2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

An Efficient Algorithm for Dempster's Completion of Block–Circulant Covariance Matrices

Francesca P. Carli, Augusto Ferrante, Michele Pavon, and Giorgio Picci

Abstract— **The present paper deals with maximum entropy completion of partially specified banded block–circulant matrices. This problem has many applications in signal processing since circulants happen to be covariance matrices of stationary periodic processes and maximum entropy completion (i.e. the completion which has maximal determinant) is in fact maximum likelihood estimation subject to conditional independence constraints. Moreover, the maximal determinant completion has the meaning of covariance matrix of stationary reciprocal processes ([18], [20], [21]), a class of stochastic processes which extends Markov processes and is particularly useful for modeling signals indexed by space instead of time (think for example of an image). The maximum entropy completion problem for circulant matrices has been solved in [5] and some generalizations are brougth forth in [6]. The main contribution of this paper is an efficient algorithm for its solution.**

I. Introduction

In this paper (a shortened version of our journal paper [4]), we consider the problem of maximizing the determinant of a partially specified banded block–circulant matrix under the constraints that the completed matrix is positive definite and circulant (CMEP). This problem is in the framework of general covariance extension problems introduced by A. P. Dempster [11] and studied by Grone, Johnson, Sa and Wolkowicz [16] (see also [12] and [14] for the particularization to given data consistent with a banded Toeplitz structure and later extensions to generic banded algebras). Notice, however, that the linear constraint that enforces the circulant structure is not present in the Dempster setting. For the class of completion problems studied by Dempster (DMEP), it is well–known that the inverse of the maximum entropy completion has zeros in the positions corresponding to the unspecified entries in the original partially given covariance matrix, a property which, from now on, will be referred to as the *Dempster property*. A relevant fact is that, even in presence of the constraint that enforces the circulant structure, the inverse of the maximum entropy completion maintains the Dempster property. This fact, which has been first observed in [5] for given data on consecutive bands and proved in complete generality, i.e. for arbitrary missing entries in a circulant structure, in [6], is in fact non–trivial, since, for example, it does not hold true for arbitrary missing elements in a Toeplitz structure (in such a case the solution fails to satisfy the Dempster property unless the given data lie on consecutive bands centered along the main diagonal). Otherwise stated, it means that the solution of the CMEP and of the DMEP with circulant data (i.e. with data consistent with a block–circulant structure) coincide ([5], [6]). Because of the above stated properties, the maximum entropy completion becomes a crucial tool in solving the identification problem for stationary reciprocal processes on the unit circle. In fact, this kind of processes, suitable for describing random signals which live on a finite interval of the integer line (think for example of an image) are characterized by having a banded block–circulant concentration matrix whose blocks are the (matrix–valued) parameter of the model. A direct approach to model identification leads to strong nonlinearities and seems to be impracticable [5], while reformulating the problem as a maximum entropy completion problem, leads to a complete solution for which efficient algorithms may be designed. An efficient algorithm for solving the CMEP is in fact the main contribution of the present paper. Since the solutions of the CMEP and of the DMEP with circulant data coincide, all the methods available in the literature for the DMEP can, in principle, be employed. In particular, it is well–known that, if the graph associated with the specified entries is chordal ([15]), the solution of the DMEP can be expressed in closed form in terms of the principal minors of the sample covariance matrix (see [16], [1], [19], [13]). However the sparsity pattern associated with the given entries in our problem is not chordal. For non-chordal graphs the maximum entropy completion has to be computed iteratively. A straightforward application of standard optimization algorithms is too expensive for large sized problems, and several specialized algorithms have been proposed in the literature ([11], [23], [22], [8]) which deal with the general, very unstructured, setting of Dempster. In the present work, we propose a modified matricial gradient descent algorithm for the solution of the CMEP which naturally follows from the variational analysis in [5] and exploits in an essential way the circulant structure of our problem. This algorithm compares very well with the algorithms proposed in the literature for the solution of the DMEP.

Work partially supported by the Italian Ministry for Education and Resarch (MIUR) under PRIN grant "Identification and Adaptive Control of Industrial Systems".

F. Carli, A. Ferrante and G. Picci are with the Department of Information Engineering (DEI), University of Padova, via Gradenigo 6/B, 35131 Padova, Italy. carlifra@dei.unipd.it, augusto@dei.unipd.it, picci@dei.unipd.it

M. Pavon is with the Department of Pure and Applied Mathematics, University of Padova, pavon@math.unipd.it

II. Notation and preliminaries

Let $\mathbf{y}$ be a wide–sense stationary, periodic process of period N taking values in $\mathbb{R}^m$. This is equivalent to assume for its covariance matrix, say $\mathbf{\Sigma}_N$, a block–circulant symmetric

structure, i.e. $\mathbf{\Sigma}_N$ is the $N \times N$ m–block matrix

$$\mathbf{\Sigma}_N = \begin{bmatrix} \tilde{\Sigma}_0 & \tilde{\Sigma}_1^\top & \dots & \tilde{\Sigma}_\tau^\top & \dots & \tilde{\Sigma}_\tau & \dots & \tilde{\Sigma}_1 \\ \tilde{\Sigma}_1 & \tilde{\Sigma}_0 & \tilde{\Sigma}_1^\top & \ddots & \tilde{\Sigma}_\tau^\top & \dots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & & \ddots & & \tilde{\Sigma}_\tau \\ \tilde{\Sigma}_\tau & \dots & \tilde{\Sigma}_1 & \tilde{\Sigma}_0 & \tilde{\Sigma}_1^\top & \dots & \ddots & \\ \vdots & \tilde{\Sigma}_\tau & \dots & & \tilde{\Sigma}_0 & & \dots & \tilde{\Sigma}_\tau^\top \\ \tilde{\Sigma}_\tau^\top & & \ddots & & & & & \vdots \\ \vdots & \ddots & & \ddots & & \ddots & \ddots & \tilde{\Sigma}_1^\top \\ \tilde{\Sigma}_1^\top & \dots & \tilde{\Sigma}_\tau^\top & \dots & \tilde{\Sigma}_\tau & & \tilde{\Sigma}_1 & \tilde{\Sigma}_0 \end{bmatrix} \quad (1)$$

