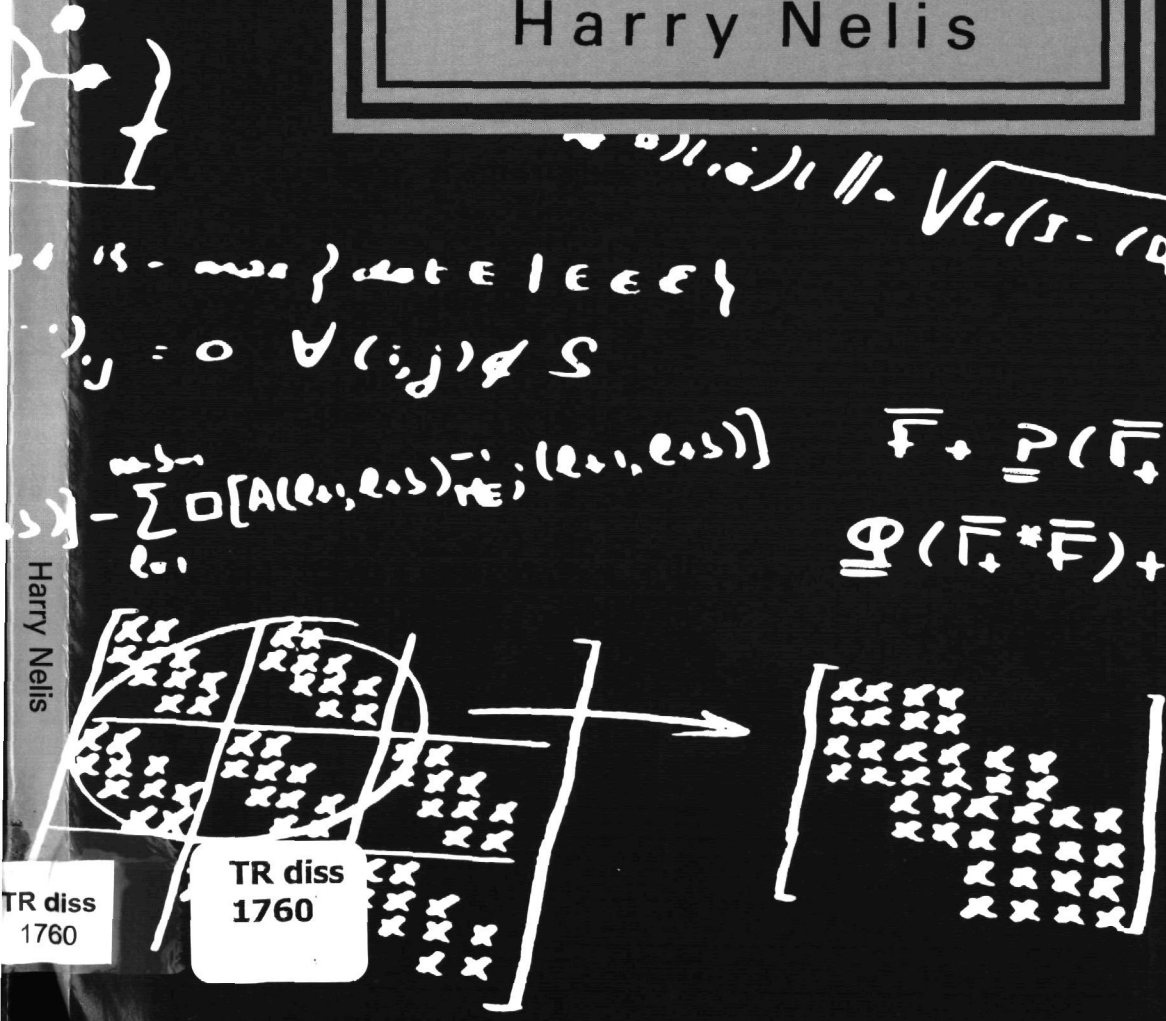
SPARSE APPROXIMATIONS OF INVERSE MATRICES

Harry Nelis

Sparse Approximations of Inverse Matrices

Harry Nelis



TR diss
1760

Sparse Approximations of Inverse Matrices

Harry Nelis

Delft University of Technology

October 1989

Sparse Approximations of Inverse Matrices

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus,
prof. drs. P.A. Schenck,
in het openbaar te verdedigen
ten overstaan van een commissie
aangewezen door het College van Dekanen
op maandag 30 oktober 1989 te 16.00 uur

door

Hendrik Willem Nelis

geboren te Haarlem
electrotechnisch ingenieur

Dit proefschrift is goedgekeurd
door de promotor prof.dr.ir. P. Dewilde

Contents

Preface

THE SOLUTION of many problems in science and engineering reduces ultimately to the inversion of a positive definite matrix. This may be a substantial burden when the matrix is large (e.g., $10,000 \times 10,000$), as for example in modeling problems. Special features of the original, physical problem may be brought in to reduce the number of computations. In this dissertation we impose structure on the inverse of the matrix — we assume that it can be approximated by a sparse matrix.

We show how techniques from inverse scattering theory such as the Wiener-Hopf factorization and the Schur algorithm can be used to determine an optimal or suboptimal sparse approximation to the inverse of a positive definite matrix. Only entries in the original matrix that correspond to nonzero entries in the approximation are used. The algorithms that are proposed have a complexity that is proportional to the number of these nonzero entries.

Chapter 2 deals with the problem of determining an optimal sparse approximation to the inverse of a positive definite matrix. We first consider the case where the sparsity pattern is monotone transitive, and show that the triangular factors of the inverse of the so-called maximum entropy extension — we use the entries in the original matrix that correspond to nonzero entries in the approximation, and (implicitly) estimate the others — are suboptimal sparse approximations in the Frobenius norm to the triangular factors of the inverse of the original matrix. When the sparsity pattern is arbitrary, this result does not hold, but of all matrices whose inverse has the desired sparsity pattern the maximum entropy extension is closest to the original matrix in

the Kullback-Leibler measure.

Chapter 3 describes a generalization of the Wiener-Hopf factorization theory to the case of general, finite dimensional, positive definite matrices that are specified on a block band. This theory provides the link between classical inverse scattering theory and matrix extension theory. It succeeds in constructing a global solution to a generalized inverse scattering problem, which turns out to be equivalent to the maximum entropy extension problem. The solution and, consequently, the triangular factors of the inverse of the maximum entropy extension are obtained by solving a set of linear equations.

Chapters 4, 5, and 6 are devoted to algorithms for computing the inverse of the maximum entropy extension. When the sparsity pattern is staircase, we use the Schur algorithm to compute the triangular factors of this matrix. The algorithm is very well suited for implementation on an array processor of the systolic or wavefront type. For general sparsity patterns we depend on iterative algorithms. As these algorithms consume much time and storage, we devise an algorithm for computing an approximation to the inverse of the maximum entropy extension for the important case where the sparsity pattern is a multiple band. The algorithm is an extension of the Schur algorithm, and computes the inverse of the maximum entropy extension of a matrix that is close to the original.

Chapter 7 is about an application: the modeling of parasitic capacitances in a large integrated circuit. It shows the power of the methods that are proposed — we are able to obtain an accurate model for a system with a large number of conductors, while recent literature considers the modeling of a few conductors already as a hard problem that is worthy of publication.

Chapter 1

Introduction

MODELING is an essential step in the analysis and design of large and complicated systems in science and engineering. A prime example is the design of dikes and drainage systems in civil engineering practice, where a detailed model is needed to calculate the flow of ground water (think of, e.g., the seepage through and below the dams). Other examples can be found in, for example, aerodynamics, geophysics, (solid state) integrated circuit design, and structural mechanics. In most cases deriving the equations that govern the phenomena is not unduly difficult, but solving them in closed form appears to be impossible. Numerical techniques such as the finite element method have been developed that provide ways of finding approximate solutions. They discretize the equations, and convert the problem into a purely algebraic one in which a matrix has to be inverted. The matrix, however, is often so large, that even with powerful computers modeling remains unfeasible.

All models are approximate, but in many cases we can suffice with a lower order, simpler model that still retains the main features of the original. Approximation in this sense is called model reduction, and it is the topic of this dissertation. The techniques that are developed approximate the inverse of a positive definite matrix by a sparse one, and can be used, for instance, in combination with the boundary element method, a finite element method for problems that are posed in integral equation form (many problems described

by partial differential equations, e.g., any problem in electromagnetics, can be reformulated in such a way). The method has the advantage that models do not have to be extended to the infinite boundaries of open systems, and that the number of variables that have to be handled is reduced. The latter is offset to some extent, because every boundary element is generally related to every other, so that the resulting matrix is dense. The inverse of the matrix (which is still very large) is the model, and to reduce the model, we approximate the inverse by a sparse matrix.

We show how techniques from classical inverse scattering theory such as the Wiener-Hopf factorization and the Schur algorithm can be used to determine an optimal or suboptimal sparse approximation to the inverse of a positive definite matrix. The algorithms that are proposed compute the inverse of the so-called maximum entropy extension of the original matrix — only entries that correspond to nonzero entries in the approximation are used, the others are implicitly estimated — or a close approximation to it, depending on the desired sparsity pattern. They have a complexity that is proportional to the number of nonzero entries in the approximation.

In this chapter we start out with a survey of the literature on matrix extensions. We proceed with an overview of the new results in this thesis, introduce the notation, and, finally, review some earlier results on maximum entropy extensions.

1.1 Background

Suppose that $A = [a_{ij}]$ is a positive definite matrix. Furthermore, suppose that it is only partially specified — it is specified only on a subset $\mathcal{S}$ of the set of index pairs $\{ (i,j) \mid i,j = 1,\ldots,n \}$ — and assume that the diagonal entries are specified. In [DG81] Dym and Gohberg studied the case where $\mathcal{S}$ *is a block band — see Figure 1.1. They showed that for this case there* is a unique positive definite extension of A — an extension of A is a matrix that coincides with A on $\mathcal{S}$ — whose determinant is maximal, and that this matrix is the unique positive definite extension whose inverse is zero on the complement of $\mathcal{S}$. Due to the analogy with the maximum entropy inequality in spectral estimation theory (see, e.g., [Bur75]), they called it the

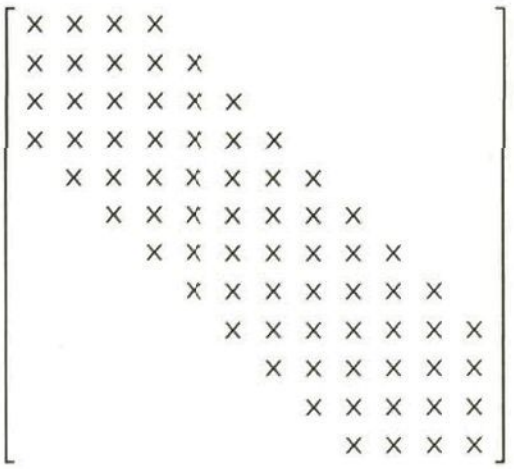


Figure 1.1: A matrix that is specified on a block band. Specified entries are marked with an '×'; unspecified entries are blank.

maximum entropy extension. In [GJSW84] Grone et al. generalized the above results to sets $\mathcal{S}$ that contain the diagonal pairs, but are arbitrary otherwise. They showed that a unique maximum entropy extension of a positive definite matrix that is specified on such a set exists, and that its inverse vanishes on the complement of the set.

The method used in [DG81] to construct the maximum entropy extension of a positive definite matrix that is specified on a block band is based on the inversion of contiguous principal submatrices of that matrix. In [DGK80] Delsarte et al. followed a similar approach to compute the triangular factors of the inverse of a completely specified, positive definite matrix, thereby in fact generalizing an algorithm known in estimation theory as the Levinson algorithm [Lev47]. The problem of computing these factors can be viewed as a time variant estimation problem, since any positive definite matrix can be interpreted as the covariance matrix of a [part of a] time varying stochastic process. Computing the triangular factors of the inverse of the maximum entropy extension of a partially specified, positive definite matrix

then corresponds to solving a partial correlation (PARCOR) problem. In [DVK78] Dewilde et al. showed how partial correlations for one dimensional, time invariant stochastic processes can be determined by using an algorithm known in interpolation theory as the Schur algorithm [Sch17]. It soon became apparent that this algorithm can be used advantageously to determine partial correlations even for time varying processes [Dep81, LK81]. This property of the Schur algorithm was exploited by Morf and Delosme [MD81] to devise an algorithm for computing the triangular factors of the inverse of a completely specified, positive definite matrix. In [DD87] Dewilde and Deprettere showed how this algorithm can be used to compute the triangular factors of the inverse of the maximum entropy extension of a positive definite matrix that is specified on a staircase band — see Figure 1.2. The

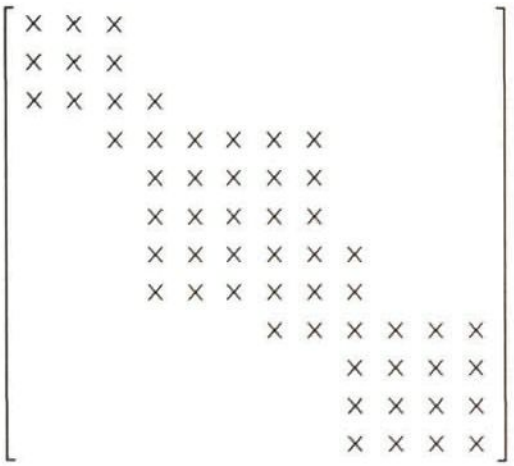


Figure 1.2: A matrix that is specified on a staircase band.

algorithm requires $O(nb^2)$ operations and $O(nb)$ storage, where n is the size of the matrix and b the average width of the band. It is very well suited for implementation on an array processor of the systolic or wavefront type. In the same paper Dewilde and Deprettere showed that, except for a diagonal bias factor, the triangular factors of the inverse of the maximum entropy

extension give an optimal approximation in the Frobenius norm (also known as the Hilbert-Schmidt norm) to the triangular factors of the inverse of the completely specified matrix.

Various iterative algorithms for computing the inverse of the maximum entropy extension of a Toeplitz-block Toeplitz matrix that is specified *on a multiple band — see Figure 1.3 — have been proposed* [LM81, LM82, RP87]. These algorithms, however, either have trouble converging or are

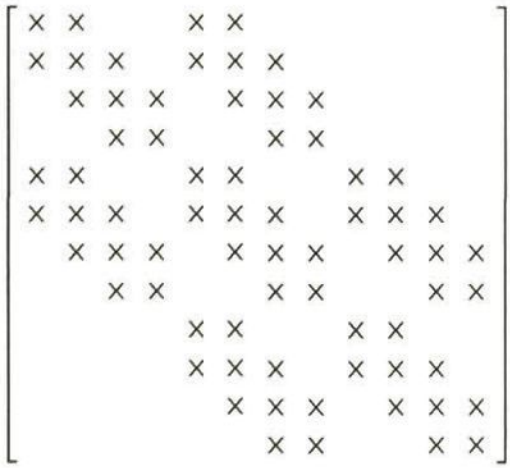


Figure 1.3: A matrix that is specified on a multiple band.

computationally very expensive. In [DD88] Dewilde and Deprettere proposed a fast algorithm for computing an approximate inverse of a general, positive definite matrix that is specified on a triple band $\mathcal{S}$. They assumed that certain conditions on the completely specified matrix are satisfied. The approximation is the inverse of a matrix that closely matches the partially specified matrix. It does not have zeros in every entry on the complement of $\mathcal{S}$, but can be represented and computed efficiently.

1.2 Contributions

Suppose that A is a positive definite matrix. Chapter 2 deals with the problem of determining an optimal sparse approximation to A^{-1} (the inverse of A) that is zero on the complement of a set $\mathcal{S}$. We first consider the case where $\mathcal{S}$ is monotone transitive — see Figure 1.4 — and show that the triangular factors of the inverse of the maximum entropy extension of the part of A with support on $\mathcal{S}$ are suboptimal sparse approximations in the Frobenius norm to the triangular factors of A^{-1}. When $\mathcal{S}$ is arbitrary, this result does not hold, but of all matrices whose inverse vanishes on the complement of $\mathcal{S}$ the maximum entropy extension is closest to A in the Kullback-Leibler measure.

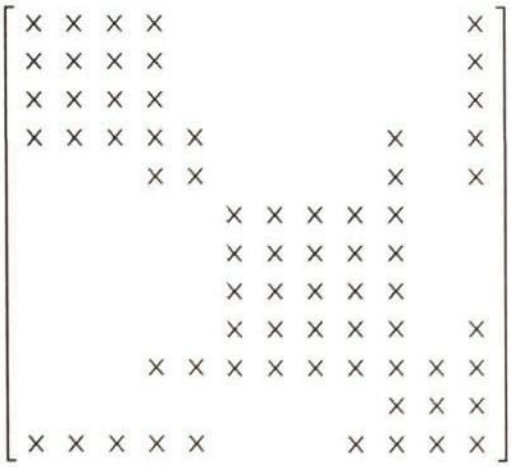


Figure 1.4: A matrix with support on a monotone transitive set. Nonzero entries are marked with an '×'; vanishing entries are blank.

Chapter 3 describes a generalization of the Wiener-Hopf factorization theory to the case of general, finite dimensional, positive definite matrices that are specified on a block band $\mathcal{S}$. This theory provides the link between classical inverse scattering theory and matrix extension theory. It succeeds in constructing a global solution to a generalized inverse scattering problem, which turns out to be equivalent to the maximum entropy extension problem.

The solution and, consequently, the triangular factors of the inverse of the maximum entropy extension are obtained by solving a set of linear equations.

