Covariance Estimation in Decomposable Gaussian Graphical Models

Ami Wiesel, *Member, IEEE*, Yonina C. Eldar, *Senior Member, IEEE*, and Alfred O. Hero III, *Fellow, IEEE*

***Abstract*—Graphical models are a framework for representing and exploiting prior conditional independence structures within distributions using graphs. In the Gaussian case, these models are directly related to the sparsity of the inverse covariance (concentration) matrix and allow for improved covariance estimation with lower computational complexity. We consider concentration estimation with the mean-squared error (MSE) as the objective, in a special type of model known as decomposable. This model includes, for example, the well known banded structure and other cases encountered in practice. Our first contribution is the derivation and analysis of the minimum variance unbiased estimator (MVUE) in decomposable graphical models. We provide a simple closed form solution to the MVUE and compare it with the classical maximum likelihood estimator (MLE) in terms of performance and complexity. Next, we extend the celebrated Stein's unbiased risk estimate (SURE) to graphical models. Using SURE, we prove that the MSE of the MVUE is always smaller or equal to that of the biased MLE, and that the MVUE itself is dominated by other approaches. In addition, we propose the use of SURE as a constructive mechanism for deriving new covariance estimators. Similarly to the classical MLE, all of our proposed estimators have simple closed form solutions but result in a significant reduction in MSE.**

***Index Terms*—Covariance estimation, graphical models, minimum variance unbiased estimation.**

I. Introduction

COVARIANCE estimation in Gaussian distributions is a classical and fundamental problem in statistical signal processing. Many applications, varying from array processing to functional genomics, rely on accurately estimated covariance matrices [1], [2]. Recent interest in inference in high dimensional settings using small sample sizes has caused the topic to rise to prominence once again. A natural approach in these settings is to incorporate additional prior knowledge in the form of structure and/or sparsity in order to ensure stable estimation. Gaussian graphical models provide a method of representing conditional independence structure among the different variables using graphs. An important property of the Gaussian distribution is that conditional independence among groups of variables is associated with sparsity in the inverse covariance. Due to the sparsity, these models allow for efficient implementation of statistical inference algorithms, e.g., belief propagation [3], [4], and the iterative proportional scaling technique [5], [6].

Over the past years, statistical graphical models have been successfully applied to speech recognition [7], [8], image processing [9], [10], sensor networks [11], computer networks [12], and other fields in signal processing. Efficient Bayesian inference in Gaussian graphical models is well established [13]–[15]. The problem of estimation of deterministic parameters have received less attention, but see the recent works on inverse covariance structure in the context of state-of-the-art array processing [16], [17].

Estimation of deterministic parameters in Gaussian graphical models is basically covariance estimation since the Gaussian distribution is completely parameterized by second order statistics. The most common approach to covariance estimation is maximum likelihood. When no prior information is available, this method yields the sample covariance matrix. It is asymptotically unbiased and efficient but does not minimize the mean-squared error (MSE) in general. Indeed, depending on the performance measure, better estimators can be obtained through regularization, shrinkage, empirical Bayes and other methods [18]–[24].

Covariance estimation in Gaussian graphical models involves estimation of the unknown covariance based on the observed realizations and prior knowledge of the conditional independence structure within the distribution [5], [6], [25], [26]. The prior information allows for better performance with lower computational complexity. Decomposable graphical models, also known as chordal or triangulated, satisfy a special structure which leads to a simple closed form expression for the maximum likelihood estimate (MLE). These models include many practical signal processing structures such as the banded concentration matrix and its variants [16], [23], [24], [27] as well as multiscale structures [9], [10].

Covariance selection is a related topic which addresses the joint problem of covariance estimation and graphical model selection. This setting is suitable to many modern applications in which the conditional independence structure is unknown and must be learned from the observations. Numerous selection methods have been recently considered for both arbitrary graphical models [28]–[31] and decomposable models [32], [33]. Clearly, these methods are intertwined with covariance

Manuscript received March 06, 2009; accepted September 28, 2009. First published November 24, 2009; current version published February 10, 2010. The work of A. Wiesel was supported by a Marie Curie Outgoing International Fellowship within the 7th European Community Framework Programme. This work was supported in part by AFOSR MURI Grant FA9550-06-1-0324. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Deniz Erdogmus.

A. Wiesel and A. O. Hero III are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: amiw@umich.edu; hero@umich.edu).

Y. C. Eldar is with the Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: yonina@ee.technion.ac.il).

Digital Object Identifier 10.1109/TSP.2009.2037350

1053-587X/$26.00 © 2010 IEEE

Authorized licensed use limited to: Technion Israel School of Technology. Downloaded on April 21,2010 at 20:35:35 UTC from IEEE Xplore. Restrictions apply.

estimation. For example, the latter is a key component of greedy stepwise selection algorithms.

In this paper, we consider covariance estimation in decomposable Gaussian graphical models with the MSE of the inverse covariance as our objective function. Except for the prior conditional independence structure, we do not assume any further knowledge on the covariance and treat it as an unknown deterministic parameter. Our main contribution is the derivation of the minimum variance unbiased estimator (MVUE) of the inverse covariance. Similarly to the MLE, the MVUE has a simple closed form solution which can be efficiently implemented in a distributed manner. Moreover, it minimizes the MSE among all unbiased estimators. We also prove that it has smaller MSE than the biased MLE. The proof is based on an extension of the celebrated Stein unbiased risk estimate (SURE) [18], [19], [34]–[37] to Gaussian graphical models. Using SURE we prove that the MVUE dominates the MLE in terms of MSE, i.e., its MSE is always smaller or equal to that of the MLE. In addition, we prove that the MVUE itself is dominated by other biased estimators. Next, we propose the use of SURE as a method for hyper-parameter tuning in existing covariance estimation approaches, e.g., the conjugate prior based methods proposed in [38], [39].

The outline of the paper is as follows. We begin in Section II defining the notation for decomposable graphical models, providing a few illustrative applications, and formulating the estimation problem. In Section III, we review the classical MLE approach and derive the finite-sample MVUE. Next, in Section IV we consider SURE and its applications to covariance estimation. While our estimators have lower MSE, they require more samples in order to ensure positive semidefiniteness. This issue is addressed in Section V. We evaluate the performance of the different estimators using numerical simulations in Section VI, and conclude in Section VII.

The following notation is used. Boldface upper case letters denote matrices, boldface lower case letters denote column vectors, and standard lower case letters denote scalars. We use indices in the subscript $[\mathbf{x}]_a$ or $[\mathbf{X}]_{a,b}$ to denote sub-vectors or sub-matrices, respectively, and $[\mathbf{X}]_{a,:}$ denotes the sub-matrix formed by the ath rows in $\mathbf{X}$. Where possible, we omit the brackets and use $\mathbf{x}_a$ or $\mathbf{X}_{a,b}$ instead. The superscripts $(\cdot)^T$ and $(\cdot)^{-1}$ denote the transpose and matrix inverse, respectively. For sets a and b, the cardinality is written as $|a|$ and the set difference operator is denoted by $a \setminus b$. The operator $\mathrm{Tr}\cdot$ denotes the trace, and $||\mathbf{X}||$ denotes the Frobenius norm of a matrix $\mathbf{X}$, namely $||\mathbf{X}||^2 = \mathrm{Tr}\,(\mathbf{X}^T\mathbf{X})$, and $\mathbf{X} \succ \mathbf{0}$ means that $\mathbf{X}$ is positive definite. The zero fill-in operator $[\cdot]^0$ outputs a conformable matrix where the argument occupies its appropriate sub-block and the rest of the matrix has zero valued elements (see [5] for the exact definition of this operator).

II. Covariance Estimation in Graphical Models

In this section, we provide an introduction to decomposable Gaussian graphical models based on [5] along with a few motivating applications for their use in modern statistical signal processing. We then formulate the inverse covariance estimation problem addressed in this paper.

A. Decomposable Gaussian Graphical Models

Graphical models are intuitive characterizations of conditional independence structures within distributions. An undirected graph $\mathcal{G} = (V, E)$ is a set of nodes $V = \{1, \ldots, |V|\}$ connected by undirected edges $E = \{(i_1, j_1), \ldots (i_{|E|}, j_{|E|})\}$, where we use the convention that each node is connected to itself, i.e., $(i, i) \in E$ for all $i \in V$. Let $\mathbf{x} = [x_1, \ldots, x_p]^T$ be a zero mean random vector of length $p = |V|$ whose elements are indexed by the nodes in V. The vector $\mathbf{x}$ satisfies the Markov property with respect to $\mathcal{G}$, if for any pair of nonadjacent nodes the corresponding pair of elements in $\mathbf{x}$ are conditionally independent of the remaining elements, i.e., x_i and x_j are conditionally independent of $\mathbf{x}_r$ for any $\{i, j\} \notin E$ and $r = \{V \setminus i, j\}$

$$p(x_i, x_j|\mathbf{x}_r) = p(x_i|\mathbf{x}_r)p(x_j|\mathbf{x}_r). \tag{1}$$

Therefore, the joint distribution satisfies the following factorization:

$$p(x_i, x_j, \mathbf{x}_r) = \frac{p(x_i, \mathbf{x}_r)p(x_j, \mathbf{x}_r)}{p(\mathbf{x}_r)}. \tag{2}$$

