Joint Inverse Covariances 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 problem of joint estimation of structured inverse covariance matrices. We assume the structure is unknown and perform the estimation using groups of measurements coming from populations with different covariances. Given that the inverse covariances 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 estimation of the inverse covariances. We propose a novel optimization algorithm discovering and exploring the underlying structure and provide its efficient implementation. Numerical simulations are presented to illustrate the performance benefits of the proposed algorithm.**

***Index Terms*—Structured inverse covariance estimation, joint inverse covariance estimation, graphical models.**

I. Introduction

Large scale covariance and inverse covariance estimation using a small number of measurements is a fundamental problem in modern multivariate statistics. In many applications, e.g., linear array processing, climatology, spectroscopy, and longitudinal data analysis statistical properties of variables can be related due to natural physical features of the systems. Such relations often imply linear structure in the corresponding population covariance or inverse covariance matrices.

In this paper we focus on structured inverse covariance (concentration) matrix estimation. There are plenty of examples of inverse covariance estimation with linear structure. A partial list includes banded [1–3], circulant [4, 5], sparse (graphical) models [6, 7], etc. An important common feature of these works is that they consider a single and static environment where the structure of the true concentration 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 [8–11]. This is, for example, the case in covariance estimation for classification across multiple classes [12]. A related problem addresses tracking a time varying covariance throughout a stream of data [13, 14], where it is assumed that the structure changes at a slower rate than the covariances themselves [15]. Here too, it is natural to divide this stream of data into independent blocks of measurements. In [16] we considered a similar setting, when the covariance matrices share the same linear structure.

Our goal is to first rigorously state the problem of joint concentration matrices estimation with linear structure and derive the lower performance bound for any unbiased estimator. Secondly, we propose and analyze a new algorithm of learning and exploring this structure to improve estimation of the concentration matrices. More exactly, given a few groups of measurements having different covariance matrices each, our target is to determine the underlying low dimensional linear space containing or approximately containing the concentration matrices of all the groups. The discovered subspace can be further used to improve the concentration 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 cases, or specific patterns like in banded, circulant and other models. We propose a new generic algorithm based on joint negative log-likelihood minimization penalized by the dimensionality of the subspace containing the inverse covariances. In [16] we approached the problem of low dimensional covariance estimation using the Truncated SVD (TSVD) technique applied to the sample covariance matrices (SCM-s) of the groups of measurements. The reason we do not use the TSVD approach straight forwardly in the case of concentrations is that it would require inversion of the SCM-s, which becomes problematic when the number of samples in groups is relatively small. The algorithm we propose avoids this restriction and demonstrates good performance even when the amount of measurements is insufficient.

The rest of the text is organized as following: first we introduce notations, state the problem and illustrate examples. Then we derive the lower performance bound, propose our Joint Inverse Covariance Estimation (JICE) algorithm and provide its efficient implementation. In the end of the paper we provide numerical simulations demonstrating the advantages of the proposed algorithm.

Given $p \in \mathbb{N}$, denote by $\mathcal{S}(p)$ the $l = \frac{p(p+1)}{2}$ dimensional linear space of $p \times p$ symmetric real 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 $\mathbf{M}^\dagger$. If $\mathbf{M}$ is symmetric, we write $\mathbf{M} \succ 0$ when it is positive definite. 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, $\|\cdot\|_2$ - the spectral and $\|\cdot\|_*$ - the nuclear (trace) norms of matrices. For any symmetric matrix $\mathbf{S}$, $\mathbf{s} = \mathrm{vech}(\mathbf{S})$ is a vector obtained by stacking the columns of the lower triangular part of $\mathbf{S}$ into a single column. 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 $\mathbb{R}^l$. 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 and Examples

Consider a heterogeneous Gaussian model, namely, assume we are given $K \geqslant l = \frac{p(p+1)}{2}$ groups of real p dimensional normal random vectors

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

with n i.i.d. (independent and identically distributed) samples in each group and covariances

$$\mathbf{Q}_k = \mathbb{E}[\mathbf{x}_k \mathbf{x}_k^T], \quad k = 1, \ldots, K. \tag{2}$$

We assume that the inverse covariances

$$\mathbf{T}_k = \mathbf{Q}_k^{-1} \succ 0, \quad k = 1, \ldots, K, \tag{3}$$

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

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

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

Let us list a few common affine subspaces which naturally appear in typical signal processing applications.