Chapters 4, 5, and 6 are devoted to algorithms for computing the inverse of the maximum entropy extension when $\mathcal{S}$ is a staircase band, an arbitrary set, and a multiple band respectively. The last case is of major interest — multiple bands arise in problems with multi dimensional geometries, for example, when the boundary element method is used to model a three dimensional system and in two dimensional spectral estimation (see, e.g., [Nin89] and [LM81]). In contrast to the case where $\mathcal{S}$ is staircase, however, no closed solution to the extension problem exists, and we have to be satisfied with an approximate solution. We devise an extension of the Schur algorithm for computing such an approximation, provided that certain conditions are satisfied. It computes the inverse of the maximum entropy extension of a partially specified matrix that is close to the original and specified on the same set $\mathcal{S}$, and requires $O(nc^2)$ operations and $O(nc)$ storage, where n is the size of the matrix and c the average number of elements in the set per row of the matrix.

Chapter 7 is about an application: the modeling of parasitic capacitances in a large integrated circuit. It shows the power of the methods that are proposed — we are able to obtain an accurate model for a system with a large number of conductors, while recent literature considers the modeling of a few conductors already as a hard problem that is worthy of publication.

1.3 Notation and Basic Concepts

We denote matrices by italic uppercase letters. Matrices have complex entries, unless we state otherwise. We denote the (i, j) entry of A by $(A)_{ij}$ or a_{ij}, the complex-conjugate transpose of A by A^*, the kth power of A by A^k, and the direct sum of A and B by $A \oplus B$. The symbols 0 and I denote the zero matrix and the identity matrix. Their size is defined by the context.

$\underline{\underline{\mathrm{P}}}$ and $\underline{\underline{\mathrm{Q}}}$ denote the operators that project a matrix on its upper and lower triangular part respectively. $\underline{\underline{\mathrm{P}}}_0$ is the projection operator on the

diagonal. For example, for $A = [a_{ij}]$, $i, j = 1, \ldots, n$,

$$(\underline{\underline{\mathrm{P}}}A)_{ij} = \begin{cases} a_{ij} & \text{if } i \leq j; \\ 0 & \text{otherwise,} \end{cases}$$

$$(\underline{\underline{\mathrm{Q}}}A)_{ij} = \begin{cases} a_{ij} & \text{if } i \geq j; \\ 0 & \text{otherwise,} \end{cases}$$

and

$$(\underline{\underline{\mathrm{P}}}_0 A)_{ij} = \begin{cases} a_{ij} & \text{if } i = j; \\ 0 & \text{otherwise.} \end{cases}$$

The symbols $\det A$, $\log A$, $\mathrm{tr}A$, and $\|A\|$ denote the determinant, the natural logarithm, the trace, and the Frobenius norm of A. The condition number $\kappa(A)$ is defined as $\kappa(A) = \|A\|\|A^{-1}\|$.

A matrix S is called *contractive* if $I - S^*S$ is positive definite; $[\Gamma\ \Delta]$ is called *admissible* if (1) Γ and Δ are upper triangular, (2) Γ is invertible, and (3) $\Gamma^{-1}\Delta$ is contractive.

If A is positive definite, then L_A and M_A refer to the (unique) upper triangular matrices with positive diagonal entries that satisfy $A = L_A L_A^* = M_A^* M_A$. We always assume that the diagonal entries in A are equal to one, which does not impair generality, because we can always convert A to $(\underline{\underline{\mathrm{P}}}_0 A)^{-\frac{1}{2}} A (\underline{\underline{\mathrm{P}}}_0 A)^{-\frac{1}{2}}$. We next attach to A (with $\underline{\underline{\mathrm{P}}}_0 A = I$) the *impedance matrix* $G_A = I + 2(\underline{\underline{\mathrm{P}}} - \underline{\underline{\mathrm{P}}}_0)A$ and the *scattering matrix* $S_A = (G_A + I)^{-1}(G_A - I)$. G_A and S_A are related via the Cayley transformation, and the following properties are equivalent:

- $A = \frac{1}{2}(G_A + G_A^*)$ is positive definite;
- S_A is contractive;
- $\left[\frac{1}{2}(G_A + I)\ \frac{1}{2}(G_A - I)\right]$ is admissible.

If A is only partially specified, then $\mathcal{E}$ and A_{ME} denote the set of positive definite extensions and the maximum entropy extension of A respectively.

A subset $\mathcal{S}$ of the set of index pairs $\{(i,j) \mid i, j = 1, \ldots, n\}$ is called

a

- *block band* if it is of the form $\{(i,j) \mid |i-j| \le b\}$ — see Figure 1.1;
- *staircase band* if for all $i \le k \le l \le j$ $[i \ge k \ge l \ge j]$ $(i,j) \in \mathcal{S}$ implies that $(k,l) \in \mathcal{S}$ — see Figure 1.2;
- *multiple band* if it is of the type shown in Figure 1.3.
- *monotone transitive set* if for all $i \le j \le k$ $[i \ge j \ge k]$ $(i,j) \in \mathcal{S}$ and $(i,k) \in \mathcal{S}$ implies that $(j,k) \in \mathcal{S}$ — see Figure 1.4;

We always assume that $\mathcal{S}$ contains the diagonal pairs (i,i).

The symbol J denotes the matrix

$$J = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}.$$

A matrix Θ is *J-unitary* if $\Theta^* J \Theta = J$. An *elementary* J-unitary matrix is a rank-two correction to the matrix $I \oplus I$ of the form

$$\begin{array}{c} \\ \\ \\ \\ i \\ \\ \\ \\ \\ j \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} & & & i & & j & \end{array} \\ \left[\begin{array}{ccccccc} 1 & & & \vdots & & \vdots & \\ & \ddots & & \vdots & & \vdots & 0 \\ & & 1 & \vdots & & \vdots & \\ \cdots & \cdots & \cdots & \mathrm{ch} & \cdots & \mathrm{sh} & \cdots \\ & & & \vdots & 1 & \vdots & \\ & & & \vdots & \ddots & \vdots & \\ & & & \vdots & & 1 & \vdots \\ \cdots & \cdots & \cdots & \mathrm{sh}^* & \cdots & \mathrm{ch} & \cdots \\ & & & \vdots & & \vdots & 1 \\ & 0 & & \vdots & & \vdots & \ddots \\ & & & \vdots & & \vdots & 1 \end{array}\right] \end{array},$$

where $1 \le i \le n$, $n+1 \le j \le 2n$, n is the size of the identity, $\mathrm{ch} = 1/\sqrt{1-|\rho|^2}$, $\mathrm{sh} = \rho/\sqrt{1-|\rho|^2}$, and $|\rho| < 1$. We denote this matrix by $\Theta(i,j,\rho)$. Clearly, $\Theta(i,j,\rho)^* J \Theta(i,j,\rho) = J$.

We denote block matrices by bold uppercase letters. If we partition A as a block matrix, then we denote the block matrix by $\mathbf{A}$. Conversely, if we interpret $\mathbf{A}$ as a matrix with scalar entries, then we denote it by A. If $\mathbf{A}$ is positive definite, then $\mathbf{L_A}$ and $\mathbf{M_A}$ refer to the (unique) upper triangular block matrices with upper triangular diagonal blocks with positive diagonal entries that satisfy $\mathbf{A} = \mathbf{L_A}\mathbf{L_A^*} = \mathbf{M_A^*}\mathbf{M_A}$. The symbol $\mathbf{A}(i,j)$ denotes the principal submatrix of $\mathbf{A}$ that lies in the rows and columns indexed by $i,\ldots,j$. $\square\,[\mathbf{A};(i,j)]$ is the block matrix that satisfies $\left(\square\,[\mathbf{A};(i,j)]\right)(i,j) = \mathbf{A}$, and is zero otherwise — it is an embedding of $\mathbf{A}$ in a zero matrix. Its size is defined by the context.

To ease the notation, we suppress as much of the subscripts as possible. For example, we write L_{ME} instead of $L_{A_{ME}}$ whenever its identity is clear.

1.4 Some Previous Results

Finally, we review three important results on maximum entropy extensions that have appeared in the literature. We start out with the well-known inequality of Hadamard.

Theorem 1.1 (Hadamard's inequality) *Let $A = [a_{ij}]$, $i,j = 1,\ldots,n$, be positive definite. Then,*

$$\det A \leq \prod_{i=1}^{n} a_{ii}.$$

Moreover, equality holds if, and only if, A is diagonal.

We restate this result as follows. Suppose that A is positive definite, and that only its diagonal entries are specified. Then, there is a unique matrix B in $\mathcal{E}$ (the set of positive definite extensions of A) such that

$$\det B = \max\{\det E \mid E \in \mathcal{E}\}.$$

Moreover, B is the unique positive definite extension whose inverse is diagonal (since B is diagonal). Thus, the following result of Dym and Gohberg is a generalization of Hadamard's inequality.

Theorem 1.2 ([DG81]) *Let $A = [a_{ij}]$ be a positive definite matrix that is specified on a block band $\mathcal{S} = \{ (i,j) \mid |i-j| \leq b \}$. Then, there is a unique matrix B in $\mathcal{E}$ such that*

$$\det B = \max\{ \det E \mid E \in \mathcal{E} \}.$$

Moreover, B is the unique positive definite extension whose inverse satisfies

$$(B^{-1})_{ij} = 0 \quad \forall\, (i,j) \notin \mathcal{S}.$$

The following result of Grone et al. in turn is a generalization of the result of Dym and Gohberg.

Theorem 1.3 ([GJSW84]) *Let A be a positive definite matrix that is specified on a set $\mathcal{S}$ that contains the diagonal pairs, but is arbitrary otherwise. Then, there is a unique matrix B in $\mathcal{E}$ such that*

$$\det B = \max\{ \det E \mid E \in \mathcal{E} \}.$$

Moreover, B is the unique positive definite extension whose inverse satisfies

$$(B^{-1})_{ij} = 0 \quad \forall\, (i,j) \notin \mathcal{S}.$$

We give the proof of Theorem 1.3 in Chapter 5. As mentioned above, we call B the maximum entropy extension of A and denote it by A_{ME}.

Chapter 2

Optimal Sparse Approximations

SUPPOSE THAT A is positive definite. In this chapter we address the problem of determining an optimal sparse approximation to A^{-1} that is zero on the complement of a set $\mathcal{S}$. More precisely, we seek the upper triangular matrix F that is such that $\|I - FL\|$ is minimal and F^*F is zero on the complement of $\mathcal{S}$ — F is a triangular factor of the approximation. We first consider the case where $\mathcal{S}$ is monotone transitive. Using the fact that for such a set F must vanish on the upper triangular part of the complement of $\mathcal{S}$, we show that the minimum occurs for $F = (D_{ME}^{-1}D)L_{ME}^{-1}$, where D is defined as $\underline{\underline{P}}_0 L$ (the diagonal of L), etc., and the maximum entropy extension is based on the part of A with support on $\mathcal{S}$. When $\mathcal{S}$ is arbitrary, this argument and result do not hold. We next show, however, that of all matrices whose inverse vanishes on the complement of $\mathcal{S}$ A_{ME} is closest to A in the Kullback-Leibler measure.

2.1 Monotone Transitive Sets

The notion of monotone transitive sets was introduced by Rose [Ros72], who studied the problem of fill-ins in the triangular factors of sparse, positive

definite matrices, and showed that if a positive definite matrix C is zero on the complement of a set $\mathcal{S}$, then its triangular factor M vanishes on the upper triangular part of the complement of the set, irrespective of the values of the entries in C, if, and only if, $\mathcal{S}$ is monotone transitive. Hence, if F is upper triangular, and F^*F vanishes on the complement of a monotone transitive set, then F must be zero on the upper triangular part of the complement of the set. Clearly, block bands and staircase bands are monotone transitive.

With $\mathcal{S}$ Rose associated a graph $\mathcal{G}$ that has a vertex for every index i and an undirected edge between vertex i and j if $(i,j) \in \mathcal{S}$. He proved that there is a permutation matrix P such that the support of PCP^* is monotone transitive if, and only if, every cycle in $\mathcal{G}$ that consists of four edges or more contains a chord, i.e., an edge joining two nonconsecutive vertices in the cycle. Graphs with this property are said to be *chordal* or *triangulated.* For example, if C is as shown in Figure 2.1, then the cycle $2-7-10-5-2$ has four edges yet no chord — we cannot permute C to a matrix with monotone transitive support. If, however, we partition C into blocks of size 4×4, then the graphs of the principal submatrices $\mathbf{C}(1,2)$ and $\mathbf{C}(2,3)$ are chordal — see Figure 2.2. Arranging the rows and columns of $\mathbf{C}(1,2)$ according to the order $1-5-2-6-3-7-4-8$, we see that the support of the permuted matrix is staircase and hence monotone transitive.

Chordal graphs also have import on the existence of positive definite extensions of partially specified, not necessarily positive definite Hermitian matrices. In [GJSW84] Grone et al. showed that if $\mathcal{G}$ is chordal, and all completely specified principal submatrices with support on $\mathcal{S}$ are positive definite, then positive definite extensions necessarily exist, and if $\mathcal{G}$ is not chordal, then they do not in general exist.

The following results are generalizations of the approximation theorems in [DD87] to the case of monotone transitive sets. $\mathcal{F}$ denotes the set of upper triangular matrices that vanish on the upper triangular part of the complement of $\mathcal{S}$. D_A is defined as $D_A = \underline{\mathrm{P}}_0 L_A$.

Lemma 2.1 *Suppose that $A = [a_{ij}]$, $i,j = 1,\ldots,n$, is a positive definite matrix that is specified on a monotone transitive set $\mathcal{S}$, and let $F \in \mathcal{F}$*

$C =$

	1	2	3	4	5	6	7	8	9	10	11	12
1	×	×			×	×						
2	×	×	×		×	×	×					
3		×	×	×		×	×	×				
4			×	×			×	×				
5	×	×			×	×			×	×		
6	×	×	×		×	×	×		×	×	×	
7		×	×	×		×	×	×		×	×	×
8			×	×			×	×			×	×
9					×	×			×	×		
10					×	×	×		×	×	×	
11						×	×	×		×	×	×
12							×	×			×	×

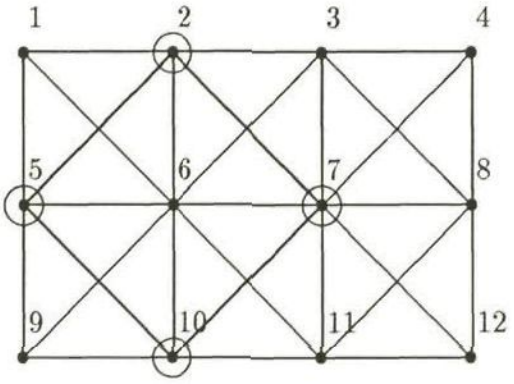


Figure 2.1: A matrix with support on a multiple band and its graph. Nonzero entries are marked with an '×'; vanishing entries are blank.

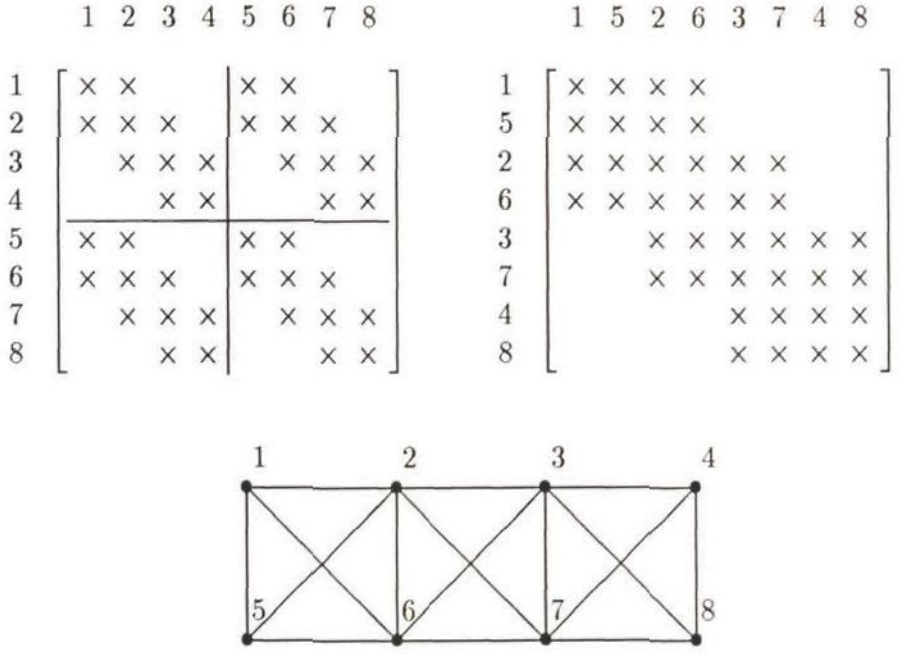


Figure 2.2: The principal submatrix $\mathbf{C}(1,2)$, the permuted matrix, and their graph.

and $H \in \mathcal{F}$. *Then,*

$$\underline{\underline{\mathrm{P}}}_0(FAH^*) = \underline{\underline{\mathrm{P}}}_0(FA_{ME}H^*).$$

Proof Suppose that $B = A - A_{ME}$. Because for all $i \leq j \leq k$ $(i,j) \in \mathcal{S}$ and $(i,k) \in \mathcal{S}$ implies that $(j,k) \in \mathcal{S}$ and hence $(k,j) \in \mathcal{S}$ — $\mathcal{S}$ is monotone transitive — and B vanishes on $\mathcal{S}$, we have that for all (i,j) in the upper triangular part of $\mathcal{S}$

$$(FB)_{ij} = \sum_{k=i}^{n} f_{ik} b_{kj} = 0.$$

Because H is upper triangular and zero on the upper triangular part of the complement of $\mathcal{S}$, it follows that

$$\underline{\mathrm{P}}_0(FBH^*) = 0$$

and hence

$$\underline{\mathrm{P}}_0(FAH^*) = \underline{\mathrm{P}}_0(FA_{ME}H^*).$$

■

Lemma 2.2 *Let A and F be as defined in Lemma 2.1. Then,*

1. $\underline{\mathrm{P}}_0(L_{ME}^{-1}AL_{ME}^{-*}) = I;$
2. $\underline{\mathrm{P}}_0\left(FA(D_{ME}^{-1}L_{ME}^{-1})^*\right) = \underline{\mathrm{P}}_0F.$

Proof Because A_{ME}^{-1} vanishes on the complement of $\mathcal{S}$, and $\mathcal{S}$ is monotone transitive, we have that $L_{ME}^{-1} \in \mathcal{F}$, and the first equality in the lemma follows directly from Lemma 2.1. Furthermore, by the same lemma,

$$\begin{aligned}\underline{\mathrm{P}}_0\left(FA(D_{ME}^{-1}L_{ME}^{-1})^*\right) &= \underline{\mathrm{P}}_0\left(FA_{ME}(D_{ME}^{-1}L_{ME}^{-1})^*\right)\\ &= \underline{\mathrm{P}}_0(FL_{ME}D_{ME}^{-*}),\end{aligned}$$

which evaluates to $\underline{\mathrm{P}}_0F$, as F and L_{ME} are upper triangular and D_{ME}^{-*} is diagonal. ■

The second assertion in Lemma 2.2 can be interpreted as a reproducing property — taking the 'generalized inner product' $[F, D_{ME}^{-1}L_{ME}^{-1}]_A = \underline{\mathrm{P}}_0(FA(D_{ME}^{-1}L_{ME}^{-1})^*)$ (the inner product is a matrix, not a scalar) reproduces the diagonal of F.

