Joint Covariance Estimation with Mutual Linear Structure

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

***Abstract*—We consider the joint estimation of structured covariance matrices. We assume the structure is unknown and perform the estimation using heterogeneous training sets. More precisely, we are given groups of measurements coming from centered normal populations with different covariance matrices. Assuming that all these covariance matrices span a low dimensional affine subspace in the space of symmetric matrices, our aim is to determine this structure. It is then utilized to improve the covariance estimation. We provide an algorithm discovering and exploring the underlying covariance structure and analyze its error bounds. Numerical simulations are presented to illustrate the performance benefits of the proposed algorithm.**

***Index Terms*—Structured covariance estimation, joint covariance estimation.**

I. Introduction

Estimation of covariance matrices is a basic problem in multivariate statistics. It arises in diverse applications such as signal processing [1], genomics [2], financial mathematics [3] and others. Given n copies of a p dimensional real random vector with bounded second moments, the most natural estimator of their covariance is the Sample Covariance Matrix (SCM), being a consistent estimator due to the law of large numbers.

High dimensional covariance estimation is the setting in which the number of measurements n is comparable to the dimension p. In this scenario, the performance of the SCM deteriorates and the approximation becomes poor [4, 5]. The most popular way to improve the behavior of an estimator in such situation is to introduce prior knowledge on the covariance matrix. The particular form of prior knowledge may differ: it may be given as an a prior distribution over the possible parameter values, as in Bayesian framework, or as an edge case, be given as a structure constraint.

In this article we focus on covariance estimation when the true covariance matrix is assumed to possess a linear structure. There are plenty of examples of such settings in the literature. In many engineering applications the physical properties of the measurements impose natural linear constraints on their covariance. A partial list of examples include Toeplitz [6–8], circulant [9, 10], sparse and banded [11, 12], persymmetric [13], factor models [14], permutation invariant [15] including proper complex and proper quaternion covariances represented as real matrices, etc.

An important common feature of the papers listed above is that they consider a single and static environment where the structure of the true covariance matrix, or at least the class of structures, as in sparse case, is known in advance. Often, this is not the case and techniques are needed to learn the structure from the observations. A typical approach is to consider multiple datasets sharing a similar structure but non homogeneous environments [16]. This is, for example, the case in covariance estimation for classification across multiple classes [17]. A related problem addresses the problem of tracking a time varying covariance throughout a single stream of data, where it is assumed that the structure changes at a slower rate than the covariances themselves [18]. Here too, it is natural to divide this stream of data into independent blocks of measurements. From a different perspective, the authors of [19, 20] assumed the populations covariances were picked at random from a Wishart distribution with unknown parameters and sought for their center.

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

Our goal is to introduce and analyze an algorithm for learning the underlying affine structure of a family of covariance matrices, which can also be adapted to a time-varying environment. More exactly, given a few groups of Gaussian measurements having different covariance matrices each, our target is to determine the underlying low-dimensional linear space containing all the covariances. The discovered subspace can be further used to improve the covariance estimation by projecting any unconstrained estimator on it. Most of the previous works considered particular cases of this method, e.g. factor models, entry-wise linear structures like in sparse and banded models, or specific patterns like in Toeplitz, circulant, proper and other models. Our method suggests to treat the SCM of the heterogeneous populations as vectors in the space of matrices and is based on application of principal component analysis to learn their low-dimensional structure. It generalizes the previous approaches and proposes a generic way to determine and utilize the joint linear structure of covariances for better estimation.

The paper is organized as following: first we introduce notations, state the problem and provide examples motivating the work. Then we derive a lower performance bound, propose our new algorithm and outline the derivation of its upper error bound. In the end of the paper we provide numerical simulations demonstrating the advantages of the proposed algorithm.