We refer the reader to [9] for an introduction to circulants; an extension of some relevant results for the block–case can be found, for example, in [6]. Here we just recall that the class of circulants is closed under sum, product, inverse and transpose. Moreover, all circulants commute and are simultaneously diagonalized by the Fourier matrix and thus have explicitly given eigenvalues and eigenvectors.

Hereafter, let $\mathfrak{S}_N$ denote the vector space of *symmetric* matrices with $N \times N$ square blocks of dimension $m \times m$. Moreover, let $\mathcal{I}_b$ be the set of pairs of indices consistent with a banded–symmetric block–circulant structure of bandwidth n, i.e. the set of the (i, j)'s such that

$$\begin{aligned} &\text{if } |i-j| \le mn \;\Rightarrow\; (i,j) \in \mathcal{I}_b \\ &\text{if } (i,j) \in \mathcal{I}_b \;\Rightarrow\; (j,i) \in \mathcal{I}_b \\ &\text{if } (i,j) \in \mathcal{I}_b \;\Rightarrow\; \Big((j+m)_{\text{mod mN}}, (i+m)_{\text{mod mN}}\Big) \in \mathcal{I}_b\,. \end{aligned}$$

We will denote by $\mathcal{I}_b^c$ the complement of $\mathcal{I}_b$, i.e. the set $\mathcal{I}_b^c = [0, \dots, N-1] \times [0, \dots, N-1] \backslash \mathcal{I}_b$. Finally, recall that the differential entropy $H(p)$ of a zero-mean Gaussian probability density function p with covariance matrix $\mathbf{\Sigma}_N$ is given by

$$H(p) = \frac{1}{2} \log(\det \mathbf{\Sigma}_N) + \frac{1}{2} n \left(1 + \log(2\pi)\right).$$

We are interested in the following *covariance extension problem for block–circulant matrices (CMEP)*

$$\max \{\det \mathbf{\Sigma}_N \mid \mathbf{\Sigma}_N \in \mathfrak{S}_N,\ \mathbf{\Sigma}_N > 0\} \quad (2a)$$

subject to :

$$e_i \mathbf{\Sigma}_N e_j^\top = r_{ij},\ \text{for } (i,j) \in \mathcal{I}_b \text{ and } r_{ij} \in \mathcal{R}_b \quad (2b)$$

$$\mathbf{\Sigma}_N \text{ is block–circulant} \quad (2c)$$

where e_k is the zero row–vector with a 1 in the k–th position, $e_k = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}$, and $\mathcal{R}_b$ denotes the set of the given data consistent with a banded-symmetric block-circulant structure, i.e.

$$\mathcal{R}_b := \big\{ r_{ij} \in \mathbb{R} \,|(i,j) \in \mathcal{I}_b, r_{ij} = r_{ji}, \; r_{ij} = r_{(j+m)_{\text{mod mN}}(i+m)_{\text{mod mN}}} \big\}.$$

(In our case the r_{ij}'s represent the given entries of a partially specified (block–circulant) sample covariance matrix, say $\mathbf{R}_N$).

If we remove the constraint (2c), we get the *covariance selection problem* studied by A. P. Dempster, which sometimes will be also referred to as the DMEP.

The first question to be addressed is *feasibility* of the CMEP. In [5] a sufficient condition on the data for the existence of a positive definite block–circulant completion has been provided while a characterization of the set of all positive definite completions of a partially specified block–circulant matrix has been derived in [6]. A necessary and sufficient condition for the feasibility of the CMEP for unitary bandwidth and block–size is provided in a forthcoming, more complete version of this paper [4].

III. Algorithms for the covariance selection problem

In this Section, we briefly discuss some of the algorithms proposed in the literature for the covariance selection problem. To this aim, it seems natural to describe the pattern of the specified entries of an $Nm \times Nm$ partial symmetric matrix $M = (m_{ij})$ by an undirected graph of Nm vertices which has an edge joining vertex i and vertex j if and only if the entry m_{ij} is specified. Since the diagonal entries are all assumed to be specified, we ignore loops at the vertices. The undirected graph will be denoted by $\mathcal{G} = (V, E)$ where V is the vertex set and E is the edge set which consists of unordered pairs of distinct vertices. We assume that the reader is familiar with some basic notions of graph theory and just recall some definitions which will be useful throughout this Section (a standard reference is [15]). In particular, we say that an undirected graph is *chordal* if every cycle of length greater than three has a chord, i.e. an edge joining two nonconsecutive vertices of the cycle. Moreover, a *clique* of $\mathcal{G}$ is any maximal subset of vertices that is complete in $\mathcal{G}$ (i.e. such that every pair of vertices is adjacent). Finally, we define the complementary graph of $\mathcal{G} = (V, E)$ as the graph $\tilde{\mathcal{G}}$ with vertex set V and edge set $\tilde{E}$ with the property that $(u, v) \in \tilde{E}$ if and only if $u \neq v$ and $(u, v) \notin E$.