Let Π denote the operator that projects a matrix on $\mathcal{F}$ with respect to the inner product $\langle B, C\rangle_A = \mathrm{tr}(BAC^*)$, and let $\|B\|_A = \sqrt{\langle B, B\rangle_A}$.

Lemma 2.3 *Let A be as defined in Lemma 2.1, F upper triangular, and $H \in \mathcal{F}$. Then,*

$$\underline{\mathrm{P}}_0(FAH^*) = \underline{\mathrm{P}}_0(\Pi(F)AH^*).$$

Proof Suppose that $K = F - \Pi F$. From the definition of Π we have

$$\langle K, H \rangle_A = \mathrm{tr}(KAH^*) = 0 \quad \forall\, H \in \mathcal{F}.$$

Specializing H to E_{ij}, for all (i,j) in the upper triangular part of $\mathcal{S}$, where

$$\left(E_{ij}\right)_{kl} = \begin{cases} 1 & \text{if } (k,l) = (i,j); \\ 0 & \text{otherwise,} \end{cases}$$

we obtain that KA is zero on the upper triangular part of $\mathcal{S}$. Hence, KA is equal to $N + O$, for some strictly lower triangular N and an upper triangular O that vanishes on the upper triangular part of $\mathcal{S}$.

Because NH^* is strictly lower triangular, and OH^* vanishes on the diagonal — O is upper triangular and zero on the upper triangular part of $\mathcal{S}$, and H vanishes on the upper triangular part of the complement of $\mathcal{S}$ — it follows that

$$\underline{\underline{\mathrm{P}}}_0(KAH^*) = \underline{\underline{\mathrm{P}}}_0(NH^*) + \underline{\underline{\mathrm{P}}}_0(OH^*) = 0$$

and hence

$$\underline{\underline{\mathrm{P}}}_0(FAH^*) = \underline{\underline{\mathrm{P}}}_0(\Pi(F)AH^*).$$

■

Theorem 2.1 *Suppose that A is positive definite, and let $\mathcal{S}$ be a monotone transitive set. Furthermore, let $\mathcal{F}$ be the set of upper triangular matrices that vanish on the upper triangular part of the complement of $\mathcal{S}$. Then,*

1. $\min\{\, \|I - FL\| \mid F \in \mathcal{F} \}$ *occurs for* $F = (D_{ME}^{-1}D)L_{ME}^{-1}$, *where the maximum entropy extension is based on the part of A with support on $\mathcal{S}$;*

2. $\|I - \left((D_{ME}^{-1}D)L_{ME}^{-1}\right) L\| = \sqrt{\mathrm{tr}\left(I - (D_{ME}^{-1}D)^2\right)}$.

Proof Suppose that $F \in \mathcal{F}$. Because by Lemma 2.3

$$\underline{\underline{P}}_0\left(\Pi(L^{-1})AF^*\right) = \underline{\underline{P}}_0(D^*F^*),$$

and by a proof similar to the proof of Lemma 2.2(2)

$$\underline{\underline{P}}_0\left(\left((D_{ME}^{-1}D)L_{ME}^{-1}\right)AF^*\right) = \underline{\underline{P}}_0(D^*F^*),$$

we have

$$\mathrm{tr}\left(\Pi(L^{-1})AF^*\right) = \mathrm{tr}\left(\left((D_{ME}^{-1}D)L_{ME}^{-1}\right)AF^*\right) \quad \forall\, F \in \mathcal{F}.$$

Since A is invertible, it readily follows that

$$\Pi L^{-1} = (D_{ME}^{-1}D)L_{ME}^{-1}.$$

Hence,

$$\min\{\, \|L^{-1} - F\|_A \mid F \in \mathcal{F}\,\}$$

and, consequently,

$$\min\{\, \|I - FL\| \mid F \in \mathcal{F}\,\}$$

— $\|L^{-1} - F\|_A = \|I - FL\|$ — occur for $F = (D_{ME}^{-1}D)L_{ME}^{-1}$.

Because by Lemma 2.2(1)

$$\underline{\underline{P}}_0(L_{ME}^{-1}AL_{ME}^{-*}) = I,$$

we have

$$\mathrm{tr}\left(\left((D_{ME}^{-1}D)L_{ME}^{-1}\right)A\left((D_{ME}^{-1}D)L_{ME}^{-1}\right)^*\right) = \mathrm{tr}\left((D_{ME}^{-1}D)^2\right),$$

so that

$$\begin{aligned}\|I - \left((D_{ME}^{-1}D)L_{ME}^{-1}\right)L\| &= \sqrt{\mathrm{tr}\left(I - 2(D_{ME}^{-1}D)^2 + (D_{ME}^{-1}D)^2\right)} \\ &= \sqrt{\mathrm{tr}\left(I - (D_{ME}^{-1}D)^2\right)}.\end{aligned}$$

■

We close this section with a simple example.

Example 2.1 Suppose that

$$A = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix},$$

where α lies between -1 and 1, and let $\mathcal{S} = \{(1,1), (2,2)\}$. Then,

$$L = \begin{bmatrix} \sqrt{1-\alpha^2} & \alpha \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} \sqrt{1-\alpha^2} & 0 \\ 0 & 1 \end{bmatrix},$$

and the maximum entropy extension of the part of A with support on $\mathcal{S}$ is given by

$$A_{ME} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Furthermore,

$$L_{ME} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad D_{ME} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and to find the minimum of $\{ \|I - FL\| \mid F \in \mathcal{F} \}$, where

$$F = \begin{bmatrix} f & 0 \\ 0 & g \end{bmatrix}$$

and $\mathcal{F}$ is the set of diagonal matrices, we minimize

$$\|I - FL\| = \sqrt{\left(1 - f\sqrt{1-\alpha^2}\right)^2 + (f\alpha)^2 + (1-g)^2}.$$

In the minimum the partial derivatives with respect to f and g must vanish, which implies that $f = \sqrt{1-\alpha^2}$ and $g = 1$ — the minimum occurs for

$$F = \begin{bmatrix} \sqrt{1-\alpha^2} & 0 \\ 0 & 1 \end{bmatrix},$$

which, indeed, is equal to $(D_{ME}^{-1}D)L_{ME}^{-1}$. The minimum itself is given by

$$\|I - \left((D_{ME}^{-1}D)L_{ME}^{-1}\right) L\| = |\alpha|.$$

■

2.2 Arbitrary Sets

In contrast to the monotone transitive case, F does not necessarily vanish on the upper triangular part of the complement of $\mathcal{S}$ when $\mathcal{S}$ is arbitrary. Moreover, in the general case $F = (D_{ME}^{-1}D)L_{ME}^{-1}$, the optimal solution when $\mathcal{S}$ is monotone transitive, does not satisfy the constraint that F^*F is zero on the complement of $\mathcal{S}$. Another norm to measure the distance between two positive definite matrices A and B is

$$\lambda(A,B) = \frac{1}{n}\Big(\mathrm{tr}(AB^{-1}) - \log\det(AB^{-1})\Big) - 1,$$

where n is the size of A and B. This norm was introduced by Lev-Ari et al. in [LPK89], and is known as the Kullback-Leibler measure. It is related to the norm

$$\mu(A,B) = \|I - L_B^{-1}L_A\|^2$$

in the following way.

Proposition 2.1 *Let A and B be positive definite. Then,*

$$\lambda(A,B) = \frac{1}{n}\mu(A,B) + 2\lambda(D_A, D_B),$$

where n is the size of A and B.

Proof Because similar matrices have equal traces,

$$\begin{aligned}\mu(A,B) &= \mathrm{tr}(I - 2L_B^{-1}L_A + L_B^{-1}AL_B^{-*}) \\ &= \mathrm{tr}(AB^{-1}) - 2\mathrm{tr}(D_AD_B^{-1}) + n.\end{aligned}$$

Furthermore,

$$\begin{aligned}\lambda(A,B) &= \frac{1}{n}\Big(\mathrm{tr}(AB^{-1}) - \log\det(AB^{-1})\Big) - 1 \\ &= \frac{1}{n}\Big(\mathrm{tr}(AB^{-1}) - 2\log\det(D_AD_B^{-1})\Big) - 1,\end{aligned}$$

and the proof of the lemma follows readily. ■

The Kullback-Leibler measure obeys a useful triangle equality. $\mathcal{B}$ denotes the set of positive definite matrices whose inverse is zero on the complement of $\mathcal{S}$.

Lemma 2.4 ([LPK89]) *Let A be a positive definite matrix that is specified on a set $\mathcal{S}$ that contains the diagonal pairs, but is arbitrary otherwise. Then, for all $B \in \mathcal{B}$*

$$\lambda(A, B) = \lambda(A, A_{ME}) + \lambda(A_{ME}, B).$$

Proof Suppose that E is an extension of A. Because $A - E$ vanishes on $\mathcal{S}$, and B^{-1} on the complement of $\mathcal{S}$, we have

$$\underline{\underline{\mathrm{P}}}_0\left((A - E)B^{-1}\right) = 0$$

and hence

$$\underline{\underline{\mathrm{P}}}_0(AB^{-1}) = \underline{\underline{\mathrm{P}}}_0(EB^{-1}).$$

Substituting E by A_{ME}, and subsequently B^{-1} by A_{ME}^{-1}, we obtain

$$\underline{\underline{\mathrm{P}}}_0(A_{ME}B^{-1}) = \underline{\underline{\mathrm{P}}}_0(AB^{-1})$$

and

$$\underline{\underline{\mathrm{P}}}_0(AA_{ME}^{-1}) = I.$$

Therefore,

$$\begin{aligned} n\left(\lambda(A, A_{ME}) + \lambda(A_{ME}, B)\right) &= \mathrm{tr}(AA_{ME}^{-1}) - \log\det(AA_{ME}^{-1}) - n + \\ &\quad \mathrm{tr}(A_{ME}B^{-1}) - \log\det(A_{ME}B^{-1}) - n \\ &= \mathrm{tr}(AB^{-1}) - \log\det(AB^{-1}) - n \\ &= n\lambda(A, B). \end{aligned}$$

■

Lemma 2.4 implies that A_{ME} is the element in $\mathcal{B}$ that is closest to A. We summarize this result in the following theorem.

Theorem 2.2 ([LPK89]) *Suppose that A is positive definite, let $\mathcal{S}$ be a set that contains the diagonal pairs, but is arbitrary otherwise, and let $\mathcal{B}$ be the set of positive definite matrices whose inverse vanishes on the complement of $\mathcal{S}$. Furthermore, assume that*

$$\lambda(A,B) = \frac{1}{n}\left(\mathrm{tr}(AB^{-1}) - \log\det(AB^{-1})\right) - 1,$$

where n is the size of A and B. Then,

$$\min\{\,\lambda(A,B) \mid B \in \mathcal{B}\,\}$$

occurs for $B = A_{ME}$, where the maximum entropy extension is based on the part of A with support on $\mathcal{S}$.

We proceed with an example.

Example 2.2 Suppose that

$$A = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix},$$

where α lies between -1 and 1, and let $\mathcal{S} = \{(1,1),(2,2)\}$. Then, the maximum entropy extension of the part of A with support on $\mathcal{S}$ is given by

$$A_{ME} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and to find the minimum of $\{\,\lambda(A,B) \mid B \in \mathcal{B}\,\}$, where

$$B = \begin{bmatrix} b & 0 \\ 0 & c \end{bmatrix},$$

$b > 0$ and $c > 0$, and $\mathcal{B}$ is the set of diagonal matrices with positive diagonal entries, we minimize

$$\lambda(A,B) = \frac{1}{2}\left(\left(\frac{1}{b}+\frac{1}{c}\right) - \log\left((1-\alpha^2)\frac{1}{bc}\right)\right) - 1.$$

Setting the partial derivatives with respect to b and c to zero, we obtain that $b = 1$ and $c \approx 1$ — the minimum occurs for

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

which, indeed, is equal to A_{ME}. The minimum itself is given by

$$\lambda(A, A_{ME}) = -\frac{1}{2}\log(1 - \alpha^2).$$

■

For future reference, we state two more results. The first is a generalization of Lemma 2.2(1) to arbitrary sets (instead of diagonals, however, it deals with traces). The second gives an expression for $\|I - L_{ME}^{-1}L\|$.

Lemma 2.5 *Let A be as defined in Theorem 2.2. Then,*

$$\mathrm{tr}(L_{ME}^{-1}AL_{ME}^{-*}) = \mathrm{tr}I.$$

Proof By the fourth equality in the proof of Lemma 2.4

$$\underline{\underline{P}}_0(AA_{ME}^{-1}) = I.$$

Because similar matrices have equal traces, it follows that

$$\mathrm{tr}(L_{ME}^{-1}AL_{ME}^{-*}) = \mathrm{tr}(AA_{ME}^{-1}) = \mathrm{tr}I.$$

■

Proposition 2.2 *Let A be as defined in Theorem 2.2. Then,*

$$\|I - L_{ME}^{-1}L\| = \sqrt{2\mathrm{tr}(I - D_{ME}^{-1}D)}.$$

Proof Because

$$\|I - L_{ME}^{-1}L\| = \sqrt{\mathrm{tr}(I - 2L_{ME}^{-1}L + L_{ME}^{-1}AL_{ME}^{-*})},$$

the proof of the lemma follows directly from Lemma 2.5. ■

2.3 Concluding Remarks

In many applications A is specified only on $\mathcal{S}$ (or we just do not want to compute the entries on the complement of $\mathcal{S}$). Consequently, $F = (D_{ME}^{-1}D)L_{ME}^{-1}$, the matrix that minimizes $\|I - FL\|$ when $\mathcal{S}$ is monotone transitive, is uncomputable — $D = \underset{\equiv 0}{\mathrm{P}} L$, and L is unknown, because A is only partially specified. From Proposition 2.2 and Theorem 2.1(2), however, we see that

$$\|I - L_{ME}^{-1}L\|^2 - \|I - \left((D_{ME}^{-1}D)L_{ME}^{-1}\right)L\|^2 = \mathrm{tr}\left((I - D_{ME}^{-1}D)^2\right),$$

which tends to zero as $D_{ME}^{-1}D$ approaches I. By Theorem 2.1(2) this happens when $(D_{ME}^{-1}D)L_{ME}^{-1}$ and L^{-1} are close, and it follows that in that case L_{ME}^{-1} is a suboptimal solution to

$$\min\{\, \|I - FL\| \mid F \in \mathcal{F} \,\}.$$

When $\mathcal{S}$ is arbitrary, the problem of determining F such that $\|I - FL\|$ is minimal and F^*F is zero on the complement of $\mathcal{S}$ is still an open problem. We illustrate the above with an example.