In the Gaussian setting, this factorization leads to sparsity in the concentration (inverse covariance) matrix. The multivariate Gaussian distribution is defined as

$$p(\mathbf{x}; \mathbf{K}) = c|\mathbf{K}|^{\frac{1}{2}} e^{-\frac{1}{2}\mathbf{x}^T\mathbf{K}\mathbf{x}} \tag{3}$$

where c is a constant, and $\mathbf{K} \succ \mathbf{0}$ is the concentration matrix. Its marginal distributions are also jointly Gaussian. For example, the marginal of the sub-block $\mathbf{x}_r$ is

$$p(\mathbf{x}_r) = c'|\bar{\mathbf{K}}_r|^{\frac{1}{2}} e^{-\frac{1}{2}\mathbf{x}_r^T\bar{\mathbf{K}}_r\mathbf{x}_r} \tag{4}$$

where c' is an appropriate constant and the marginal concentration matrix is

$$\bar{\mathbf{K}}_r = ([\mathbf{K}^{-1}]_{r,r})^{-1}. \tag{5}$$

Together with (2) this implies that

$$\mathbf{K} = [\bar{\mathbf{K}}_{ir}]^0 + [\bar{\mathbf{K}}_{jr}]^0 - [\bar{\mathbf{K}}_r]^0 \tag{6}$$

$$|\mathbf{K}| = \frac{|\bar{\mathbf{K}}_{ir}||\bar{\mathbf{K}}_{jr}|}{|\bar{\mathbf{K}}_r|} \tag{7}$$

where $\bar{\mathbf{K}}_{ir}$, $\bar{\mathbf{K}}_{jr}$ and $\bar{\mathbf{K}}_r$ are the marginal concentrations of $\{x_i, \mathbf{x}_r\}, \{x_j, \mathbf{x}_r\}$ and $\{\mathbf{x}_r\}$, respectively, and are defined in a similar manner to (5). All of the matrices in the right hand side of (6) have a zero value in the $\{i, j\}$th position, and therefore

$$[\mathbf{K}]_{i,j} = 0 \quad \text{for all} \quad \{i, j\} \notin E. \tag{8}$$

This property is the core of Gaussian graphical models: the concentration matrix $\mathbf{K}$ has a sparsity pattern which represents the topology of the conditional independence graph.

Decomposable models (also known as chordal or triangulated models) are a special type of graphical model in which the conditional independence graphs satisfy an appealing structure. A decomposable graph can be divided into an ordered sequence of fully connected subgraphs known as cliques and denoted by

Authorized licensed use limited to: Technion Israel School of Technology. Downloaded on April 21,2010 at 20:35:35 UTC from IEEE Xplore. Restrictions apply.

$C_1, \ldots, C_K$. These ordered cliques are coupled through separators

$$S_j = (C_1 \cup C_2 \cup \cdots \cup C_{j-1}) \cap C_j \tag{9}$$

for $j = 2, \ldots, K$, and satisfy the *running intersection property*: for all $j \geq 2$ there is a $k < j$ such that $S_j \subseteq C_k$. When these conditions hold, we can extend the elegant structure in (2) to cliques and separators

$$p(\mathbf{x}; \mathbf{K}) = \frac{\prod_{k=1}^{K} p\left(\mathbf{x}_{C_k}; \bar{\mathbf{K}}_k\right)}{\prod_{k=2}^{K} p\left(\mathbf{x}_{S_k}; \bar{\mathbf{K}}_{[k]}\right)} \tag{10}$$

where

$$\begin{aligned} \bar{\mathbf{K}}_k &= \left([\mathbf{K}^{-1}]_{C_k, C_k}\right)^{-1} \\ \bar{\mathbf{K}}_{[k]} &= \left([\mathbf{K}^{-1}]_{S_k, S_k}\right)^{-1} \end{aligned} \tag{11}$$

are the marginal concentrations of the cliques and separators. Similarly, (6) can be extended to

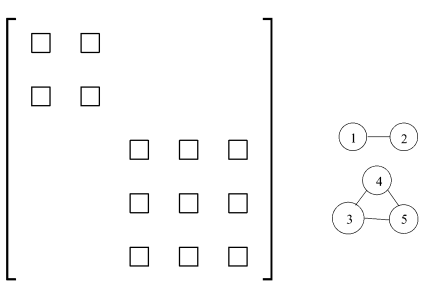
$$\mathbf{K} = \sum_{k=1}^{K} [\bar{\mathbf{K}}_k]^0 - \sum_{k=2}^{K} \left[\bar{\mathbf{K}}_{[k]}\right]^0. \tag{12}$$

For later use, we denote the cardinalities by $|C_k| = c_k$ and $|S_k| = s_k$, and define the set of decomposable concentration matrices as

$$\mathcal{K} = \{\mathbf{K} : \mathbf{K} \succ \mathbf{0}, \mathbf{K} \text{ satisfies (12)}\}. \tag{13}$$

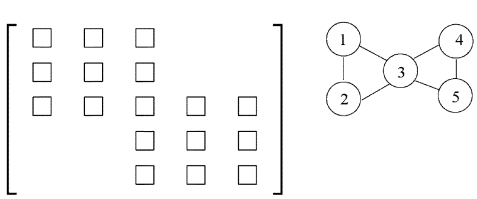
Decomposable graphical models appear in many signal processing applications. We now review the following few representative examples.

- *Diagonal or Block Diagonal:* A trivial graphical model is the diagonal or block diagonal model, in which the cliques are non-overlapping. For example, the following matrix has two cliques $C_1 = \{1, 2\}$ and $C_2 = \{3, 4, 5\}$



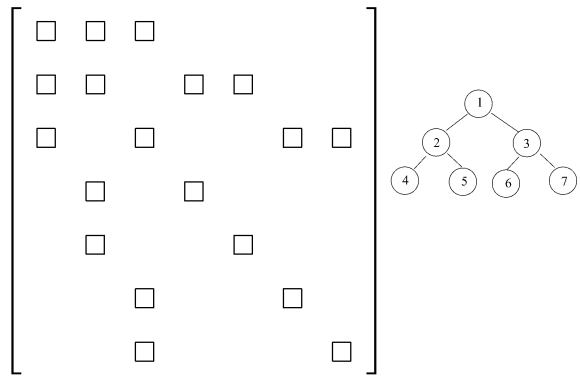
(14)

- *Two Coupled Blocks*: The simplest nontrivial decomposable graphical model is the two coupled blocks. For example, the following matrix has two cliques $C_1 = \{1, 2, 3\}$ and $C_2 = \{3, 4, 5\}$ coupled through $S_2 = \{3\}$:



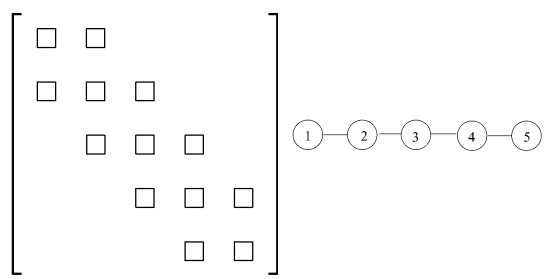
(15)

- *Multiscale*: A common graphical model in image processing is based on the multiscale (multiresolution) framework. Here, the decomposable graph is a tree of nodes (or cliques) [9], [10]:



(16)

- *Banding*: Another frequently used decomposable graphical model is the L'th order banded structure in which only the $L+1$ principal diagonals of $\mathbf{K}$ have nonzero elements. For example, the following matrix is banded with $L = 2$:



(17)

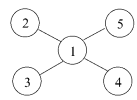
and has four cliques $C_1 = \{1, 2\}, C_2 = \{2, 3\}, C_3 = \{3, 4\}$ and $C_4 = \{4, 5\}$. It is appropriate whenever the indices of the multivariate represent physical quantities such as time or space, and the underlying assumption is that distant variables are conditionally independent of closer variables. A special case of this structure is the stationary autoregressive (AR) model which leads to a banded Toeplitz matrix. The more general banded graphical model corresponds to a non-stationary autoregressive process. It was recently shown that this structure is a good model for state-of-the-art radar systems [16] (see also [27]). A natural extension of the Lth banded model is differential banding in which multiple band lengths are utilized. It is straightforward to show that the corresponding graph is still decomposable with cliques of different cardinalities. This form was empirically validated to be a reasonable model in call center management in operations research [39].

- *Arrow (Star)*: Another common decomposable model takes the form of an arrow motif in the concentration matrix. This structure is appropriate when there is a single common global sub-block and numerous local sub-blocks which are conditionally independent given the global variables. For

Authorized licensed use limited to: Technion Israel School of Technology. Downloaded on April 21,2010 at 20:35:35 UTC from IEEE Xplore. Restrictions apply.

example, the following concentration matrix specifies an arrow graphical model:

$$\begin{bmatrix} \square & \square & \square & \square & \square \\ \square & \square & & & \\ \square & & \square & & \\ \square & & & \square & \\ \square & & & & \square \end{bmatrix} \qquad (18)$$



with cliques $C_1 = \{1,2\}, C_2 = \{1,3\}, C_3 = \{1,4\}$ and $C_4 = \{1,5\}$. A typical signal processing application is a wireless network in which the global node is the access point and the local nodes are the terminals. Other applications of these models are discussed in [40].

B. Problem Formulation

We are now ready to state the problem addressed in this paper. Let $\mathbf{x}$ be a length-p zero mean Gaussian random vector, with concentration matrix $\mathbf{K} \in \mathcal{K}$ as in (13). Given n independent realizations of $\mathbf{x}$ denoted by $\{\mathbf{x}[i]\}_{i=1}^n$, and knowledge of the conditional independence structure, our goal is to derive an estimate $\hat{\mathbf{K}}$ of $\mathbf{K}$ with minimum MSE, where the MSE is defined as

$$\mathrm{MSE}\,(\mathbf{K}) = E\{\|\hat{\mathbf{K}} - \mathbf{K}\|^2\}. \qquad (19)$$

Here the norm is the matrix Frobenius norm. The MSE in (19) is a function of the unknown parameter $\mathbf{K}$ and cannot be minimized directly. This dependency is the main difficulty in minimum MSE estimation of deterministic parameters, in contrast to the Bayesian framework in which the MSE is a function of the distribution of $\mathbf{K}$ but not of $\mathbf{K}$ itself. More details on this issue can be found in [41] and [42].

