2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

Optimal Covariance Selection for Estimation Using Graphical Models

Sergey Vichik and Yaakov Oshman

***Abstract*— We consider a problem encountered when trying to estimate a Gaussian random field using a distributed estimation approach based on Gaussian graphical models. Because of constraints imposed by estimation tools used in Gaussian graphical models, the a priori covariance of the random field is constrained to embed conditional independence constraints among a significant number of variables. The problem is, then: given the (unconstrained) a priori covariance of the random field, and the conditional independence constraints, how should one select the constrained covariance, optimally representing the (given) a priori covariance, but also satisfying the constraints? In 1972, Dempster provided a solution, optimal in the maximum likelihood sense, to the above problem. Since then, many works have used Dempster's optimal covariance, but none has addressed the issue of suitability of this covariance for Bayesian estimation problems. We prove that Dempster's covariance is not optimal in most minimum mean squared error (MMSE) estimation problems. We also propose a method for finding the MMSE optimal covariance, and study its properties. We then illustrate the analytical results via a numerical example, that demonstrates the estimation performance advantage gained by using the optimal covariance vs Dempster's covariance. The numerical example also shows that, for the particular estimation scenario examined, Dempster's covariance violates the necessary conditions for optimality.**

I. Introduction

Consider the Bayesian problem of estimating the random variable X from a noisy measurement $Z = X + V$, where $V \sim N(0, R)$ and $X \sim N(0, S)$. In the classical problem formulation, the a priori covariance of X is known, and has a general, unconstrained, form. Often, however, some constraints must be imposed on the components of X, resulting in constraints on the a priori covariance. This can happen, e.g., when dealing with problems of large dimension, where it is difficult or altogether impractical to employ the full a priori covariance of X due to limited computation resources or limited enrolling data, or when using estimation algorithms that impose constraints on the covariance structure. In such cases, some constraints must be imposed on the covariance, resulting in a modified, but tractable, estimation problem. As a trivial example, we can assume that some elements of X are uncorrelated, which is equivalent to requiring that the corresponding entries of the (constrained) covariance of X vanish. In these cases the problem becomes: given the full (unconstrained) a priori covariance, and the constraints imposed on the components of X, how should one compute a constrained covariance of X, such that 1) it is closest (in some well defined sense) to the full (unconstrained) a priori covariance, yet 2) it fully satisfies the constraints?

In 1972 Dempster [1] studied the model selection (or covariance selection) problem. In his seminal work Dempster presents an effective method for constrained covariance selection, using reduction of the number of model parameters. The estimated (constrained) covariance was shown to be optimal (i.e., to best approximate the original covariance, while simultaneously satisfying the constraints) in the maximum likelihood sense. While seeking to reduce the number of parameters to be estimated, Dempster showed that when some components of the random vector X are forced to be conditionally independent, the maximum likelihood covariance takes a specific form, that facilitates the reduction of the number of parameters.

In the following years the field of graphical model methods has been founded. A graphical model is a probabilistic model, in which a graph is used to denote the conditional independence structure between random variables. Dawid and Lauritzen [2] showed that cases associated with conditional independence properties can be effectively represented by Gaussian graphical models. In such cases, Dempster's results can be applied.

Subsequently to the introduction of Dempster's original work [1], a number of algorithms were developed to find Dempster's covariance in an efficient way [3]–[7]. Other works studied a related problem of finding the covariance structure (selecting the independent variables) [8], [9].

In estimation applications Dempster's maximum likelihood covariance was used in a number of speech recognition algorithms [10]. In [11] the authors use this covariance for estimation using a generalization of the Kalman filter to tree applications. The original Dempster covariance was used, or implicitly assumed, in algorithms for solving estimation problems using graphical methods, such as [12]–[15].

Perhaps surprisingly, none of the aforementioned works has addressed the following fundamental question: is Dempster's covariance, developed for the maximum likelihood modeling problem, also optimal for a Bayesian estimation problem? Although many criteria exist, estimation algorithms' performance is commonly measured by the engineering community by the mean squared error (MSE) criterion. Whereas for the Gaussian linear, non-constrained case it is well known that the MSE and the maximum likelihood optimization criteria are equivalent, this is not true in general. Therefore, for more complex cases, different methods may yield different results. Indeed, a similar question was studied

S. Vichik (vserg@tx.technion.ac.il) and Y. Oshman (yaakov.oshman@technion.ac.il) are with the Department of Aerospace Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

978-1-4577-0079-8/11/$26.00 ©2011 AACC

by Eldar [16], who showed that, for low signal to noise ratio (SNR) cases, MSE optimization yields results that are different from those obtained by minimum variance optimization.

