Omega}', \Delta\mathbf{\Omega}' \in \mathcal{L}', \alpha \in [0,1]$. Since $'$ is a tension-compression map, the maximal and minimal distance changes are known and we can use an inequality

$$\underline{\lambda}\left\|\widehat{\mathbf{\Theta}}^{-1} - \mathbf{\Theta}_0^{-1}\right\|_F \leq \left\|\widehat{\mathbf{\Theta}}'^{-1} - \mathbf{\Theta}_0'^{-1}\right\|_F \tag{9}$$

to establish high probability bounds for inverse Tyler's estimator given the respective bounds on its $'$-image.

Our strategy is therefore to show that the linear term is close to zero and the quadratic term is strongly convex with high probability. These together will provide the desired result. First, consider the

linear term in (8), using Bernstein's concentration inequalities it can be shown that it sharply concentrates around its mean value zero, more exactly we have

Lemma 1. *Uniformly over* $\mathbf{U}' \in \mathcal{L}'$

$$\mathbb{P}\left(\left|\nabla \widehat{f}_{\mathbf{I}}(\mathbf{U}')\right| \geq \sqrt{2}tp \left\|\mathbf{U}'\right\|_F\right) \leq 2\exp\left(\frac{-nt^2}{2(1+1.7t)}\right), \forall t \geq 0. \tag{10}$$

Using matrix concentration inequalities, we can shown that with exponentially in n decaying probability the empirical Hessian in the second term of (8), considered as a quadratic form, is close to its mean

$$H_{\mathbf{\Omega}}(\mathbf{U}) = \mathrm{Tr}\left((\mathbf{\Omega}^{-1}\mathbf{U})^2\right) - p\mathbb{E}\left[\left(\frac{\mathbf{x}^H\mathbf{U}\mathbf{x}}{\mathbf{x}^H\mathbf{\Omega}\mathbf{x}}\right)^2\right].$$

When $\mathbf{\Omega}$ is not far from $\mathbf{\Omega}_0, H_{\mathbf{\Omega}}(\mathbf{U})$ only slightly differs from its value at the true point $H_{\mathbf{\Omega}_0}(\mathbf{U})$

$$\begin{aligned} H_{\mathbf{\Omega}_0}(\mathbf{U}) &= \frac{p\mathrm{Tr}\left((\mathbf{\Theta}_0\mathbf{U})^2\right) - \left(\mathrm{Tr}\left(\mathbf{\Theta}_0\mathbf{U}\right)\right)^2}{p+1} \\ &= \frac{p\left\|\mathbf{\Theta}_0^{1/2}\mathbf{U}\mathbf{\Theta}_0^{1/2}\right\|_F^2 - \left(\mathrm{Tr}\left(\mathbf{\Theta}_0^{1/2}\mathbf{U}\mathbf{\Theta}_0^{1/2}\right)\right)^2}{p+1}. \end{aligned}$$

Note that $H_{\mathbf{\Omega}_0}(\mathbf{U})$ coincides with the Fisher Information Matrix (FIM), also viewed as a quadratic form. The exact statements quantifying the behavior of the derivatives at hand can be found in [26]. Here we only describe qualitatively the main properties of $H_{\mathbf{\Omega}_0}(\mathbf{U})$. It can be easily shown that $H_{\mathbf{\Omega}_0}(\mathbf{U})$ has a one-dimensional kernel spanned by $\mathbf{\Omega}_0$. The mission of the trace constraint in (5) is to eliminate this direction and make the Hessian strongly convex in the vicinity of $\mathbf{\Omega}_0$ with high probability. The following simple lemma will help us to further quantify the strong convexity properties of the $H_{\mathbf{\Omega}_0}(\mathbf{U})$

Lemma 2. *Let* $\mathcal{V}$ *be a Euclidean space and* $\mathcal{S} \subset \mathcal{V}$ *be its subspace of codimension one with normal vector* $\mathbf{n}$*, then for any* $\mathbf{v} \in \mathcal{V}$

$$\sin^2(\angle\mathbf{u},\mathbf{v}) \geq \cos^2(\angle\mathbf{n},\mathbf{v}) = \frac{(\mathbf{n},\mathbf{v})}{\|\mathbf{n}\|\,\|\mathbf{v}\|}, \forall \mathbf{u} \in \mathcal{S}.$$