Example 2.3 Suppose that

$$A = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix},$$

where α lies between -1 and 1, and let $\mathcal{S} = \{\,(1,1),(2,2)\,\}$. Then,

$$L = \begin{bmatrix} \sqrt{1-\alpha^2} & \alpha \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} \sqrt{1-\alpha^2} & 0 \\ 0 & 1 \end{bmatrix},$$

the maximum entropy extension of the part of A with support on $\mathcal{S}$ is given by

$$A_{ME} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and

$$L_{ME} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad D_{ME} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Furthermore,

$$\|I - \left((D_{ME}^{-1}D)L_{ME}^{-1}\right) L\| = |\alpha|$$

and

$$\|I - L_{ME}^{-1}L\| = \sqrt{2\left(1 - \sqrt{1-\alpha^2}\right)},$$

and it follows that

$$\|I - L_{ME}^{-1}L\|^2 - \|I - \left((D_{ME}^{-1}D)L_{ME}^{-1}\right) L\|^2 = \left(1 - \sqrt{1-\alpha^2}\right)^2,$$

which tends to zero as α approaches zero — the suboptimal solution L_{ME}^{-1} gets better when

$$(D_{ME}^{-1}D)L_{ME}^{-1} = \begin{bmatrix} \sqrt{1-\alpha^2} & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad L^{-1} = \begin{bmatrix} \frac{1}{\sqrt{1-\alpha^2}} & \frac{-\alpha}{\sqrt{1-\alpha^2}} \\ 0 & 1 \end{bmatrix}$$

become closer. ■

The following chapters (Chapters 3–6) are devoted to algorithms for computing L_{ME}^{-1}. In the next chapter we show that if $\mathcal{S}$ is a block band, then L_{ME}^{-1} *can be found by solving a set of linear equations. When* $\mathcal{S}$ *is staircase, we* can obtain this factor by using the Schur algorithm, which is described in the subsequent chapter. For the general case we depend on iterative algorithms for computing A_{ME}^{-1}. Two of them are described in the following chapter. As these algorithms consume much time and storage, the last chapter describes an algorithm for computing an approximation to A_{ME}^{-1} when $\mathcal{S}$ is a multiple band.

Chapter 3

The Wiener-Hopf Factorization

SUPPOSE THAT A is a positive definite matrix that is specified on a block band $\mathcal{S}$. In this chapter we show that L_{ME}^{-*} and M_{ME}^{-1} can be found by solving a set of linear equations. We present a generalization of the Wiener-Hopf factorization theory to the case of general, finite dimensional, positive definite matrices that are specified on a block band, and derive a global solution to a generalized inverse scattering problem, which turns out to be equivalent to the maximum entropy extension problem. The solution and, consequently, the triangular factors of the inverse of the maximum entropy extension are obtained by solving a set of linear equations. We also give a linear fractional description of all contractive extensions of S (the scattering matrix related to A).

We embed our matrices in doubly infinite ones — in fact, doubly infinite zero and identity matrices — and think of them as operators on l^2, the Hilbert space of doubly infinite sequences with quadratic norm. Operations with matrices of this type are bounded, and coincide with well-defined actions of operators on l^2. We embed A in a doubly infinite identity matrix, and

denote the embedding by $\bar{A}$:

$$(\bar{A})_{ij} = \begin{cases} a_{ij} & \text{if } i,j = 1,\ldots,n; \\ \delta_{ij} & \text{otherwise,} \end{cases}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j; \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, we embed G, L, M and A_{ME} and G_{ME}, L_{ME}, M_{ME} in doubly infinite identity matrices. S and S_{ME} are embedded in doubly infinite zero matrices. The symbols $\bar{0}$ and $\bar{I}$ denote the doubly infinite zero matrix and the doubly infinite identity matrix. The symbol $\bar{J}$ denotes the matrix

$$\bar{J} = \begin{bmatrix} \bar{I} & \bar{0} \\ \bar{0} & -\bar{I} \end{bmatrix}.$$

With the notation developed so far we have $\bar{A} = \frac{1}{2}(\bar{G} + \bar{G}^*) = \bar{L}\bar{L}^* = \bar{M}^*\bar{M}$ and $\bar{S} = (\bar{G} + \bar{I})^{-1}(\bar{G} - \bar{I})$. Similar relations hold for the doubly infinite matrices related to A_{ME}.

We write $\bar{A}$ as a power series of a unitary shift matrix $\bar{Z}$:

$$\bar{A} = \sum_{k=-\infty}^{\infty} \bar{A}_k \bar{Z}^k,$$

where the $\bar{A}_k$'s are diagonal matrices and

$$\bar{Z} = \begin{bmatrix} \ddots & \ddots & \ddots & & \\ \ddots & \boxed{0} & 1 & 0 & \\ \ddots & 0 & 0 & 1 & \ddots \\ & 0 & 0 & 0 & \ddots \\ & & \ddots & \ddots & \ddots \end{bmatrix}$$

(the box marks the 11-entry of the matrix). For convenience' sake, Table 3.1 gives the directions in which the entries of a matrix are shifted when it is

	$\bar{I}$	$\bar{Z}$	$\bar{Z}^*$
$\bar{I}$	$\cdot$	$\rightarrow$	$\leftarrow$
$\bar{Z}$	$\uparrow$	$\nearrow$	$\nwarrow$
$\bar{Z}^*$	$\downarrow$	$\searrow$	$\swarrow$

Table 3.1: Directions in which the entries of a matrix are shifted when it is pre- or postmultiplied by $\bar{Z}$.

pre- or postmultiplied by $\bar{Z}$. For example, the $(2,3)$ entry gives the result of premultiplication by $\bar{Z}$ and postmultiplication by $\bar{Z}^*$.

We denote the space of bounded operators on l^2 by $\mathcal{L}$. The symbols $\mathcal{H}$ and $\mathcal{K}$ denote the subspaces of $\mathcal{L}$ whose operators have power series representations with vanishing coefficients of strictly negative and strictly positive powers of $\bar{Z}$ respectively. An operator in $\mathcal{H}$ is called upper triangular, one in $\mathcal{K}$ lower triangular. In this chapter $\underline{\underline{P}}$ and $\underline{\underline{Q}}$ denote the operators that project an operator in $\mathcal{L}$ on $\mathcal{H}$ and $\mathcal{K}$ respectively. For the case of general operators on l^2 these definitions have to be extended, but that is not necessary here.

3.1 The Factorization

First, some remarks. The maximum entropy property in Lemma 3.1, Theorem 3.1, and Corollary 3.1 — they deal with matrices related to A_{ME} — is not essential. We can state these results for any positive definite extension of A, but have chosen for this approach to economize on notation. Furthermore, all propositions for embeddings also hold for the corresponding finite dimensional matrices.

Lemma 3.1 *Let A be a positive definite matrix that is specified on a set $\mathcal{S}$ that contains the diagonal pairs, but is arbitrary otherwise. Then, there are unique matrices $\bar{B}$, $\bar{C}$, $\bar{D}$, $\bar{E}$ such that*

$$\begin{aligned}\begin{bmatrix} \bar{I} & \bar{S}_{ME} \\ \bar{S}_{ME}^* & \bar{I} \end{bmatrix} &= \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{C} & \bar{I} \end{bmatrix}^{-*} \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{C} & \bar{I} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \bar{I} & \bar{D} \\ \bar{0} & \bar{E} \end{bmatrix}^{-*} \begin{bmatrix} \bar{I} & \bar{D} \\ \bar{0} & \bar{E} \end{bmatrix}^{-1},\end{aligned} \tag{3.1}$$

where (1) $\bar{B}$, $\bar{E}$ and $\bar{C}$, $\bar{D}$ are embeddings of finite dimensional matrices in doubly infinite identity and zero matrices respectively, (2) $\bar{B} \in \mathcal{K}$ and $\bar{E} \in \mathcal{H}$, and (3) $\bar{B}$ and $\bar{E}$ have positive diagonal entries.

Proof Because S_{ME} is contractive, there are unique matrices B, C, D, E such that

$$\begin{aligned}\begin{bmatrix} I & S_{ME} \\ S_{ME}^* & I \end{bmatrix}^{-1} &= \begin{bmatrix} B & 0 \\ C & I \end{bmatrix} \begin{bmatrix} B & 0 \\ C & I \end{bmatrix}^* \\ &= \begin{bmatrix} I & D \\ 0 & E \end{bmatrix} \begin{bmatrix} I & D \\ 0 & E \end{bmatrix}^*,\end{aligned} \tag{3.2}$$

where B is lower triangular, E upper triangular, and B and E have positive diagonal entries. When we embed S_{ME} and C and D in doubly infinite zero matrices and B and E in doubly infinite identity matrices, we obtain

$$\begin{aligned}\begin{bmatrix} \bar{I} & \bar{S}_{ME} \\ \bar{S}_{ME}^* & \bar{I} \end{bmatrix}^{-1} &= \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{C} & \bar{I} \end{bmatrix} \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{C} & \bar{I} \end{bmatrix}^* \\ &= \begin{bmatrix} \bar{I} & \bar{D} \\ \bar{0} & \bar{E} \end{bmatrix} \begin{bmatrix} \bar{I} & \bar{D} \\ \bar{0} & \bar{E} \end{bmatrix}^*.\end{aligned}$$

The inverses of the factors in this equation reduce to the inverses of the related finite dimensional matrices, and, consequently, are lower and upper triangular respectively (with positive diagonal entries). We next prove the

uniqueness of the first factorization — the uniqueness of the second is proved in a similar way.

Suppose that

$$\begin{bmatrix} \bar{I} & \bar{S}_{ME} \\ \bar{S}_{ME}^* & \bar{I} \end{bmatrix}^{-1} = \bar{F}\bar{F}^* = \bar{H}\bar{H}^*,$$

where both $\bar{F}$ and $\bar{H}$ and their inverses are lower triangular with positive diagonal entries. Because $\bar{H}^{-1}\bar{F} = \bar{H}^*\bar{F}^{-*}$, and $\bar{H}^{-1}\bar{F}$ and $\bar{H}^*\bar{F}^{-*}$ are lower and upper triangular respectively, we have that $\bar{H}^{-1}\bar{F}$ is diagonal. In fact, $\bar{H}^{-1}\bar{F}$ is equal to $\bar{I}$ — its diagonal entries are positive, because they evaluate to the products of the diagonal entries in $\bar{H}^{-1}$ and $\bar{F}$, which have positive diagonal entries, and can only be equal to one, since $\bar{H}^{-1}\bar{F} = (\bar{H}^{-1}\bar{F})^{-*}$ — so that $\bar{F} = \bar{H}$. The above suffices to prove uniqueness for our case, where infinite dimensional matrices are embeddings of finite dimensional matrices in doubly infinite zero or identity matrices. ■

A more general case is treated in [RR85].

Theorem 3.1 *Suppose that A is a positive definite matrix that is specified on a set $\mathcal{S}$ that contains the diagonal pairs, but is arbitrary otherwise, and let $\bar{B}$, $\bar{C}$, $\bar{D}$, $\bar{E}$ be as defined in Equation 3.1, that is,*

$$\begin{aligned} \begin{bmatrix} \bar{I} & \bar{S}_{ME} \\ \bar{S}_{ME}^* & \bar{I} \end{bmatrix} &= \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{C} & \bar{I} \end{bmatrix}^{-*} \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{C} & \bar{I} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \bar{I} & \bar{D} \\ \bar{0} & \bar{E} \end{bmatrix}^{-*} \begin{bmatrix} \bar{I} & \bar{D} \\ \bar{0} & \bar{E} \end{bmatrix}^{-1}, \end{aligned}$$

where (1) $\bar{B}$, $\bar{E}$ and $\bar{C}$, $\bar{D}$ are embeddings of finite dimensional matrices in doubly infinite identity and zero matrices respectively, (2) $\bar{B} \in \mathcal{K}$ and $\bar{E} \in \mathcal{H}$, and (3) $\bar{B}$ and $\bar{E}$ have positive diagonal entries. Furthermore, assume that

$$\bar{\Theta} = \begin{bmatrix} \bar{B} & \bar{D} \\ \bar{C} & \bar{E} \end{bmatrix}. \tag{3.3}$$

Then, $\bar{\Theta}$

1. is $\bar{J}$-unitary;

2. can be decomposed as

$$\bar{\Theta} = \begin{bmatrix} \frac{1}{2}(\bar{G}_{ME} + \bar{I})^* & -\frac{1}{2}(\bar{G}_{ME} - \bar{I}) \\ -\frac{1}{2}(\bar{G}_{ME} - \bar{I})^* & \frac{1}{2}(\bar{G}_{ME} + \bar{I}) \end{bmatrix} \begin{bmatrix} \bar{L}_{ME}^{-*} & \bar{0} \\ \bar{0} & \bar{M}_{ME}^{-1} \end{bmatrix}.$$

Proof The first assertion follows from the various relationships between $\bar{B}$, $\bar{C}$, $\bar{D}$, $\bar{E}$. By the same relations, and using those between $\bar{G}_{ME}$ and $\bar{S}_{ME}$, we obtain

$$\bar{\Theta} = \begin{bmatrix} \bar{I} & \bar{D}\bar{E}^{-1} \\ \bar{C}\bar{B}^{-1} & \bar{I} \end{bmatrix} \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{0} & \bar{E} \end{bmatrix} = \begin{bmatrix} \bar{I} & -\bar{S}_{ME} \\ -\bar{S}_{ME}^* & \bar{I} \end{bmatrix} \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{0} & \bar{E} \end{bmatrix} =$$

$$\begin{bmatrix} \frac{1}{2}(\bar{G}_{ME} + \bar{I})^* & -\frac{1}{2}(\bar{G}_{ME} - \bar{I}) \\ -\frac{1}{2}(\bar{G}_{ME} - \bar{I})^* & \frac{1}{2}(\bar{G}_{ME} + \bar{I}) \end{bmatrix} \begin{bmatrix} 2(\bar{G}_{ME} + \bar{I})^{-*}\bar{B} & 0 \\ 0 & 2(\bar{G}_{ME} + \bar{I})^{-1}\bar{E} \end{bmatrix},$$

and because $\frac{1}{2}(\bar{G}_{ME} + \bar{I})\bar{B}^{-*}$ is upper triangular with positive diagonal entries, and

$$\left(\frac{1}{2}(\bar{G}_{ME} + \bar{I})\bar{B}^{-*}\right)\left(\frac{1}{2}(\bar{G}_{ME} + \bar{I})\bar{B}^{-*}\right)^* =$$

$$\frac{1}{2}(\bar{G}_{ME} + \bar{I})(\bar{I} - \bar{S}_{ME}\bar{S}_{ME}^*)\frac{1}{2}(\bar{G}_{ME} + \bar{I})^* = \frac{1}{2}(\bar{G}_{ME} + \bar{G}_{ME}^*) = \bar{A}_{ME},$$

we have $2(\bar{G}_{ME} + \bar{I})^{-*}\bar{B} = \bar{L}_{ME}^{-*}$. In a similar way, $2(\bar{G}_{ME} + \bar{I})^{-1}\bar{E} = \bar{M}_{ME}^{-1}$. ■

Corollary 3.1 *Let $\bar{\Theta}$ be as defined in Equation 3.3. Then,*

$$[\bar{I}\ \bar{I}]\bar{\Theta} = [(\bar{B} + \bar{C})\,(\bar{D} + \bar{E})] = [\bar{L}_{ME}^{-*}\ \bar{M}_{ME}^{-1}].$$

Proof The proof of the corollary follows directly from Theorem 3.1(2). ■

In the following results we use the fact that if A is specified on a staircase band $\mathcal{S}$, then the related scattering matrix S is specified on the strictly upper triangular part of $\mathcal{S}$, and vice-versa. This can be seen from the relations

$$S = (G + I)^{-1}(G - I)$$

and

$$G = (I + S)(I - S)^{-1},$$

from which it follows that entries in S on the strictly upper triangular part of $\mathcal{S}$ are exclusively dependent on entries in G on that same part of $\mathcal{S}$, and conversely. Furthermore, as G and G_{ME} coincide on the strictly upper triangular part of $\mathcal{S}$, this also implies that S and S_{ME} coincide on that part of the band.