Due to the difficulty of minimum MSE estimation, it is customary to restrict attention to unbiased estimators. For this purpose, the MSE is decomposed into its squared bias and variance components defined as

$$\mathrm{MSE}(\mathbf{K}) = \mathrm{BIAS}^2(\mathbf{K}) + \mathrm{VAR}(\mathbf{K}) \qquad (20)$$

where

$$\begin{aligned} \mathrm{BIAS}^2(\mathbf{K}) &= \|E\{\hat{\mathbf{K}}\} - \mathbf{K}\|^2 \\ \mathrm{VAR}(\mathbf{K}) &= E\{\|\hat{\mathbf{K}} - E\{\hat{\mathbf{K}}\}\|^2\}. \end{aligned} \qquad (21)$$

We call $\hat{\mathbf{K}}$ an unbiased estimator if $\mathrm{BIAS}(\mathbf{K}) = 0$. Although the variance may also depend on $\mathbf{K}$, in many cases an estimate exists that is asymptotically unbiased and minimizes the MSE.

Our choice of MSE of $\mathbf{K}$ as a performance measure requires further elaboration. There are numerous competing metrics which could have been adopted: MSE of $\mathbf{K}^{-1}$ [21]; weighted norms [43]; Stein's loss [19], [34]; and others. Each of these measures will lead to different estimators. Following [20], [23], [24], and [39], we focus on the MSE of the inverse covariance due to the following reasons. Graphical models specify the structure of the concentration matrix rather than the structure of the covariance matrix so that the concentration is more intuitive. Furthermore, the concentration is the natural parameter of the exponential family of multivariate Gaussian distributions [5]. The concentration matrix is parameterized by the free variables associated with the cliques, and has zero values elsewhere. In contrast, the covariance matrix is not a natural parameter of the exponential family. Since it captures unconditioned dependency between variables, the covariance matrix generally has nonzero entries in locations where the entries of the concentration matrix are zero, i.e., entries that correspond to links between cliques. When included in the performance measure, these nonzero entries mask the behavior of the free variables within the graphical model. Finally, we remark that several of the results in this paper, such as the SURE identity, can also be applied to other performance measures.

III. Maximum Likelihood and Minimum Variance Unbiased Estimation

In this section, we review the classical MLE approach to inverse covariance estimation in decomposable Gaussian graphical models and then derive the MVUE estimator.

A. Maximum Likelihood Estimation

We begin with a short review of the sufficient statistics and the MLE of the concentration matrix in multivariate Gaussian models. For a more detailed treatment the reader is referred to [5].

When no prior information is available and the model consists of single clique $C_1 = \{1,\ldots,p\}$, the model is said to be saturated. In this case, a minimal sufficient statistic for estimating $\mathbf{K}$ is the sample covariance matrix

$$\bar{\mathbf{S}} = \sum_{i=1}^{n} \mathbf{x}[i]\mathbf{x}^T[i]. \qquad (22)$$

When $n \geq p$, its distribution is Wishart with n degrees of freedom and natural parameter $\mathbf{K}$

$$p(\bar{\mathbf{S}};\mathbf{K}) = \mathcal{W}(\bar{\mathbf{S}};\mathbf{K}) = c''|\bar{\mathbf{S}}|^{\frac{n-p-1}{2}}|\mathbf{K}|^{\frac{n}{2}}e^{-\frac{1}{2}\,\mathrm{Tr}\,\{\mathbf{K}\bar{\mathbf{S}}\}} \qquad (23)$$

where c'' is a constant and the support set is $\bar{\mathbf{S}} \succ \mathbf{0}$. In graphical models, it is *a priori* known that $\mathbf{K}$ has zero values outside the cliques, and the complete sample covariance is no longer necessary. The sub-blocks associated with the cliques, which are denoted by

$$\mathbf{S}_k = [\mathbf{S}]_{C_k,C_k}, \quad k = 1,\ldots,K \qquad (24)$$

are sufficient. For convenience, we will also define the sub-blocks of the separators (which are contained within the cliques)

$$\mathbf{S}_{[k]} = [\mathbf{S}]_{S_k,S_k}, \quad k = 2,\ldots,K. \qquad (25)$$

Authorized licensed use limited to: Technion Israel School of Technology. Downloaded on April 21,2010 at 20:35:35 UTC from IEEE Xplore. Restrictions apply.

Assuming that $n \geq \max_k c_k$, the marginal distributions of these sub-matrices are also Wishart distributed

$$p(\mathbf{S}_k; \bar{\mathbf{K}}_k) = \mathcal{W}(\mathbf{S}_k; \bar{\mathbf{K}}_k)$$
$$p\left(\mathbf{S}_{[k]}; \bar{\mathbf{K}}_{[k]}\right) = \mathcal{W}\left(\mathbf{S}_{[k]}; \bar{\mathbf{K}}_{[k]}\right) \tag{26}$$

with the marginal concentration matrices defined in (11). It is customary to define the incomplete sample covariance $\mathbf{S}$ as a $p \times p$ matrix which agrees with $\bar{\mathbf{S}}$ in the clique locations, and has unspecified values elsewhere. Similarly to (10), the distribution of this incomplete sample covariance is [26]

$$p(\mathbf{S}; \mathbf{K}) = \frac{\prod_{k=1}^{K} p(\mathbf{S}_k; \bar{\mathbf{K}}_k)}{\prod_{k=2}^{K} p(\mathbf{S}_{[k]}; \bar{\mathbf{K}}_{[k]})}. \tag{27}$$

The MLE of $\mathbf{K}$ is defined as[1]

$$\hat{\mathbf{K}}_{\mathrm{MLE}} = \arg\max_{\mathbf{K} \succ \mathbf{0}} \log p(\mathbf{S}; \mathbf{K}) \tag{28}$$

and has the following closed form solution:

$$\hat{\mathbf{K}}_{\mathrm{MLE}} = \sum_{k=1}^{K} \left[n\mathbf{S}_k^{-1}\right]^0 - \sum_{k=2}^{K} \left[n\mathbf{S}_{[k]}^{-1}\right]^0. \tag{29}$$

This estimator exists with probability one if and only if $n \geq \max_k c_k$. It is positive definite and locally consistent in the sense that the local and global versions of the cliques agree with each other

$$\left[\hat{\mathbf{K}}_{\mathrm{MLE}}^{-1}\right]_{C_k, C_k} = \frac{1}{n}\mathbf{S}_k, \quad k = 1, \ldots, K. \tag{30}$$

Both of these properties suggest that the MLE in a decomposable model performs as if the model was block diagonal with non-overlapping cliques C_k.

In general, the MLE is a biased estimator and does not minimize the MSE. One of the main motivations for the MLE is that asymptotically in n it is an MVUE. Therefore, we now address the finite sample MVUE in decomposable graphical models. Interestingly, we will show that the MVUE does not behave as if the model was block diagonal and improves performance by taking into account the coupling between the cliques. We will also prove that it dominates the MLE, specifically its MSE is smaller for all possible values of $\mathbf{K}$.

B. Minimum Variance Unbiased Estimation

For finite sample size, the MVUE is provided in the following theorem.

Theorem 1: The minimum variance unbiased estimator (MVUE) of $\mathbf{K} \in \mathcal{K}$ given the incomplete sample covariance matrix $\mathbf{S}$ is

$$\hat{\mathbf{K}}_{\mathrm{MVUE}} = \sum_{k=1}^{K} \left[(n - c_k - 1)\mathbf{S}_k^{-1}\right]^0 - \sum_{k=2}^{K} \left[(n - s_k - 1)\mathbf{S}_{[k]}^{-1}\right]^0. \tag{31}$$

It exists when the local sample covariances are all invertible, i.e., with probability one if $n \geq \max_k c_k$.

[1]An alternative but equivalent definition is $\hat{\mathbf{K}}_{\mathrm{MLE}} = \arg\max_{\mathbf{K} \in \mathcal{K}} \log p(\bar{\mathbf{S}}; \mathbf{K})$.

Proof: Using the inverse moments of the marginal Wishart distribution, the linearity of the expectation, and the identity in (12), it is easy to verify that the estimator is unbiased. It is a function of the minimal sufficient statistic (the incomplete sample covariance), and is therefore the MVUE. A different constructive proof based on the general MVUE for exponential family distributions is provided in Appendix I. ■

Theorem 1 specifies the MVUE of $\mathbf{K}$ in decomposable graphical models. The estimator is similar in structure to the MLE in (29) and it is easy to see that asymptotically in n the two estimators are equivalent. Its computational complexity is exactly the same as the MLE, and involves the inversion of the local sample covariances. In the saturated case the MVUE is a scaled version of the MLE. In many signal processing applications (e.g., principal component analysis) the overall performance is indifferent to a change in scaling of the covariance. In decomposable graphical models, the MVUE is not a simple rescaling of the MLE and it can have improved performance with almost no additional cost in computational complexity.