In this work we study the problem of optimal covariance selection in the MSE sense for a Bayesian estimation problem solved using Gaussian graphical models. We show that if an estimation algorithm uses a covariance that is constrained by the special conditional independence structure, the optimal value of the MSE criterion thus achieved is superior to that achieved by the matrix used in all previous works. In fact, we show that, for most Bayesian estimation problems, Dempster's covariance is not an optimal constrained covariance. We propose a method for computing the optimal covariance matrix, and study its properties.

The remainder of this paper is organized as follows. In Section II we present some mathematical background. This is followed by a definition of the problem in Section III. In Section IV we define the optimal covariance in terms of the solution of an optimization problem, and study its properties. In Section V we prove that Dempster's covariance is not optimal in most Bayesian estimation problems of the type dealt with herein. Section VI illustrates the analytical results via a numerical simulation that demonstrates the benefits of using the optimal covariance instead of Dempster's covariance. Concluding remarks are offered in the last section.

II. Theoretical Background

In this section we provide some theoretical background, and briefly review Dempster's covariance selection problem and its solution. We start with a presentation of some facts from graph theory and Gaussian graphical models.

A. Gaussian Graphical Models

Let the pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph, where $\mathcal{V}$ is a group of vertices and $\mathcal{E}$ is a group of edges that connect some (or all) of the vertices from $\mathcal{V}$. Let $C \in \mathcal{V}$ be a group of vertices. C is called a clique if there exist edges in $\mathcal{E}$ that connect every two vertices in C. Let C and B be two cliques in $\mathcal{G}$. The intersection of C and B, $S = C \cap B$, is called separator. For neighboring cliques a separator defines all vertices that separate one clique from another.

Consider now the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. For notation simplicity, we extend the definition of $\mathcal{E}$ to include edges from every node to itself (the diagonal entries in an adjacency matrix). Let $X \in \mathbb{R}^{|\mathcal{V}|}$ be a Gaussian random vector, such that every component of X is associated with a corresponding entry of $\mathcal{V}$, and let f_X be the pdf of X. The pair of $\mathcal{G}$ and f_X constitutes a Gaussian graphical model.

For sets of vertices $x, y, z \subseteq \mathcal{V}$, we say that x and y are separated by z in $\mathcal{G}$ if every path from x to y includes members from z. Separation in a Gaussian graphical model is associated with conditional independence, defined as follows.

Definition 1 (Conditional Independence [17]): Let X, Y, and Z be jointly distributed Gaussian random variables. Then X and Y are conditionally independent given Z, if the joint probability density function (pdf) can be factored as follows:

$$f_{X,Y,Z}(x, y, z) = g_1(x, z) g_2(y, z) \tag{1}$$

We say that x is separated from y by z in $\mathcal{G}$ iff the random vectors associated with x, y, and z, denoted by X, Y, and Z, respectively, satisfy that X is independent of Y given Z [17].

B. The Covariance Selection Problem

Let $X \sim N(\mu, S)$ be a Gaussian random vector, where S is its true (dense) unconstrained covariance, and let $\mathcal{G}$ be a graph we want to associate with X. The association of $\mathcal{G}$ with X imposes constraints among the components of X, such that the (unconstrained) covariance S needs to be approximated by a constrained covariance P that 1) best approximates S in some well defined sense, and 2) satisfies the constraints imposed on the components of X. To investigate this approximation problem, we recall the following useful lemma, proved in [18].

Lemma 1 (Inverse covariance structure): Let P be the covariance of the random vector X associated with the graphical model $(\mathcal{G}, f_X)$. Then, the inverse covariance matrix of X, P^{-1}, is nonzero only in entries (i, j) such that there is an edge in $\mathcal{E}$ from vertex i to vertex j.

Thus, according to Lemma 1, we need to find an approximation P of S, such that P^{-1} has zeroes according to the structure of $\mathcal{G}$.

The problem of finding P from S is known as the covariance selection problem. The first thorough analysis and a partial solution of this problem were given by Dempster in [1]. The article does not address graphical models, but it uses zeroes in the inverse covariance matrix for reducing the number of free parameters in the covariance estimation problem (model learning). Dempster proved in [1] a number of results that are very important in the context of the problem treated in this paper. We summarize these results herewith.

Let $\mathcal{I}$ be the set of indices corresponding to all zero entries of P^{-1}, and let $\mathcal{J}$ be the set of indices of all other entries of P^{-1}. Then the following properties hold.