- **Diagonal**: The simplest example of a structured concentration matrix is a diagonal matrix. This is often the case when the noise vectors are uncorrelated or can be assumed such with great precision. In this case $r = p$. Note that the diagonal structure remains unaltered under inversion, thus making diagonal concentration equivalent to the diagonal covariance case, considered in [16].
- **Banded**: It is often reasonable to assume that the non-neighboring elements of a normal random vector are conditionally independent given all the other elements. Specifically, claiming that i-th element of the random vector is conditionally independent on the h-th if $|i-h| > b$ leads to the b-banded inverse covariance structure. The subspace of symmetric b-banded matrices constitutes an $r = \frac{(2p-b)(b+1)}{2}$ dimensional subspace inside $\mathcal{S}(p)$. Banded inverse covariance matrices are ubiquitous in graphical models, [6, 7].
- **Circulant**: The next common type of structured concentration matrices are symmetric circulant matrices, defined as

$$\mathbf{T} = \begin{pmatrix} t_1 & t_2 & t_3 & \dots & t_p \\ t_p & t_1 & t_2 & \dots & t_{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_2 & t_3 & t_4 & \dots & t_1 \end{pmatrix}, \tag{4}$$

  with the natural symmetry conditions such as $t_p = t_2$, etc. Such matrices are typically used as approximations to Toeplitz matrices which are associated with signals obeying periodic stochastic properties, e.g. the yearly variation of temperature in a particular location. A special case of such processes are classical stationary processes, which are ubiquitous in engineering, [4, 5]. Symmetric circulant matrices constitute an $r = p/2$ dimensional subspace if p is even and $(p+1)/2$ if it is odd. Interestingly, like in the diagonal case, this structure does not change under inversion, [17], making the estimation problems in covariances and concentrations analogous, [16].
- **Sparse**: Sparse inverse covariance models generalize banded structures and are very common. In multivariate Gaussian distributions zero entries of the concentration reveal conditional independences. When graph representation is utilized to express the relations between the variates, zeros in inverse covariance are translated to missing edges in the graph making it sparse. Recently, Gaussian graphical models have attracted considerable attention due to developments in biology, medicine, neuroscience, compressed sensing and many other areas, [6, 7, 18, 19]. An important property of the sparse graphical models is that they do not usually assume the graph structure known in advance and one of the purposes of most estimation algorithms is to define this structure.

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

$$\mathbf{t}_k = \operatorname{vech}(\mathbf{T}_k), \quad k = 1, \dots, K, \tag{5}$$

$$\mathbf{Y} = [\mathbf{t}_1, \dots, \mathbf{t}_K]. \tag{6}$$

Using these notation, the prior subspace knowledge discussed above is equivalent to a low-rank constraint

$$\mathbf{Y} = \mathbf{U}\mathbf{Z}, \tag{7}$$

where $\mathbf{U} \in \mathbb{R}^{l \times r}$ and $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_K] \in \mathbb{R}^{r \times K}$. Essentially our problem reduces to estimation of $\mathbf{Y}$ assuming it is low-rank. In the analysis we will assume r is known in advance, but the algorithm we propose recovers it from the data.

III. Lower Performance Bounds

Before addressing possible solutions for the above inverse 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

$$\mathbf{MSE} = \mathbb{E}\left[\left\|\widehat{\mathbf{Y}} - \mathbf{Y}\right\|_F^2\right]. \tag{8}$$

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 $\mathbf{x}_k^i$ into a single vector

$$\mathbf{x}^i = \begin{pmatrix} \mathbf{x}_1^i \\ \vdots \\ \mathbf{x}_K^i \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{T}^{-1}), \quad i = 1, \dots, n. \tag{9}$$

where the block-diagonal matrix $\mathbf{T}$ is defined as

$$\begin{aligned} \mathbf{T}(\mathbf{U}, \mathbf{Z}) &= \operatorname{diag}\{\mathbf{T}_1, \dots, \mathbf{T}_k\} \\ &= \operatorname{diag}\{\operatorname{mat}(\mathbf{U}\mathbf{z}_1), \dots, \operatorname{mat}(\mathbf{U}\mathbf{z}_K)\}. \end{aligned} \tag{10}$$