Recall that the MLE requires only $n \geq \max_k c_k$ samples in order for it to exist and to be positive definite. This is not true for the MVUE, which may require more samples to ensure positive semidefiniteness. For example, consider a simple $K = 2$ cliques model. Using the matrix inversion formula for partitioned matrices [44, p. 572], it can be verified that

$$\left[\hat{\mathbf{K}}_{\mathrm{MVUE}}^{-1}\right]_{S_2, S_2} = \frac{1}{n - p - 1}[\mathbf{S}]_{S_2, S_2}. \tag{32}$$

A necessary condition for positive definiteness with probability one of $\hat{\mathbf{K}}_{\mathrm{MVUE}}$ is $[\hat{\mathbf{K}}_{\mathrm{MVUE}}^{-1}]_{S_2, S_2} \succ \mathbf{0}$ which is equivalent to $n > p + 1$. Thus, although $n \geq \max_k c_k$ suffices for existence, the MVUE lacks this fundamental positive definite property unless $n > p + 1$. Identity (32) may suggest that the MVUE is locally consistent but it can be verified that this is not true, i.e.,

$$\left[\hat{\mathbf{K}}_{\mathrm{MVUE}}^{-1}\right]_{C_k, C_k} \neq \frac{1}{n - p - 1}\mathbf{S}_k \tag{33}$$

for $k = 1, 2$. Evidently, in contrast to the MLE, the MVUE does not behave as if the model were block diagonal and it accounts for the coupling between the cliques.

The MVUE minimizes the MSE over the class of unbiased estimators. This is an important property but it does not ensure optimality over all estimators, whether biased or unbiased. In the next section, we prove that the MVUE actually dominates the biased MLE in terms of MSE performance.

IV. Stein's Unbiased Risk Estimate (SURE)

SURE provides an unbiased approximation of the MSE. The SURE approach was originally applied to the estimation of a Gaussian mean parameter [18]. It was generalized to the Wishart distribution in [19], [34], [43] and later extended to estimating the natural parameters of any exponential family distribution in [35]–[37]. The following theorem extends these results to decomposable Gaussian graphical models.

Theorem 2: Let $\mathbf{S}$ be an incomplete sample covariance matrix associated with a decomposable graphical model, and assume that $n \geq \max_k c_k$. Let $\mathbf{H}(\mathbf{S})$ be a differentiable matrix function

Authorized licensed use limited to: Technion Israel School of Technology. Downloaded on April 21,2010 at 20:35:35 UTC from IEEE Xplore. Restrictions apply.

of $\mathbf{S}$ which satisfies the technical conditions in (69) for each of its elements. Then

$$E\{\mathrm{Tr}\,\{\mathbf{H}(\mathbf{S})\mathbf{K}\}\}=E\{\mathrm{Tr}\,\{\mathbf{H}(\mathbf{S})\hat{\mathbf{K}}_{\mathrm{MVUE}}+2\boldsymbol{\nabla}\mathbf{H}(\mathbf{S})\}\} \quad (34)$$

where $\hat{\mathbf{K}}_{\mathrm{MVUE}}$ is defined in (31), and the differential operator is a $p \times p$ matrix with elements

$$[\boldsymbol{\nabla}]_{ij} = \begin{cases} \frac{\partial}{\partial \mathbf{S}_{ij}}, & i=j \\ \frac{1}{2}\frac{\partial}{\partial \mathbf{S}_{ij}}, & i \neq j, \{i,j\} \in E \\ 0, & \text{else.} \end{cases} \quad (35)$$

Proof: The identity is a special case of the general SURE in exponential family distributions [37]. The only difference is the compact representation using decomposable matrix notations which account for the symmetry and the conditional independence. More details are provided in Appendix I. ■

For later use, we note that the differential operator defined in (35) can be expressed as

$$\boldsymbol{\nabla} = \sum_{k=1}^{K}[\boldsymbol{\nabla}_k]^0 - \sum_{k=2}^{K}\left[\boldsymbol{\nabla}_{[k]}\right]^0 \quad (36)$$

where $\boldsymbol{\nabla}_k = [\boldsymbol{\nabla}]_{C_k,C_k}$ and $\boldsymbol{\nabla}_{[k]} = [\boldsymbol{\nabla}]_{S_k,S_k}$ are the $c_k \times c_k$ and $s_k \times s_k$ differential operators within the saturated cliques and the saturated separators, respectively.

In the following subsections, we apply SURE to derive and analyze the MSE performance of several estimators.

A. MVUE Dominates MLE

Our first application of Theorem 2 is to prove that the MLE is inadmissible and dominated by the MVUE.

Theorem 3: The MVUE in (31) dominates the MLE in (29) in terms of MSE

$$E\{\|\hat{\mathbf{K}}_{\mathrm{MVUE}} - \mathbf{K}\|^2\} \leq E\{\|\hat{\mathbf{K}}_{\mathrm{MLE}} - \mathbf{K}\|^2\} \quad (37)$$

for all $\mathbf{K}$ in the set $\mathcal{K}$ defined in (13).

Proof: The difference in MSEs is

$$\begin{aligned}\delta &= E\{\|\hat{\mathbf{K}}_{\mathrm{MVUE}} - \mathbf{K}\|^2\} - E\{\|\hat{\mathbf{K}}_{\mathrm{MLE}} - \mathbf{K}\|^2\} \\ &= E\{\|\hat{\mathbf{K}}_{\mathrm{MVUE}}\|^2 - \|\hat{\mathbf{K}}_{\mathrm{MLE}}\|^2 - 2\,\mathrm{Tr}\,\{\mathbf{HK}\}\} \\ &= E\{-\|\mathbf{H}\|^2 - 4\,\mathrm{Tr}\,\{\boldsymbol{\nabla}\mathbf{H}\}\}\end{aligned} \quad (38)$$

where we applied Theorem 2 with

$$\begin{aligned}\mathbf{H} &= \hat{\mathbf{K}}_{\mathrm{MVUE}} - \hat{\mathbf{K}}_{\mathrm{MLE}} \\ &= -\left[\sum_{k=1}^{K}(c_k+1)\left[\mathbf{S}_k^{-1}\right]^0 - \sum_{k=2}^{K}(s_k+1)\left[\mathbf{S}_{[k]}^{-1}\right]^0\right].\end{aligned} \quad (39)$$

Therefore, in order to prove that $\delta \leq 0$ it is sufficient to show that

$$\mathrm{Tr}\,\{\boldsymbol{\nabla}\mathbf{H}\} \geq 0. \quad (40)$$

From [19, (5.4)iii]

$$\mathrm{Tr}\,\left\{\boldsymbol{\nabla}_k\mathbf{S}_k^{-1}\right\} = -\frac{1}{2}\,\mathrm{Tr}\,\left\{\mathbf{S}_k^{-2}\right\} - \frac{1}{2}\mathrm{Tr}^2\left\{\mathbf{S}_k^{-1}\right\}$$

$$\mathrm{Tr}\,\left\{\boldsymbol{\nabla}_{[k]}\mathbf{S}_{[k]}^{-1}\right\} = -\frac{1}{2}\,\mathrm{Tr}\,\left\{\mathbf{S}_{[k]}^{-2}\right\} - \frac{1}{2}\mathrm{Tr}^2\left\{\mathbf{S}_{[k]}^{-1}\right\}. \quad (41)$$

Therefore

$$\begin{aligned}\mathrm{Tr}\,\{\boldsymbol{\nabla}\mathbf{H}\} &= \frac{1}{2}\sum_{k=1}^{K}(c_k+1)\left[\,\mathrm{Tr}\,\left\{\mathbf{S}_k^{-2}\right\} + \mathrm{Tr}^2\left\{\mathbf{S}_k^{-1}\right\}\right] \\ &\quad - \frac{1}{2}\sum_{k=2}^{K}(s_k+1)\left[\,\mathrm{Tr}\,\left\{\mathbf{S}_{[k]}^{-2}\right\} + \mathrm{Tr}^2\left\{\mathbf{S}_{[k]}^{-1}\right\}\right] \\ &\geq \frac{1}{2}(s_k+1)\left[\sum_{k=2}^{K}\mathrm{Tr}\,\left\{\mathbf{S}_k^{-2}\right\} - \mathrm{Tr}\,\left\{\mathbf{S}_{[k]}^{-2}\right\}\right. \\ &\quad \left. + \mathrm{Tr}^2\left\{\mathbf{S}_k^{-1}\right\} - \mathrm{Tr}^2\left\{\mathbf{S}_{[k]}^{-1}\right\}\right] \geq 0\end{aligned} \quad (42)$$

where we have used $c_k \geq s_k$ for $k = 2, \ldots, K$ and

$$\mathrm{Tr}\,\left\{\mathbf{S}_k^{-1}\right\} \geq \mathrm{Tr}\,\left\{\mathbf{S}_{[k]}^{-1}\right\}, \quad k = 2, \ldots, K \quad (43)$$

$$\mathrm{Tr}\,\left\{\mathbf{S}_k^{-2}\right\} \geq \mathrm{Tr}\,\left\{\mathbf{S}_{[k]}^{-2}\right\}, \quad k = 2, \ldots, K. \quad (44)$$

The last two inequalities follow from Appendix II. ■

B. MVUE is Inadmissible

We now use SURE to prove that the MVUE is inadmissible since its MSE can be improved upon by another biased estimator.