1) The entries of P associated with the index set $\mathcal{J}$ are sufficient statistics for the problem of determining the covariance under the constraint $P^{-1}(\mathcal{I}) = 0$.
2) $P(\mathcal{J}) = S(\mathcal{J})$.
3) The matrix P is a maximum likelihood estimate of the covariance of X under the constraint $P^{-1}(\mathcal{I}) = 0$. Thus, P is the solution of the following minimization problem:

$$\min_{P \in \{P^{-1}(\mathcal{I})=0\}} L(P) \tag{2a}$$

where

$$L(P) \triangleq \sum_{i}^{N} \log \left(\frac{(2\pi)^{-\frac{n}{2}}}{\det(P)^{\frac{1}{2}}} e^{\left[-\frac{1}{2} X_i P^{-1} X_i\right]} \right) \tag{2b}$$

and X_i is a single sample out of N samples available for the covariance estimation.

In other words, Dempster proved that, given some data about the real covariance of X, the $\mathcal{J}$ entries of the maximum likelihood estimate of the covariance under the constraint $P^{-1}(\mathcal{I}) = 0$ are equal to the corresponding entries of the unconstrained covariance. The other entries are chosen to satisfy the $P^{-1}(\mathcal{I}) = 0$ constraint.

Although there exist many algorithms to compute the Dempster covariance from the original covariance, a closed form expression given in [2, p. 1306] is useful for further analysis and understanding how the computation is done:

$$F = \sum_{C \in \mathcal{C}} \left[F^C\right]^0 - \sum_{S \in \mathcal{S}} \left[F^S\right]^0 \tag{3}$$

where $F = P^{-1}$, $\mathcal{C}$ is the group of all cliques in the graph $\mathcal{G}$, $\mathcal{S}$ is the group of all separators, and F^C and F^S represent the inverse covariance matrices of the corresponding subgroups. The $[A]^0$ operator appends zeros to the matrix A to give it the correct dimensions.

III. Problem definition

Let $X \sim N(\mu, S)$ be an unobserved Gaussian random vector, and let Z be a vector of noisy observations of X, such that $Z = HX + V$, where $V \sim N(0, R)$ and H is the observation matrix. Assuming that X is associated with a given graphical model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we need to estimate X given Z using the graphical model $\mathcal{G}$. Thus, the covariance P associated with the graphical model must have a structure as stated in Lemma 1.

We can state the problem in equivalent terms without using the graphical model machinery, but still abiding by its constraints. Thus, the estimation problem may be posed in a conventional matrix form, with the constrained covariance that is dictated by the graphical model. To do that, assume, for the moment, that P is the a priori covariance of X. Then, the MMSE estimate of X from Z is given by the following equations:

$$\hat{X} = \hat{P}_p(H^T R^{-1} Z + P^{-1}\mu) \tag{4a}$$

$$\hat{P}_p = (P^{-1} + H^T R^{-1} H)^{-1} \tag{4b}$$

where $\hat{P}_p$ is the (a posteriori) covariance of $\hat{X}$. Obviously, when $P = S$, that is, when the true covariance of X is used, equations (4) yield the optimal estimate (in the MMSE sense). However, because of the constraint on the covariance, we cannot use S. Thus, for any $P \neq S$ we can expect (4) to yield a suboptimal solution.

Our problem is to find the optimal constrained covariance P, that yields the MMSE estimate of X, that is

$$P = \arg \min_{P \in \{P^{-1}(\mathcal{I})=0\}} \operatorname{tr}(E\tilde{X}\tilde{X}^T) \tag{5}$$

where $\tilde{X} \triangleq \hat{X} - X$ is the estimation error of X.

IV. The Optimal Covariance

Dempster's result $P(\mathcal{J}) = S(\mathcal{J})$ means that the nonzero cross-covariances in P are unchanged relative to the true covariance, S. The usage of the constrained covariance, P, in the estimation problem, may give rise to non-optimal performance, because it means that although some edges have been removed from the graphical model, no compensation has been incorporated and new measurements are processed using wrongly correlated components of X. We thus seek for a solution that takes this model reduction into account.