We denote by $\mathcal{S}(p)$ the $l = \frac{p(p+1)}{2}$ dimensional linear space of $p \times p$ symmetric real matrices and by $\mathcal{P}(p) \subset \mathcal{S}(p)$ the closed cone of positive semi-definite matrices. $\mathbf{I}_d$ stands for the $d \times d$ identity matrix. For a matrix $\mathbf{M}$ its Moore-Penrose generalized inverse is denoted by $\mathbf{M}^\dagger$. For any two matrices $\mathbf{M}$ and $\mathbf{P}$ we denote by $\mathbf{M} \otimes \mathbf{P}$ their tensor (Kronecker) product. $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|_2$ - the spectral norm of matrices, and $\|\cdot\|$ - the Euclidean norm of vectors. For any symmetric matrix $\mathbf{S}$, we denote by $\mathbf{s} = \mathrm{vech}(\mathbf{S})$ a vector obtained by stacking the columns of the lower triangular part of $\mathbf{S}$ into a single column vector. In addition, given an l dimensional column vector $\mathbf{m}$ we denote by $\mathrm{mat}(\mathbf{m})$ the inverse operator constructing a $p \times p$ symmetric matrix such that $\mathrm{vech}(\mathrm{mat}(\mathbf{m})) = \mathbf{m}$. Due to this natural linear bijection below we often consider subsets of $\mathcal{S}(p)$ as subsets of the column space $\mathbb{R}^l$ without specifying this explicitly. In addition, let $\mathrm{vec}(\mathbf{S})$ be a p^2 dimensional vector obtained by stacking the columns of $\mathbf{S}$, and denote by $\mathcal{I}$ its indices corresponding to the related elements of $\mathrm{vech}(\mathbf{S})$.

II. Problem Formulation

We focus on the heterogeneous measurements model, namely assume we are given $K \geq l = \frac{p(p+1)}{2}$ groups of real p dimensional normal random vectors

$$\mathbf{x}_k^i \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k), i = 1, \ldots, n,\ k = 1, \ldots, K, \tag{1}$$

with n independent samples in each group. We assume that

$$\mathbf{Q}_k \in \mathcal{P}(p), k = 1, \ldots, K, \tag{2}$$

belong to an r dimensional affine subspace of $\mathcal{S}(p)$. Our main goal is to estimate this subspace and use it to improve the covariance

estimation. In the analysis we will assume r is known in advance, but we will also explain how to determine it from the data.

Let us list the most common types of affine covariance structures.

- **Diagonal**: The simplest example of a structured covariance is a diagonal matrix. This is often the case when the noise vectors are uncorrelated or can be assumed uncorrelated with great precision. In this case $r = p$.
- **Banded**: A natural approach to covariance modeling is to quantify the statistical relation using the notion of independence or correlation, which corresponds to sparsity in the covariance matrix [11, 12]. Assuming that i-th element of the random vector is uncorrelated with the h-th if $|i-h| > b$ leads to b-banded structure, also known as time varying moving average models. The subspace of symmetric b-banded matrices constitutes an $r = \frac{(2p-b)(b+1)}{2}$ dimensional subspace inside $\mathcal{S}(p)$.
- **Circulant**: The next common type of structured covariance matrices are symmetric circulant matrices, defined as

$$\mathbf{Q} = \begin{pmatrix} q_1 & q_2 & q_3 & \dots & q_p \\ q_p & q_1 & q_2 & \dots & q_{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_2 & q_3 & q_4 & \dots & q_1 \end{pmatrix}, \tag{3}$$

  with the natural symmetry conditions such as $q_p = q_2$, etc. Such matrices are typically used as approximations to Toeplitz matrices which are associated with signals that obey periodic stochastic properties for example the yearly variation of temperature in a particular location. A special case of such processes are the classical stationary processes, which are ubiquitous in engineering, [9, 10]. Symmetric circulant matrices belong to an $r = p/2$ dimensional subspace if p is even and $(p+1)/2$ if it is odd.
- **Toeplitz**: A natural generalization of circulant are Toeplitz matrices. In stationary time series, the covariance between the i-th and the h-th components depend only on the the difference $|i-h|$. This kind of processes is encountered very often in many engineering areas including statistical signal processing, radar imaging, target detection, speech recognition, and communications systems, [7, 8, 21, 22]. In the Toeplitz case $r = p$
- **Proper Complex**: Many physical processes can be conveniently described in terms of complex signals. The most frequently appearing model of complex Gaussian noise is circularly symmetric [23]. Such noise is completely characterized by its mean and rotation invariant hermitian covariance matrix $\mathbf{Q}^C$. We denote centered proper complex distributions as

$$\mathbf{x} \sim \mathcal{CN}(\mathbf{0}, \mathbf{Q}^C).$$

  The real representation of the covariance reads as

$$\mathbf{Q}^R = \frac{1}{2}\begin{pmatrix} \mathrm{Re}(\mathbf{Q}^C) & -\mathrm{Im}(\mathbf{Q}^C) \\ \mathrm{Im}(\mathbf{Q}^C) & \mathrm{Re}(\mathbf{Q}^C) \end{pmatrix}. \tag{4}$$

  We see that $\mathbf{Q}^R$ possess a simple linear structure. Matrices of dimension $p \times p$, possessing such structure constitute an $r = p/2$ dimensional subspace of $\mathcal{S}(p)$.