Theorem 4: The estimator

$$\hat{\mathbf{K}}_{\mathrm{BE}} = \hat{\mathbf{K}}_{\mathrm{MVUE}} - \frac{1}{\mathrm{Tr}\,\{\mathbf{S}\}}\mathbf{I} \quad (45)$$

dominates the MVUE in (31) in terms of MSE

$$E\left\{\left\|\hat{\mathbf{K}}_{\mathrm{BE}} - \mathbf{K}\right\|^2\right\} \leq E\left\{\left\|\hat{\mathbf{K}}_{\mathrm{MVUE}} - \mathbf{K}\right\|^2\right\} \quad (46)$$

for all $\mathbf{K}$ in the set $\mathcal{K}$ defined in (13).

Proof: The difference in MSEs is

$$\begin{aligned}\delta &= E\left\{\left\|\hat{\mathbf{K}}_{\mathrm{BE}} - \mathbf{K}\right\|^2\right\} - E\left\{\left\|\hat{\mathbf{K}}_{\mathrm{MVUE}} - \mathbf{K}\right\|^2\right\} \\ &= E\left\{-\frac{2\,\mathrm{Tr}\,\left\{\hat{\mathbf{K}}_{\mathrm{MVUE}}\right\}}{\mathrm{Tr}\,\{\mathbf{S}\}} + \frac{p}{\mathrm{Tr}^2\{\mathbf{S}\}} + 2\,\mathrm{Tr}\,\left\{\frac{\mathbf{K}}{\mathrm{Tr}\,\{\mathbf{S}\}}\right\}\right\} \\ &= E\left\{\frac{p}{\mathrm{Tr}^2\{\mathbf{S}\}} + 4\,\mathrm{Tr}\,\left\{\boldsymbol{\nabla}\frac{1}{\mathrm{Tr}\,\{\mathbf{S}\}}\mathbf{I}\right\}\right\}\end{aligned} \quad (47)$$

where we applied Theorem 2 with $\mathbf{H} = 1/(\,\mathrm{Tr}\,\{\mathbf{S}\})\mathbf{I}$. Using the identity

$$\mathrm{Tr}\,\left\{\boldsymbol{\nabla}\frac{1}{\mathrm{Tr}\,\{\mathbf{S}\}}\mathbf{I}\right\} = \sum_i \frac{\partial}{\partial[\mathbf{S}]_{i,i}}\frac{1}{\mathrm{Tr}\,\{\mathbf{S}\}} = -\frac{p}{\mathrm{Tr}^2\{\mathbf{S}\}} \quad (48)$$

we have that

$$\delta = E\left\{-\frac{3p}{\mathrm{Tr}^2\{\mathbf{S}\}}\right\} \leq 0 \quad (49)$$

completing the proof. ■

Authorized licensed use limited to: Technion Israel School of Technology. Downloaded on April 21,2010 at 20:35:35 UTC from IEEE Xplore. Restrictions apply.

Theorem 4 proves the inadmissibility of the MLE and MVUE in any decomposable graphical model. This contribution extends the results in [40], [45], [46]. The specific form of $\hat{\mathbf{K}}_{\mathrm{BE}}$ is not of great importance and has been chosen for simplicity. It is based on a similar Efron–Morris type estimator derived for saturated models in [20].

Finally, it is worth mentioning that Theorem 4 is an example of the well known Stein's phenomenon in which the simultaneous estimation of multiple unrelated parameters can be more accurate than estimating them separately. Indeed, the simplest case of decomposable models is the diagonal (or block diagonal) inverse covariance matrix in which $\mathbf{S}_k$ are statistically independent of each other and depend on different parameters. Theorem 4 establishes that inverse covariance estimation can be improved by global shrinkage.

C. SURE-Based Parameter Tuning

The main application of SURE in signal processing is parameter tuning [36], [47], [48]. Thus, we now illustrate how automatic parameter tuning in decomposable graphical models can utilize SURE.

Consider a class of estimators parameterized by one or more variables. For simplicity, we restrict ourselves to a special class of estimators with one design parameter

$$\hat{\mathbf{K}}_d = \sum_{k=1}^{K} \left[(n-c_k-1-d)\mathbf{S}_k^{-1}\right]^0 - \sum_{k=2}^{K} \left[(n-s_k-1-d)\mathbf{S}_{[k]}^{-1}\right]^0 \tag{50}$$

parameterized by d. Given this class of estimators, we would like to find the value d which minimizes the MSE

$$\min_d E\{\|\hat{\mathbf{K}}_d - \mathbf{K}\|^2\} \tag{51}$$

or excluding constant terms

$$\min_d E\{\|\hat{\mathbf{K}}_d\|^2 - 2\,\mathrm{Tr}\,\{\hat{\mathbf{K}}_d\mathbf{K}\}\}. \tag{52}$$

Solving (52) is impossible as the expectation and the second term in the objective depend on $\mathbf{K}$, which is unknown. Instead, we propose to use the SURE result in Theorem 2 and replace the unknown MSE with its unbiased estimate

$$\min_d \|\hat{\mathbf{K}}_d\|^2 - 2\,\mathrm{Tr}\,\{\hat{\mathbf{K}}_d\mathbf{K}_{\mathrm{MVUE}} + 2\nabla\hat{\mathbf{K}}_d\}. \tag{53}$$

Substitution of $\hat{\mathbf{K}}_d$ from (50) and excluding constant terms yields

$$\min_d d^2\|\mathbf{D}\|^2 + 4d\,\mathrm{Tr}\,\{\nabla\mathbf{D}\} \tag{54}$$

where

$$\mathbf{D} = \sum_{k=1}^{K} \left[\mathbf{S}_k^{-1}\right]^0 - \sum_{k=2}^{K} \left[\mathbf{S}_{[k]}^{-1}\right]^0. \tag{55}$$

Finally, solving for d results in

$$d = -2\frac{\mathrm{Tr}\,\{\nabla\mathbf{D}\}}{\|\mathbf{D}\|^2} = -2\frac{\sum_{k=1}^{K}\mathrm{Tr}\,\{\nabla_k\mathbf{S}_k^{-1}\} - \sum_{k=2}^{K}\mathrm{Tr}\,\left\{\nabla_{[k]}\mathbf{S}_{[k]}^{-1}\right\}}{\|\mathbf{D}\|^2} \tag{56}$$

where the derivatives are defined in (41).

Simulation results presented in Section VI show promising performance gains. While we adopted a particularly simple class of estimators in (50), it is likely that more advanced estimator classes can be treated as well. For example, state-of-the-art methods for covariance estimation in decomposable graphical models involve the use of Bayesian methods and conjugate priors [38], [39]. These distributions depend on tuning parameters that must be chosen beforehand or estimated from the available data. Currently, these parameters are chosen through cross validation, or empirical Bayes methods. It would be interesting to examine the use of SURE as an alternative.

V. Positive Part Estimators

In the previous sections we proposed estimators which dominate the MLE in terms of MSE. The conditions guaranteeing their existence are similar to those of the MLE, however they may require more samples in order to be positive semidefinite. For small sample size, we propose to project these estimators onto the set of decomposable positive semi-definite matrices $\mathcal{K}$ in (13). We prove that this projection results in legitimate positive semidefinite estimators with better or equal MSE performance.

Let $\hat{\mathbf{K}}$ be a given estimator of $\mathbf{K}$. Define $\widetilde{\mathbf{K}}$ as the projection of $\hat{\mathbf{K}}$ onto the set $\mathcal{K}$ in (13)

$$\widetilde{\mathbf{K}} = \arg\min_{\widetilde{\mathbf{K}}\in\mathcal{K}} \|\widetilde{\mathbf{K}} - \hat{\mathbf{K}}\|^2. \tag{57}$$

The optimization (57) can be expressed as a semidefinite program (SDP). Therefore, the projected estimator $\widetilde{\mathbf{K}}$ can be efficiently computed using standard SDP optimization packages, e.g., [49]. The following theorem states that the projected estimator reduces the error with probability one.

Theorem 5: Let $\hat{\mathbf{K}}$ be a given estimator of $\mathbf{K} \in \mathcal{K}$ and define $\widetilde{\mathbf{K}}$ as its projection in (57). Then

$$\|\widetilde{\mathbf{K}} - \mathbf{K}\|^2 \le \|\hat{\mathbf{K}} - \mathbf{K}\|^2 \tag{58}$$

with probability one for all $\mathbf{K}$ in the set $\mathcal{K}$ in (13).

Proof: The proof is based on the convexity of the set $\mathcal{K}$ in (13) and the classical theorem of projection onto convex sets (POCS). POCS states that [50]

$$\mathrm{Tr}\,\{(\hat{\mathbf{K}} - \widetilde{\mathbf{K}})^T(\mathbf{K} - \widetilde{\mathbf{K}})\} \le 0 \tag{59}$$

for every $\mathbf{K} \in \mathcal{K}$. Adding and subtracting $\mathbf{K}$ in the first parenthesis yields

$$\|\mathbf{K} - \widetilde{\mathbf{K}}\|^2 \le \mathrm{Tr}\,\{(\mathbf{K} - \hat{\mathbf{K}})^T(\mathbf{K} - \widetilde{\mathbf{K}})\}. \tag{60}$$

Application of the Cauchy Schwartz inequality results in

$$\|\mathbf{K} - \widetilde{\mathbf{K}}\|^2 \le \|\mathbf{K} - \hat{\mathbf{K}}\|\|\mathbf{K} - \widetilde{\mathbf{K}}\| \tag{61}$$

Authorized licensed use limited to: Technion Israel School of Technology. Downloaded on April 21,2010 at 20:35:35 UTC from IEEE Xplore. Restrictions apply.

and therefore

$$\|\mathbf{K} - \widetilde{\mathbf{K}}\| \leq \|\mathbf{K} - \hat{\mathbf{K}}\|. \tag{62}$$