A. Estimation Error Covariance

Using (4a) The estimation error is

$$\begin{aligned}\tilde{X} &= \hat{P}_p(H^T R^{-1}(HX + V) + P^{-1}\mu) - X \\ &= (\hat{P}_p H^T R^{-1} H - I)X + \hat{P}_p H^T R^{-1} V + \hat{P}_p P^{-1}\mu.\end{aligned} \tag{6}$$

Thus, the mean of the estimation error is

$$\begin{aligned}E(\tilde{X}) &= (\hat{P}_p H^T R^{-1} H - I)\mu + \hat{P}_p P^{-1}\mu \\ &= (\hat{P}_p(P^{-1} + H^T R^{-1} H) - I)\mu = 0\end{aligned} \tag{7}$$

and its covariance becomes

$$\begin{aligned}E(\tilde{X}\tilde{X}^T) =&(\hat{P}_p H^T R^{-1} H - I)S(H^T R^{-1} H \hat{P}_p - I) \\ &+ \hat{P}_p H^T R^{-1} H \hat{P}_p.\end{aligned} \tag{8}$$

Notice that, as could be expected, the estimation error covariance does not depend on μ. Furthermore, this result holds for every constrained covariance P, not only for $P = S$ (since (8) does not depend on μ). Hence, without loss of generality, we will assume $\mu = 0$ in the sequel.

B. Covariance Optimization Problem

Denoting $K = \hat{P}_p^{-1}$, we use (8) to derive an optimization problem whose solution yields the optimal constrained covariance:

$$\begin{aligned}&\min_K \operatorname{tr}\big[(K^{-1}H^T R^{-1} H - I)S(H^T R^{-1} H K^{-1} - I) \\ &\qquad + K^{-1} H^T R^{-1} H K^{-1}\big] \\ &\text{s.t. } K = H^T R^{-1} H + \sum_{e_i \in \mathcal{E}} \gamma_{e_i} C_{e_i}\end{aligned} \tag{9}$$

where $\mathcal{E}$ is the set of edges of the graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, including self-edges (a self-edge is an edge from every node to itself), and C_{e_i} is the connectivity matrix associated with edge e_i. For example, if e_i is the edge connecting nodes 1 and 2, C_{e_i} is an $|\mathcal{V}| \times |\mathcal{V}|$ matrix whose only nonzero entries are ones at the $(1, 2)$ and $(2, 1)$ positions. $\{\gamma_{e_i}\}_{i=1}^{|\mathcal{E}|}$ are optimization variables.

Remark 1: Notice that the matrix K is set to satisfy an affinity structural constraint (it is an affine combination of all connectivity matrices of the graph $\mathcal{G}$). That the optimization domain is an affine set bears practical importance for the optimization procedure [19].

For effective optimization, we define

$$B \triangleq H^T R^{-1} H S H^T R^{-1} H + H^T R^{-1} H \tag{10a}$$

$$D \triangleq S H^T R^{-1} H. \tag{10b}$$

In terms of the matrices B, D, the optimization problem can be rewritten as

$$\min_K J \triangleq \operatorname{tr}\left[K^{-1} B K^{-1} - D K^{-1} - K^{-1} D^T + S)\right]$$

$$\text{s.t. } K = H^T R^{-1} H + \sum_{e_i \in \mathcal{E}}^{l} \gamma_{e_i} C_{e_i}. \tag{11}$$

C. Necessary Conditions for Optimality

Let K_{opt} be the optimum matrix. Because the optimization domain is affine, we can study the properties of this optimization problem on lines defined as $K_{\text{opt}} + kC_i$, where we denote $C_i = C_{e_i}$ and k is a free parameter. The conditions for optimality are

$$\frac{\partial J}{\partial k} = 0 \quad \forall C_i, \quad K = K_{\text{opt}} + kC_i. \tag{12}$$

In the sequel we derive an expression for the derivative.

For any $K = K_{\text{opt}} + kC_i$ and a symmetric matrix C_i we have

$$\frac{\partial J}{\partial k} = \frac{\partial \operatorname{tr}\left[K^{-1}BK^{-1}\right]}{\partial k} - 2\frac{\partial \operatorname{tr}\left[DK^{-1}\right]}{\partial k} \tag{13}$$

To compute the first term on the right-hand side (RHS) of (13) we use the matrix chain rule [20, Eq. (126)]

$$\frac{\partial \operatorname{tr}\left[K^{-1}BK^{-1}\right]}{\partial k} = \operatorname{tr}\left\{\left(\frac{\partial \operatorname{tr}\left[K^{-1}BK^{-1}\right]}{\partial K}\right)^T \frac{\partial K}{\partial k}\right\} \tag{14}$$