The Jacobian matrix of this parametrization reads as

$$\begin{aligned} \mathbf{J} = \frac{\partial \mathbf{T}}{\partial(\mathbf{U}, \mathbf{Z})} &= \begin{pmatrix} \frac{\partial \mathbf{t}_1}{\partial \mathbf{U}} & \frac{\partial \mathbf{t}_1}{\partial \mathbf{z}_1} & 0 & \dots & 0 \\ \frac{\partial \mathbf{t}_2}{\partial \mathbf{U}} & 0 & \frac{\partial \mathbf{t}_2}{\partial \mathbf{z}_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{t}_K}{\partial \mathbf{U}} & 0 & 0 & \dots & \frac{\partial \mathbf{t}_K}{\partial \mathbf{z}_K} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{z}_1^T \otimes \mathbf{I}_l & \mathbf{U} & 0 & \dots & 0 \\ \mathbf{z}_2^T \otimes \mathbf{I}_l & 0 & \mathbf{U} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{z}_K^T \otimes \mathbf{I}_l & 0 & 0 & \dots & \mathbf{U} \end{pmatrix} \in \mathbb{R}^{lK \times (lr + Kr)}, \end{aligned} \tag{11}$$

where we have used the following notation:

$$\frac{\partial \mathbf{t}_k}{\partial \mathbf{U}} = \left[\frac{\partial \mathbf{t}_k}{\partial \mathbf{u}_1} \; \frac{\partial \mathbf{t}_k}{\partial \mathbf{u}_2} \; \dots \; \frac{\partial \mathbf{t}_k}{\partial \mathbf{u}_r}\right], \tag{12}$$

and the formulae

$$\frac{\partial \mathbf{t}_k}{\partial \mathbf{u}_j} = \frac{\partial \mathbf{U}\mathbf{z}_k}{\partial \mathbf{u}_j} = z_k^j \mathbf{I}_l, \quad \frac{\partial \mathbf{t}_k}{\partial \mathbf{z}_k} = \frac{\partial \mathbf{U}\mathbf{z}_k}{\partial \mathbf{z}_k} = \mathbf{U}. \tag{13}$$

Since at most r of the vectors $\mathbf{z}_1, \dots, \mathbf{z}_K$ are linearly independent, at most $lr + (K-r)r$ rows of the matrix $\mathbf{J}$ can be linearly independent and, therefore, in this case we obtain

$$\operatorname{rank} \mathbf{J} = lr + Kr - r^2 \leqslant \min[lK, lr + Kr], \tag{14}$$

reflecting the fact that the parametrization of $\mathbf{T}$ 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 in [20]. Given n i.i.d. samples $\mathbf{x}^i, i = 1, \dots, n$, we obtain

$$\mathbf{CRB} = \frac{1}{n} \mathbf{J}\, \mathbf{FIM}(\mathbf{U}, \mathbf{Z})^{\dagger} \mathbf{J}^T. \tag{15}$$

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

$$\mathbf{FIM}(\mathbf{U}, \mathbf{Z}) = \frac{1}{2} \mathbf{J}^T \operatorname{diag}\left\{\left[\mathbf{T}_k^{-1} \otimes \mathbf{T}_k^{-1}\right]_{\mathcal{I},\mathcal{I}}\right\} \mathbf{J}, \tag{16}$$

where $[\mathbf{M}]_{\mathcal{I},\mathcal{I}}$ is the square submatrix of $\mathbf{M}$ corresponding to the index set $\mathcal{I}$, defined in the notations section. The bound on the $\mathbf{MSE}$ is therefore given by

$$\begin{aligned}\mathbf{MSE} &\geqslant \mathrm{Tr}\left(\mathbf{CRB}\right) = \frac{1}{n}\mathrm{Tr}\left(\mathbf{FIM}(\mathbf{U},\mathbf{Z})^{\dagger}\mathbf{J}^T\mathbf{J}\right) \\ &= \frac{2}{n}\mathrm{Tr}\left(\left[\mathbf{J}^T\mathrm{diag}\left\{\left[\mathbf{T}_k^{-1}\otimes\mathbf{T}_k^{-1}\right]_{\mathcal{I},\mathcal{I}}\right\}\mathbf{J}\right]^{\dagger}\mathbf{J}^T\mathbf{J}\right). \quad (17)\end{aligned}$$