In the following it will be convenient to use a single matrix notation for the multiple covariances

$$\mathbf{Y} = [\mathrm{vech}(\mathbf{Q}_1) \dots \mathrm{vech}(\mathbf{Q}_K)]. \tag{5}$$

Using this notation, the prior subspace knowledge is equivalent to a low rank constraint

$$\mathbf{Y} = \mathbf{U}\mathbf{Z} + \mathbf{u}_0 \cdot [1, \dots, 1], \tag{6}$$

where $\mathbf{U} \in \mathbb{R}^{l \times r}$ is a basis of the r dimensional subspace spanned by $\mathbf{Q}_1, \dots, \mathbf{Q}_K$, $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_K] \in \mathbb{R}^{r \times K}$ and $\mathbf{u}_0$ is the intercept vector. Essentially our problem reduces to estimation of $\mathbf{Y}$ assuming it is low-rank.

III. Lower Performance Bounds

Before addressing possible solutions for the above covariance structure estimation problem, it is instructive to examine the inherent performance bounds. For this purpose, we use the Cramer-Rao Bound ($\mathbf{CRB}$) to lower bound the Mean Squared Error ($\mathbf{MSE}$) of any unbiased estimator $\widehat{\mathbf{Y}}$ of $\mathbf{Y}$, defined as

$$\mathbf{MSE} = \mathbb{E}\left[\left\|\widehat{\mathbf{Y}} - \mathbf{Y}\right\|_F^2\right]. \tag{7}$$

The $\mathbf{MSE}$ is bounded from below by the trace of the corresponding $\mathbf{CRB}$ matrix. To compute this matrix, for each i we stack the measurements $\mathbf{x}_k^i$ from (1) into a single vector

$$\mathbf{x}^i = \begin{pmatrix} \mathbf{x}_1^i \\ \vdots \\ \mathbf{x}_K^i \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), i = 1, \dots, n. \tag{8}$$

where

$$\begin{aligned} \mathbf{Q}(\mathbf{U}, \mathbf{Z}) &= \mathrm{diag}\left\{\mathbf{Q}_1, \dots, \mathbf{Q}_k\right\} \\ &= \mathrm{diag}\left\{\mathrm{mat}\left(\mathbf{u}_0 + \mathbf{U}\mathbf{z}_1\right), \dots, \mathrm{mat}\left(\mathbf{u}_0 + \mathbf{U}\mathbf{z}_K\right)\right\}. \end{aligned} \tag{9}$$

For simplicity in this section we assume $\mathbf{u}_0$ and r are both known. The Jacobian matrix of this parametrization reads as

$$\begin{aligned} \mathbf{J} = \frac{\partial \mathbf{Q}}{\partial(\mathbf{U}, \mathbf{Z})} &= \begin{pmatrix} \frac{\partial \mathbf{q}_1}{\partial \mathbf{U}} & \frac{\partial \mathbf{q}_1}{\partial \mathbf{z}_1} & 0 & \dots & 0 \\ \frac{\partial \mathbf{q}_2}{\partial \mathbf{U}} & 0 & \frac{\partial \mathbf{q}_2}{\partial \mathbf{z}_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{q}_K}{\partial \mathbf{U}} & 0 & 0 & \dots & \frac{\partial \mathbf{q}_K}{\partial \mathbf{z}_K} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{z}_1^T \otimes \mathbf{I}_l & \mathbf{U} & 0 & \dots & 0 \\ \mathbf{z}_2^T \otimes \mathbf{I}_l & 0 & \mathbf{U} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}_K^T \otimes \mathbf{I}_l & 0 & 0 & \dots & \mathbf{U} \end{pmatrix} \in \mathbb{R}^{lK \times (lr + Kr)}, \end{aligned} \tag{10}$$

where we have used the following notation:

$$\frac{\partial \mathbf{q}_k}{\partial \mathbf{U}} = \begin{bmatrix} \frac{\partial \mathbf{q}_k}{\partial \mathbf{u}_1} & \frac{\partial \mathbf{q}_k}{\partial \mathbf{u}_2} & \dots & \frac{\partial \mathbf{q}_k}{\partial \mathbf{u}_r} \end{bmatrix}, \tag{11}$$

and the formulas

$$\frac{\partial \mathbf{q}_k}{\partial \mathbf{u}_j} = \frac{\partial \mathbf{U}\mathbf{z}_k}{\partial \mathbf{u}_j} = z_k^j \mathbf{I}_l, \quad \frac{\partial \mathbf{q}_k}{\partial \mathbf{z}_k} = \frac{\partial \mathbf{U}\mathbf{z}_k}{\partial \mathbf{z}_k} = \mathbf{U}. \tag{12}$$