Now, [20, Eq. (114)]

$$\begin{aligned}\frac{\partial \operatorname{tr}\left[K^{-1}BK^{-1}\right]}{\partial K} & \\ &= -B^{-1}KK^{-1}BK^{-1}(I+I)K^{-1}BK^{-1} \\ &= -2K^{-2}BK^{-1},\end{aligned} \tag{15}$$

thus

$$\begin{aligned}\frac{\partial \operatorname{tr}\left[K^{-1}BK^{-1}\right]}{\partial k} &= \operatorname{tr}\left[-2\left(K^{-1}BK^{-2}\right)^T C_i\right] \\ &= \operatorname{tr}\left[-2K^{-2}BK^{-1}C_i\right].\end{aligned} \tag{16}$$

For the second term on the RHS of (13) we have

$$\frac{\partial \operatorname{tr}\left[DK^{-1}\right]}{\partial k} = \operatorname{tr}\left\{\left(\frac{\partial \operatorname{tr}\left[DK^{-1}\right]}{\partial K}\right)^T \frac{\partial K}{\partial k}\right\} \tag{17}$$

where [20, Eq. (113)]

$$\frac{\partial \operatorname{tr}\left[DK^{-1}\right]}{\partial K} = -K^{-1}D^T K^{-1}, \tag{18}$$

therefore

$$\frac{\partial \operatorname{tr}\left[DK^{-1}\right]}{\partial k} = \operatorname{tr}\left[-K^{-1}DK^{-1}C_i\right]. \tag{19}$$

Using both (16) and (19) in (13) finally yields

$$\begin{aligned}\frac{\partial J}{\partial k} &= 2\operatorname{tr}\left[K^{-1}(D - K^{-1}B)K^{-1}C_i\right], \\ K &= K_{\text{opt}} + kC_i.\end{aligned} \tag{20}$$

Remark 2: In the constrained case, treated in this paper, the derivative has to vanish just along the directions C_i from (11). On the other hand, in the unconstrained case, this derivative has to vanish along any direction, yielding the condition $D - K^{-1}B = 0$. Thus $K = BD^{-1} = H^T R^{-1} H + S^{-1}$, which is the familiar unconstrained optimal covariance (4b).

V. Non-Optimality of Dempster's Covariance

We prove that, in most cases, Dempster's covariance is not an optimal solution of the aforementioned optimization problem (hence its use in MMSE estimation problems is suboptimal, at best). Before formally stating and proving this result, however, we need the following two technical lemmas.

Lemma 2: Let P be a covariance matrix approximation of the original (dense) covariance matrix S, such that $P \neq S$. Denote by $\mathcal{M}$ the set of all real symmetric matrices in $\mathbb{R}^{m,m}$ with bounded entries, and define the set $\chi \subset \mathcal{M}$ as

$$\chi \triangleq \{R \in \mathcal{M} \mid \frac{\partial J}{\partial k} = 0\} \tag{21}$$

where, for all $R \in \chi$, the derivative $\frac{\partial J}{\partial k}$ is computed according to Eq. (20). Let $N = \frac{m(m+1)}{2}$ and notice that there exists a one-to-one correspondence between $\mathcal{M}$ and $\mathbb{R}^N$. Let λ denote the Lebesgue measure in $\mathbb{R}^N$ [21, page 176]. Then, $\lambda(\chi) = 0$.

Proof: We begin by examining the value of $\frac{\partial J}{\partial k}$ as a function of a single entry of the matrix R, when all other entries of R are fixed. Changing the derivative by varying the value of a single entry of R is equivalent to a change in $\frac{\partial J}{\partial k}$ along a single coordinate in $\mathbb{R}^N$.

Using the definitions of B and D from Eqs. (10), Eq. (20) can be rewritten as follows

$$\begin{aligned}\frac{\partial J}{\partial k} =& 2\operatorname{tr}\left[(H^T R^{-1} H + P^{-1})^{-1} C_i (H^T R^{-1} H + P^{-1})^{-1}\right. \\ & \times \left\{S H^T R^{-1} H - (H^T R^{-1} H + P^{-1})^{-1}\right. \\ & \left.\left.\times\ (H^T R^{-1} H + S^{-1}) S H^T R^{-1} H\right\}\right]\end{aligned} \tag{22}$$

for all $C_i, e_i \in \mathcal{E}$. Since both P and S are constant, $\frac{\partial J}{\partial k}$ is a rational function of each of the entries of the matrix R, when all other entries are fixed. This rational function is not constant, because $\frac{\partial J}{\partial k} \to 0$ when $R \to 0$, and $\frac{\partial J}{\partial k} \to \infty$ when $H^T R^{-1} H \to -P^{-1}$. Therefore, according to the fundamental theorem of algebra, it vanishes at a finite number of (isolated) points.