If the graph of the specified entries is *chordal*, the covariance selection problem admits a closed form solution in terms of the principal minors of the sample covariance matrix (see [16], [1], [19], [13]). However, the graph associated with a banded circulant sparsity pattern is never chordal. We have therefore to resort to *iterative algorithms*. For the applications we have in mind, we are dealing with vector–valued processes possibly defined on a quite large interval. A straightforward application of standard optimization algorithms is too expensive for problems of such a size, and several specialized algorithms have been proposed in the literature ([11], [23], [22], [8]) which deals with the general, very unstructured, setting of the DMEP. In his early work ([11]), Dempster himself proposed two iterative algorithms which however are very demanding from a computational point of view. Two popular methods are those proposed by T. P. Speed and H. T. Kiiveri in [22], which we now briefly discuss.

a) First algorithm: As mentioned in the Introduction, for the class of problems studied by Dempster, the inverse of

the unique completion which maximizes the entropy functional has the property to be zero in the complementary positions of those fixed in $\mathbf{\Sigma}_N$. Thus, a rather natural procedure to compute the solution of the covariance selection problem seems to be the following: iterate maintaing the elements of $\mathbf{\Sigma}_N$ in $\mathcal{I}_b$ at the desired value (i.e. equal to the corresponding elements in the sample covariance matrix) while forcing the elements of $\mathbf{\Sigma}_N^{-1}$ in $\mathcal{I}_b^c$ to zero. In particular, the entries of $\mathbf{\Sigma}_N^{-1}$ in $\mathcal{I}_b^c$ are forced to zero $|\tilde{c}_t|$ at time, where $|\cdot|$ denotes the cardinality of the t–th clique, $\tilde{c}_t$, in the complementary graph $\tilde{\mathcal{G}}$. The algorithm reads as follows.

Algorithm 1 First algorithm (Speed and Kiiveri [22])

Compute all the cliques $\tilde{c}_t$ in the complementary graph $\tilde{\mathcal{G}}$
Initialize $\mathbf{\Sigma}_N^{(0)} = \mathbf{R}_N$;
while some stopping criterion is satisfyied **do**
  for all the cliques $\tilde{c}_t$ in $\tilde{\mathcal{G}}$ **do**

$$\mathbf{\Sigma}_N^{(t)} = \mathbf{\Sigma}_N^{(t-1)} + \phi\left(\mathbf{\Sigma}_N^{(t-1)}\right)$$

  end for
end while

where $\phi\left(\mathbf{\Sigma}_N^{(t-1)}\right)$ is the $mN \times mN$ zero matrix which equals

$$\left\{\text{diag}\left[\left((\mathbf{\Sigma}_N^{(t-1)})^{-1}\right)_{\tilde{c}_t}\right]^{-1}\right\}^{-1} - \left[\left((\mathbf{\Sigma}_N^{(t-1)})^{-1}\right)_{\tilde{c}_t}\right]^{-1}$$

in the positions corresponding to the current clique $\tilde{c}_t$ (given a $Nm \times Nm$ matrix M and a set $a \subset [1, \ldots, Nm]$, M_a denotes the submatrix with entries $\{m_{ij} : i, j \in a\}$). In this first algorithm, every cycle consists of as many steps as the cliques in the complementary graph $\tilde{\mathcal{G}}$ (the graph associated to the elements in $\mathcal{I}_b^c$). At each step, only the elements in $\mathbf{\Sigma}_N$ corresponding to the current clique $\tilde{c}_t$ (i.e. only a subset of the entries in $\mathcal{I}_b^c$) are modified in such a way to set the elements of $\mathbf{\Sigma}_N^{-1}$ in the corresponding positions to the desired zero–value. Through the iterations the elements in $\mathbf{\Sigma}_N^{(t)}$ are fixed over $\mathcal{I}_b$ while the elements of $\left(\mathbf{\Sigma}_N^{(t)}\right)^{-1}$ vary over $\mathcal{I}_b^c$.

b) Second algorithm: The role of $\mathbf{\Sigma}_N$ and $\mathbf{\Sigma}_N^{-1}$ can also be swapped, yielding a second iterative procedure, which is the analogous of iterative proportional scaling (IPS) for contingency tables [17]. If we denote by $\varphi\left(\mathbf{\Sigma}_N^{(t-1)}\right)$ the $mN \times mN$ zero matrix which equals

$$((\mathbf{R}_N)_{c_t})^{-1} - \left(\left(\mathbf{\Sigma}_N^{(t-1)}\right)_{c_t}\right)^{-1}$$

in the positions corresponding to the current clique c_t in $\mathcal{G}$ (the graph associated with the given entries), the procedure can be formally described as in Algorithm 2. Every cycle consists of as many steps as the cliques in $\mathcal{G}$. At each step, only the elements in $\mathbf{\Sigma}_N^{-1}$ corresponding to the current clique c_t (i.e. only a subset of the entries in $\mathcal{I}_b$) are modified in such a way to set the elements of $\mathbf{\Sigma}_N$ in the corresponding positions to the desired value, namely equal to the sample covariance $\mathbf{R}_N$. Through the iterations, the elements in $\left(\mathbf{\Sigma}_N^{(t)}\right)^{-1}$ are fixed over $\mathcal{I}_b^c$, while the elements of $\mathbf{\Sigma}_N^{(t)}$ vary over $\mathcal{I}_b$.