Lemma 3.2 *Let A be a positive definite matrix that is specified on a staircase band $\mathcal{S}$. Then, the maximum entropy extension of*

$$\begin{bmatrix} I & S \\ S^* & I \end{bmatrix}$$

(in which the entries in S on the upper triangular part of the complement of $\mathcal{S}$ are unspecified) is equal to

$$\begin{bmatrix} I & S_{ME} \\ S_{ME}^* & I \end{bmatrix}.$$

Proof Let B and C be as defined in Equation 3.2. By Corollary 3.1 we have $B + C = L_{ME}^{-*}$. Because B is lower triangular, and C is equal to $-S_{ME}^* B$ and hence strictly lower triangular — S_{ME}^* is strictly lower triangular — it follows that

$$\det B = \det L_{ME}^{-*}.$$

From Equation 3.2 we obtain

$$\det \begin{bmatrix} I & S_{ME} \\ S_{ME}^* & I \end{bmatrix} = |\det B|^{-2}$$

and hence, by the first equation,

$$\det \begin{bmatrix} I & S_{ME} \\ S_{ME}^* & I \end{bmatrix} = \det A_{ME}.$$

The chain of equalities leading to this result is independent of the maximum entropy property. For every positive definite extension K of A we have likewise that

$$\det \begin{bmatrix} I & S_K \\ S_K^* & I \end{bmatrix} = \det K,$$

and it follows that

$$\max\{ \det \begin{bmatrix} I & S_K \\ S_K^* & I \end{bmatrix} \mid K \in \mathcal{E} \} = \max\{ \det K \mid K \in \mathcal{E} \} = \det A_{ME} =$$

$$\det \begin{bmatrix} I & S_{ME} \\ S_{ME}^* & I \end{bmatrix}.$$

■

Lemma 3.3 *Suppose that $A = [a_{ij}]$ is a positive definite matrix that is specified on a block band $\mathcal{S} = \{ (i,j) \mid |i-j| \leq b \}$, and let $\bar{B}$, $\bar{C}$, $\bar{D}$, $\bar{E}$ be as defined in Equation 3.1. Then,*

$$\begin{array}{llll} \bar{B} \in \mathcal{K} & & \bar{Z}^b \bar{B} & \in \mathcal{H} \\ \bar{C} \in \bar{Z}^{-1}\mathcal{K} & \text{and} & \bar{Z}^b \bar{C} & \in \mathcal{H} \\ \bar{D} \in \bar{Z}\mathcal{H} & & \bar{Z}^{-b} \bar{D} & \in \mathcal{K} \\ \bar{E} \in \mathcal{H} & & \bar{Z}^{-b} \bar{E} & \in \mathcal{K} \end{array}.$$

Proof The left half of the inclusions has already been shown. By Lemma 3.2 the maximum entropy extension of

$$\begin{bmatrix} I & S \\ S^* & I \end{bmatrix}$$

is equal to

$$\begin{bmatrix} I & S_{ME} \\ S_{ME}^* & I \end{bmatrix},$$

and by Equation 3.2

$$\begin{bmatrix} I & S_{ME} \\ S_{ME}^* & I \end{bmatrix}^{-1} = \begin{bmatrix} BB^* & BC^* \\ CB^* & I+CC^* \end{bmatrix} = \begin{bmatrix} I+DD^* & DE^* \\ ED^* & EE^* \end{bmatrix}.$$

From Theorem 1.3 we obtain that BC^* and, consequently, as B is lower triangular, C^* vanish on the upper triangular part of the complement of $\mathcal{S}$, which implies that $\bar{C} \in \bar{Z}^{-b}\mathcal{H}$ and hence $\bar{Z}^b\bar{C} \in \mathcal{H}$. Furthermore, it follows that $I + C^*C$ is zero on the complement of $\mathcal{S}$, and because $I + C^*C = B^*B$, we conclude that B vanishes on the lower triangular part of the complement of the set, so that $\bar{Z}^b\bar{B} \in \mathcal{H}$. Symmetric properties are true for $\bar{D}$ and $\bar{E}$. ■

Theorem 3.2 *Suppose that $A = [a_{ij}]$ is a positive definite matrix that is specified on a block band $\mathcal{S} = \{(i,j) \mid |i-j| \leq b\}$, and let $\bar{\Gamma}_+$ be derived from the related scattering matrix S as $\bar{\Gamma}_+ = \underline{\underline{P}}(\bar{Z}^b\bar{S}^*)$. Then,*

1. *the generalized Wiener-Hopf equations*

$$\begin{cases} \bar{F} + \underline{\underline{P}}(\bar{\Gamma}_+\bar{H}) = \bar{F}_0^{-*} \\ \underline{\underline{Q}}(\bar{\Gamma}_+^*\bar{F}) + \bar{H} = \bar{0} \end{cases} \tag{3.4}$$

$$\begin{cases} \bar{K} + \underline{\underline{P}}(\bar{\Gamma}_+\bar{N}) = \bar{0} \\ \underline{\underline{Q}}(\bar{\Gamma}_+^*\bar{K}) + \bar{N} = \bar{N}_0^{-*} \end{cases}$$

have a unique solution $\{\bar{F}, \bar{H}, \bar{K}, \bar{N}\}$, where (a) $\bar{F}$, $\bar{N}$ and $\bar{H}$, $\bar{K}$ are embeddings of finite dimensional matrices in doubly infinite identity and zero matrices respectively and (b) $\bar{F}$ and $\bar{N}$ have positive diagonal entries ($\bar{F}_0$ and $\bar{N}_0$ are the constant terms in the power series decompositions of $\bar{F}$ and $\bar{N}$);

2. *the $\bar{B}$, $\bar{C}$, $\bar{D}$, $\bar{E}$ defined by*

$$\bar{B} = \bar{N}, \quad \bar{C} = \bar{Z}^{*b}\bar{K}, \quad \bar{D} = \bar{H}\bar{Z}^b, \quad \bar{E} = \bar{Z}^{*b}\bar{F}\bar{Z}^b$$

satisfy Equation 3.1:

$$\begin{bmatrix} \bar{I} & \bar{S}_{ME} \\ \bar{S}^*_{ME} & \bar{I} \end{bmatrix} = \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{C} & \bar{I} \end{bmatrix}^{-*} \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{C} & \bar{I} \end{bmatrix}^{-1} = \begin{bmatrix} \bar{I} & \bar{D} \\ \bar{0} & \bar{E} \end{bmatrix}^{-*} \begin{bmatrix} \bar{I} & \bar{D} \\ \bar{0} & \bar{E} \end{bmatrix}^{-1},$$

where (a) $\bar{B}$, $\bar{E}$ and $\bar{C}$, $\bar{D}$ are embeddings of finite dimensional matrices in doubly infinite identity and zero matrices respectively, (b) $\bar{B} \in \mathcal{K}$ and $\bar{E} \in \mathcal{H}$, and (c) $\bar{B}$ and $\bar{E}$ have positive diagonal entries.

Proof Suppose that $\bar{\Gamma} = \bar{Z}^b \bar{S}^*_{ME}$ and $\bar{\Gamma}_- = (\bar{I} - \underline{\underline{P}})(\bar{Z}^b \bar{S}^*_{ME})$. Because $\bar{\Gamma}_+ = \underline{\underline{P}}(\bar{Z}^b \bar{S}^*) = \underline{\underline{P}}(\bar{Z}^b \bar{S}^*_{ME})$ — $\bar{S}$ and $\bar{S}_{ME}$ coincide on the first b upper diagonals — we have $\bar{\Gamma} = \bar{\Gamma}_- + \bar{\Gamma}_+$. From Equation 3.1 we find, for example, by direct verification, that

$$\begin{aligned} \begin{bmatrix} \bar{I} & \bar{\Gamma} \\ \bar{\Gamma}^* & \bar{I} \end{bmatrix} &= \begin{bmatrix} \bar{0} & \bar{Z}^b \\ \bar{I} & \bar{0} \end{bmatrix} \begin{bmatrix} \bar{I} & \bar{S}_{ME} \\ \bar{S}^*_{ME} & \bar{I} \end{bmatrix} \begin{bmatrix} \bar{0} & \bar{I} \\ \bar{Z}^{*b} & \bar{0} \end{bmatrix} \\ &= \begin{bmatrix} \bar{0} & \bar{Z}^b \\ \bar{I} & \bar{0} \end{bmatrix} \begin{bmatrix} \bar{I} & \bar{D} \\ \bar{0} & \bar{E} \end{bmatrix}^{-*} \begin{bmatrix} \bar{0} & \bar{I} \\ \bar{Z}^{*b} & \bar{0} \end{bmatrix} \\ &\quad \begin{bmatrix} \bar{0} & \bar{Z}^b \\ \bar{I} & \bar{0} \end{bmatrix} \begin{bmatrix} \bar{I} & \bar{D} \\ \bar{0} & \bar{E} \end{bmatrix}^{-1} \begin{bmatrix} \bar{0} & \bar{I} \\ \bar{Z}^{*b} & \bar{0} \end{bmatrix} \\ &= \begin{bmatrix} \bar{Z}^b \bar{E} \bar{Z}^{*b} & \bar{0} \\ \bar{D} \bar{Z}^{*b} & \bar{I} \end{bmatrix}^{-*} \begin{bmatrix} \bar{Z}^b \bar{E} \bar{Z}^{*b} & \bar{0} \\ \bar{D} \bar{Z}^{*b} & \bar{I} \end{bmatrix}^{-1} \end{aligned}$$

and

$$\begin{aligned} \begin{bmatrix} \bar{I} & \bar{\Gamma} \\ \bar{\Gamma}^* & \bar{I} \end{bmatrix} &= \begin{bmatrix} \bar{0} & \bar{Z}^b \\ \bar{I} & \bar{0} \end{bmatrix} \begin{bmatrix} \bar{I} & \bar{S}_{ME} \\ \bar{S}^*_{ME} & \bar{I} \end{bmatrix} \begin{bmatrix} \bar{0} & \bar{I} \\ \bar{Z}^{*b} & \bar{0} \end{bmatrix} \\ &= \begin{bmatrix} \bar{0} & \bar{Z}^b \\ \bar{I} & \bar{0} \end{bmatrix} \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{C} & \bar{I} \end{bmatrix}^{-*} \begin{bmatrix} \bar{0} & \bar{I} \\ \bar{Z}^{*b} & \bar{0} \end{bmatrix} \\ &\quad \begin{bmatrix} \bar{0} & \bar{Z}^b \\ \bar{I} & \bar{0} \end{bmatrix} \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{C} & \bar{I} \end{bmatrix}^{-1} \begin{bmatrix} \bar{0} & \bar{I} \\ \bar{Z}^{*b} & \bar{0} \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} \bar{I} & \bar{Z}^b\bar{C} \\ \bar{0} & \bar{B} \end{bmatrix}^{-*} \begin{bmatrix} \bar{I} & \bar{Z}^b\bar{C} \\ \bar{0} & \bar{B} \end{bmatrix}^{-1},$$

and with

$$\bar{F} = \bar{Z}^b\bar{E}\bar{Z}^{*b}, \quad \bar{H} = \bar{D}\bar{Z}^{*b}, \quad \bar{K} = \bar{Z}^b\bar{C}, \quad \bar{N} = \bar{B}$$

we have

$$\begin{bmatrix} \bar{I} & \bar{\Gamma} \\ \bar{\Gamma}^* & \bar{I} \end{bmatrix} \begin{bmatrix} \bar{F} \\ \bar{H} \end{bmatrix} = \begin{bmatrix} \bar{F}^{-*} \\ \bar{0} \end{bmatrix} \tag{3.5}$$

$$\begin{bmatrix} \bar{I} & \bar{\Gamma} \\ \bar{\Gamma}^* & \bar{I} \end{bmatrix} \begin{bmatrix} \bar{K} \\ \bar{N} \end{bmatrix} = \begin{bmatrix} \bar{0} \\ \bar{N}^{-*} \end{bmatrix}.$$

Projecting the equations on the first rows of Equation 3.5 on $\mathcal{H}$, and those on the second rows on $\mathcal{K}$, we obtain

$$\begin{cases} \underline{\underline{P}}\bar{F} + \underline{\underline{P}}(\bar{\Gamma}\ \bar{H}) = \underline{\underline{P}}\bar{F}^{-*} \\ \underline{\underline{Q}}(\bar{\Gamma}^*\bar{F}) + \underline{\underline{Q}}\bar{H} = \bar{0} \end{cases} \tag{3.6}$$

$$\begin{cases} \underline{\underline{P}}\bar{K} + \underline{\underline{P}}(\bar{\Gamma}\bar{N}) = \bar{0} \\ \underline{\underline{Q}}(\bar{\Gamma}^*\bar{K}) + \underline{\underline{Q}}\bar{N} = \underline{\underline{Q}}\bar{N}^{-*} \end{cases}.$$

From the definition of $\bar{F}$, $\bar{H}$, $\bar{K}$, $\bar{N}$ and Lemma 3.3 we have $\bar{F} \in \mathcal{H}$, $\bar{H} \in \mathcal{K}$, $\bar{K} \in \mathcal{H}$, $\bar{N} \in \mathcal{K}$, so that

$$\underline{\underline{P}}\bar{F} = \bar{F}, \quad \underline{\underline{Q}}\bar{H} = \bar{H}, \quad \underline{\underline{P}}\bar{K} = \bar{K}, \quad \underline{\underline{Q}}\bar{N} = \bar{N}.$$

Furthermore,

$$\begin{aligned} \underline{\underline{P}}(\bar{\Gamma}\bar{H}) &= \underline{\underline{P}}\left((\bar{\Gamma}_- + \bar{\Gamma}_+)\bar{H}\right) = \underline{\underline{P}}(\bar{\Gamma}_+\bar{H}) \\ \underline{\underline{Q}}(\bar{\Gamma}^*\bar{F}) &= \underline{\underline{Q}}\left((\bar{\Gamma}_-^* + \bar{\Gamma}_+^*)\bar{F}\right) = \underline{\underline{Q}}(\bar{\Gamma}_+^*\bar{F}) \end{aligned}$$

and $\underline{\underline{P}}\bar{F}^{-*} = \bar{F}_0^{-*}$, and similarly,

$$\begin{aligned} \underline{\underline{P}}(\bar{\Gamma}\bar{N}) &= \underline{\underline{P}}\left((\bar{\Gamma}_- + \bar{\Gamma}_+)\bar{N}\right) = \underline{\underline{P}}(\bar{\Gamma}_+\bar{N}) \\ \underline{\underline{Q}}(\bar{\Gamma}^*\bar{K}) &= \underline{\underline{Q}}\left((\bar{\Gamma}_-^* + \bar{\Gamma}_+^*)\bar{K}\right) = \underline{\underline{Q}}(\bar{\Gamma}_+^*\bar{K}) \end{aligned}$$

and $\underline{\underline{Q}}\bar{N}^{-*} = \bar{N}_0^{-*}$. Substituting these identities in Equation 3.6, we obtain Equation 3.4. Uniqueness follows by remarking that Equation 3.4 is actually equivalent to Equation 3.5, which in turn is equivalent to Equation 3.1. ■

The Wiener-Hopf technique of Theorem 3.2 is due to Dym and Gohberg [DG79] (see also [DG88]). A general scheme for dealing with contractive and positive extension problems, which covers the Dym-Gohberg approach, is described in [GKW89a, GKW89b] and [Woe89].

We next show how Theorem 3.2 leads to a set of linear equations that is based on a Hankel matrix whose entries are equal to the diagonals of $\bar{\Gamma}_+$ (or, equivalently, those of $\bar{S}$).

Corollary 3.2 *Suppose that A and $\bar{\Gamma}_+ = \underline{\underline{P}}(\bar{Z}^b \bar{S}^*)$ are as defined in Theorem 3.2, and let*

$$\bar{\Gamma}_+ = \sum_{k=0}^{m} \bar{\Gamma}_k \bar{Z}^k$$

and

$$\mathbf{\Gamma} = \begin{bmatrix} \bar{\Gamma}_0 & \bar{\Gamma}_1 \bar{Z} & \cdots & \bar{\Gamma}_m \bar{Z}^m \\ \bar{\Gamma}_1 \bar{Z} & & \iddots & \\ \vdots & \iddots & & \\ \bar{\Gamma}_m \bar{Z}^m & & & 0 \end{bmatrix},$$

where $m = b - 1$ (and $\bar{\Gamma}_0 = \bar{S}_b^*, \bar{\Gamma}_1 \bar{Z} = \bar{Z}\bar{S}_{b-1}^*, \ldots, \bar{\Gamma}_m \bar{Z}^m = \bar{Z}^{b-1}\bar{S}_1^*$). Then,

1. *$\mathbf{\Gamma}$ is contractive;*
2. *the sets of linear equations*

$$\begin{bmatrix} \mathbf{I} & \mathbf{\Gamma} \\ \mathbf{\Gamma}^* & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{F} \\ \mathbf{H} \end{bmatrix} = \begin{bmatrix} \mathbf{F}_0 \\ \mathbf{0} \end{bmatrix} \tag{3.7}$$

$$\begin{bmatrix} \mathbf{I} & \mathbf{\Gamma} \\ \mathbf{\Gamma}^* & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{K} \\ \mathbf{N} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{N}_0 \end{bmatrix}$$

have a unique solution $\{\mathbf{F}, \mathbf{H}, \mathbf{K}, \mathbf{N}\}$, *where (a)*

$$\mathbf{F} = \begin{bmatrix} \bar{F}_0 \\ \bar{F}_1 \bar{Z} \\ \vdots \\ \bar{F}_m \bar{Z}^m \end{bmatrix}, \quad \mathbf{F}_0 = \begin{bmatrix} \bar{F}_0^{-*} \\ \bar{0} \\ \vdots \\ \bar{0} \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} \bar{H}_0 \\ \bar{H}_{-1} \bar{Z}^* \\ \vdots \\ \bar{H}_{-m} \bar{Z}^{*m} \end{bmatrix},$$

$$\mathbf{K} = \begin{bmatrix} \bar{K}_0 \\ \bar{K}_1 \bar{Z} \\ \vdots \\ \bar{K}_m \bar{Z}^m \end{bmatrix}, \quad \mathbf{N} = \begin{bmatrix} \bar{N}_0 \\ \bar{N}_{-1} \bar{Z}^* \\ \vdots \\ \bar{N}_{-m} \bar{Z}^{*m} \end{bmatrix}, \quad \mathbf{N}_0 = \begin{bmatrix} \bar{N}_0^{-*} \\ \bar{0} \\ \vdots \\ \bar{0} \end{bmatrix},$$

(b) the $\bar{F}_k$*'s,* $\bar{N}_k$*'s and* $\bar{H}_k$*'s,* $\bar{K}_k$*'s are embeddings of finite dimensional diagonal matrices in doubly infinite identity and zero matrices respectively, and (c)* $\bar{F}_0$ *and* $\bar{N}_0$ *have positive diagonal entries;*

3. *the* $\bar{B}$, $\bar{C}$, $\bar{D}$, $\bar{E}$ *defined by*

$$\bar{B} = \bar{N}, \quad \bar{C} = \bar{Z}^{*b} \bar{K}, \quad \bar{D} = \bar{H} \bar{Z}^b, \quad \bar{E} = \bar{Z}^{*b} \bar{F} \bar{Z}^b,$$

where

$$\bar{N} = \sum_{k=-m}^{0} \bar{N}_k \bar{Z}^k, \quad \bar{K} = \sum_{k=0}^{m} \bar{K}_k \bar{Z}^k, \quad \bar{H} = \sum_{k=-m}^{0} \bar{H}_k \bar{Z}^k, \quad \bar{F} = \sum_{k=0}^{m} \bar{F}_k \bar{Z}^k,$$

satisfy Equation 3.1.