Now, because $\chi \subset \mathcal{M}$, it can be covered by a bounded interval in $\mathbb{R}^N$. This interval can be made to have an arbitrarily small size along one of the coordinates, because it is required to cover only the isolated points where $\frac{\partial J}{\partial k}$ vanishes. Thus, the size of the interval is arbitrarily small, rendering the Lebesgue measure of χ in $\mathbb{R}^N$ zero [22]. ■

Consider now the probability space $\{\mathcal{M}, \mathcal{F}, \mathbb{P}\}$, where $\mathcal{F}$ is the σ-algebra defined on $\mathcal{M}$ and $\mathbb{P}$ is an absolutely continuous probability measure defined on $\mathcal{F}$. Let $\mathcal{M}'$ denote the set of all positive definite matrices in $\mathcal{M}$, and notice that all elements of $\mathcal{M}'$ can function as measurement noise covariance matrices. The next lemma then addresses the probability of randomly selecting a measurement noise covariance matrix from $\mathcal{M}'$.

Lemma 3: Assuming that $\mathbb{P}$ is an absolutely continuous probability measure on $\mathcal{F}$, it can be defined such that $\mathbb{P}(\mathcal{M}') > 0$.

Proof: Since $\mathbb{P}$ is an absolutely continuous measure, it is dominated by the Lebesgue measure. Hence, we need to show that $\lambda(\mathcal{M}') > 0$. To that end, we study the interval that covers the vicinity of the identity matrix I, which is a single element of $\mathcal{M}'$. It can be easily shown, that when adding ϵ (in a symmetric way) to any entry of the matrix I, its non-trivial eigenvalues become $1 \pm \epsilon$. Hence, all symmetric matrices thus formed by changing I along one of the coordinates remain in $\mathcal{M}'$, as long as $|\epsilon| < 1$. Therefore, the size of the interval covering the vicinity of I along any coordinate is nonzero, rendering the Lebesgue measure of the set of positive definite matrices in the vicinity of the identity matrix strictly positive. ■

Having the two technical lemmas on hand, we now state the main result of this section.

Theorem 1: Let P be a covariance matrix approximation of the original (dense) covariance matrix S, such that $P \neq S$. Then, for any measurement noise covariance matrix $R \in \mathcal{M}'$ we have $\mathbb{P}(\frac{\partial J}{\partial k} = 0 \mid R \in \mathcal{M}') = 0$.

Proof: For any $R \in \mathcal{M}'$ the probability that $\frac{\partial J}{\partial k}$ vanishes can be expressed as

$$\mathbb{P}(\chi \mid \mathcal{M}') = \frac{\mathbb{P}(\chi \cap \mathcal{M}')}{\mathbb{P}(\mathcal{M}')} \tag{23}$$

Now, $\mathbb{P}(\chi \cap \mathcal{M}') < \mathbb{P}(\chi) = 0$ because $\lambda(\chi) = 0$ and $\mathbb{P}$ is absolutely continuous. Selecting $\mathbb{P}$ such that $\mathbb{P}(\mathcal{M}') > 0$, proven possible by Lemma 3, thus yields the theorem. ■

Remark 3: Theorem 1 holds for every approximation of the covariance matrix S, that does not explicitly take into account the covariance of the measurement noise. Thus, it shows that (subject to the theorem's assumptions) Dempster's classical approximation of S, in particular, does not constitute an optimal constrained covariance for any particular MMSE estimation problem (with a given measurement noise covariance), as it violates the necessary conditions for optimality with probability 1.

VI. Simulation Study

To illustrate the analytical results and demonstrate the estimation performance obtained using the optimal constrained covariance vs Dempster's covariance, we use the following example. The covariance S has ones along its diagonal and 0.9 as its off-diagonal entries. We test the 5, 8, and 14-dimensional cases. The estimation performance measure is taken to be the RMS criterion, which is the square root of the MSE criterion. The tested graph is a chain graph—a single thread connecting all vertices from 1 to n.

The RMS values are calculated by using (8) with either Dempster's covariance or the optimal constrained covariance, computed via numerically solving the optimization problem (11). For reference, the values corresponding to the true covariance, S, are also computed. All RMS values have been corroborated by a direct calculation using estimation error realizations obtained in a 100,000-run Monte Carlo (MC) simulation study with random measurement samples.