Algorithm 2 Second algorithm (Speed and Kiiveri [22])

Compute all the cliques c_t in $\mathcal{G}$
Initialize $\mathbf{\Sigma}_N^{(0)} = I_{Nm}$;
while some stopping criterion is satisfyied **do**
  for all the cliques c_t in $\mathcal{G}$ **do**

$$\left(\mathbf{\Sigma}_N^{(t)}\right)^{-1} = \left(\mathbf{\Sigma}_N^{(t-1)}\right)^{-1} + \varphi\left(\mathbf{\Sigma}_N^{(t-1)}\right)$$

  end for
end while

The crucial step in both algorithms involves going from $\mathbf{\Sigma}_N^{(t-1)}$ (respectively, from $\left(\mathbf{\Sigma}_N^{(t-1)}\right)^{-1}$) to $\mathbf{\Sigma}_N^{(t)}$ ($\left(\mathbf{\Sigma}_N^{(t)}\right)^{-1}$). We refer the reader to [22] for more details about this and for a convergence proof. An intuitive justification is also possible if one recognizes that the two procedures can be interpreted as a sequence of $\mathcal{I}$–projection ([7]) on suitable "subspaces" in the spirit of the von Neumann's alternating projection theorem.

c) Comparison between the two algorithms: The choice of which algorithm is to be preferred in any application is very much dependent on the number and size of the cliques in $\mathcal{G}$ and $\tilde{\mathcal{G}}$. In our setting, the complexity of the graph associated with the given entries depends on the bandwidth mn. In particular, for a bandwidth not too large with respect to the completion size (which is the case we are interested in) the complexity of the graph associated with the given data $\mathcal{G}$ is far lower than the complexity of its complementary (which, for small bandwidth, is almost complete). It turns out that, for small bandwidth, the second algorithm (which, from now on, will be referred to as IPS) runs faster than the first, and thus has to be preferred (see Table I for a numerical comparison of the execution times of the two algorithms for $N = 30$ and bandwidth varying between 2 and 8).

d) Covariance selection via chordal embedding: In [8], Dahl, Vanderberghe and Roychowdhury propose a new technique to improve the efficiency of the Newton's method for the covariance selection problem based on chordal embedding: the given sparsity pattern is embedded in a chordal one for which they provide efficient techniques for computing the gradient and the Hessian. The complexity of the method is dominated by the cost of forming and solving a system of linear equations in which the number of unknowns depends on the number of nonzero entries added in the chordal embedding. For a circulant sparsity pattern, it is easy to check that the number of nonzero elements added in the chordal embedding is quite large. Hence the method does not seem to be effective. A preconditioned conjugate gradient method has also been proposed in ([8]), but the comparison is not brought forth in the present paper.

n	First algorithm # of cl. (max. cl. sz.)	First algorithm CPU time [s]	Second algorithm # of cl. (max. cl. sz.)	Second algorithm CPU time [s]
2	4608(10)	9.7877	30(3)	0.4109
3	2406(7)	4.1515	30(4)	0.1783
4	1241(6)	1.9419	30(5)	0.3153
5	706(5)	1.0525	30(6)	0.5535
6	445(4)	0.6258	30(7)	0.9854
7	295(3)	0.4145	30(8)	1.7477
8	175(3)	0.2480	30(9)	3.0665

TABLE I: Execution time of the first and second algorithm for $N = 30$, $m = 1$, bandwidth $n = \{2, \dots, 8\}$.

IV. Matricial Gradient Descent Algorithm

Let $\mathbf{U}_N$ denote the block–circulant "shift" matrix with $N \times N$ blocks, whose first block–row is given by $[0\,, I_m\,, 0\,, \dots\,, 0]$, $\mathbf{T}_n \in \mathfrak{S}_{n+1}$ the block–Toeplitz matrix of *boundary data* whose first block–row is given by $\left[\Sigma_0,\, \Sigma_1^\top,\, \dots,\, \Sigma_n^\top\right]$ and E_n the $N \times (n+1)$ block matrix given by $E_n = \left[I_{n+1}, \mathbf{0}_{n+1,N-n-1}\right]^\top$, where $\mathbf{0}_{h,k}$ denotes the $h \times k$ zero–matrix. Following [5], the maximum entropy problem for banded block-circulant matrices (CMEP) can be written as

$$\max\left\{\log\det \mathbf{\Sigma}_N \mid \mathbf{\Sigma}_N \in \mathfrak{S}_N,\ \mathbf{\Sigma}_N > 0\right\} \tag{3a}$$

$$\text{subject to :}$$

$$E_n^\top \mathbf{\Sigma}_N E_n = \mathbf{T}_n, \tag{3b}$$

$$\mathbf{U}_N^\top \mathbf{\Sigma}_N \mathbf{U}_N = \mathbf{\Sigma}_N. \tag{3c}$$

where we have exploited the invariance of block–circulants under the similarity $\mathbf{C}_N \mapsto \mathbf{U}_N^\top \mathbf{C}_N \mathbf{U}_N$. Problem (3) is a *convex optimization problem* since we are minimizing a strictly convex function on the intersection of a convex cone (minus the zero matrix) with a linear manifold. Hence, we are dealing with a *convex optimization problem*. We shall solve this problem by resorting to *duality theory*. To this aim, consider the linear map

$$\begin{array}{rccl} A\ : & \mathfrak{S}_{n+1} \times \mathfrak{S}_N & \to & \mathfrak{S}_N \\ & (\Lambda, \Theta) & \mapsto & E_n \Lambda E_n^\top + \mathbf{U}_N \Theta \mathbf{U}_N^\top - \Theta \end{array}$$

and define the set

$$\begin{aligned} \mathcal{L}_+ :=& \{(\Lambda, \Theta) \in (\mathfrak{S}_{n+1} \times \mathfrak{S}_N) \mid (\Lambda, \Theta) \in (\ker(A))^\perp, \\ & \left(E_n \Lambda E_n^\top + \mathbf{U}_N \Theta \mathbf{U}_N^\top - \Theta\right) > 0\}. \end{aligned} \tag{4}$$