Consider its subspace $\mathcal{L} \subset \mathcal{S}(p)$ defined by the condition $\mathrm{Tr}\,(\mathbf{U}) = 0$. Apply to this subspace the following linear transformation:

$$\mathcal{L}' = \mathbf{\Theta}_0^{1/2}\mathcal{L}\mathbf{\Theta}_0^{1/2}. \tag{11}$$

$\mathbf{\Omega}_0$ is a normal vector of $\mathcal{L}'$, since $(\mathbf{\Theta}_0^{1/2}\mathbf{U}\mathbf{\Theta}_0^{1/2}, \mathbf{\Omega}_0) = \mathrm{Tr}\,(\mathbf{U}) = 0, \forall \mathbf{U} \in \mathcal{L}$. Recall that

$$\begin{aligned} &\mathrm{Tr}\left(\mathbf{\Theta}_0^{1/2}\mathbf{U}\mathbf{\Theta}_0^{1/2}\right) = (\mathbf{\Theta}_0^{1/2}\mathbf{U}\mathbf{\Theta}_0^{1/2}, \mathbf{I}) \\ &= \sqrt{p}\left\|\mathbf{\Theta}_0^{1/2}\mathbf{U}\mathbf{\Theta}_0^{1/2}\right\|_F \cos(\angle\mathbf{\Theta}_0^{1/2}\mathbf{U}\mathbf{\Theta}_0^{1/2}, \mathbf{I}), \end{aligned}$$

and set $\mathbf{n} = \mathbf{\Omega}_0, \mathbf{v} = \mathbf{I}$ to apply Lemma 2 and get

$$\begin{aligned} &H_{\mathbf{\Omega}_0}(\mathbf{U}) = \frac{p}{p+1}\sin^2(\angle\mathbf{\Theta}_0^{1/2}\mathbf{U}\mathbf{\Theta}_0^{1/2}, \mathbf{I})\left\|\mathbf{\Theta}_0^{1/2}\mathbf{U}\mathbf{\Theta}_0^{1/2}\right\|_F^2 \\ &\geq \frac{p}{p+1}\cos^2(\angle\mathbf{I}, \mathbf{\Omega}_0)\left\|\mathbf{\Theta}_0^{1/2}\mathbf{U}\mathbf{\Theta}_0^{1/2}\right\|_F^2, \forall \mathbf{U} \in \mathcal{L}. \end{aligned}$$

We define the coherence from the identity

$$1-c^2(\mathbf{\Omega}_0) = \cos^2(\angle\mathbf{I}, \mathbf{\Omega}_0) = \left(\frac{\mathrm{Tr}\,(\mathbf{\Omega}_0)}{\|\mathbf{I}\|_F\,\|\mathbf{\Omega}_0\|_F}\right)^2 = \left(\frac{\mathrm{Tr}\,(\mathbf{\Omega}_0)}{\sqrt{p}\,\|\mathbf{\Omega}_0\|_F}\right)^2,$$

as was already introduced in (2). This identity basically says that the convexity properties of the Hessian restricted to $\mathcal{L}$ near the true parameter are defined by the coherence $c(\mathbf{\Omega}_0)$ naturally appearing in the bound (6).

IV. Symmetric Tyler Estimator and its Error Bound

When high-dimensional settings are concerned, the performance of an estimator can be improved by assuming some prior knowledge on the true parameter value. Indeed, such prior structure reduces the number of degrees of freedom in the model and allows more accurate estimation with a small number of samples. Prior knowledge on the structure can originate from the physics of the underlying phenomena or from similar datasets, see e.g. [19–21]. Many covariance structures are easily represented in a convex form, such as Toeplitz, sparse, banded and others, [16, 19–22]. Another important example of linear structure is group symmetry, which has played an important role in statistics since 1940-s, see a broad exposition in [32].