The results are presented in Table I. The two leftmost columns show the normalized RMS values obtained when using Dempster's covariance and the optimal constrained covariance, respectively. These values are computed via normalizing the RMS values corresponding to the constrained covariances by the RMS values corresponding to the true covariance, S (hence, values closer to 1 are better). For reference, the normalizing RMS values, corresponding to the true covariance, are shown in the rightmost column. As can be observed from Table I, the estimation quality benefits from using the optimal covariance. The improvement increases with problem dimension, from a 2% improvement for $|\mathcal{V}| = 5$, to a 14% improvement for $|\mathcal{V}| = 14$.

TABLE I

RMS estimation error using Dempster's covariance and the optimal constrained covariance.

$\lvert\mathcal{V}\rvert$	Dempster's	Optimal	True
	covariance	covariance	covariance
	(normalized)	(normalized)	(not normalized)
5	1.025	1.006	1.088
8	1.074	1.011	1.231
14	1.153	1.019	1.452

We next demonstrate that, as expected by theory, Dempster's covariance is not optimal for the estimation problem considered. We do this by showing that it violates the necessary conditions for optimality. Table II shows the norms of the gradients of the cost function J from (20), computed using both Dempster's covariance and the optimal constrained covariance. As can be observed from Table II, when Dempster's covariance is used, the gradient of the cost function clearly never vanishes, demonstrating that this covariance is not optimal (in the MMSE sense). In contradistinction, using the numerically obtained optimal constrained covariance, the corresponding cost gradient norm is practically zero in all tested cases.

Finally, Fig. 1 shows the dependence of the estimation error on the measurement noise covariance, which we take to be $RI_{|\mathcal{V}|}$ where $I_{|\mathcal{V}|}$ is the identity matrix of dimension $|\mathcal{V}|$. The figure compares the increment of RMS estimation error relative to values obtained with the true (unconstrained) covariance, of the optimal and Dempster's covariances. The true covariance, S, is identical to that used in the previous example, and we set $|\mathcal{V}| = 5$. As can be seen from Fig. 1, for low measurement noise, the information embedded in the measurements dominates the a priori information and, therefore, there is no difference between both methods. For larger measurement noise values, the optimal method approaches the results obtained with the true covariance, whereas the results obtained with Dempster's covariance are significantly worse.

TABLE II

Cost function gradient norm for Dempster's covariance and the optimal constrained covariance

$\lvert\mathcal{V}\rvert$	Dempster's	Optimal
	covariance	covariance
5	0.0426	1.28e-14
8	0.0297	1.02e-10
14	0.182	4.01e-10

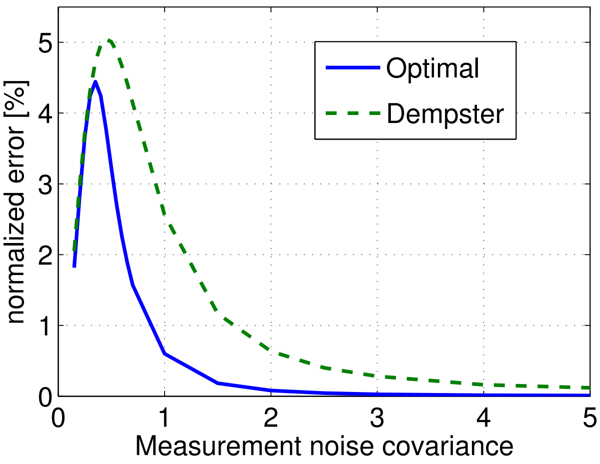

Fig. 1. Increment of RMS estimation error (relative to unconstrained covariance) vs measurement noise intensity. Solid line: optimal covariance, dashed line: Dempster's covariance.

VII. Conclusions

We have investigated the problem of covariance selection in graphical models that are used in distributed Bayesian estimation problems. This problem arises when trying to approximate the full a priori covariance, associated with the given Gaussian random field, by a constrained covariance, incorporating conditional independence constraints imposed on the problem to facilitate the use of Gaussian graphical models machinery. The importance of optimally reducing the a priori covariance in estimation applications involving graphical models stems from the fact that it allows a significant model reduction, thus enabling the use of algorithms suitable for specific graph structure, such as trees, while, at the same time, maintaining close-to-ideal estimation performance.

A well known covariance selection method, previously suggested by Dempster, finds the constrained covariance closest to the original (dense) covariance in the maximum likelihood sense, while complying with the imposed conditional independence structural constraints. However, this method is not geared for the problem at hand, because it does not address the Bayesian estimation goal of minimizing the mean squared estimation error criterion, nor does it take into account the effect of the measurement noise on the (a posteriori) estimation error covariance.