Note that

$$\mathrm{rank}\,(\mathbf{J}) = lr + Kr - r^2 \leq \min[lK, lr + Kr], \tag{13}$$

reflecting the fact that the parametrization of $\mathbf{Q}$ or $\mathbf{Y}$ by the pair $(\mathbf{U}, \mathbf{Z})$ is unidentifiable. Indeed for any invertible matrix $\mathbf{A}$, the pair $(\mathbf{U}\mathbf{A}, \mathbf{A}^{-1}\mathbf{Z})$ fits as good. Due to this ambiguity the matrix $\mathbf{FIM}(\mathbf{U}, \mathbf{Z})$ is singular and in order to compute the $\mathbf{CRB}$ we use the Moore-Penrose pseudo-inverse of $\mathbf{FIM}(\mathbf{U}, \mathbf{Z})$ instead of inverse, as justified by [24]. Given n i.i.d. samples $\mathbf{x}^i, i = 1, \dots, n$, we obtain

$$\mathbf{CRB} = \frac{1}{n}\mathbf{J}\mathbf{FIM}(\mathbf{U}, \mathbf{Z})^\dagger \mathbf{J}^T. \tag{14}$$

For the Gaussian population the matrix $\mathbf{FIM}(\mathbf{U}, \mathbf{Z})$ is given by

$$\mathbf{FIM}(\mathbf{U}, \mathbf{Z}) = \frac{1}{2}\mathbf{J}^T \mathrm{diag}\left\{\left[\mathbf{Q}_k^{-1} \otimes \mathbf{Q}_k^{-1}\right]_{\mathcal{I},\mathcal{I}}\right\}\mathbf{J}, \tag{15}$$

where $[\mathbf{M}]_{\mathcal{I},\mathcal{I}}$ is the square submatrix of $\mathbf{M}$ corresponding to the indices from $\mathcal{I}$. The bound on the $\mathbf{MSE}$ is therefore given by

$$\begin{aligned}\mathbf{MSE} &\geq \mathrm{Tr}\left(\mathbf{CRB}\right) = \frac{1}{n}\mathrm{Tr}\left(\mathbf{FIM}(\mathbf{U},\mathbf{Z})^{\dagger}\mathbf{J}^T\mathbf{J}\right)\\ &= \frac{2}{n}\mathrm{Tr}\left(\left[\mathbf{J}^T\mathrm{diag}\left\{\left[\mathbf{Q}_k^{-1}\otimes\mathbf{Q}_k^{-1}\right]_{\mathcal{I},\mathcal{I}}\right\}\mathbf{J}\right]^{\dagger}\mathbf{J}^T\mathbf{J}\right). \qquad (16)\end{aligned}$$

To get more insight on this expression we bound it from below. Denote $\underline{\lambda} = \min_k \lambda_p(\mathbf{Q}_k)$, where $\lambda_p(\mathbf{M})$ is the minimal eigenvalue of $\mathbf{M}$ to get a bound

$$\begin{aligned}\mathbf{MSE} &\geq \frac{2\underline{\lambda}^2}{n}\mathrm{Tr}\left(\left[\mathbf{J}^T\mathbf{J}\right]^{\dagger}\mathbf{J}^T\mathbf{J}\right)\\ &= \frac{2\underline{\lambda}^2}{n}\mathrm{rank}\left(\mathbf{J}\right) = \frac{2\underline{\lambda}^2}{n}(lr + Kr - r^2). \qquad (17)\end{aligned}$$

The dependence on the model parameters here is similar to that obtained by [25] for the problem of low-rank matrix reconstruction. An important quantity is the marginal $\mathbf{MSE}$ per one matrix $\mathbf{Q}_k$, which is proportional to

$$\frac{\mathbf{MSE}}{K} \sim \frac{lr - r^2}{Kn} + \frac{r}{n}. \qquad (18)$$

IV. Algorithm

In this section we present our algorithm for joint estimation of the covariances $\mathbf{Q}_1,\ldots,\mathbf{Q}_K$ utilizing the representation (6) of $\mathbf{Y}$. For this purpose consider the SCM of the k-th group of measurements

$$\mathbf{S}_k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^k\mathbf{x}_i^{kT}, \qquad (19)$$

and denote

$$\mathbf{s}_k = \mathrm{vech}\left(\mathbf{S}_k\right). \qquad (20)$$

Compute the SCM average

$$\widehat{\mathbf{u}}_0 = \frac{1}{K}\sum_{k=1}^{K}\mathbf{s}_k, \qquad (21)$$

and consider the matrix

$$\widehat{\mathbf{Y}}' = [\mathbf{s}_1 - \widehat{\mathbf{u}}_0,\ldots,\mathbf{s}_K - \widehat{\mathbf{u}}_0], \qquad (22)$$