To get more intuition on the dependence of the $\mathbf{MSE}$ on the model parameters, we bound it from below. Denote

$$\underline{\lambda} = \min_k\left[\left\|\mathbf{T}_k^{-1}\right\|_2^{-1}\right], \quad (18)$$

to get the bound

$$\begin{aligned}\mathbf{MSE} &\geqslant \frac{2\underline{\lambda}^2}{n}\mathrm{Tr}\left(\left[\mathbf{J}^T\mathbf{J}\right]^{\dagger}\mathbf{J}^T\mathbf{J}\right) \\ &= \frac{2\underline{\lambda}^2}{n}\mathrm{rank}\,\mathbf{J} = \frac{2\underline{\lambda}^2}{n}(lr+Kr-r^2). \quad (19)\end{aligned}$$

As expected, the dependence on the model parameters is similar to that obtained in [16] for the joint structured covariance estimation and in [21] for the problem of low-rank matrix reconstruction.

IV. JICE Algorithm

We now proceed to the Joint Inverse Covariance Estimation (JICE) algorithm of the inverse covariances $\mathbf{T}_1,\ldots,\mathbf{T}_K$, utilizing the representation (7) of $\mathbf{Y}$. The most natural brute force approach would be to form the K SCM-s

$$\mathbf{S}_k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^k\mathbf{x}_i^{kT}, \quad k=1,\ldots,K, \quad (20)$$

invert them and fit by a subspace of small dimension. This can be done, for example, by the means of Principal Component Analysis (PCA). Such approach, applied to the SCM-s (not to their inverses), was proposed in [16] to treat the problem of joint covariance estimation. When the number of samples n in each group is smaller than or even close to the dimension p (the scenario we are mostly interested in), the inversion of the SCM-s would be impossible due to rank deficiency or would approximate the true inverse covariances poorly, thus causing the proposed PCA algorithm to fail. Instead, we propose a rather different optimization algorithm based on regularized likelihood maximization and its efficient implementation.

A. The Basic Algorithm

Recall that our aim is to estimate the true inverse covariances, while simultaneously trying to keep the dimension of the space spanned by the estimators small. For this purpose we suggest the following regularized average log-likelihood optimization program

$$\begin{aligned}&[\widehat{\mathbf{T}}_1,\ldots,\widehat{\mathbf{T}}_K] \\ &= \underset{\widetilde{\mathbf{T}}_1,\ldots,\widetilde{\mathbf{T}}_K}{\mathrm{argmin}}\sum_{k=1}^{K}\left[-\log|\widetilde{\mathbf{T}}_k| + \mathrm{Tr}\left(\mathbf{S}_k\widetilde{\mathbf{T}}_k\right)\right] + \eta\left\|\widetilde{\mathbf{Y}}\right\|_*, \quad (21)\end{aligned}$$

where

$$\widetilde{\mathbf{Y}} = [\mathrm{vech}\left(\widetilde{\mathbf{T}}_1\right),\ldots,\mathrm{vech}\left(\widetilde{\mathbf{T}}_K\right)], \quad (22)$$

and η is a regularization parameter chosen as

$$\eta = c\frac{K}{n}, \quad (23)$$

where the detailed derivation of the appropriate value of the constant c is postponed to the full paper due to lack of space. Below we use cross-validation technique to establish the best fit of c numerically.

Program (21) aims at minimizing the average negative log-likelihood of the K groups of measurements and simultaneously enforces joint low dimensional structure on the concentration matrices. The latter is achieved by reducing the nuclear norm of $\widetilde{\mathbf{Y}}$. Nuclear norm is a convex envelope of the counting measure on singular values of a matrix. Therefore, intuitively the purpose of the penalty term in (21) is to decrease the number of non-zero singular values, [22]. Both terms of the program are convex, thus guarantying success of any standard numerical solver, such as CVX, [23, 24].

B. Bias Removal

The estimator defined by (21) suffers from bias introduced by the nuclear norm regularization. We suggest an additional step aimed at removing this bias. Denote the vectorized solution of (21) by

$$\widehat{\mathbf{Y}} = \left[\mathrm{vech}\left(\widehat{\mathbf{T}}_1\right),\ldots,\mathrm{vech}\left(\widehat{\mathbf{T}}_1\right)\right].$$