Explicitly addressing the MMSE estimation problem properties, we have formulated an optimization problem, distinctly different than the problem solved by Dempster, the solution of which provides the optimal constrained covariance that yields the minimal a posteriori estimation error covariance complying with the conditional independence constraints. In addition, we have proved that Dempster's method computes a constrained covariance that does not yield optimal performance in most estimation problems of the type addressed herein.

A simple numerical example is used to demonstrate the performance advantage of using the optimal constrained covariance, computed by solving the optimization problem formulated in this paper, over using Dempster's covariance. The example also illustrates the fact, formally proved herein, that Dempster's covariance is not optimal in a particular estimation problem, by numerically showing that this constrained covariance violates the necessary conditions for optimality.

References

[1] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, March 1972.

[2] A. Dawid and S. Lauritzen, "Hyper Markov laws in the statistical analysis of decomposable graphical models," *The annals of Statistics*, vol. 21, pp. 1272–1317, 1993.

[3] N. Wermuth and E. Scheidt, "Algorithm as 105: Fitting a covariance selection model to a matrix," *Applied Statistics*, vol. 26, no. 1, pp. 88–92, 1977. [Online]. Available: http://www.jstor.org/stable/2346883

[4] J. Huang, N. Liu, M. Pourahmadi, and L. Liu, "Covariance matrix selection and estimation via penalised normal likelihood," *Biometrika*, vol. 93, no. 1, pp. 85–98, 2006.

[5] D. Cox and N. Wermuth, "An approximation to maximum likelihood estimates in reduced models," *Biometrika*, vol. 77, no. 4, pp. 747–761, 1990.

[6] J. Dahl, V. Roychowdhury, and L. Vandenberghe, "Maximum likelihood estimation of gaussian graphical models: numerical implementation and topology selection," *UCLA preprint*, 2005.

[7] T. P. Speed and H. T. Kiiveri, "Gaussian markov distributions over finite graphs," *The annals of Statistics*, vol. 14, no. 1, pp. 138–150, Mar 1986.

[8] M. Drton and M. Perlman, "A SINful approach to Gaussian graphical model selection," *Journal of Statistical Planning and Inference*, vol. 138, no. 4, pp. 1179–1200, 2008.

[9] O. Banerjee, L. E. Ghaoui, A. d'Aspremont, and G. Natsoulis, "Convex optimization techniques for fitting sparse gaussian graphical models," in *ICML '06: Proceedings of the 23rd international conference on Machine learning*. New York, NY, USA: ACM, 2006, pp. 89–96.

[10] S. Chen and R. Gopinath, "Model Selection in Acoustic Modeling," in *Sixth European Conference on Speech Communication and Technology*. ISCA, 1999.

[11] H. Huang and N. Cressie, "Multiscale graphical modeling in space: applications to command and control," *Spatial Statistics: Methodological Aspects and Applications*, pp. 83–113, 2001.

[12] E. B. Sudderth, M. J. Wainwright, and A. S. Willsky, "Embedded trees: Estimation of gaussian processes on graphs with cycles," in *IEEE Transactions on Signal Processing*, vol. 52, no. 11, November 2004, pp. 3136–3150.

[13] K. H. Plarre and P. R. Kumar, "Extended message passing algorithm for inference in loopy gaussian graphical models," *Ad Hoc Networks*, vol. 2, pp. 153–169, 2004.

[14] V. Delouille, R. Neelamani, and R. Baraniuk, "Robust distributed estimation using the embedded subgraphs algorithm," *IEEE Transactions on Signal Processing*, vol. 54, no. 8, pp. 2998–3010, August 2006.

[15] V. Chandrasekaran, J. Johnson, and A. Willsky, "Estimation in gaussian graphical models using tractable subgraphs: A walk-sum analysis," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1916–1930, May 2008.

[16] Y. C. Eldar and A. V. Oppenheim, "Covariance shaping least-squares estimation," *IEEE Transactions on Signal Processing*, vol. 51, no. 3, pp. 686–697, 2003.

[17] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, 1988.

[18] S. L. Lauritzen, *Graphical Models*, ser. Oxford Statistical Science Series. UK: Clarendon Press, 1996, no. 17.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[20] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*. Technical University of Denmark, October 2008, version 20081110, http://www2.imm.dtu.dk/pubdb/p.php?3274.

[21] P. Billingsley, *Probability and Measure*, 2nd ed. John Wiley & sons, 1986.

[22] A. J.Weir, *Lebesgue integration and measure*. Cambridge University Press, 1973, vol. 1.