Definition 2. *Let* $\mathcal{G}$ *be a finite unitary matrix group (a finite set containing the identity matrix and closed under matrix multiplication and inversion). Given a set* $\mathcal{V} \subset \mathcal{S}(p)$ *we denote by*

$$\mathcal{V}^{\mathcal{G}} = \{\mathbf{M} \in \mathcal{V} \colon g^H\mathbf{M}g = \mathbf{M}, \forall g \in \mathcal{G}\}.$$

its subset invariant under conjugation by $\mathcal{G}$*. We say* $\mathcal{V}$ *is group symmetric if* $\mathcal{V} = \mathcal{V}^{\mathcal{G}}$ *for some group* $\mathcal{G}$*.*

In the recent paper [24] we have shown that such constraints are g-convex sets, which together with the g-convexity of the target (7), [23], allowed us to extend the work of [32] and to introduce a group symmetric version of Tyler's estimator, which we refer to as STyler. The scale invariant STyler estimator is given by the following fixed point equation

$$\widehat{\mathbf{\Theta}}^{\mathcal{G}} = \frac{p}{n|G|}\sum_{i=1}^{n}\sum_{g\in\mathcal{G}}\frac{(g\mathbf{x}_i)(g\mathbf{x}_i)^H}{\mathbf{x}_i^H[\widehat{\mathbf{\Theta}}^{\mathcal{G}}]^{-1}\mathbf{x}_i}, \tag{12}$$

which can be viewed as applying the standard Tyler fixed point equation (1) to the original data plus additional synthetically rotated copies of it. Alternatively, for the purposes of our analysis, this is the solution of a program similar to (5) with additional group symmetric constraints. Interestingly, the group symmetry structure is known to be equivalent to a (possibly, block) diagonal structure in a certain basis, [25]. For example, consider a well known circulant covariance model, which is invariant under the action of the shift matrix

$$\mathbf{\Pi} = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \tag{13}$$

and all its powers $\mathbf{\Pi}^j, j = 1, \dots, p$, forming a cyclic group. It is known that rotation by the FFT matrix

$$\mathbf{Q}_c = \frac{1}{\sqrt{p}}\begin{pmatrix} 1 & 1 & \dots & 1 \\ w_0 & w_1 & \dots & w_{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ w_0^{p-2} & w_1^{p-2} & \dots & w_{p-1}^{p-2} \\ w_0^{p-1} & w_1^{p-1} & \dots & w_{p-1}^{p-1} \end{pmatrix}, \tag{14}$$

where $w_j = e^{2\pi ij/p}$ are the complex roots of unity, makes every circulant matrix diagonal.

Other examples of group symmetry include important cases of permutation-invariant [32], persymmetric [33] covariance matrices and others. For each of these structures there exists a basis in which the covariance is block diagonal. The specific sizes and multiplicities of the blocks can be calculated using the tools of finite group representation theory, [34, 35]. For any group $\mathcal{G}$, denote by $r(\mathcal{G})$ the number of non-zero elements in the corresponding block diagonal representation of $\mathcal{G}$-invariant matrices, e.g., $r(\mathcal{G}) = p$ in the circular

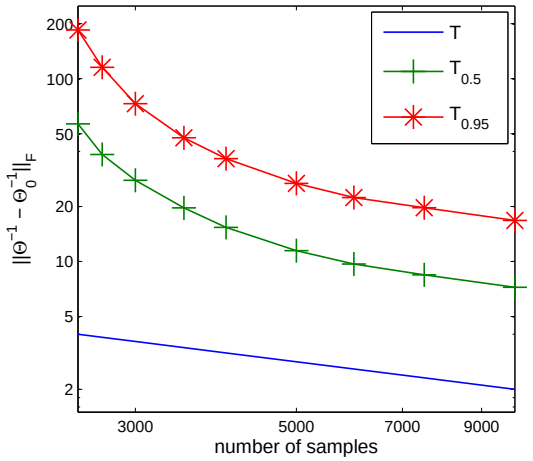

Fig. 1. Tyler's estimator performance bounds, $p = 50$, $\mathbf{\Theta}_0 = \mathbf{I}$.

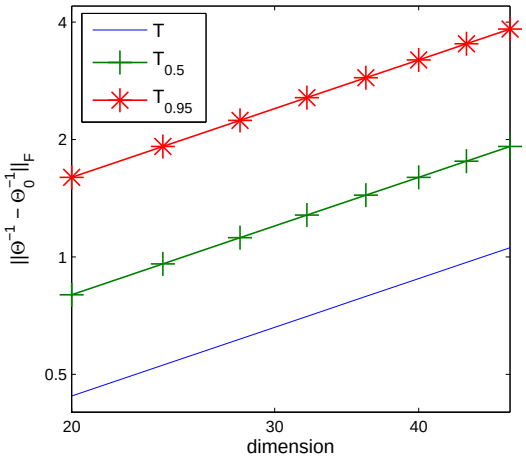

Fig. 2. Tyler's estimator performance bounds, $n = 2500, \mathbf{\Theta}_0(p) = \mathbf{I}_p$.

case. In addition, note that the distribution $\mathcal{U}(\mathbf{I})$ in invariant under rotation of the basis. These properties allow us to obtain a group symmetric variant of Lemma 1 with a tighter bound.