Analogously to the original parametrization (7), consider the decomposition

$$\widehat{\mathbf{Y}} = \widehat{\mathbf{U}}\widehat{\mathbf{Z}}, \quad (24)$$

where $\widehat{\mathbf{U}}\in\mathbb{R}^{l\times s}$ has orthonormal columns and $\widehat{\mathbf{Z}}\in\mathbb{R}^{s\times K}$. Such decomposition naturally suggests treating the columns of $\widehat{\mathbf{U}}$ as the basis vectors of the approximate low dimensional subspace of $\mathcal{S}(p)$. We get an improved concentration matrices estimator by minimizing the average negative log-likelihood over the subspace spanned by $\widehat{\mathbf{U}}$

$$\begin{aligned}&[\widehat{\mathbf{T}}'_1,\ldots,\widehat{\mathbf{T}}'_K] \\ &\quad= \underset{\widetilde{\mathbf{T}}_1,\ldots,\widetilde{\mathbf{T}}_K\subset\mathrm{span}\left(\widehat{\mathbf{U}}\right)}{\mathrm{argmin}}\sum_{k=1}^{K}\left[-\log|\widetilde{\mathbf{T}}_k| + \mathrm{Tr}\left(\mathbf{S}_k\widetilde{\mathbf{T}}_k\right)\right]. \quad (25)\end{aligned}$$

Remarkably, this additional bias removal step requires solution of a simple convex optimization problem. Thus, the algorithm decouples into two convex programs, which can be treated using any off-the-shelf numerical solver. Below we compare both algorithms using numerical experiments.

C. ADMM Implementation

Following [25], we propose an efficient way to solve (21) based on the Alternating Direction Method of Multipliers (ADMM). The idea behind the ADMM is to optimize the augmented target

$$\begin{aligned}\min_{\widetilde{\mathbf{T}}_1,\ldots,\widetilde{\mathbf{T}}_K}&\sum_{k=1}^{K}\left[-\log|\widetilde{\mathbf{T}}_k| + \mathrm{Tr}\left(\mathbf{S}_k\widetilde{\mathbf{T}}_k\right)\right] + \eta\left\|\widetilde{\mathbf{Y}}\right\|_* \\ &+ \frac{\rho}{2}\sum_{k=1}^{K}\left\|\mathrm{vech}\left(\widetilde{\mathbf{T}}_k\right) - \widetilde{\mathbf{Y}}_{:,k}\right\|^2, \quad (26)\end{aligned}$$

$$\text{s.t.} \qquad \mathrm{vech}\left(\widetilde{\mathbf{T}}_k\right) = \widetilde{\mathbf{Y}}_{:,k}, \quad k=1,\ldots,K. \quad (27)$$

Here $\widetilde{\mathbf{Y}}_{:,k}$ denotes the k-th column of $\widetilde{\mathbf{Y}}$ and the constraints enforce consensus between the variables. The solution is obtained via introduction of dual variables $\mathbf{U}_k\in\mathbb{R}^l$ and follows the iterative scheme (we replace iteration indices by arrows to simplify notations)

$$\begin{aligned}&\widetilde{\mathbf{T}}_k \leftarrow \underset{\overline{\mathbf{T}}_k}{\mathrm{argmin}} -\log|\overline{\mathbf{T}}_k| + \mathrm{Tr}\left(\mathbf{S}_k\overline{\mathbf{T}}_k\right) \\ &+ \frac{\rho}{2}\left\|\mathrm{vech}\left(\overline{\mathbf{T}}_k\right) - \widetilde{\mathbf{Y}}_{:,k} + \mathbf{U}_k\right\|^2, k=1,\ldots,K, \quad (28)\\ &\widetilde{\mathbf{Y}} \leftarrow \underset{\bar{\mathbf{Y}}}{\mathrm{argmin}}\,\eta\left\|\bar{\mathbf{Y}}\right\|_* + \frac{\rho}{2}\left\|\mathrm{vech}\left(\widetilde{\mathbf{T}}_k\right) - \bar{\mathbf{Y}}_{:,k} + \mathbf{U}_k\right\|^2, \quad (29)\\ &\mathbf{U}_k \leftarrow \mathbf{U}_k + \mathrm{vech}\left(\widetilde{\mathbf{T}}_k\right) - \widetilde{\mathbf{Y}}_{:,k}.\end{aligned}$$

The main advantage of the proposed ADMM technique is that both updates in (28) and (29) have easily computed closed form solutions. Indeed, the first order optimality conditions for the $\widetilde{\mathbf{T}}_k$ update target (28) yield

$$\rho\overline{\mathbf{T}}_k - \overline{\mathbf{T}}_k^{-1} = \rho\left(\mathrm{mat}\left(\widetilde{\mathbf{Y}}_{:,k} - \mathbf{U}_k\right)\right) - \mathbf{S}_k. \tag{30}$$