Proof The proof of the theorem follows when we split Equations 3.4 and 3.5 into separate equations for each power of $\bar{Z}$. The operator that corresponds to $\bar{\Gamma}$ in Equation 3.5 is contractive — $\bar{\Gamma}$ is contractive. Because $\mathbf{\Gamma}$ is a block in this operator, it follows that $\mathbf{\Gamma}$ is contractive as well. This in turn implies that Equation 3.7 has a unique solution. ■

Because we have embedded our matrices in (doubly infinite) zero and identity matrices, Equation 3.7 reduces to a finite number of equations.

Corollary 3.3 *Let A and $\bar{\Theta}$ be as defined in Theorem 3.2 and Equation 3.3. Then, $\bar{\Theta}$ and, consequently, $\bar{L}_{ME}^{-*}$ and $\bar{M}_{ME}^{-1}$ are uniquely determined by Equation 3.7.*

Proof The proof of the corollary follows directly from Corollaries 3.2 and 3.1. ■

Corollaries 3.2 and 3.3 imply that to determine the factorization

$$\begin{aligned}\begin{bmatrix} \bar{I} & \bar{S}_{ME} \\ \bar{S}_{ME}^{*} & \bar{I} \end{bmatrix} &= \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{C} & \bar{I} \end{bmatrix}^{-*} \begin{bmatrix} \bar{B} & \bar{0} \\ \bar{C} & \bar{I} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \bar{I} & \bar{D} \\ \bar{0} & \bar{E} \end{bmatrix}^{-*} \begin{bmatrix} \bar{I} & \bar{D} \\ \bar{0} & \bar{E} \end{bmatrix}^{-1},\end{aligned}$$

where (1) $\bar{B}$, $\bar{E}$ and $\bar{C}$, $\bar{D}$ are embeddings of finite dimensional matrices in doubly infinite identity and zero matrices respectively, (2) $\bar{B} \in \mathcal{K}$ and $\bar{E} \in \mathcal{H}$, and (3) $\bar{B}$ and $\bar{E}$ have positive diagonal entries, and, consequently, to obtain

$$\bar{\Theta} = \begin{bmatrix} \bar{B} & \bar{D} \\ \bar{C} & \bar{E} \end{bmatrix}$$

and $\bar{L}_{ME}^{-*}$ and $\bar{M}_{ME}^{-1}$, we only need to know the first b upper diagonals of $\bar{S}_{ME}$ (which are equal to those of $\bar{S}$). We illustrate the above with a simple example.

Example 3.1 Suppose that

$$A = \begin{bmatrix} 1 & \alpha & ? \\ \alpha & 1 & \alpha \\ ? & \alpha & 1 \end{bmatrix},$$

where α lies between -1 and 1, and let $\mathcal{S} = \{ (i,j) \mid |i-j| \leq 1 \}$. Then,

$$S = \begin{bmatrix} 0 & \alpha & ? \\ 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \bar{\Gamma}_+ = \begin{bmatrix} \ddots & \ddots & \ddots & & \\ \ddots & \boxed{\alpha} & 0 & 0 & \\ \ddots & 0 & \alpha & 0 & \ddots \\ & 0 & 0 & 0 & \ddots \\ & & \ddots & \ddots & \ddots \end{bmatrix},$$

and the second set of equations in Equation 3.7 reduces to

$$\begin{bmatrix} 1 & 0 & 0 & \alpha & 0 & 0 \\ 0 & 1 & 0 & 0 & \alpha & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \alpha & 0 & 0 & 1 & 0 & 0 \\ 0 & \alpha & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} k_1 & 0 & 0 \\ 0 & k_2 & 0 \\ 0 & 0 & k_3 \\ n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ n_1^{-*} & 0 & 0 \\ 0 & n_2^{-*} & 0 \\ 0 & 0 & n_3^{-*} \end{bmatrix},$$

which has a unique solution, because the coefficient matrix is positive definite — $-1 < \alpha < 1$. It follows that

$$N = \begin{bmatrix} \frac{1}{\sqrt{1-\alpha^2}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{1-\alpha^2}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad K = \begin{bmatrix} \frac{-\alpha}{\sqrt{1-\alpha^2}} & 0 & 0 \\ 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

and, indeed, $\bar{N} + \bar{Z}^* \bar{K}$ is equal to the embedding of

$$L_{ME}^{-*} = \begin{bmatrix} \frac{1}{\sqrt{1-\alpha^2}} & 0 & 0 \\ \frac{-\alpha}{\sqrt{1-\alpha^2}} & \frac{1}{\sqrt{1-\alpha^2}} & 0 \\ 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} & 1 \end{bmatrix}.$$

A_{ME}^{-1} and A_{ME} evaluate to

$$A_{ME}^{-1} = \begin{bmatrix} \frac{1}{1-\alpha^2} & \frac{-\alpha}{1-\alpha^2} & 0 \\ \frac{-\alpha}{1-\alpha^2} & \frac{1+\alpha^2}{1-\alpha^2} & \frac{-\alpha}{1-\alpha^2} \\ 0 & \frac{-\alpha}{1-\alpha^2} & \frac{1}{1-\alpha^2} \end{bmatrix} \quad \text{and} \quad A_{ME} = \begin{bmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{bmatrix}.$$

■

The $\bar{\Theta}$ matrix occuring in the LIS (Lossless Inverse Scattering) theorem of [DD87] and the one occuring in [DG88] are related in the following way: the former is given by

$$\begin{bmatrix} \bar{B} & \bar{D} \\ \bar{C} & \bar{E} \end{bmatrix},$$

the latter by

$$\begin{bmatrix} \bar{N} & \bar{H} \\ \bar{K} & \bar{F} \end{bmatrix} = \begin{bmatrix} \bar{I} & \bar{0} \\ \bar{0} & \bar{Z}^b \end{bmatrix} \begin{bmatrix} \bar{B} & \bar{D} \\ \bar{C} & \bar{E} \end{bmatrix} \begin{bmatrix} \bar{I} & \bar{0} \\ \bar{0} & \bar{Z}^{*b} \end{bmatrix},$$

where $\bar{B}$, $\bar{C}$, $\bar{D}$, $\bar{E}$ and $\bar{F}$, $\bar{H}$, $\bar{K}$, $\bar{N}$ are as defined in Theorem 3.2.

3.2 A Linear Fractional Description

We now show that if $\mathcal{S}$ is a block band, then

$$\bar{\Theta} = \begin{bmatrix} \bar{B} & \bar{D} \\ \bar{C} & \bar{E} \end{bmatrix}$$

Next, suppose that $\bar{\Gamma}^{-1}\bar{\Delta} = \bar{S}_l \bar{Z}^{b+1}$, where $\bar{S}_l$ is upper triangular. Then,

$$[\bar{I}\ \bar{S}]\bar{\Theta} = \bar{\Gamma}[\bar{I}\ (\bar{\Gamma}^{-1}\bar{\Delta})] = \bar{\Gamma}[\bar{I}\ (\bar{S}_l \bar{Z}^{b+1})].$$

Since $\bar{\Theta}$ and $\bar{Z}$ are $\bar{J}$-unitary and unitary respectively, we have

$$\bar{I} - \bar{S}\bar{S}^* = \bar{\Gamma}\left(\bar{I} - (\bar{\Gamma}^{-1}\bar{\Delta})(\bar{\Gamma}^{-1}\bar{\Delta})^*\right)\bar{\Gamma}^* = \bar{\Gamma}\,(\bar{I} - \bar{S}_l \bar{S}_l^*)\,\bar{\Gamma}^*,$$

and as $\bar{S}$ is contractive and $\bar{\Gamma}$ invertible, we conclude that $\bar{\Gamma}^{-1}\bar{\Delta}$ and $\bar{S}_l$ are contractive. Equation 3.8 now follows by direct calculation.

Because the construction given above is dependent only on the entries in S on the strictly upper triangular part of $\mathcal{S}$, the last assertion in the theorem also holds. ■

We close this section with an example.

Example 3.2 Suppose that

$$A = \begin{bmatrix} 1 & \alpha & ? \\ \alpha & 1 & \alpha \\ ? & \alpha & 1 \end{bmatrix},$$

where α lies between -1 and 1, and let $\mathcal{S} = \{(i,j) \mid |i-j| \leq 1\}$. Then,

$$S = \begin{bmatrix} 0 & \alpha & ? \\ 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$B = \begin{bmatrix} \dfrac{1}{\sqrt{1-\alpha^2}} & 0 & 0 \\ 0 & \dfrac{1}{\sqrt{1-\alpha^2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 & 0 \\ \dfrac{-\alpha}{\sqrt{1-\alpha^2}} & 0 & 0 \\ 0 & \dfrac{-\alpha}{\sqrt{1-\alpha^2}} & 0 \end{bmatrix},$$

$$D = \begin{bmatrix} 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} & 0 \\ 0 & 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} \\ 0 & 0 & 0 \end{bmatrix}, \quad E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{1-\alpha^2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{1-\alpha^2}} \end{bmatrix}.$$

With S_l a contractive, upper triangular matrix of size 3×3 and β the $(1,1)$ entry of this matrix Equation 3.8 reduces to

$$S = \begin{bmatrix} 0 & \alpha & \beta(1-\alpha^2) \\ 0 & 0 & \alpha \\ 0 & 0 & 0 \end{bmatrix}.$$

When S_l ranges over all contractive, upper triangular matrices, β varies between -1 and 1, and we obtain all contractive extensions of S. Using the relations between S and the impedance matrix G, we find that all positive definite extensions of A are given by

$$A = \begin{bmatrix} 1 & \alpha & \alpha^2+\beta(1-\alpha^2) \\ \alpha & 1 & \alpha \\ \alpha^2+\beta(1-\alpha^2) & \alpha & 1 \end{bmatrix}, \quad \text{where} \quad -1 < \beta < 1.$$

The maximum entropy extension results when $\beta = 0$. ■

Chapter 4

The Schur Algorithm

SUPPOSE THAT A is a positive definite matrix that is specified on a staircase band $\mathcal{S}$. In this case we can use the Schur algorithm presented in [DD87] to compute L_{ME}^{-*} and M_{ME}^{-1}. The algorithm requires $O(nb^2)$ operations and $O(nb)$ storage, where n is the size of the matrix and b the average width of the band, and is very well suited for implementation on an array processor of the systolic or wavefront type. In this chapter we review the algorithm, and give an estimate of $\|I - L_{ME}^{-1}L\|$.

4.1 The Algorithm

The Schur algorithm is based on the following results of Dewilde and Deprettere.

Theorem 4.1 ([DD87]) *Suppose that $A = [a_{ij}]$, $i, j = 1, \ldots, n$, is a positive definite matrix that is specified on a staircase band $\mathcal{S}$, and let $\Gamma = \frac{1}{2}(G + I)$ and $\Delta = \frac{1}{2}(G - I)$. Then,*

1. *there are elementary J-unitary matrices $\Theta_1, \ldots, \Theta_m$ such that*

$$[\Gamma\ \Delta]\Theta_1 \cdots \Theta_m = [\Gamma_m\ \Delta_m], \tag{4.1}$$

where Γ_m and Δ_m are upper triangular and Δ_m vanishes on the upper triangular part of $\mathcal{S}$;

2. *the product* $\Theta = \Theta_1 \cdots \Theta_m$ *can be decomposed as*

$$\Theta = \begin{bmatrix} \frac{1}{2}(G_{ME}+I)^* & -\frac{1}{2}(G_{ME}-I) \\ -\frac{1}{2}(G_{ME}-I)^* & \frac{1}{2}(G_{ME}+I) \end{bmatrix} \begin{bmatrix} L_{ME}^{-*} & 0 \\ 0 & M_{ME}^{-1} \end{bmatrix}. \quad (4.2)$$

Proof The existence of $\Theta_1, \ldots, \Theta_m$ is proved by induction. To begin with, $[\Gamma\ \Delta]$ is admissible — Γ and Δ are upper triangular, Γ is invertible, and $\Gamma^{-1}\Delta$ contractive, because $\Gamma\Gamma^* - \Delta\Delta^*$ is equal to A and hence positive definite and Γ is invertible — and Δ is zero on the upper triangular part of a staircase band, namely the diagonal.

Now, suppose that for some k

1. $[\Gamma_{k-1}\ \Delta_{k-1}]$ is admissible;
2. Δ_{k-1} vanishes on the upper triangular part of some staircase band $\mathcal{S}_{k-1}$.

Assume that $\mathcal{S}_k = \mathcal{S}_{k-1} \cup \{(i,j),(j,i)\}$, where (i,j) is such that $\mathcal{S}_k$ is staircase, and let $i \leq j$ and

$$\Theta_k = \Theta(i, n+j, \rho_{ij}) \text{ with } \rho_{ij} = -\frac{(\delta_{k-1})_{ij}}{(\gamma_{k-1})_{ii}},$$

where $|\rho_{ij}| < 1$, because $(\delta_{k-1})_{ij}/(\gamma_{k-1})_{ii}$ is the (i,j) entry of $\Gamma_{k-1}^{-1}\Delta_{k-1}$, which is contractive, as we have assumed that $[\Gamma_{k-1}\ \Delta_{k-1}]$ is admissible, and the modulus of an entry of a contractive matrix is smaller than one. Furthermore, let

$$[\Gamma_{k-1}\ \Delta_{k-1}]\Theta_k = [\Gamma_k\ \Delta_k].$$

Then,

1. $[\Gamma_k\ \Delta_k]$ is admissible — Γ_k and Δ_k are upper triangular, Γ_k is invertible, and $\Gamma_k^{-1}\Delta_k$ contractive, because $\Gamma_k\Gamma_k^* - \Delta_k\Delta_k^*$ is equal to

$$\begin{aligned} [\Gamma_k\ \Delta_k]J[\Gamma_k\ \Delta_k]^* &= [\Gamma_{k-1}\ \Delta_{k-1}]\Theta_k J \Theta_k^* [\Gamma_{k-1}\ \Delta_{k-1}]^* \\ &= [\Gamma_{k-1}\ \Delta_{k-1}]J[\Gamma_{k-1}\ \Delta_{k-1}]^* \\ &= \Gamma_{k-1}\Gamma_{k-1}^* - \Delta_{k-1}\Delta_{k-1}^* \end{aligned}$$

and hence positive definite, as we have assumed that $[\Gamma_{k-1}\ \Delta_{k-1}]$ is admissible, and Γ_k is invertible;

2. Δ_k is zero on the upper triangular part of $\mathcal{S}_k$ — $(\delta_{k-1})_{ij}$ is eliminated, and no fill-ins are produced.

We proceed with the second part of the theorem. Because all Θ_k's are J-unitary, the product $\Theta = \Theta_1 \cdots \Theta_m$ is J-unitary, and by a proof similar to the proof of Theorem 3.1(2), we can decompose it as

$$\Theta = \begin{bmatrix} \frac{1}{2}(G_B + I)^* & -\frac{1}{2}(G_B - I) \\ -\frac{1}{2}(G_B - I)^* & \frac{1}{2}(G_B + I) \end{bmatrix} \begin{bmatrix} L_B^{-*} & 0 \\ 0 & M_B^{-1} \end{bmatrix},$$

where B is some positive definite matrix, so that

$$[I\ I]\Theta = [L_B^{-*}\ M_B^{-1}].$$

With the help of this equation and by direct calculation it is easy to verify that L_B^{-*} and M_B^{-1} vanish on the lower and upper triangular part of the complement of $\mathcal{S}$ respectively.

By Equation 4.1

$$[\frac{1}{2}(G + I)\ \frac{1}{2}(G - I)]\Theta = [\Gamma_m\ \Delta_m],$$

and by the above decomposition of Θ

$$[\frac{1}{2}(G_B + I)\ \frac{1}{2}(G_B - I)]\Theta = [L_B\ 0].$$

Subtracting this equation from the previous one, and using the second equality, we find

$$\frac{1}{2}(G - G_B)[L_B^{-*}\ M_B^{-1}] = [(\Gamma_m - L_B)\ \Delta_m]$$

and hence

$$\frac{1}{2}(G - G_B) = \Delta_m M_B.$$

Since Δ_m is upper triangular and zero on the upper triangular part of $\mathcal{S}$, and M_B is upper triangular, it follows that $\Delta_m M_B$ vanishes on the upper triangular part of $\mathcal{S}$, and hence that G and G_B coincide on that part of the band. For this reason, and because L_B^{-*} and M_B^{-1} vanish on the lower and upper triangular part of the complement of $\mathcal{S}$ respectively, we conclude that $B = A_{ME}$. ■

The Θ matrix in Theorem 4.1 is the same as the one in Theorem 3.1.

Corollary 4.1 ([DD87]) *Let Θ be as defined in Theorem 4.1. Then,*

$$[I\ I]\Theta = [L_{ME}^{-*}\ M_{ME}^{-1}]. \tag{4.3}$$

Proof The proof of the corollary follows directly from Equation 4.2. ■

The Schur algorithm computes L_{ME}^{-*} and M_{ME}^{-1} in two steps. First, it computes the elementary J-unitary matrices $\Theta_1, \ldots, \Theta_m$ defined in Equation 4.1. Starting from $[\Gamma\ \Delta]$, it determines Θ_1 so that δ_{12} is eliminated when $[\Gamma\ \Delta]$ is postmultiplied by Θ_1:

$$\Theta_1 = \Theta(1, n+2, \rho_{12}) \text{ with } \rho_{12} = -\frac{\delta_{12}}{\gamma_{11}}.$$

The result is $[\Gamma\ \Delta]\Theta_1 = [\Gamma_1\ \Delta_1]$, where Γ_1 is upper triangular, Δ_1 strictly upper triangular, and $(\delta_1)_{12} = 0$. Next, it computes Θ_2 so that $(\delta_1)_{23}$ is eliminated when $[\Gamma_1\ \Delta_1]$ is postmultiplied by Θ_2:

$$\Theta_2 = \Theta(2, n+3, \rho_{23}) \text{ with } \rho_{23} = -\frac{(\delta_1)_{23}}{(\gamma_1)_{22}}.$$

The ρ_{ij}'s are called *reflection coefficients* and will be generically attached to A. When for some k the entries on the first upper diagonal of Δ_k have been eliminated, Θ_{k+1} is determined so as to eliminate the first entry on the second upper diagonal, etc. The recursion ends when for some m all entries in Δ_m on the upper triangular part of $\mathcal{S}$ are zero. Next, Equation 4.3 is evaluated — *$[I\ I]$ instead of $[\Gamma\ \Delta]$ is postmultiplied by $\Theta_1 \cdots \Theta_m$. We proceed with an* example.