$\mathcal{L}_+$ is an open, convex subset of $(\ker(A))^\perp$. The *Lagrangian function* is given by

$$\begin{aligned} L(\mathbf{\Sigma}_N, \Lambda, \Theta) \ =& -\operatorname{tr}\log \mathbf{\Sigma}_N + \operatorname{tr}\left(E_n \Lambda E_n^\top \mathbf{\Sigma}_N\right) - \operatorname{tr}\left(\Lambda \mathbf{T}_n\right) \\ & + \operatorname{tr}\left(\mathbf{U}_N \Theta \mathbf{U}_N^\top \mathbf{\Sigma}_N\right) - \operatorname{tr}\left(\Theta \mathbf{\Sigma}_N\right) \end{aligned}$$

and its first variation (at $\mathbf{\Sigma}_N$ in direction $\delta\mathbf{\Sigma}_N \in \mathfrak{S}_N$) is

$$\begin{aligned} \delta L(\mathbf{\Sigma}_N, \Lambda, \Theta; \delta\mathbf{\Sigma}_N) =& -\operatorname{tr}\left(\mathbf{\Sigma}_N^{-1} \delta\mathbf{\Sigma}_N\right) + \operatorname{tr}\left(E_n \Lambda E_n^\top \delta\mathbf{\Sigma}_N\right) \\ & + \operatorname{tr}\left(\left(\mathbf{U}_N \Theta \mathbf{U}_N^\top - \Theta\right) \delta\mathbf{\Sigma}_N\right). \end{aligned}$$

Thus $\delta L(\mathbf{\Sigma}_N, \Lambda, \Theta; \delta\mathbf{\Sigma}_N) = 0,\ \forall \delta\mathbf{\Sigma}_N \in \mathfrak{S}_N$ if and only if

$$\mathbf{\Sigma}_N^{-1} = E_n \Lambda E_n^\top + \mathbf{U}_N \Theta \mathbf{U}_N^\top - \Theta.$$

It follows that, for each fixed pair $(\Lambda, \Theta) \in \mathcal{L}_+$, the unique $\mathbf{\Sigma}_N^o$ minimizing the Lagrangian over $\mathfrak{S}_{N,+} := \{\mathbf{\Sigma}_N \in \mathfrak{S}_N,\ \mathbf{\Sigma}_N > 0\}$ is

$$\mathbf{\Sigma}_N^o = \left(E_n \Lambda E_n^\top + \mathbf{U}_N \Theta \mathbf{U}_N^\top - \Theta\right)^{-1}. \tag{5}$$

Moreover, the Lagrangian computed for $\mathbf{\Sigma}_N = \mathbf{\Sigma}_N^o$ results

$$\begin{aligned} L(\mathbf{\Sigma}_N^o, \Lambda, \Theta) =& \operatorname{tr}\log\left(E_n \Lambda E_n^\top + \mathbf{U}_N \Theta \mathbf{U}_N^\top - \Theta\right) \\ & + \operatorname{tr} I_{mN} - \operatorname{tr}\left(\Lambda \mathbf{T}_n\right). \end{aligned}$$

This is a strictly concave function on $\mathcal{L}_+$ whose maximization is the *dual problem* of (CMEP). We can equivalently consider the convex problem

$$\min\left\{J(\Lambda, \Theta), (\Lambda, \Theta) \in \mathcal{L}_+\right\}, \tag{6}$$

where J is given by

$$J(\Lambda, \Theta) = \operatorname{tr}\left(\Lambda \mathbf{T}_n\right) - \operatorname{tr}\log\left(E_n \Lambda E_n^\top + \mathbf{U}_N \Theta \mathbf{U}_N^\top - \Theta\right).$$

It can be shown ([5, Theorem 6.1]) that the function J admits a unique minimum point $\left(\bar\Lambda, \bar\Theta\right)$ in $\mathcal{L}_+$.

In this section we propose a *modified gradient descent algorithm with backtracking line search* (see, e.g., [2, Ch. 9]) for the numerical solution of the dual problem (6). This task requires some care because we are working in a *matricial space*. Let $\pi_{\mathfrak{C}_N}$ denote the orthogonal projection onto the linear subspace of symmetric, block-circulant matrices $\mathfrak{C}_N$. Before proceeding, we need two preliminary lemmas. We refer the reader to [5] for the proof of these statements.

Lemma 4.1: Let $\Lambda \in \mathfrak{S}_{n+1}$ be the matrix

$$\Lambda = \begin{bmatrix} \Lambda_{00} & \Lambda_{01} & \dots & \Lambda_{0n} \\ \Lambda_{01}^\top & \Lambda_{11} & \dots & \Lambda_{1n} \\ \vdots & & \ddots & \vdots \\ \Lambda_{0n}^\top & \Lambda_{1n}^\top & \dots & \Lambda_{nn} \end{bmatrix}.$$

The orthogonal projection of $E_n \Lambda E_n^\top$ onto $\mathfrak{C}_N$, say Π_Λ, is given by

$$\Pi_\Lambda := \pi_{\mathfrak{C}_N}\left(E_n \bar\Lambda E_n^\top\right) = \begin{bmatrix} \Pi_0 & \Pi_1^\top & \Pi_2^\top & \dots & \Pi_1 \\ \Pi_1 & \Pi_0 & \Pi_1^\top & \dots & \Pi_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \Pi_2^\top & \dots & \Pi_1 & \Pi_0 & \Pi_1^\top \\ \Pi_1^\top & \Pi_2^\top & \dots & \Pi_1 & \Pi_0 \end{bmatrix}$$

with

$$\Pi_0 = \frac{1}{N}\left(\Lambda_{00} + \Lambda_{11} + \dots + \Lambda_{nn}\right), \tag{7a}$$

$$\Pi_1 = \frac{1}{N}\left(\Lambda_{01} + \Lambda_{12} + \dots + \Lambda_{n-1,n}\right)^\top, \tag{7b}$$

$$\vdots$$

$$\Pi_n = \frac{1}{N}\Lambda_{0n}^\top, \tag{7c}$$

while $\Pi_i = 0$, forall i in the interval $n+1 \le i \le N-n-1$.