The SVD of $\widehat{\mathbf{Y}}'$ reads as

$$\widehat{\mathbf{Y}}' = \widehat{\mathbf{U}}\begin{pmatrix}\widehat{\mathbf{\Sigma}} & 0\\ 0 & \widehat{\mathbf{\Sigma}}_n\end{pmatrix}\widehat{\mathbf{W}}^T = [\widehat{\mathbf{U}}_1\widehat{\mathbf{U}}_2]\begin{pmatrix}\widehat{\mathbf{\Sigma}} & 0\\ 0 & \widehat{\mathbf{\Sigma}}_n\end{pmatrix}[\widehat{\mathbf{W}}_1\widehat{\mathbf{W}}_2]^T, \qquad (23)$$

where the singular values are sorted in the decreasing order and $\widehat{\mathbf{\Sigma}} \in \mathbb{R}^{r\times r}$. We propose to use the matrix

$$\widetilde{\mathbf{Y}}' = \widehat{\mathbf{U}}_1\widehat{\mathbf{\Sigma}}\widehat{\mathbf{W}}_1^T, \qquad (24)$$

as an estimator of

$$\mathbf{Y}' = \mathbf{U}\mathbf{Z}. \qquad (25)$$

This approach is based on Eckart-Young theorem and we refer to it as Truncated SVD (TSVD) method [26]. Finally, for the estimator of $\mathbf{Y}$ we have

$$\widetilde{\mathbf{Y}} = \widetilde{\mathbf{Y}}' + \widehat{\mathbf{u}}_0\cdot[1,\ldots,1]. \qquad (26)$$

In the real world settings the rank r is rarely known in advance and one needs to estimate it from the data before applying the TSVD technique. It is intuitive to think about rank estimation, followed by TSVD, simply as thresholding of the data singular values. A large variety of thresholding techniques exist, e.g. hard thresholding, see [27] and references therein. We propose another method based on taking the largest singular values carrying a fixed percentage of the power of signal depending on the Signal to Noise Ratio

$$\mathbf{SNR} = \frac{\|\mathbf{Y}'\|_F^2}{\mathbb{E}\left[\left\|\widehat{\mathbf{Y}}' - \mathbf{Y}'\right\|_F^2\right]}. \qquad (27)$$

In our work as a rule of thumb we took a $\frac{\mathbf{SNR}+r/l}{\mathbf{SNR}+1}$ power threshold, when the $\mathbf{SNR}$ was known or its estimation otherwise. Due to lack of space we postpone the detailed explanation of this rule for the full paper.

A. Upper Performance Bound

In this section we outline the performance analysis of the proposed TSVD algorithm assuming r and $\mathbf{u}_0$ are known. Introduce the SVD of $\mathbf{Y}'$

$$\mathbf{Y}' = [\mathbf{U}_1\mathbf{U}_2]\begin{pmatrix}\mathbf{\Sigma} & 0\\ 0 & 0\end{pmatrix}[\mathbf{W}_1\mathbf{W}_2]^T, \mathbf{\Sigma}\in\mathbb{R}^{r\times r}, \qquad (28)$$

and denote by $\mathbf{R}$ the zero mean noise matrix

$$\mathbf{R} = \widehat{\mathbf{Y}}' - \mathbf{Y}'. \qquad (29)$$

Theorem 1. *When r and $\mathbf{u}_0$ are known,*

$$\left\|\widetilde{\mathbf{Y}} - \mathbf{Y}\right\|_F \leq \sqrt{r}\left\|\mathbf{R}\right\|_2\left(\frac{\sqrt{2}}{\sigma_r(\mathbf{\Sigma})}\left(\sigma_1(\mathbf{\Sigma}) + 1 + \left\|\mathbf{R}\right\|_2\right) + 1\right). \qquad (30)$$