Lemma 3. *Let $\mathbf{\Theta}_0 \in \mathcal{P}(p)^{\mathcal{G}}$, then uniformly over $\mathbf{U}' \in \mathcal{L}'^{\mathcal{G}}$*

$$\mathbb{P}\left(\left|\nabla \widehat{f}_{\mathbf{I}}(\mathbf{U}')\right| \geq \sqrt{2r(\mathcal{G})}t \left\|\mathbf{U}'\right\|_F\right) \leq 2\exp\left(\frac{-nt^2}{2(1+\frac{1.2tp}{\sqrt{r(\mathcal{G})}})}\right).$$

This result basically reduces the dimension of the estimation problem by the factor of $\frac{\sqrt{r(\mathcal{G})}}{p}$, and, thus, improves the performance bounds stated in Theorem 1. More exactly, we obtain

Theorem 2. *Assume we are given $n > p$ i.i.d. copies of $\mathbf{x} \sim \mathcal{U}(\mathbf{\Theta}_0), \mathbf{\Theta}_0 \in \mathcal{P}(p)^{\mathcal{G}}$, then for $\theta \geq 0$ with probability at least*

$$1 - 2\exp\left(\frac{-\theta^2}{2(1+1.2\frac{\theta p}{\sqrt{r(\mathcal{G})n}})}\right) - 2p^2\exp\left(-\frac{n(1-c^2(\mathbf{\Omega}_0))}{80\ln(7p)(1+\frac{1}{p})}\right)\left(1+\frac{8\cdot 10^3(1+\frac{1}{p})^4}{n(1-c^2(\mathbf{\Omega}_0))^4}\right)$$

STyler estimator (12) scaled by the condition $\mathrm{Tr}\left([\widehat{\mathbf{\Theta}}^{\mathcal{G}}]^{-1}\right) = \mathrm{Tr}\left(\mathbf{\Theta}_0^{-1}\right)$ satisfies

$$\left\|[\widehat{\mathbf{\Theta}}^{\mathcal{G}}]^{-1} - \mathbf{\Theta}_0^{-1}\right\|_F \leq \frac{15\theta}{\lambda_{\min}(\mathbf{\Theta}_0)(1-c^2(\mathbf{\Omega}_0))}\sqrt{\frac{r(\mathcal{G})}{n}}. \quad (15)$$

V. Numerical Results

In this section we provide numerical simulations supporting our analysis. Figure 1 compares the behavior of Tyler's estimator with its 0.95 and 0.5-probability bounds for $p = 50, \mathbf{\Theta}_0 = \mathbf{I}$. Figure 2 verifies the dependence of the performance on the dimension.

Figure 3 compares the behavior of the Tyler's estimator with the STyler in the circulant model. As we have explained above in this case the sparse structure is diagonal and $r(\mathcal{G}) = p$, which is verified by the empirical MSEs. For this experiment we set $p = 8$.

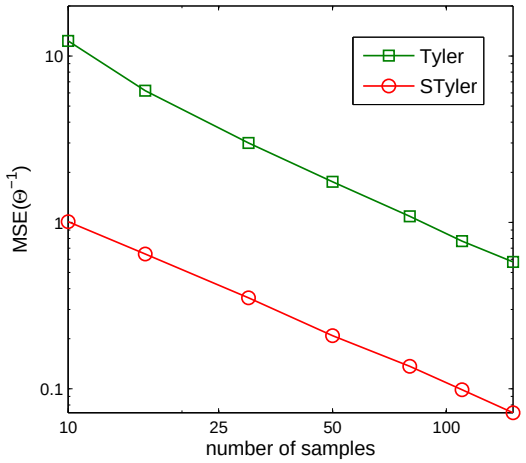

Fig. 3. STyler's performance in the circulant case.

References

[1] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of empirical finance*, vol. 10, no. 5, pp. 603–621, 2003.