Take the orthogonal eigenvalue decomposition of the symmetric right-hand side

$$\rho\left(\mathrm{mat}\left(\widetilde{\mathbf{Y}}_{:,k} - \mathbf{U}_k\right)\right) - \mathbf{S}_k = \mathbf{D}_k\mathbf{\Lambda}_k\mathbf{D}_k^T, \tag{31}$$

and multiply (30) by $\mathbf{D}_k^T$ on the left and by $\mathbf{D}_k$ on the right to get

$$\rho\underline{\mathbf{T}}_k - \underline{\mathbf{T}}_k^{-1} = \mathbf{\Lambda}_k, \tag{32}$$

where $\underline{\mathbf{T}}_k = \mathbf{D}_k^T\overline{\mathbf{T}}_k\mathbf{D}_k$. We can now construct a diagonal solution of this equation by solving the p quadratic equations involving the diagonal entries of $\underline{\mathbf{T}}_k$ to obtain

$$\underline{\mathbf{T}}_k = \frac{1}{2\rho}\left(\mathbf{\Lambda}_k + \sqrt{\mathbf{\Lambda}_k^2 + 4\rho\mathbf{I}}\right). \tag{33}$$

The corresponding iteration becomes

$$\widetilde{\mathbf{T}}_k \leftarrow \frac{1}{2\rho}\mathbf{D}_k^T\left(\mathbf{\Lambda}_k + \sqrt{\mathbf{\Lambda}_k^2 + 4\rho\mathbf{I}}\right)\mathbf{D}_k, \tag{34}$$

which is guaranteed to be a positive definite matrix.

In order to derive a closed form solution for the second iterative step (29), we introduce the matrix singular value soft thresholding operator. Given a matrix $\mathbf{M}$ with the SVD

$$\mathbf{M} = \mathbf{U}\begin{pmatrix}\sigma_1 & 0 & \dots\\ 0 & \sigma_2 & \dots\\ \vdots & \vdots & \ddots\end{pmatrix}\mathbf{V}^T, \tag{35}$$

the singular values soft thresholding operator is defined as

$$S_\varepsilon(\mathbf{M}) = \mathbf{U}\begin{pmatrix}\max(\sigma_1-\varepsilon,0) & 0 & \dots\\ 0 & \max(\sigma_2-\varepsilon,0) & \dots\\ \vdots & \vdots & \ddots\end{pmatrix}\mathbf{V}^T. \tag{36}$$

It can be easily shown that the solution to (29) is given by

$$\widetilde{\mathbf{Y}} \leftarrow S_{\eta/\rho}\left(\left[\mathrm{vech}\left(\widetilde{\mathbf{T}}_1\right) + \mathbf{U}_1, \dots, \mathrm{vech}\left(\widetilde{\mathbf{T}}_K\right) + \mathbf{U}_K\right]\right). \tag{37}$$

Note that the proposed iterative algorithm involves two stages on each iteration. The first performs the $\widetilde{\mathbf{T}}_k$ updates which can be done in parallel, and the second gathers the new $\widetilde{\mathbf{T}}_k$ values and updates $\widetilde{\mathbf{Y}}$. ADMM algorithms are known for their rapid convergence since the early 70-x, the details on their performance analysis, iterative implementations, techniques of choosing the tuning parameter ρ and further references can be found in [25]. In our implementation we always used $\rho = 1$.

V. Numerical Simulations

For our numerical experiment we took the banded structure model model with $p = 6$ and $b = 1$, which implies that $r = 11$. The $K = 50$ true concentration matrices were generated in the following way. We took a 6×6 matrix $\mathbf{M}$ with uniformly $[0, 1]$ distributed entries on the main diagonal and the first upper subdiagonal. Then we constructed a symmetric $\mathbf{M}' = \mathbf{M} + \mathbf{M}^T$ and checked whether it is positive definite or not. If yes, it was added to the set of inverse covariances, if not, an additional trial was taken. The regularization parameter η fitting was achieved using the cross-validation technique. Figure 1 shows the empirical, averaged over 10^3 experiments, **MSE**-s of the proposed basic algorithm (JICE) and its bias removing modification (JICE_BR) as functions of n. For comparison we also plot the