Proof. Use the triangle inequality, Weyl's and Wedin's theorems (4.11 and 4.2 from [28]) to get

$$\begin{aligned}&\left\|\widetilde{\mathbf{Y}} - \mathbf{Y}\right\|_F = \left\|\widetilde{\mathbf{Y}}' - \mathbf{Y}'\right\|_F \leq \left\|\widehat{\mathbf{U}}_1\widehat{\mathbf{\Sigma}}\widehat{\mathbf{W}}_1 - \mathbf{U}_1\widehat{\mathbf{\Sigma}}\widehat{\mathbf{W}}_1\right\|_F\\ &+ \left\|\mathbf{U}_1\widehat{\mathbf{\Sigma}}\widehat{\mathbf{W}}_1 - \mathbf{U}_1\mathbf{\Sigma}\widehat{\mathbf{W}}_1\right\|_F + \left\|\mathbf{U}_1\mathbf{\Sigma}\widehat{\mathbf{W}}_1 - \mathbf{U}_1\mathbf{\Sigma}\mathbf{W}_1\right\|_F\\ &\leq \left\|\widehat{\mathbf{\Sigma}}\right\|_2\left\|\widehat{\mathbf{U}}_1 - \mathbf{U}_1\right\|_F + \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_F + \sigma_1(\mathbf{\Sigma})\left\|\widehat{\mathbf{W}}_1 - \mathbf{W}_1\right\|_F\\ &\leq \left(\left\|\widehat{\mathbf{\Sigma}}\right\|_2 + \sigma_1(\mathbf{\Sigma})\right)\sqrt{2r}\frac{\|\mathbf{R}\|_2}{\sigma_r(\mathbf{\Sigma})} + \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_F\\ &\leq \sqrt{2r}\frac{1+\sigma_1(\mathbf{\Sigma})}{\sigma_r(\mathbf{\Sigma})}\left\|\mathbf{R}\right\|_2 + \left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_2\left(\sqrt{r} + \frac{\sqrt{2r}}{\sigma_r(\mathbf{\Sigma})}\left\|\mathbf{R}\right\|_2\right)\\ &\leq \sqrt{r}\left\|\mathbf{R}\right\|_2\left(\frac{\sqrt{2}}{\sigma_r(\mathbf{\Sigma})}\left(\sigma_1(\mathbf{\Sigma}) + 1 + \left\|\mathbf{R}\right\|_2\right) + 1\right). \qquad (31)\end{aligned}$$

□

As expected, Theorem 1 shows that the Frobenius norm or the error is proportional to the squared root of the rank rather than the dimension. Intuitively, this captures the correct dependence on the model parameters. A rigorous analysis of the proposed TSVD algorithm requires much stronger tools of Random Matrix Theory. We postpone it to the full publication.

V. Numerical Simulations

For our first experiment we took a Toeplitz model with $p = 10, \mathbf{U}_0 = \mathbf{I}_p$. The true covariances were positive definite matrices generated as $\mathbf{Q}_k = \mathbf{U}_0 + \sum_{j=1}^{r} z_k^j\mathbf{D}_j/\left\|\mathbf{D}_j\right\|_F$, where $\mathbf{D}_j$ has ones on the j-th and $-j$-th subdiagonals and zeros otherwise, $r = p-1$ and z_k^j were i.i.d. uniformly distributed over the interval $[-\frac{1}{2},\frac{1}{2}]$. Figure 1 shows the dependence of the $\mathbf{MSE}$ on n when $K = l$. In the unknown r case we took $\alpha = \frac{\mathbf{SNR}+r/l}{\mathbf{SNR}+1}$ power threshold for our TSVD algorithm. For comparison we also plot the $\mathbf{MSE}$-s of the SCM and its projection onto the known subspace structure and the true $\mathbf{CRB}$ bound given by (16). In the second experiment we set $n = 100$ fixed and explored the dependence of the $\mathbf{MSE}$ on the

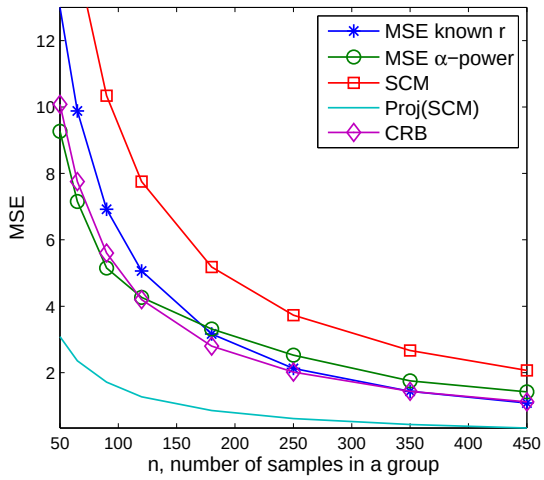

Fig. 1. TSVD algorithm performance.