Since all of the above inequalities apply to any realization of the random matrix $\hat{\mathbf{K}}$, (58) holds with probability one. ■

When solving (57) is too computationally expensive, we can relax the constraint set and consider the projection onto the semidefinite cone

$$\min_{\widetilde{\mathbf{K}} \succeq \mathbf{0}} \|\hat{\mathbf{K}} - \widetilde{\mathbf{K}}\|^2. \tag{63}$$

Similarly to Theorem 5, the semidefinite cone $\{\widetilde{\mathbf{K}} \succeq \mathbf{0}\}$ is a convex set and the solution to (63) dominates $\hat{\mathbf{K}}$ in terms of MSE. Its main advantage is that it satisfies a simple closed form. Let

$$\hat{\mathbf{K}} = \mathbf{U}\mathbf{D}\mathbf{U}^T \tag{64}$$

be the eigenvalue decomposition of $\hat{\mathbf{K}}$ where $\mathbf{U}$ is a unitary matrix and $\mathbf{D}$ is a diagonal matrix with the eigenvalues $[\mathbf{D}]_{ii}$. Then, the projected estimator is equal to

$$\hat{\mathbf{K}}_+ = \mathbf{U}\mathbf{D}_+\mathbf{U}^T \tag{65}$$

where $\mathbf{D}_+$ is a diagonal matrix with the elements $[\mathbf{D}_+]_{ii} = \max\{[\mathbf{D}]_{ii}, 0\}$. Due to its similarity to the positive-part shrinkage estimator in James-Stein regression, we refer to (65) as the positive-part estimator.

VI. Numerical Results

Here we present results of numerical experiments in order to illustrate the performance of the above estimators. A standard benchmark used for testing (inverse) covariance estimation and covariance selection is the call center data set [23], [29], [39]. Our goal is to demonstrate estimation precision rather than model selection accuracy. Therefore, we estimate the true call center covariance matrix using fixed decomposable models as proposed and discussed in [39]. Next, we artificially generate n independent and identically distributed realizations of jointly Gaussian vectors which follow the true covariance structure. We repeat this procedure 100 times and report the average performance over these independent trials. We use the three decomposable graphical models analyzed in [39].

1) *Two Coupled Cliques:* $C_1 = \{1, \ldots, 70\}$ and $C_2 = \{61, \ldots, 100\}$.
2) *Banding:* A non-stationary autoregressive model with $p = 239$ and cliques $C_k = \{j, \ldots, j+L\}$ for $j = 1, \ldots, j-p$ with an empirically validated bandwidth of $L = 20$.
3) *Differential Banding:* An empirically validated and refined banding model in which the first 58 cliques have a bandwidth of $L = 14$ and in the following cliques the bandwidth is equal to $L = 4$.

Throughout the simulations, we compare the performance of three estimators: the MLE in (29), the MVUE in (31), and the SURE-based estimator in (50) with d given by (56). For each realization, we compute the estimators and check their semidefiniteness. When an estimator is not positive semidefinite, we resort to its positive-part projection defined in (63). In Figs. 1–3 we present the normalized MSE defined as $\|\hat{\mathbf{K}} - \mathbf{K}\|^2/\|\mathbf{K}\|^2$ as a function of the sample size n.

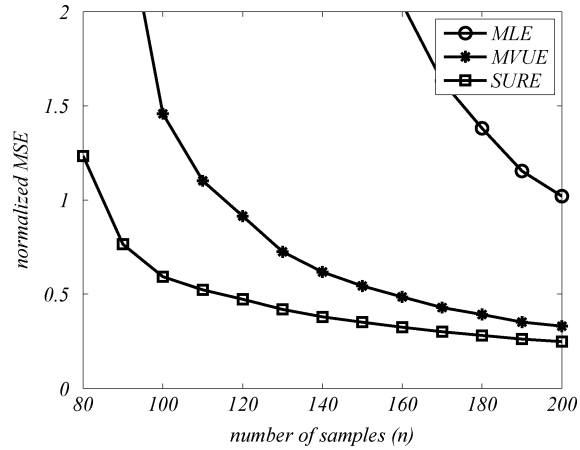


Fig. 1. Two cliques model: significant MSE improvement with respect to MLE.

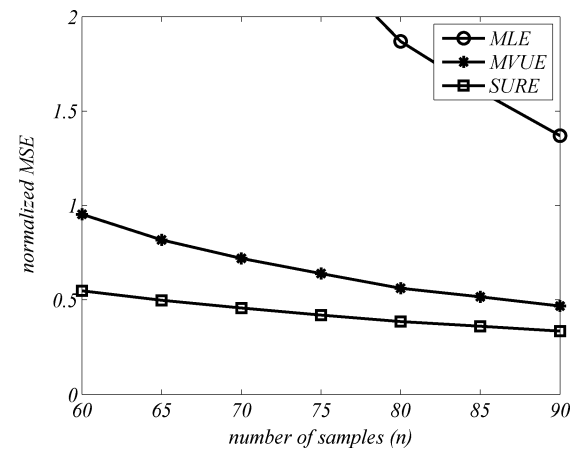


Fig. 2. Banding model: significant MSE improvement with respect to MLE.

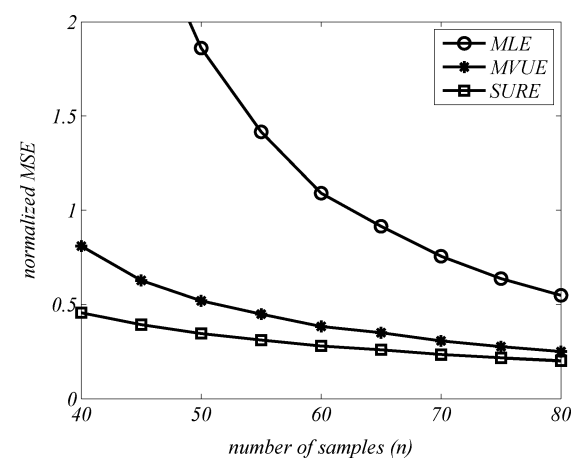


Fig. 3. Differential banding model: significant MSE improvement with respect to MLE.

It is easy to see the significant MSE performance advantage of the MVUE and the SURE based estimators of $\mathbf{K}$ as compared to the MLE. The gain is most significant when the

Authorized licensed use limited to: Technion Israel School of Technology. Downloaded on April 21,2010 at 20:35:35 UTC from IEEE Xplore. Restrictions apply.

number of samples is small. In this regime, the MLE performs poorly and is actually worse than the zero estimator, i.e., $\hat{\mathbf{K}} = \mathbf{0}$ which ignores the observations altogether, whereas the newly proposed estimators provide reasonable performance. In small sample sizes the MVUE and SURE based estimators had to be adjusted using their positive part variants. Simulation results (not reported) suggest that the improvement in MSE due to the positive part adjustment is negligible.

VII. Conclusion and Future Work

In this paper, we suggested several alternatives to the MLE for concentration estimation in decomposable graphical models. We derived the MVUE and further proposed two biased estimators that have lower MSE than the MLE. The suggested estimators have simple closed form solutions and their computational complexity is similar to the MLE. In addition, we generalized SURE to decomposable graphical models.

Throughout this work, we assumed that the graphical model is decomposable and illustrated our results for practical signal processing examples, e.g., banded and arrow structured concentration matrices. Moreover, any conditional independence graph can be approximated as decomposable using available graph theoretical tools. An important challenge for future work is the extension of our results to non-decomposable graphs.

Appendix I
MVUE and SURE in the Exponential Family

A natural exponential family is defined as

$$f(\mathbf{u};\boldsymbol{\theta}) = k(\mathbf{u})e^{\boldsymbol{\theta}^T\mathbf{u}-\psi(\boldsymbol{\theta})}. \tag{66}$$

Its natural parameter is $\boldsymbol{\theta}$ and $\mathbf{u} \in \mathcal{U}$ is a complete sufficient statistic. Under mild technical conditions, the MVUE for estimating $\boldsymbol{\theta}$ given $\mathbf{u}$ is [35]

$$\hat{\boldsymbol{\theta}}_{\mathrm{MVUE}} = -\frac{\partial}{\partial \mathbf{u}}\log(k(\mathbf{u})) \tag{67}$$

and the SURE identity is [35]–[37]

$$\begin{aligned} E\{h(\mathbf{u})\cdot[\boldsymbol{\theta}]_i\} &= E\left\{h(\mathbf{u})\left[-\frac{\partial\log(k(\mathbf{u}))}{\partial[\mathbf{u}]_i}\right] - \frac{\partial h(\mathbf{u})}{\partial[\mathbf{u}]_i}\right\} \\ &= E\left\{h(\mathbf{u})[\hat{\boldsymbol{\theta}}_{\mathrm{MVUE}}]_i - \frac{\partial h(\mathbf{u})}{\partial[\mathbf{u}]_i}\right\}. \end{aligned} \tag{68}$$

The technical conditions for the validity of (67) and (68) are [37, Th. 2.1]

$$\begin{aligned} &\mathcal{U} \text{ is a finite union of open connected sets;} \\ &h(\mathbf{u}) \text{ is an indefinite integral of } \frac{\partial h(\mathbf{u})}{\partial[\mathbf{u}]_i}; \\ &E\left\{\left|\frac{\partial h(\mathbf{u})}{\partial[\mathbf{u}]_i}\right|\right\} < \infty; \\ &E\left\{\left|\left[-\frac{\partial\log(k(\mathbf{u}))}{\partial[\mathbf{u}]_i} - [\boldsymbol{\theta}]_i\right]\frac{\partial h_i(\mathbf{u})}{\partial[\mathbf{u}]_i}\right|\right\} < \infty \\ &\text{if length}(\boldsymbol{\theta}) > 1. \end{aligned} \tag{69}$$