Example 4.1 Suppose that

$$A = \begin{bmatrix} 1 & \alpha & ? \\ \alpha & 1 & \alpha \\ ? & \alpha & 1 \end{bmatrix},$$

where α lies between -1 and 1, and let $\mathcal{S} = \{(i,j) \mid |i-j| \leq 1\}$. Then,

$$[\Gamma\ \Delta] = \left[\begin{array}{ccc|ccc} 1 & \alpha & ? & 0 & \alpha & ? \\ 0 & 1 & \alpha & 0 & 0 & \alpha \\ 0 & 0 & 1 & 0 & 0 & 0 \end{array}\right],$$

and to eliminate the entries on the first upper diagonal of Δ, we postmultiply $[\Gamma\ \Delta]$ by

$$\Theta_1 = \left[\begin{array}{ccc|ccc} \frac{1}{\sqrt{1-\alpha^2}} & 0 & 0 & 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ \hline \frac{-\alpha}{\sqrt{1-\alpha^2}} & 0 & 0 & 0 & \frac{1}{\sqrt{1-\alpha^2}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

and

$$\Theta_2 = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{1-\alpha^2}} & 0 & 0 & 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} & 0 & 0 & 0 & \frac{1}{\sqrt{1-\alpha^2}} \end{array}\right].$$

The product $\Theta = \Theta_1\Theta_2$ evaluates to

$$\Theta = \left[\begin{array}{ccc|ccc} \frac{1}{\sqrt{1-\alpha^2}} & 0 & 0 & 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} & 0 \\ 0 & \frac{1}{\sqrt{1-\alpha^2}} & 0 & 0 & 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ \frac{-\alpha}{\sqrt{1-\alpha^2}} & 0 & 0 & 0 & \frac{1}{\sqrt{1-\alpha^2}} & 0 \\ 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} & 0 & 0 & 0 & \frac{1}{\sqrt{1-\alpha^2}} \end{array}\right].$$

$[I\ I]\Theta$ is equal to

$$\left[\begin{array}{ccc|ccc} \frac{1}{\sqrt{1-\alpha^2}} & 0 & 0 & 1 & \frac{-\alpha}{\sqrt{1-\alpha^2}} & 0 \\ \frac{-\alpha}{\sqrt{1-\alpha^2}} & \frac{1}{\sqrt{1-\alpha^2}} & 0 & 0 & \frac{1}{\sqrt{1-\alpha^2}} & \frac{-\alpha}{\sqrt{1-\alpha^2}} \\ 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} & 1 & 0 & 0 & \frac{1}{\sqrt{1-\alpha^2}} \end{array}\right],$$

so that

$$L_{ME}^{-*} = \left[\begin{array}{ccc} \frac{1}{\sqrt{1-\alpha^2}} & 0 & 0 \\ \frac{-\alpha}{\sqrt{1-\alpha^2}} & \frac{1}{\sqrt{1-\alpha^2}} & 0 \\ 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} & 1 \end{array}\right]$$

and

$$M_{ME}^{-1} = \left[\begin{array}{ccc} 1 & \frac{-\alpha}{\sqrt{1-\alpha^2}} & 0 \\ 0 & \frac{1}{\sqrt{1-\alpha^2}} & \frac{-\alpha}{\sqrt{1-\alpha^2}} \\ 0 & 0 & \frac{1}{\sqrt{1-\alpha^2}} \end{array}\right].$$

A_{ME}^{-1} and A_{ME} are given by

$$A_{ME}^{-1} = \begin{bmatrix} \frac{1}{1-\alpha^2} & \frac{-\alpha}{1-\alpha^2} & 0 \\ \frac{-\alpha}{1-\alpha^2} & \frac{1+\alpha^2}{1-\alpha^2} & \frac{-\alpha}{1-\alpha^2} \\ 0 & \frac{-\alpha}{1-\alpha^2} & \frac{1}{1-\alpha^2} \end{bmatrix} \quad \text{and} \quad A_{ME} = \begin{bmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{bmatrix}.$$

■

It is easy to verify that the Schur algorithm requires $O(nb^2)$ operations and $O(nb)$ storage, where n is the size of A and b the average width of $\mathcal{S}$. A flow-graph representation of the algorithm is shown in Figure 4.1, where A is 5×5 and $\mathcal{S} = \{(i,j) \mid |i-j| \leq 2\}$. A node denotes a virtual processor, a box a delay, and a '·' a 'don't care'. The operation of a node is as follows: from the first two input values, x_1 and y_1 say (with $|x_1| > |y_1|$), it computes $\rho = -y_1/x_1$ and

$$[x_1 \; y_1] \frac{1}{\sqrt{1-|\rho|^2}} \begin{bmatrix} 1 & \rho \\ \rho^* & 1 \end{bmatrix} = [\sqrt{x_1^2 - y_1^2} \; 0],$$

for all subsequent input values x_k and y_k

$$[x_k \; y_k] \frac{1}{\sqrt{1-|\rho|^2}} \begin{bmatrix} 1 & \rho \\ \rho^* & 1 \end{bmatrix} = [x_k' \; y_k'].$$

The algorithm is very well suited for implementation on a systolic or wavefront array processor. A node can be realized as a pipelined CORDIC (Coordinate Rotation Digital Computer) device (see [Vol59, Wal71, LHDB88]). The one described in the last reference is capable of achieving a throughput of 10^7–10^8 floating point operations per second, and measures about 1 square centimeter.

We cannot use the Schur algorithm to compute L_{ME}^{-*} and M_{ME}^{-1} when A is specified on a set other than a staircase band. If, however, there is a permutation matrix P such that $B = PAP^*$ is specified on a staircase band — in Chapter 6 we shall encounter matrices of this type — then we use the following theorem to compute A_{ME}^{-1} from B_{ME}^{-1}.

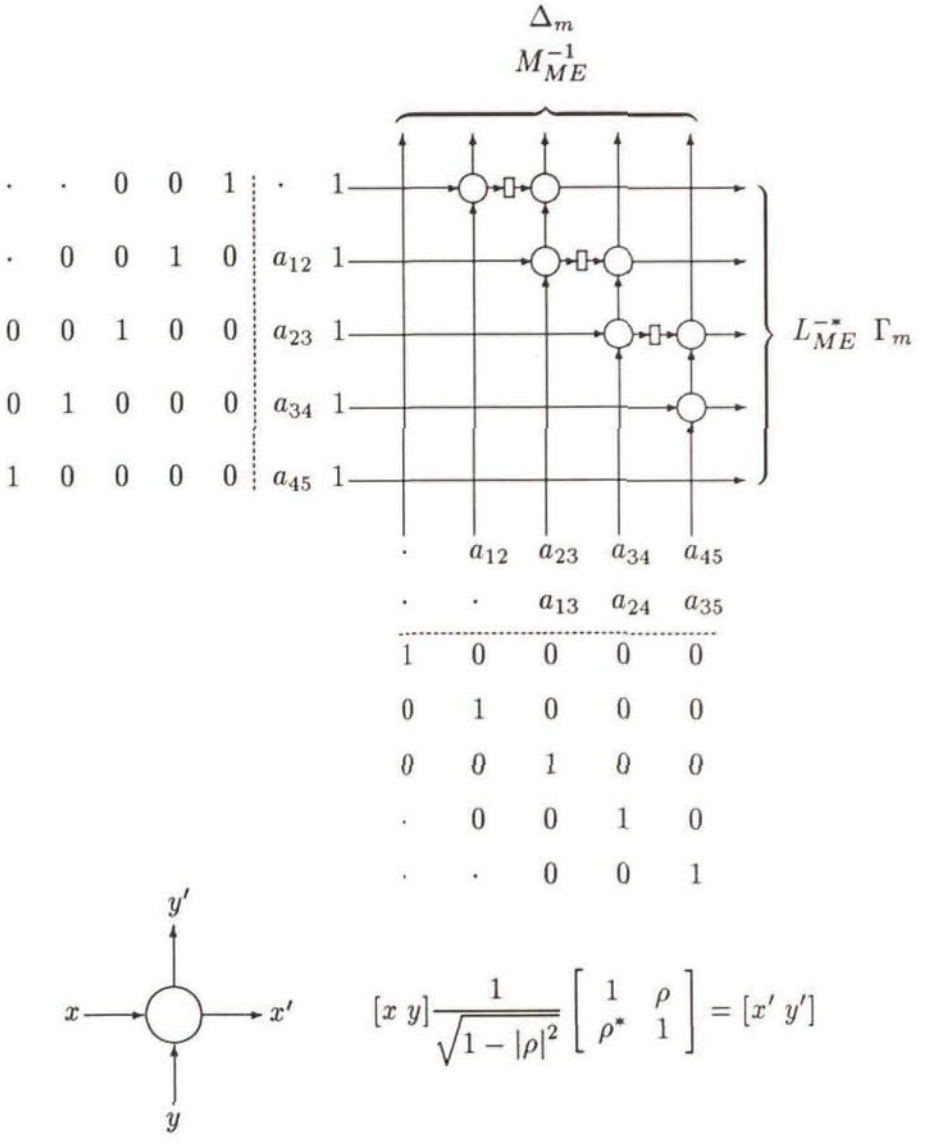


Figure 4.1: Flow-graph representation of the Schur algorithm.

Theorem 4.2 *Suppose that A is a positive definite matrix that is specified on a set $\mathcal{S}$ that contains the diagonal pairs, but is arbitrary otherwise, and let $B = PAP^*$, where P is a permutation matrix. Then, $A_{ME}^{-1} = P^* B_{ME}^{-1} P$.*

Proof The maximum entropy extension is the (unique) positive definite extension whose determinant is maximal. Because $\det B = \det P \det A (\det P)^*$, and $\det P$ is constant, it follows that $B_{ME} = PA_{ME}P^*$, so that $A_{ME}^{-1} = P^* B_{ME}^{-1} P$. ■

Because B is specified on a staircase band, we determine the triangular factors of B_{ME}^{-1} by using the Schur algorithm, which requires $O(nb^2)$ operations and $O(nb)$ storage, where n is the size of the matrix and b the average width of the band. The computation of B_{ME}^{-1} from its triangular factors demands the same amount of resources — the triangular factors of B_{ME}^{-1} are banded. For these reasons, and because we obtain B and A_{ME}^{-1} by reindexing the entries of A and B_{ME}^{-1}, which asks almost no effort, we can compute A_{ME}^{-1} in $O(nb^2)$ operations and with $O(nb)$ storage.

Remember that we have assumed that the diagonal entries in A are equal to one. If this is not the case, then we normalize A to $C = (\underline{\underline{\mathrm{P}}}_0 A)^{-\frac{1}{2}} A (\underline{\underline{\mathrm{P}}}_0 A)^{-\frac{1}{2}}$, compute $L_{C_{ME}}^{-*}$ and $M_{C_{ME}}^{-1}$ by using the Schur algorithm, and by an argument similar to the one used in the proof of Theorem 4.2, determine $L_{A_{ME}}^{-*}$ as $(\underline{\underline{\mathrm{P}}}_0 A)^{-\frac{1}{2}} L_{C_{ME}}^{-*}$ and $M_{A_{ME}}^{-1}$ as $(\underline{\underline{\mathrm{P}}}_0 A)^{-\frac{1}{2}} M_{C_{ME}}^{-1}$.

4.2 Error Analysis

We return to the case where $\mathcal{S}$ is staircase and the diagonal entries in A are equal to one. The following theorem gives an estimate of $\|I - L_{ME}^{-1} L\|$. D_A is defined as $\underline{\underline{\mathrm{P}}}_0 L_A$.

Theorem 4.3 *Suppose that $A = [a_{ij}]$, $i, j = 1, \ldots, n$, is a positive definite matrix that is specified on a staircase band $\mathcal{S}$, and let ρ_{ij} be the ijth reflection coefficient of A. Furthermore, let $b_i = \max\{ j \mid (i,j) \in \mathcal{S} \}$. Then,*

$$\|I - L_{ME}^{-1} L\| = \sqrt{2n - 2 \sum_{i=1}^{n} \prod_{j=b_i+1}^{n} \sqrt{1 - |\rho_{ij}|^2}}.$$

Proof By Proposition 2.2

$$\|I - L_{ME}^{-1}L\| = \sqrt{2\mathrm{tr}(I - D_{ME}^{-1}D)},$$

and with the help of Equation 4.3 and by direct calculation it is easy to verify that

$$(L_{ME}^{-1})_{ii} = \prod_{j=i+1}^{b_i} \frac{1}{\sqrt{1 - |\rho_{ij}|^2}}.$$

If $\mathcal{S}$ were equal to the full set $\{(i,j) \mid i,j = 1,\ldots,n\}$, then A_{ME} would be equal to A, so that

$$(L^{-1})_{ii} = \prod_{j=i+1}^{n} \frac{1}{\sqrt{1 - |\rho_{ij}|^2}}$$

(for $j = i+1,\ldots,b_i$ the ρ_{ij}'s are the same as in the previous equation) and hence

$$(L_{ME}^{-1}L)_{ii} = \prod_{j=b_i+1}^{n} \sqrt{1 - |\rho_{ij}|^2},$$

and the proof of the theorem follows readily. ■

We express $\|I - \left((D_{ME}^{-1}D)L_{ME}^{-1}\right) L\|$, the distance between L^{-1} and its optimal sparse approximation — see Chapter 2 — in a way similar to Theorem 4.3:

$$\|I - \left((D_{ME}^{-1}D)L_{ME}^{-1}\right) L\| = \sqrt{n - \sum_{i=1}^{n} \prod_{j=b_i+1}^{n} \left(1 - |\rho_{ij}|^2\right)}.$$

With increasing b_i's $\|I - L_{ME}^{-1}L\|$ and $\|I - \left((D_{ME}^{-1}D)L_{ME}^{-1}\right) L\|$ tend monotonely to zero.

We next show that if $\|I - L_B^{-1}L_A\|$ is small, then A is close to B, and similarly for A^{-1} and B^{-1}.

Theorem 4.4 *Suppose that A and B are positive definite, and let*

$$\|I - L_B^{-1}L_A\| < \epsilon,$$

where $\epsilon < 1$. *Then,*

$$\|A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}}\| < 2\epsilon + O(\epsilon^2)$$

and

$$\|A^{\frac{1}{2}}(A^{-1}-B^{-1})A^{\frac{1}{2}}\| < 2\epsilon + O(\epsilon^2).$$

Proof Because the QR factorization of $A^{-\frac{1}{2}}$ is unique, and

$$A^{-\frac{1}{2}}A^{-\frac{1}{2}} = L_A^{-*}L_A^{-1},$$

we have

$$A^{-\frac{1}{2}} = QL_A^{-1},$$

where Q is unitary. Hence, as the Frobenius norm is invariant under unitary transformations,

$$\|A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}}\| = \|L_A^{-1}(A-B)L_A^{-*}\|.$$

Now, suppose that $I - L_B^{-1}L_A = E$. Then, $L_B = L_A(I-E)^{-1}$, $\|E\| < \epsilon < 1$, and because $\|E\| < 1$,

$$(I-E)^{-1} = \sum_{k=0}^{\infty} E^k,$$

and it follows that

$$\begin{aligned}
\|A^{-\frac{1}{2}}(A-B)A^{-\frac{1}{2}}\| &= \|L_A^{-1}(A-B)L_A^{-*}\| \\
&= \|I - L_A^{-1}BL_A^{-*}\| \\
&= \|I - (I-E)^{-1}(I-E)^{-*}\| \\
&= \|I - \sum_{k=0}^{\infty} E^k \left(\sum_{k=0}^{\infty} E^k\right)^*\| \\
&< 2\epsilon + O(\epsilon^2).
\end{aligned}$$

In a similar way,

$$\begin{aligned} \|A^{\frac{1}{2}}(A^{-1} - B^{-1})A^{\frac{1}{2}}\| &= \|L_A^*(A^{-1} - B^{-1})L_A\| \\ &= \|I - L_A^* B^{-1} L_A\| \\ &= \|I - (I - E)^*(I - E)\| \\ &< 2\epsilon + O(\epsilon^2). \end{aligned}$$

■

Theorem 4.4 implies that $\|A^{-\frac{1}{2}}(A - A_{ME})A^{-\frac{1}{2}}\|$ and $\|A^{\frac{1}{2}}(A^{-1} - A_{ME}^{-1})A^{\frac{1}{2}}\|$ are essentially twice $\|I - L_{ME}^{-1}L\|$ (of which an estimate is given in Theorem 4.3).

Chapter 5

Iterative Algorithms

SUPPOSE THAT A is a positive definite matrix that is specified on a set $\mathcal{S}$ that contains the diagonal pairs, but is arbitrary otherwise. In this case we depend on iterative algorithms for computing A_{ME}^{-1}. They consume much time and storage — in each iteration step a matrix (of the same size as A) has to be inverted — and for that reason, have little value in practice. In this chapter we suffice with sketching their basic idea.