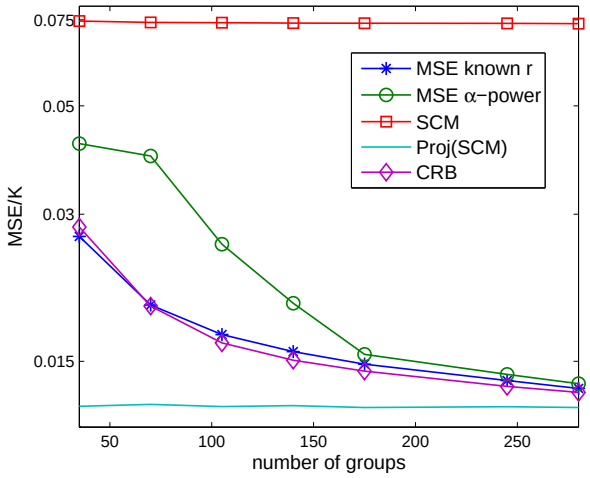

Fig. 2. Marginal TSVD algorithm performance, $n = 100$.

number of groups K in the same setting as before. Figure 2 verifies that the marginal **MSE** depends on K as predicted by formula (18).

For the second experiment we considered the problem of tracking a time-varying covariance in complex populations. We used the Data Generating Process (DGP) of Patton and Sheppard, [29], which allows for dynamically changing covariances in the spirit of a multivariate GARCH-type model, [30, 31]. One of variations of this DGP suggests the following data model:

$$\mathbf{x}_t = \mathbf{H}_t^{1/2}\mathbf{y}_t, t = 1, \dots Kn, \tag{32}$$

where we assumed the generating data to be proper complex, $\mathbf{y}_t \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$ and defined the hermitian time-varying covariance $\mathbf{H}_t$ to change according to the law

$$\widehat{\mathbf{H}}_t = (1-\beta)\mathbf{H}_{t-1} + \beta\mathbf{M}_t\mathbf{M}_t^H, \tag{33}$$

$$\mathbf{H}_t = \frac{\widehat{\mathbf{H}}_t}{\left\|\widehat{\mathbf{H}}_t\right\|_F}, t = 1, \dots Kn. \tag{34}$$

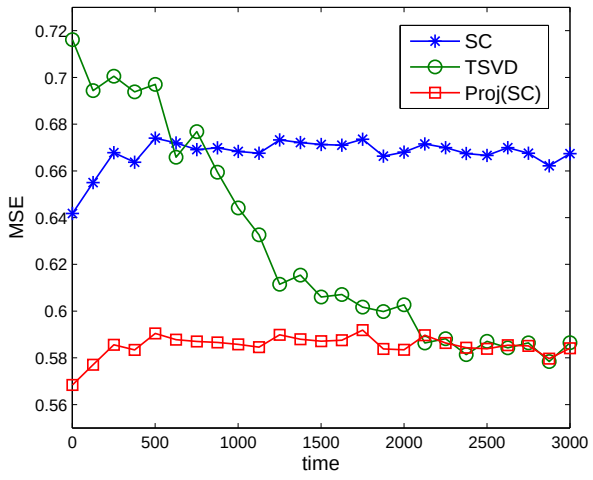

Fig. 3. Learning the low-dimensional subspace with time.

Here $\mathbf{M}_t$ are random $p \times p$ matrices with i.i.d. standard normally distributed entries, $\mathbf{H}_0$ is arbitrary positive-definite hermitian and $\beta \in [0, 1]$.

The low-dimensional structure appearing in this setting is due to properness of the covariances (see (4)). In order to explore it, the obtained complex data was represented as double-sized real measurements. Each n clock ticks we formed the SCM $\mathbf{S}_{\frac{t}{n}}$ of the last n measurements, where t was the last time count. Then we concatenated the vector vech $\left(\mathbf{S}_{\frac{t}{n}}\right)$ to the matrix $\mathbf{Y}_{\frac{t}{n}-1} \in \mathbb{R}^{l\times(\frac{t}{n}-1)}$ of growing size to obtain $\mathbf{Y}_{\frac{t}{n}}$ and applied our TSVD algorithm to it. Thus, our structure knowledge was updated every n ticks, and we expected the error of the covariance estimation to decrease with time. We performed the experiment with $p = 4, K = 100, n = 30, \beta = 0.01$ and used 90%-power threshold to discover the underlying low-dimensional structure. Figure 3 compares the temporal behavior of the **MSE**s of the SCM, TSVD applied to it and the projection of SCM on the subspace spanned by proper covariances. The **MSE**s were obtained by averaging the squared errors over 10000 iterations.