Lemma 4.2: Let $M \in \mathfrak{S}_N$. Then $M \in (\mathfrak{C}_N)^\perp$ if and only if $\exists\, S \in \mathfrak{S}_N$ s.t. $M = \mathbf{U}_N S \mathbf{U}_N^\top - S$.

Consider the functional

$$\bar J(\Lambda) := \operatorname{tr}\left(\Lambda \mathbf{T}_n\right) - \operatorname{tr}\log\left\{\pi_{\mathfrak{C}_N}\left(E_n \Lambda E_n^\top\right)\right\} \tag{8}$$

whose gradient $\nabla_\Lambda \bar{J}(\Lambda)$ is given by

$$\nabla_\Lambda \bar{J}(\Lambda) = -E_n^\top \left[\pi_{\mathfrak{C}_N}\left(E_n \Lambda E_n^\top\right)\right]^{-1} E_n + \mathbf{T}_n \,.$$

The proposed algorithm is as follows.

Algorithm 3 Matricial gradient descent algorithm

Given a starting point $\Lambda \in \text{dom}\, \bar{J}$, $\alpha \in (0, 0.5)$, $\beta \in (0,1)$
while $\left\|\nabla_\Lambda \bar{J}(\Lambda)\right\|_2 > \eta$ **do**
  $\Delta\Lambda := -\nabla_\Lambda \bar{J}(\Lambda)$
  $t := 1$
  while $\bar{J}(\Lambda + t\Delta\Lambda) > \bar{J}(\Lambda) + \alpha t \,\text{tr}\left\{\nabla_\Lambda \bar{J}(\Lambda)^\top \Delta\Lambda\right\}$ **do**
    $t := \beta t$
  end while
  $\Lambda := \Lambda + t\Delta\Lambda$
end while

Theorem 4.1: Algorithm 3 is a gradient descent algorithm restricted to the subspace

$$\left\{(\Lambda, \Theta) \mid \pi_{\mathfrak{C}_N^\perp}\left(E_n \Lambda E_n^\top\right) = -\left(U_N \Theta U_N^\top - \Theta\right)\right\}. \quad (9)$$

Proof: Let $(\bar{\Lambda}, \bar{\Theta})$ be the unique minimum point of the functional J on $\mathcal{L}_+$. We know that $(\bar{\Lambda}, \bar{\Theta})$ are such that $\Sigma^o = E_n \bar{\Lambda} E_n^\top + U_N \bar{\Theta} U_N^\top - \bar{\Theta}$ is circulant. Thus one can think of restricting the search for the solution of the optimization problem to the set

$$\left\{(\Lambda, \Theta) \mid \left(E_n \Lambda E_n^\top + U_N \Theta U_N^\top - \Theta\right) \text{ is circulant}\right\}$$

i.e. to the set

$$\left\{(\Lambda, \Theta) \mid \pi_{\mathfrak{C}_N^\perp}\left(E_n \Lambda E_n^\top + U_N \Theta U_N^\top - \Theta\right) = 0\right\}.$$

Since $\left(U_N \Theta U_N^\top - \Theta\right) \in \mathfrak{C}_N^\perp$ (see Lemma 4.2), the latter can be written as

$$\left\{(\Lambda, \Theta) \mid \pi_{\mathfrak{C}_N^\perp}\left(E_n \Lambda E_n^\top\right) = -\left(U_N \Theta U_N^\top - \Theta\right)\right\}.$$

If we compute the dual function J on the set (9) we obtain

$$\begin{aligned} & J(\Lambda, \Theta)\,|_{\left\{(\Lambda,\Theta)\,|\,\pi_{\mathfrak{C}_N^\perp}(E_n \Lambda E_n^\top) = -\left(U_N \Theta U_N^\top - \Theta\right)\right\}} \\ &= \text{tr}\,(\Lambda T_n) - \text{tr}\log\left(E_n \Lambda E_n^\top + U_N \Theta U_N^\top - \Theta^\top\right) \\ &= \text{tr}\,(\Lambda T_n) - \text{tr}\log\left(E_n \Lambda E_n^\top - \pi_{\mathfrak{C}_N^\perp}\left(E_n \Lambda E_n^\top\right)\right) \\ &= \text{tr}\,(\Lambda T_n) - \text{tr}\log\left(\pi_{\mathfrak{C}_N}\left(E_n \Lambda E_n^\top\right)\right) \end{aligned}$$

which is the modified functional defined above. Thus the proposed algorithm is nothing but a gradient descent algorithm in which the search of the minimum point has been restricted to the subspace where the optimal solution is known to be, i.e. to the subspace (9). ∎

A. Initialization

The following Theorem is useful to provide a good starting point for the iterative procedure 3. We refer the reader to [4] for a complete proof of the statement.