The distribution of the incomplete sample covariance $\mathbf{S}$ in (27) belongs to the natural exponential family. The variable $\mathbf{x}$ is a vector of the specified elements in the upper triangular part of $\mathbf{S}$ with a scaling of 2 in the off diagonal to account for the symmetry. The natural parameter $\boldsymbol{\theta}$ is a vector with the nonzero elements in the upper triangular part of $\mathbf{K}$. Therefore, the MVUE can be derived by plugging (23) into (27)

$$\log(k(\mathbf{S})) = \sum_{k=1}^{K}\log|\mathbf{S}_k|^{\frac{n-c_k-1}{2}} - \sum_{k=1}^{K}\log|\mathbf{S}_k||^{\frac{n-s_k-1}{2}} \tag{70}$$

and applying (67) to obtain

$$\begin{aligned} \hat{\mathbf{K}}_{\mathrm{MVUE}} &= 2\boldsymbol{\nabla}\log(k(\mathbf{S})) \\ &= \sum_{k=1}^{K}\left[(n-c_k-1)\mathbf{S}_k^{-1}\right]^0 - \sum_{k=2}^{K}[(n-s_k-1)\mathbf{S}_{[k]}^{-1}]^0 \end{aligned} \tag{71}$$

where $\boldsymbol{\nabla}$ is a symmetric matrix version of the derivative in (67) which accounts for the two factors in the off diagonal elements (see [19] and [34] for more details on this operator and its application to various matrix functions).

Similarly, the decomposable SURE identity in Theorem 2 is a matrix representation of (68) using a matrix function $\mathbf{H}(\mathbf{S})$ and notations which account for the symmetry factors and take the derivatives only with respect to the specified elements in $\mathbf{S}$.

Appendix II
Technical Inequalities

For simplicity, we partition the submatrix of the kth clique as

$$\mathbf{S}_k = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \tag{72}$$

where $\mathbf{S}_{[k]} = \mathbf{C}$ is the intersection with the separator.

Proof of (43): Using the partitioned matrix inverse

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Delta}^{-1} & -\boldsymbol{\Delta}^{-1}\mathbf{B}\mathbf{C}^{-1} \\ -\mathbf{C}^{-1}\mathbf{B}^T\boldsymbol{\Delta}^{-1} & \mathbf{C}^{-1}+\mathbf{Z} \end{bmatrix} \tag{73}$$

where $\boldsymbol{\Delta} = \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T$ and $\mathbf{Z} = \mathbf{C}^{-1}\mathbf{B}^T\boldsymbol{\Delta}^{-1}\mathbf{B}\mathbf{C}^{-1}$. Therefore

$$\mathrm{Tr}\,\{\mathbf{S}_k^{-1}\} = \mathrm{Tr}\,\{\boldsymbol{\Delta}^{-1}\} + \mathrm{Tr}\,\{\mathbf{C}^{-1}\} + \mathrm{Tr}\,\{\mathbf{Z}\} \geq \mathrm{Tr}\,\{\mathbf{C}^{-1}\} \tag{74}$$

where the last inequality is due to the positive semidefiniteness of $\mathbf{S}_k \succeq \mathbf{0}$ and its Schur complement $\boldsymbol{\Delta} \succeq \mathbf{0}$.

Proof of (44): Using the partitioned matrix inverse once again, we obtain

$$\begin{aligned} \mathrm{Tr}\,\{\mathbf{S}_k^{-2}\} &= \mathrm{Tr}\,\{\boldsymbol{\Delta}^{-2}\} + 2\mathrm{Tr}\,\{\boldsymbol{\Delta}^{-1}\mathbf{B}\mathbf{C}^{-2}\mathbf{B}^T\boldsymbol{\Delta}^{-1}\} \\ &\quad + \mathrm{Tr}\,\{\mathbf{C}^{-2}\} + 2\,\mathrm{Tr}\,\left\{\mathbf{C}^{-\frac{1}{2}}\mathbf{Z}\mathbf{C}^{-\frac{1}{2}}\right\} \\ &\quad + \mathrm{Tr}\,\{\mathbf{Z}^2\} \geq \mathrm{Tr}\,\{\mathbf{C}^{-2}\}. \end{aligned} \tag{75}$$

Authorized licensed use limited to: Technion Israel School of Technology. Downloaded on April 21,2010 at 20:35:35 UTC from IEEE Xplore. Restrictions apply.

ACKNOWLEDGMENT

The authors would like to thank C. M. Carvalho for providing the call center dataset, and the anonymous reviewers for their constructive comments.

REFERENCES

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.
[2] E. R. Dougherty, A. Datta, and C. Sima, "Research issues in genomic signal processing," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 46–68, Nov. 2005.
[3] Y. Weiss and W. T. Freeman, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology," *Neural Computation*, vol. 13, no. 10, pp. 2173–2200, 2001.
[4] O. Shental, P. H. Siegel, J. K. Wolf, D. Bickson, and D. Dolev, "Gaussian belief propagation solver for systems of linear equations," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2008, pp. 1863–1867.
[5] S. L. Lauritzen, *Graphical Models*. Oxford, U.K.: Oxford University Press, 1996.
[6] A. P. Dempster, "Covariance selection," in *Biometrics*, 1972, vol. 28, pp. 157–175.
[7] J. A. Bilmes, "Factored sparse inverse covariance matrices," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2000, pp. 1009–1012.
[8] J. A. Bilmes and C. Bartels, "Graphical model architectures for speech recognition," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 89–100, Sep. 2005.
[9] A. Willsky, "Multiresolution Markov models for signal and image processing," *Proc. IEEE*, vol. 90, no. 8, pp. 1396–1458, Aug. 2002.
[10] M. J. Choi and A. S. Willsky, "Multiscale Gaussian graphical models and algorithms for large-scale inference," in *Proc. IEEE/SP 14th Workshop Statistical Signal Process. (SSP)*, 2007, pp. 229–233.
[11] M. Cetin, L. Chen, J. W. Fisher, A. T. Ihler, R. L. Moses, M. J. Wainwright, and A. S. Willsky, "Distributed fusion in sensor networks: A graphical models perspective," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 42–55, Jul. 2006.
[12] A. Wiesel and A. O. Hero III, "Decomposable principal component analysis," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4369–4377, Nov. 2009.
[13] E. B. Sudderth, M. J. Wainwright, and A. S. Willsky, "Embedded trees: Estimation of Gaussian processes on graphs with cycles," *IEEE Trans. Signal Process.*, vol. 52, no. 11, pp. 3136–3150, Nov. 2004.
[14] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, "Walk-sums and belief propagation in Gaussian graphical models," *J. Mach. Learn. Res.*, vol. 7, pp. 2031–2064, Oct. 2006.
[15] V. Chandrasekaran, J. K. Johnson, and A. S. Willsky, "Estimation in Gaussian graphical models using tractable subgraphs: A walk-sum analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1916–1930, May 2008.
[16] Y. I. Abramovich, N. K. Spencer, and M. D. E. Turley, "Time-varying autoregressive (TVAR) models for multiple radar observations," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1298–1311, Apr. 2007.
[17] Y. I. Abramovich, B. A. Johnson, and N. K. Spencer, "Two-dimensional multivariate parametric models for radar applications—Part II: Maximum-entropy extensions for hermitian-block matrices," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5527–5539, Nov. 2008.
[18] C. Stein, "Estimation of a covariance matrix," presented at the Rietz Lecture, 39th Annual Meeting IMS, Atlanta, GA, 1975.
[19] L. R. Haff, "Empirical bayes estimation of the multivariate normal covariance matrix," *Annals Statistics*, vol. 8, no. 3, pp. 586–597, 1980.
[20] H. Tsukuma and Y. Konno, "On improved estimation of normal precision matrix and discriminant coefficients," *J. Multivariate Anal.*, vol. 97, pp. 1477–1500, 2006.
[21] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivariate Anal.*, vol. 88, no. 2, pp. 365–411, Feb. 2004.
[22] R. Yang and J. O. Berger, "Estimation of a covariance matrix using the reference prior," *Annals Statistics*, vol. 22, pp. 195–211, 1994.
[23] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Annals Statistics*, vol. 36, no. 1, pp. 199–227, 2008.
[24] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *Annals Statististics*, vol. 36, no. 6, pp. 2577–2604, 2008.
[25] T. P. Speed and H. T. Kiiveri, "Gaussian Markov distributions over finite graphs," *Annals Statistics*, vol. 14, no. 1, pp. 138–150, 1986.
[26] A. P. Dawid and S. L. Lauritzen, "Hyper markov laws in the statistical analysis of decomposable graphical models," *Annals Statistics*, vol. 21, no. 3, pp. 1272–1317, 1993.
[27] A. Kavcic and J. M. F. Moura, "Matrices with banded inverses: Inversion algorithms and factorization of Gauss-Markov processes," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1495–1509, Jul. 2000.
[28] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the L}ASSO," *Biostat*, vol. 9, no. 3, pp. 432–441, Jul. 2008.
[29] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.
[30] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation," *J. Mach. Learn. Res.*, vol. 9, pp. 485–516, Mar. 2008.
[31] A. Anandkumar, L. Tong, and A. Swami, "Detection of Gauss Markov random fields with nearest-neighbor dependency," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 816–827, Feb. 2009.
[32] A. Deshpande, M. N. Garofalakis, and M. I. Jordan, "Efficient stepwise selection in decomposable models," in *Proc. 17th Conf. Uncertainty Artif. Intell. (UAI'01)*, San Francisco, CA, 2001, pp. 128–135.
[33] B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West, "Experiments in stochastic computation for high-dimensional graphical models," *Statistical Sci.*, vol. 20, pp. 388–400, 2005.
[34] C. Stein, "Lectures on the theory of estimation of many parameters," in *Zapiski Nauchnykh Seminarov Leningradskogo Otdeleniya Matematicheskogo Instituta im. V. A. Steklova AN SSSR*, 1977, vol. 74, pp. 4–65.
[35] H. M. Hudson, "A natural identity for exponential families with applications in multiparameter estimation," *Annals Statistics*, vol. 6, no. 3, pp. 473–484, 1978.
[36] Y. C. Eldar, "Generalized SURE for exponential families: Applications to regularization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 471–481, Feb. 2009.
[37] J. P. Chou, "An identity for multidimensional continuous exponential families and its applications," *J. Multivariate Anal.*, vol. 24, pp. 129–142, 1988.
[38] G. Letac and H. Massam, "Wishart distributions for decomposable graphs," *Annals Statistics*, vol. 35, no. 3, pp. 1278–1323, 2007.
[39] B. Rajaratnam, H. Massam, and C. M. Carvalho, "Flexible covariance estimation in graphical Gaussian models," *Annals Statistics*, vol. 36, no. 6, pp. 2818–2849, 2007.
[40] D. Sun and X. Sun, "Estimation of the multivariate normal precision and covariance matrices in a star-shape model," *Annals Inst. Statistical Math.*, vol. 57, no. 3, pp. 455–484, Sep. 2005.
[41] S. Kay and Y. C. Eldar, "Rethinking biased estimation," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 133–136, Mar. 2008.
[42] Y. C. Eldar, "Rethinking biased estimation: Improving maximum likelihood and the Cramer-Rao bound," *Foundations Trends Signal Process.*, vol. 1, no. 4, pp. 305–449, 2007.
[43] L. R. Haff, "Minimax estimators for a multinormal precision matrix," *J. Multivariate Anal.*, vol. 7, pp. 374–385, 1977.
[44] S. M. Kay, *Fundamentals of Statistical Signal Processing—Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[45] Y. Konno, "Inadmissibility of the maximum likelihood estimator of normal covariance matrices with the lattice conditional independence," *J. Multivariate Anal.*, vol. 79, pp. 33–51, 2001.
[46] X. Sun and D. Sun, "Estimation of a multivariate normal covariance matrix with staircase pattern data," *Annals Inst. Statistical Math.*, vol. 59, no. 2, pp. 211–233, Jun. 2007.
[47] M. O. Ulfarsson and V. Solo, "Dimension estimation in noisy PCA with SURE and random matrix theory," *IEEE Trans. Signal Process.*, vol. 56, no. 12, pp. 5804–5816, Dec. 2008.
[48] F. Luisier, T. Blu, and M. Unser, "A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 593–606, Mar. 2007.
[49] J. F. Sturm, "Using SEDUMI 1.02, a Matlab toolbox for optimizations over symmetric cones," *Optimization Meth. Soft.*, vol. 11–12, pp. 625–653, 1999.
[50] S. Boyd and L. Vandenberghe, *Introduction to Convex Optimization With Engineering Applications*. Stanford, CA: Stanford Univ., 2003.