References

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

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

[3] 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.

[4] J. Fan, Y. Fan, and J. Lv, "High dimensional covariance matrix estimation using a factor model," *Journal of Econometrics*, vol. 147, no. 1, pp. 186–197, 2008.

[5] 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.

[6] D. L. Snyder, J. A. O'Sullivan, and M. I. Miller, "The use of maximum likelihood estimation for forming images of diffuse

radar targets from delay-Doppler data," *IEEE Transactions on Information Theory*, vol. 35, no. 3, pp. 536–548, 1989.

[7] 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.

[8] W. J. Roberts and Y. Ephraim, "Hidden Markov modeling of speech using Toeplitz covariance matrices," *Speech Communication*, vol. 31, no. 1, pp. 1–14, 2000.

[9] A. Dembo, C. Mallows, and L. Shepp, "Embedding nonnegative definite Toeplitz matrices in nonnegative definite circulant matrices, with application to covariance estimation," *IEEE Transactions on Information Theory*, vol. 35, no. 6, pp. 1206–1212, 1989.

[10] T. T. Cai, Z. Ren, and H. H. Zhou, "Optimal rates of convergence for estimating Toeplitz covariance matrices," *Probability Theory and Related Fields*, vol. 156, no. 1-2, pp. 101–143, 2013.

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

[12] E. Levina, A. Rothman, J. Zhu *et al.*, "Sparse estimation of large covariance matrices via a nested lasso penalty," *The Annals of Applied Statistics*, vol. 2, no. 1, pp. 245–263, 2008.

[13] 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.

[14] R. F. Engle, V. K. Ng, and M. Rothschild, "Asset pricing with a factor-ARCH covariance structure: Empirical estimates for treasury bills," *Journal of Econometrics*, vol. 45, no. 1, pp. 213–237, 1990.

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

[16] J. Guo, E. Levina, G. Michailidis, and J. Zhu, "Joint estimation of multiple graphical models," *Biometrika*, vol. 98, no. 1, pp. 1–15, 2011.

[17] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 373–397, 2014.

[18] A. Ahmed and E. P. Xing, "Recovering time-varying networks of dependencies in social and biological studies," *Proceedings of the National Academy of Sciences*, vol. 106, no. 29, pp. 11 878–11 883, 2009.

[19] S. Bidon, O. Besson, and J.-Y. Tourneret, "A Bayesian approach to adaptive detection in nonhomogeneous environments," *Signal Processing, IEEE Transactions on*, vol. 56, no. 1, pp. 205–217, 2008.

[20] O. Besson, S. Bidon, and J.-Y. Tourneret, "Covariance matrix estimation with heterogeneous samples," *Signal Processing, IEEE Transactions on*, vol. 56, no. 3, pp. 909–920, 2008.

[21] A. Wiesel, O. Bibi, and A. Globerson, "Time varying autoregressive moving average models for covariance estimation," *IEEE Transactions on Signal Processing*, vol. 61, no. 11, pp. 2791–2801, 2013.

[22] R. Dahlhaus, "Efficient parameter estimation for self-similar processes," *The Annals of Statistics*, pp. 1749–1766, 1989.

[23] S. M. Kay, "Fundamentals of statistical signal processing, volume i: Estimation theory (v. 1)," 1993.

[24] 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.

[25] G. Tang and A. Nehorai, "Lower bounds on the mean-squared error of low-rank matrix reconstruction," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4559–4571, 2011.

[26] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM review*, vol. 35, no. 4, pp. 551–566, 1993.

[27] D. L. Donoho and M. Gavish, "The optimal hard threshold for singular values is $4/\sqrt{3}$," *arXiv preprint arXiv:1305.5870*, 2013.

[28] G. W. Stewart and J.-G. Sun, "Matrix perturbation theory," 1990.

[29] A. J. Patton and K. Sheppard, "Evaluating volatility and correlation forecasts," pp. 801–838, 2009.

[30] T. Bollerslev, R. F. Engle, and J. M. Wooldridge, "A capital asset pricing model with time-varying covariances," *The Journal of Political Economy*, pp. 116–131, 1988.

[31] B. Hansson and P. Hordahl, "Testing the conditional CAPM using multivariate GARCH-M," *Applied Financial Economics*, vol. 8, no. 4, pp. 377–388, 1998.