Theorem 4.2: Let $\mathbf{T}_n$, the block–Toeplitz matrix of boundary data, be positive definite and let $\{\hat{\Sigma}_k, k = 0, 1, 2, \ldots\}$ with $\hat{\Sigma}_k = \Sigma_k$, $k = 0, 1, 2, \ldots, n$ be the maximum entropy (positive) extension of $\{\Sigma_0, \Sigma_1, \ldots, \Sigma_n\}$. Then, for N large enough the block–circulant matrix $\mathbf{\Sigma}_N^{(c)}$ given by

$$\begin{aligned} \text{Toepl}\big(&\hat{\Sigma}_0, \hat{\Sigma}_1^\top, \ldots, \hat{\Sigma}_n^\top, \hat{\Sigma}_{n+1}^\top, \ldots, \hat{\Sigma}_{\frac{N}{2}-1}^\top, \hat{\Sigma}_{\frac{N}{2}}^\top + \hat{\Sigma}_{\frac{N}{2}}, \\ &\hat{\Sigma}_{\frac{N}{2}-1}, \ldots, \hat{\Sigma}_{n+1}, \hat{\Sigma}_n, \ldots \hat{\Sigma}_1\big), \end{aligned}$$

for N even, and by

$$\begin{aligned} \text{Toepl}\big(&\hat{\Sigma}_0, \hat{\Sigma}_1^\top, \ldots, \hat{\Sigma}_n^\top, \hat{\Sigma}_{n+1}^\top, \ldots, \hat{\Sigma}_{\frac{N-1}{2}}^\top, \\ &\hat{\Sigma}_{\frac{N-1}{2}}, \ldots, \hat{\Sigma}_{n+1}, \hat{\Sigma}_n, \ldots, \hat{\Sigma}_1\big), \end{aligned}$$

for N odd, is a covariance matrix which for $N \to \infty$ is arbitrarily close to the $mN \times mN$ maximum entropy block-circulant extension of $\mathbf{T}_n$.

A possible way to compute the maximum entropy completion of a partially specified block–*Toeplitz* matrix is the following. The maximum entropy spectrum is given by

$$\Phi_{ME} = \left[L_n(z^{-1})\right]^{-1} \Lambda_n \left[L_n(z^{-1})\right]^{-*}$$

where $L(z^{-1})$ is the n–th Levinson–Whittle matrix polynomial of the block–Toeplitz matrix $\mathbf{T}_n$

$$L(z^{-1}) = \sum_{k=0}^{n} A_n(k) z^{-k}$$

with the $A_n(k)$'s and $\Lambda_n = \Lambda_n^\top > 0$ solutions of the Yule–Walker type equation

$$\begin{bmatrix} A_n(0) & A_n(1) & \ldots & A_n(n) \end{bmatrix} \mathbf{T}_n^\top = \begin{bmatrix} \Lambda_n & 0 & \ldots & 0 \end{bmatrix},$$

see [24], [10] and [25]. It follows that the spectral factor $W(z) := \left[L_n(z^{-1})\right]^{-1} \Lambda_n^{\frac{1}{2}}$ is given by

$$W(z) = C(zI - A)^{-1} B + D$$

with $D = \Lambda_n^{\frac{1}{2}}$,

$$C = -\begin{bmatrix} A_n(n) & A_n(n-1) & \ldots & \ldots & A_n(1) \end{bmatrix}$$

$$A = \begin{bmatrix} 0 & I_m & 0 & 0 & \ldots & 0 \\ 0 & 0 & I_m & 0 & \ldots & 0 \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & 0 \\ 0 & \ldots & \ldots & \ldots & 0 & I_m \\ -A_n(n) & -A_n(n-1) & \ldots & \ldots & \ldots & -A_n(1) \end{bmatrix},$$

and $B^\top = \begin{bmatrix} 0 & 0 & \ldots & 0 & \left(\Lambda_n^{\frac{1}{2}}\right)^\top \end{bmatrix}$. The positive real part of the maximum entropy spectrum is

$$\Phi_{ME,+}(z) = C(zI - A)^{-1} \bar{C}^\top + \frac{1}{2}\Sigma_0$$

where $\bar{C}^\top = APC^\top + BD^\top$, with $P = APA^\top + BB^\top$ and the maximum entropy covariance extension results

$$\hat{\Sigma}_k = CA^{k-1}\bar{C}^\top, \qquad k > n.$$

With this extension in hand, we can compute an approximation for the maximum entropy block-circulant extension as suggested by Theorem 4.2. A good starting point for our gradient descent algorithm can then be obtained from (7) assuming for Λ a Toeplitz structure.

		GDI		GDT	
N	m	# of itz.	CPU time	# of itz.	CPU time
10	5	99	0.1455	61	0.0767
20	5	212	0.4143	65	0.1270
30	5	322	0.8355	97	0.2504
40	5	432	1.4233	130	0.4285
50	5	541	2.1937	163	0.6603

TABLE II: CPU time [in sec.] for the matricial gradient descent algorithm with different initializations (identity on the left and as in Section IV-A on the right). The reported times have been computed for $n = 3$ and $m = 5$.

	$m = 5$		$m = 10$	
N	IPS	GD	IPS	GD
10	4.7048	0.0767	69.7671	0.1516
20	16.4981	0.1270	307.9596	0.4459
30	29.2779	0.2504	597.3791	0.8988
40	43.8072	0.4285	924.6431	1.4798
50	63.8069	0.6603	1341.0976	2.2052

TABLE III: Matricial gradient descent algorithm vs. iterative proportional scaling: CPU time [in sec.] for $N = [10, 20, 30, 40, 50]$, $n = 3$, $m = 5$ and $m = 10$.