[2] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.

[3] D. E. Tyler, "A distribution-free M-estimator of multivariate scatter," *The Annals of Statistics*, vol. 15, no. 1, pp. 234–251, 1987.

[4] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[5] R. A. Maronna, "Robust M-estimators of multivariate location and scatter," *The annals of statistics*, pp. 51–67, 1976.

[6] G. Frahm, "Generalized elliptical distributions: theory and applications," *Universität zu Köln*, 2004.

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

[8] E. Ollila and V. Koivunen, "Robust antenna array processing using M-estimators of pseudo-covariance," *14th IEEE Proceedings on Personal, Indoor and Mobile Radio Communications*,

vol. 3, pp. 2659–2663, 2003.

[9] Y. I. Abramovich, N. K. Spencer, and M. D. Turley, "Time-varying autoregressive (TVAR) models for multiple radar observations," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1298–1311, 2007.

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

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

[12] 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 Transactions on Signal Processing*, vol. 56, no. 1, pp. 34–48, 2008.

[13] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *Compressed sensing: theory and applications, Edited by Y. Eldar and G. Kutyniok, Cambridge University Press*, 2012.

[14] ——, "How close is the sample covariance matrix to the actual covariance matrix?" *Journal of Theoretical Probability*, vol. 25, no. 3, pp. 655–686, 2012.

[15] N. Srivastava and R. Vershynin, "Covariance estimation for distributions with 2+ε moments," *arXiv preprint arXiv:1106.2775*, 2011.

[16] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, pp. 199–227, 2008.

[17] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing l_1-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[18] Y.-H. Li and P.-C. Yeh, "An interpretation of the Moore-Penrose generalized inverse of a singular Fisher Information Matrix," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5532–5536, 2012.

[19] D. R. Fuhrmann, "Application of Toeplitz covariance estimation to adaptive beamforming and detection," *IEEE Transactions on Signal Processing*, vol. 39, no. 10, pp. 2194–2198, 1991.

[20] D. S. Pollock, "Circulant matrices and time-series analysis," *International Journal of Mathematical Education in Science and Technology*, vol. 33, no. 2, pp. 213–230, 2002.

[21] P. Stoica, P. Babu, and J. Li, "SPICE: A sparse covariance-based estimation method for array processing," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 629–638, 2011.

[22] T. T. Cai, Z. Ren, and H. H. Zhou, "Optimal rates of convergence for estimating Toeplitz covariance matrices," *Probability Theory and Related Fields*, pp. 1–43, 2012.

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

[24] I. Soloveychik and A. Wiesel, "Group symmetry and non-Gaussian covariance estimation," *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2013.

[25] K. Murota, Y. Kanno, M. Kojima, and S. Kojima, "A numerical algorithm for block-diagonal decomposition of matrix $*$-algebras with application to semidefinite programming," *Japan Journal of Industrial and Applied Mathematics*, vol. 27, no. 1, pp. 125–160, 2010.

[26] I. Soloveychik and A. Wiesel, "Performance analysis of Tyler's covariance estimator," *submitted to IEEE Transactions on Signal Processing, available: http://arxiv.org/pdf/1401.6926v4.pdf*.

[27] D. E. Tyler, "Statistical analysis for the angular central Gaussian distribution on the sphere," *Biometrika*, vol. 74, no. 3, pp. 579–589, 1987.

[28] G. Frahm and U. Jaekel, "Tyler's M-estimator, random matrix theory, and Generalized Elliptical distributions with applications to finance," Tech. Rep., 2007.

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

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

[31] J. K. Bradley and C. Guestrin, "Sample complexity of composite likelihood," *International Conference on Artificial Intelligence and Statistics*, pp. 136–160, 2012.

[32] P. Shah and V. Chandrasekaran, "Group symmetry and covariance regularization," *Electronic Journal of Statistics*, vol. 6, pp. 1600–1640, 2012.

[33] G. Pailloux, P. Forster, J. Ovarlez, and F. Pascal, "Persymmetric adaptive radar detectors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 47, no. 4, pp. 2376–2390, 2011.

[34] W. Feit, "The representation theory of finite groups," *Elsevier*, 1982.

[35] C. W. Curtis and I. Reiner, "Representation theory of finite groups and associative algebras," *American Mathematical Society*, 1962.