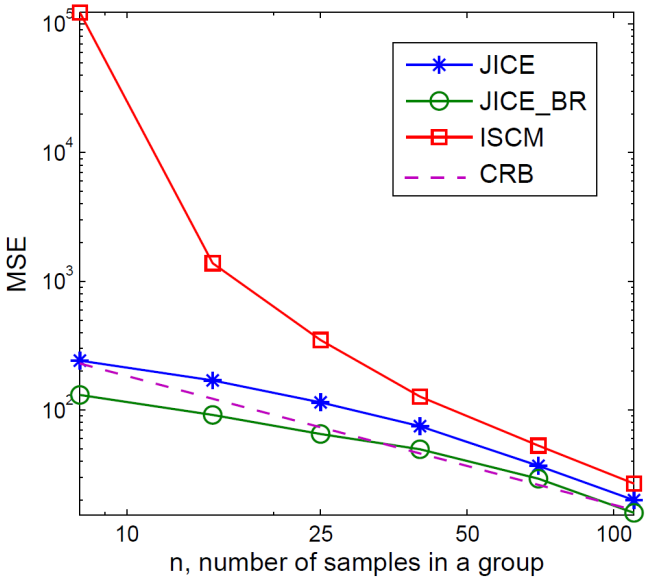

Fig. 1. JICE algorithm performance in the banded structure case, $b = 1,\ p = 6,\ l = 21,\ r = 11,\ K = 50$.

MSE-s of the inverse SCM (ISCM), its orthogonal (with respect to the scalar product defined as $(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}\left(\mathbf{A}\mathbf{B}^T\right)$) projection (Proj(ISCM)) onto the known subspace structure, and the true **CRB** bound given by (17). The graph demonstrates that when the number of samples n in each of the groups is relatively small, both JICE and JICE_BR significantly outperform the competitors. In addition, as expected, the bias removal step improves the performance, and approaches the performance of the projector when n becomes large. It is worth mentioning, that the "bias removal" step does not remove the statistical bias completely, but rather makes an attempt to achieve this. Therefore, even after the introduction of this extra step into the algorithm, the estimator may remain biased. Taking this into account, it is not surprising that the corresponding empirical **MSE** can lie below the **CRB** when the number of samples n is small.

VI. Conclusion

In this paper we consider the problem of joint inverse covariance estimation with linear structure, given heterogeneous measurements. The main challenge in this scenario is twofold. At first, the underlaying structure is to be discovered and then it should be utilized to improve the concentrations estimation. We propose a novel algorithm coupling these two stages into one optimization program and propose its efficient implementation. The main aim of our current research is to provide tight upper bound guarantees on the performance of the proposed technique.

References

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

[2] A. Kavcic and J. M. F. Moura, "Matrices with banded inverses: Inversion algorithms and factorization of Gauss-Markov processes," *IEEE transactions on Information Theory*, vol. 46, no. 4, pp. 1495–1509, 2000.

[3] A. Asif and J. M. F. Moura, "Block matrices with L-block-banded inverse: inversion algorithms," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 630–642, 2005.

[4] 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.

[5] 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.

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

[7] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *The Journal of Machine Learning Research*, vol. 9, pp. 485–516, 2008.

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

[9] 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.

[10] 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.

[11] A. Aubry, V. Carotenuto, A. De Maio, and G. Foglia, "Exploiting multiple a priori spectral models for adaptive radar detection," *Radar, Sonar & Navigation, IET*, vol. 8, no. 7, pp. 695–707, 2014.

[12] 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.

[13] 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.

[14] E. Moulines, P. Priouret, F. Roueff *et al.*, "On recursive estimation for time varying autoregressive processes," *The Annals of statistics*, vol. 33, no. 6, pp. 2610–2654, 2005.

[15] 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.

[16] I. Soloveychik and A. Wiesel, "Joint covariance estimation with mutual linear structure," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[17] ——, "Group symmetric robust covariance estimation," *arXiv preprint arXiv:1412.2345*, 2014.

[18] S. L. Lauritzen, "Graphical models," *Oxford University Press*, 1996.

[19] A. Wiesel, Y. C. Eldar, and A. O. Hero, "Covariance estimation in decomposable Gaussian graphical models," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1482–1492, 2010.

[20] 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.

[21] 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.

[22] E. J. Candes and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[23] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.0 beta," http://cvxr.com/cvx, Sep. 2013.

[24] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.

[25] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.