V. Numerical experiments

The matricial gradient descent algorithm has been implemented in Matlab. The implementation exploits the block–circulant symmetric structure (see [9] for efficient implementation of the inverse of a block–circulant). At each iteration, the algorithm requires the inversion of $\lceil \frac{N+1}{2} \rceil$ matrices of order m. It follows that the execution time increases as the completion size N and the block size m increase. Moreover, it also increases, even to a lesser amount, for increasing bandwidth n. Table II presents a comparison between the execution times for different starting point Λ_0. The gradient descent algorithm has been initialized to the normalized identity and by the procedure of Section IV-A. The proposed initialization acts effectively to reduce the number of iterations (and thus the computational time) to reach the minimum.

Finally, the gradient descent algorithm (GD) has been compared to the iterative proportional scaling procedure (IPS) by Speed and Kiiveri. The Bron–Kerbosch algorithm [3] has been employed for finding the cliques in the graph for IPS. The execution times for different completion size N and block size m are reported in Table III. It can be seen that the gradient descent algorithm runs faster than the iterative proportional scaling and the gap between the two increases as N increases. Moreover, the gap becomes much more evident as m grows, making the gradient descent algorithm more attractive for applications where the process under observation is vector–valued ($m > 1$).

VI. Conclusions

The main contribution of the present paper is an efficient algorithm to solve the maximum entropy band extension problem for block–circulant matrices. This problem has many applications in signal processing since it arises in connection with maximum likelihood estimation of periodic, and in particular quasi–Markov (or reciprocal), processes. Even if matrix completion problems have gained considerable attention in the past (think for example to the covariance extension problem for stationary processes on the integer line, i.e. for Toeplitz matrices), the maximum entropy band extension problem for block–circulant matrices has been addressed for the first time in [5]. The proposed algorithm heavily exploits the circulant structure and relies on the variational analysis brought forth in [5].

References

[1] W.W. Barrett, C. R. Johnson, and M. Lundquist. Determinantal formulation for matrix completions associated with chordal graphs. *Linear Algebra and Applications*, 121:265–289, 1989.

[2] S. Boyd and L. Vanderberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] C. Bron and J. Kerbosch. Algorithm 475: finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, 1973.

[4] F. P. Carli, A. Ferrante, M. Pavon, and G. Picci. An efficient algorithm for maximum–entropy extension of block–circulant covariance matrices. *Submitted*.

[5] F. P. Carli, A. Ferrante, M. Pavon, and G. Picci. A maximum entropy solution of the covariance extension problem for reciprocal processes. *IEEE Transactions on Automatic Control*, to appear, 2011.

[6] F. P. Carli and T. T. Georgiou. On the covariance completion problem under a circulant structure. *IEEE Transactions on Automatic Control*, 56(4):918–922, 2011.

[7] I. Csiszar. I–divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

[8] J. Dahl, L. Vanderberghe, and V. Roychowdhury. Covariance selection for non–chordal graphs via chordal embedding. *Optimization Methods and Software*, 23:501–520, 2008.

[9] P. Davis. *Circulant Matrices*. John Wiley & Sons, 1979.

[10] P. Delsarte, Y. V. Genin, and Y. G. Kamp. Orthogonal polynomial matrices on the unit circle. *IEEE Trans. Circuits Systems*, CAS 25:149–160, 1978.

[11] A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

[12] H. Dym and I. Gohberg. Extension of band matrices with band inverses. *Linear Algebra and Applications*, 36:1–24, 1981.

[13] M. Fukuda, M. Kojima, K. Murota, and K. Nakata. Exploiting sparsity in semidefinite programming via matrix completion i: general framework. *SIAM Journal in Optimization*, 11:647–674, 2000.

[14] I. Gohberg, Goldberg, and M. Kaashoek. *Classes of Linear Operators vol II*. Birkhauser, Boston, 1994.

[15] M. Golumbic. *Algorithmic Graph Theory and Perfect Graphs*. Academic Press, New York, 1980.

[16] R. Grone, C. R. Johnson, E. M. Sa, and H. Wolkowicz. Positive definite completions of partial Hermitian matrices. *Linear Algebra and Its Applications*, 58:109–124, 1984.

[17] S. J. Haberman. *The Analysis of frequancy data.* Univ. Chicago Press, 1974.

[18] B. Jamison. Reciprocal processes: the stationary Gaussian case. *Ann. Math. Stat.*, 41:1624–1630, 1970.

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

[20] B. C. Levy, R. Frezza, and A.J. Krener. Modeling and estimation of discrete-time Gaussian reciprocal processes. *IEEE Trans. Automatic Control*, AC-35(9):1013–1023, 1990.

[21] J. A. Sand. *Four papers in Stochastic Realization Theory*. PhD thesis, Dept. of Mathematics, Royal Institute of Technology (KTH), Stockholm, Sweden, 1994.

[22] T. P. Speed and H. T. Kiiveri. Gaussian markov distribution over finite graphs. *The Annals of Statistics*, 14(1):138–150, 1986.

[23] N. Wermuth and E. Scheidt. Fitting a covariance selection model to a matrix. Algorithm AS105. *Appl. Statist.*, 26:88–92, 1977.

[24] P. Whittle. On the fitting of multivariate autoregressions and the approximate spectral factorization of a spectral density matrix. *Biometrica*, 50:129–134, 1963.

[25] D.C. Youla and N.N. Kazanjian. Bauer-type factorization of positive matrices and the theory of matrix polynomials orthogonal on the unit circle. *IEEE Trans. Circuits and Systems*, CAS-25:57–65, 1978.