Ami Wiesel (S'02–M'09) received the B.Sc. and M.Sc. degrees in electrical engineering from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 2000 and 2002, respectively, and the Ph.D. degree in electrical engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 2007.

Currently, he is a Postdoctoral Fellow with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor.

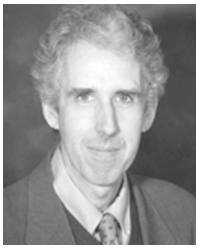
Dr. Wiesel was a recipient of the Young Author Best Paper Award for a 2006 paper in the IEEE TRANSACTIONS IN SIGNAL PROCESSING and a Student Paper Award for the 2005 Workshop on Signal Processing Advances in Wireless Communications (SPAWC) paper. He was awarded the Weinstein Study Prize in 2002, the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

Authorized licensed use limited to: Technion Israel School of Technology. Downloaded on April 21,2010 at 20:35:35 UTC from IEEE Xplore. Restrictions apply.

Yonina C. Eldar (S'98–M'02–SM'07) received the B.Sc. degree in physics and the B.Sc. degree in electrical engineering from Tel-Aviv University (TAU), Tel-Aviv, Israel, in 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2001.

From January 2002 to July 2002, she was a Postdoctoral Fellow with the Digital Signal Processing Group, MIT. She is currently an Associate Professor with the Department of Electrical Engineering, the Technion—Israel Institute of Technology, Haifa, Israel. She is also a Research Affiliate with the Research Laboratory of Electronics, MIT. Her research interests include the general areas of signal processing, statistical signal processing, and computational biology.

In 2004, Dr. Eldar was awarded the Wolf Foundation Krill Prize for Excellence in Scientific Research, in 2005 the Andre and Bella Meyer Lectureship, in 2007 the Henry Taub Prize for Excellence in Research, in 2008 the Hershel Rich Innovation Award, the Award for Women with Distinguished Contributions, the Muriel & David Jacknow Award for Excellence in Teaching, and the Technion Outstanding Lecture Award, and in 2009 the Technion's Award for Excellence in Teaching. She is a member of the IEEE Signal Processing Theory and Methods technical committee and the Bio Imaging Signal Processing technical committee, an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the *EURASIP Journal of Signal Processing*, the *SIAM Journal on Matrix Analysis and Applications*, and the *SIAM Journal on Imaging Sciences*, and on the Editorial Board of Foundations and Trends in Signal Processing.



Alfred O. Hero III (F'98) received the B.S. (*summa cum laude*) degree from Boston University, Boston, MA, in 1980 and the Ph.D. from Princeton University, Princeton, NJ, in 1984, both in electrical engineering.

Since 1984, he has been with the University of Michigan, Ann Arbor, where he is the R. Jamison and Betty Professor of Engineering. At the University of Michigan, his primary appointment is in the Department of Electrical Engineering and Computer Science and he also has appointments, by courtesy, in the Department of Biomedical Engineering and the Department of Statistics. He has held other visiting positions at LIDS Massachussets Institute of Technology (2006), Boston University (2006), I3S University of Nice, Sophia-Antipolis, France (2001), Ecole Normale Supérieure de Lyon (1999), Ecole Nationale Supérieure des Télécommunications, Paris (1999), Lucent Bell Laboratories (1999), Scientific Research Labs of the Ford Motor Company, Dearborn, Michigan (1993), Ecole Nationale Superieure des Techniques Avancees (ENSTA), Ecole Superieure d'Electricite, Paris (1990), and M.I.T. Lincoln Laboratory (1987–1989). His recent research interests have been in detection, classification, pattern analysis, and adaptive sampling for spatio-temporal data. Of particular interest are applications to network security, multi-modal sensing and tracking, biomedical imaging, and genomic signal processing.

Dr. Hero was awarded the Digiteo Chaire d'Excellence, sponsored by Digiteo Research Park in Paris, located at the Ecole Superieure d'Electricite, Gif-sur-Yvette, France, in 2008. He has served on the editorial boards of the IEEE TRANSACTIONS ON INFORMATION THEORY (1995–1998, 1999), the IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS (2004–2006), and the IEEE TRANSACTIONS ON SIGNAL PROCESSING (2002, 2004). He was Chairman of the Statistical Signal and Array Processing (SSAP) Technical Committee (1997–1998) and Treasurer of the Conference Board of the IEEE Signal Processing Society. He was Chairman for Publicity of the 1986 IEEE International Symposium on Information Theory (Ann Arbor, MI) and General Chairman of the 1995 IEEE International Conference on Acoustics, Speech, and Signal Processing (Detroit, MI). He was co-chair of the 1999 IEEE Information Theory Workshop on Detection, Estimation, Classification and Filtering (Santa Fe, NM) and the 1999 IEEE Workshop on Higher Order Statistics (Caesaria, Israel). He Chaired the 2002 NSF Workshop on Challenges in Pattern Recognition. He co-chaired the 2002 Workshop on Genomic Signal Processing and Statistics (GENSIPS). He was Vice President (Finance) of the IEEE Signal Processing Society (1999–2002). He was Chair of Commission C (Signals and Systems) of the US National Commission of the International Union of Radio Science (URSI) (1999–2002). He was member of the Signal Processing Theory and Methods (SPTM) Technical Committee of the IEEE Signal Processing Society (1999–2004). He was President of the IEEE Signal Processing Society (2006–2007) and during his term he served on the TAB Periodicals Committee (2006). He was a member of the IEEE TAB Society Review Committee (2009) and is Director-elect of IEEE for Division IX (2009). He is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE), a member of Tau Beta Pi, the American Statistical Association (ASA), the Society for Industrial and Applied Mathematics (SIAM), and the US National Commission (Commission C) of the International Union of Radio Science (URSI). He has been plenary and keynote speaker at several major conferences and received a IEEE Signal Processing Society Meritorious Service Award (1998), a IEEE Signal Processing Society Best Paper Award (1998), a IEEE Third Millenium Medal (2000) and a 2002 IEEE Signal Processing Society Distinguished Lecturership.

Authorized licensed use limited to: Technion Israel School of Technology. Downloaded on April 21,2010 at 20:35:35 UTC from IEEE Xplore. Restrictions apply.