5.1 The Algorithms

We start out with the proof of Theorem 1.3, which is reproduced here.

Theorem 1.3 ([GJSW84]) *Suppose that A is a positive definite matrix that is specified on a set $\mathcal{S}$ that contains the diagonal pairs, but is arbitrary otherwise. Then, there is a unique matrix B in $\mathcal{E}$ such that*

$$\det B = \max\{ \det E \mid E \in \mathcal{E} \}.$$

Moreover, B is the unique positive definite extension whose inverse satisfies

$$(B^{-1})_{ij} = 0 \quad \forall\, (i,j) \notin \mathcal{S}.$$

The proof of the theorem is based on the following results.

Lemma 5.1 ([GJSW84]) *Let A be as defined in Theorem 1.3. Then, the function $f(E) = \log\det E$ is strictly concave on $\mathcal{E}$.*

Proof Suppose that $B \in \mathcal{E}$ and $C \in \mathcal{E}$. Then, for all $0 < \alpha < 1$ $\alpha B + (1-\alpha)C \in \mathcal{E}$ — $\mathcal{E}$ is convex. Furthermore, there is a matrix F with $\det F = 1$ such that

$$F^* B F = H \quad \text{and} \quad F^* C F = K,$$

where H and K are diagonal (see, e.g., [GV83]), and because $\det(\alpha H + (1-\alpha)K) > \alpha \det H + (1-\alpha)\det K$, and the function $\log z$ is strictly concave, it follows that for all $0 < \alpha < 1$

$$\begin{aligned}
\log\det\Big(\alpha B + (1-\alpha)C\Big) &= \log\det\Big(F^*\Big(\alpha B + (1-\alpha)C\Big)F\Big) \\
&= \log\det\Big(\alpha H + (1-\alpha)K\Big) \\
&> \alpha \log\det H + (1-\alpha)\log\det K \\
&= \alpha \log\det B + (1-\alpha)\log\det C.
\end{aligned}$$

■

Lemma 5.2 *Let A be as defined in Theorem 1.3. Then, $f(E) = \log\det E$ is bounded on $\mathcal{E}$.*

Proof The proof of the lemma follows directly from Hadamard's inequality — see Theorem 1.1. ■

Lemma 5.3 *Let A be as defined in Theorem 1.3. Then, for $(i,j) \notin \mathcal{S}$ the partial derivative of $f(E) = \log\det E$ with respect to e_{ij} is given by*

$$\frac{\partial f(E)}{\partial e_{ij}} = (E^{-1})^*_{ij}$$

($e_{ij} = a_{ij}$ if $(i,j) \in \mathcal{S}$).

Proof We follow Lev-Ari in [Lev85]. Because $\log\det E = \mathrm{tr}\log E$ for any positive definite E, we have

$$\begin{aligned}\frac{\partial f(E)}{\partial e_{ij}} &= \frac{\partial}{\partial e_{ij}}\left(\log\det E\right) = \frac{\partial}{\partial e_{ij}}\left(\mathrm{tr}\log E\right)\\ &= \mathrm{tr}\left(\frac{\partial \log E}{\partial e_{ij}}\right) = \mathrm{tr}\left(E^{-1}\frac{\partial E}{\partial e_{ij}}\right)\\ &= (E^{-1})^{*}_{ij}.\end{aligned}$$

■

We proceed with the proof of Theorem 1.3.

Proof $\mathcal{E}$ is an open, convex set whose limit points are singular, positive semidefinite extensions. Because by Lemmas 5.1 and 5.2 $f(E)$ is strictly concave and bounded on $\mathcal{E}$, and $f(E) = -\infty$ for singular extensions, it follows that $f(E) = \log\det E$ and, consequently — $(\log z)^{-1} = e^{z}$ is strictly monotone increasing — $\det E$ have a unique global maximum on $\mathcal{E}$.

Suppose that the maximum occurs for $E = B$. Then, we have for this extension

$$\frac{\partial f(E)}{\partial e_{ij}} = 0 \quad \forall\,(i,j) \notin \mathcal{S},$$

and we obtain from Lemma 5.3 that

$$(B^{-1})_{ij} = 0 \quad \forall\,(i,j) \notin \mathcal{S}.$$

■

The proof of Theorem 1.3 suggests that to determine A_{ME}, we can maximize $f(E) = \log\det E$ over $\mathcal{E}$ by using standard techniques for (unconstrained) nonlinear optimization. We illustrate the idea by the method of steepest ascent (see, e.g., [Lue73]). $H(E)$ denotes the gradient of $f(E)$ with respect to the entries in E on the complement of $\mathcal{S}$:

$$\left(H(E)\right)_{ij} = \begin{cases} (E^{-1})^{*}_{ij} & \text{if } (i,j) \notin \mathcal{S};\\ 0 & \text{otherwise}\end{cases}$$

(it measures the departure of E^{-1} from being zero on the complement of $\mathcal{S}$). Steepest ascent starts with an initial guess for A_{ME}, that is, some positive definite extension of A (which may be difficult to find), A_0 say. Next, it computes the gradient $H_0 = H(A_0)$, the direction in which $f(E)$ increases most rapidly, and determines λ_0 such that $f(A_0 + \lambda_0 H_0)$ is maximal (and, of course, $A_0 + \lambda_0 H_0$ positive definite). This is a one dimensional optimization problem, which can be solved by, for example, Newton's method. The next guess for A_{ME} is

$$A_1 = A_0 + \lambda_0 H_0,$$

etc., and the procedure ends when for some k A_k is close enough to A_{ME}. The method is computationally very expensive — in each iteration step we have to compute H_k, which amounts to inverting A_k.

In [LPK89] Lev-Ari et al. showed that the dual of maximizing $\log\det E$ subject to $E \in \mathcal{E}$ is to minimize $\mathrm{tr}(AC) - \log\det C - n$ subject to $C \in \mathcal{C}$, where $\mathcal{C}$ is the set of positive definite matrices that vanish on the complement of $\mathcal{S}$. Hence, to compute A_{ME}^{-1}, we can minimize $g(C) = \mathrm{tr}(AC) - \log\det C$ over $\mathcal{C}$ by using, for instance, the method of steepest descent. It is readily verified that for $(i,j) \in \mathcal{S}$ the partial derivative of $g(C)$ with respect to c_{ij} is given by

$$\frac{\partial g(C)}{\partial c_{ij}} = a_{ij}^* - (C^{-1})_{ij}^*,$$

and we denote the gradient of $g(C)$ with respect to the entries in C on $\mathcal{S}$ by $K(C)$:

$$\Big(K(C)\Big)_{ij} = \begin{cases} a_{ij}^* - (C^{-1})_{ij}^* & \text{if } (i,j) \in \mathcal{S}; \\ 0 & \text{otherwise} \end{cases}$$

(it measures the departure of C^{-1} from being an extension of A). Steepest descent is analog to steepest ascent, but it goes in the opposite direction. It starts with an initial guess for A_{ME}^{-1}, that is, a matrix that vanishes on the complement of $\mathcal{S}$ (e.g., the identity), C_0 say. Next, it computes $K_0 = K(C_0)$, and determines λ_0 such that $g(C_0 - \lambda_0 K_0)$ is minimal and $C_0 - \lambda_0 K_0$ positive definite ($-K_0$ is the direction in which $g(C)$ decreases most rapidly). The next guess for A_{ME}^{-1} is

$$C_1 = C_0 - \lambda_0 K_0,$$

etc., and the method proceeds until for some k C_k is close enough to A_{ME}^{-1}. Clearly, the dual problem has the same disadvantage as the original: solving it requires a lot of time and storage.

Chapter 6

An Extension of the Schur Algorithm

SUPPOSE THAT A is a positive definite matrix that is specified on a multiple band $\mathcal{S}$. Because a multiple band is not staircase, and no permutation matrix P exists such that PAP^* is specified on such a band, we depend on iterative algorithms for computing A_{ME}^{-1}. As we have seen in the previous chapter, these algorithms consume much time and storage, and we have to be satisfied with a close approximation to A_{ME}^{-1} that can be computed efficiently. In this chapter we present an extension of the Schur algorithm for computing such an approximation, provided that certain conditions are satisfied. It computes the inverse of the maximum entropy extension of a partially specified matrix that is close to A and specified on the same set $\mathcal{S}$, and requires $O(nc^2)$ operations and $O(nc)$ storage, where n is the size of A and c the average number of elements in $\mathcal{S}$ per row of A.

6.1 The Algorithm

First, we describe the structure of the triangular factor $\mathbf{L}_{ME}^{-*}$ of the inverse of the maximum entropy extension of a positive definite block matrix that is specified on a block band.

Theorem 6.1 *Let* $\mathbf{A} = [\mathbf{a}_{kl}]$, $k,l = 1,\ldots,m$, *be a positive definite block matrix with blocks* $\mathbf{a}_{kl}$ *of size* $n_k \times n_l$ *that is specified on a block band* $\mathcal{B} = \{(k,l) \mid |k-l| \leq b\}$. *Then, the columns of* $\mathbf{L}_{ME}^{-*}$ *are such that for* $l = 1,\ldots,m-b-1$

$$\left(\mathbf{L}_{ME}^{-*}\right)_{kl} = \left(\mathbf{L}_{\mathbf{A}(l,l+b)}^{-*}\right)_{(k-l+1)1}, \qquad k = l,\ldots,l+b;$$

$$\left(\mathbf{L}_{ME}^{-*}\right)_{kl} = 0, \qquad k = l+b+1,\ldots,m,$$

and for $l = m-b,\ldots,m$

$$\left(\mathbf{L}_{ME}^{-*}\right)_{kl} = \left(\mathbf{L}_{\mathbf{A}(m-b,m)}^{-*}\right)_{(k-(m-b)+1)(l-(m-b)+1)}, \qquad k = l,\ldots,m.$$

Proof We follow Dym and Gohberg in [DG81]. Let Π_0 and Π_- denote the operators that project a matrix on $\mathcal{B}$ and the lower triangular part of the complement of $\mathcal{B}$ respectively. Furthermore, suppose that $\mathbf{X} = [\mathbf{x}_{kl}]$, $k,l = 1,\ldots,m$, is an upper triangular block matrix with blocks $\mathbf{x}_{kl}$ of size $n_k \times n_l$, and that the columns of $\mathbf{X}^{-*}$ are such that for $l = 1,\ldots,m-b-1$

$$\left(\mathbf{X}^{-*}\right)_{kl} = \left(\mathbf{L}_{\mathbf{A}(l,l+b)}^{-*}\right)_{(k-l+1)1}, \qquad k = l,\ldots,l+b;$$

$$\left(\mathbf{X}^{-*}\right)_{kl} = 0, \qquad k = l+b+1,\ldots,m,$$

and for $l = m-b,\ldots,m$

$$\left(\mathbf{X}^{-*}\right)_{kl} = \left(\mathbf{L}_{\mathbf{A}(m-b,m)}^{-*}\right)_{(k-(m-b)+1)(l-(m-b)+1)}, \qquad k = l,\ldots,m.$$

We shall prove that $\mathbf{X}^{-*} = \mathbf{L}_{ME}^{-*}$, and hence that $\mathbf{L}_{ME}^{-*}$ is as described above.

Clearly, $\mathbf{X}^{-*}$ is lower triangular with lower triangular diagonal blocks with positive diagonal entries, and $\mathbf{X}^{-*}\mathbf{X}^{-1}$ vanishes on the complement of $\mathcal{B}$. We now define

$$\mathbf{E} = \Pi_-\mathbf{E} + \Pi_0\mathbf{E} + (\Pi_-\mathbf{E})^*$$

with

$$\Pi_-\mathbf{E} = -\Pi_-\left(\Pi_0(\mathbf{A})\,\mathbf{X}^{-*}\right)\mathbf{X}^*$$

and

$$\Pi_0 \mathbf{E} = \Pi_0 \mathbf{A},$$

and show that $\mathbf{X}\mathbf{X}^* = \mathbf{E}$, an extension of $\mathbf{A}$.

It follows from the definitions of $\mathbf{E}$ and $\mathbf{X}^{-*}$ that $\mathbf{E}\mathbf{X}^{-*}$ is zero on the strictly lower triangular part of $\mathcal{B}$, and that $\mathbf{X}^{-1}\mathbf{E}\mathbf{X}^{-*}$ has identities on the diagonal. Because $\mathbf{X}^{-*}$ is lower triangular, we have

$$\begin{aligned}
\Pi_-(\mathbf{E}\mathbf{X}^{-*}) &= \Pi_-\left(\Pi_-(\mathbf{E})\,\mathbf{X}^{-*}\right) + \Pi_-\left(\Pi_0(\mathbf{E})\,\mathbf{X}^{-*}\right) + \Pi_-\left((\Pi_-\mathbf{E})^*\,\mathbf{X}^{-*}\right) \\
&= \Pi_-(\mathbf{E})\,\mathbf{X}^{-*} + \Pi_-\left(\Pi_0(\mathbf{A})\,\mathbf{X}^{-*}\right) \\
&= 0,
\end{aligned}$$

and hence that $\mathbf{E}\mathbf{X}^{-*}$ is upper triangular. Since

$$\mathbf{X}^{-1}(\mathbf{E}\mathbf{X}^{-*}) = (\mathbf{E}\mathbf{X}^{-*})^*\mathbf{X}^{-*},$$

and $\mathbf{X}^{-1}$ and $\mathbf{E}\mathbf{X}^{-*}$ are upper triangular, it follows that $\mathbf{X}^{-1}\mathbf{E}\mathbf{X}^{-*}$ is diagonal — the left-hand side of the equation is upper triangular, and the right-hand side lower triangular — and hence equal to the identity, so that $\mathbf{X}\mathbf{X}^* = \mathbf{E}$. ■

We illustrate Theorem 6.1 with a simple example.

Example 6.1 Suppose that

$$A = \begin{bmatrix} 1 & \alpha & ? \\ \alpha & 1 & \alpha \\ ? & \alpha & 1 \end{bmatrix},$$

where α is a scalar between -1 and 1, and let $\mathcal{B} = \{(k,l) \mid |k-l| \leq 1\}$. Then,

$$L^{-*}_{A(1,2)} = \begin{bmatrix} \dfrac{1}{\sqrt{1-\alpha^2}} & 0 \\ \dfrac{-\alpha}{\sqrt{1-\alpha^2}} & 1 \end{bmatrix} \quad \text{and} \quad L^{-*}_{A(2,3)} = \begin{bmatrix} \dfrac{1}{\sqrt{1-\alpha^2}} & 0 \\ \dfrac{-\alpha}{\sqrt{1-\alpha^2}} & 1 \end{bmatrix},$$

and

$$L_{ME}^{-*} = \begin{bmatrix} \frac{1}{\sqrt{1-\alpha^2}} & 0 & 0 \\ \frac{-\alpha}{\sqrt{1-\alpha^2}} & \frac{1}{\sqrt{1-\alpha^2}} & 0 \\ 0 & \frac{-\alpha}{\sqrt{1-\alpha^2}} & 1 \end{bmatrix}.$$

A_{ME}^{-1} and A_{ME} evaluate to

$$A_{ME}^{-1} = \begin{bmatrix} \frac{1}{1-\alpha^2} & \frac{-\alpha}{1-\alpha^2} & 0 \\ \frac{-\alpha}{1-\alpha^2} & \frac{1+\alpha^2}{1-\alpha^2} & \frac{-\alpha}{1-\alpha^2} \\ 0 & \frac{-\alpha}{1-\alpha^2} & \frac{1}{1-\alpha^2} \end{bmatrix} \quad \text{and} \quad A_{ME} = \begin{bmatrix} 1 & \alpha & \alpha^2 \\ \alpha & 1 & \alpha \\ \alpha^2 & \alpha & 1 \end{bmatrix}.$$

■

We return to the case where A is specified on a multiple band $\mathcal{S}$. Theorem 6.1 implies that if it were possible to partition A as $\mathbf{A} = [\mathbf{a}_{kl}]$ so that for some block band $\mathcal{B} = \{(k,l) \mid |k-l| \leq b\}$ the blocks on $\mathcal{B}$ are specified and the blocks on the complement of $\mathcal{B}$ are unspecified, then $\mathbf{L}_{ME}^{-*}$ would be found from the triangular factors of the inverses of the principal submatrices $\mathbf{A}(l, l+b)$. Clearly, we cannot partition A in this way, but it has the following property (to which we shall refer as Property $\mathcal{M}$):

> A can be partitioned as $\mathbf{A} = [\mathbf{a}_{kl}]$, $k,l = 1,\ldots,m$, so that for some block band $\mathcal{B} = \{(k,l) \mid |k-l| \leq b\}$ (1) the partially specified principal submatrices $\mathbf{A}(l, l+b)$ can be permuted to Hermitian matrices that are specified on a staircase band and (2) the blocks on the complement of $\mathcal{B}$ are unspecified.

For example, if A is as shown in Figure 6.1, then we can partition it into blocks of size 4×4, so that $\mathbf{A}(1,2)$ and $\mathbf{A}(2,3)$ can be permuted to Hermitian matrices that are specified on a staircase band — see Figure 6.2 — and the blocks on the complement of $\{(k,l) \mid |k-l| \leq 1\}$ are unspecified.

We now assume that the triangular factors of the inverses of the maximum entropy extensions of the principal submatrices $\mathbf{A}(l, l+b)$ are close

$$
A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \end{array}
\begin{array}{c}
\begin{array}{cccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \end{array} \\
\left[\begin{array}{cccc|cccc|cccc}
\times & \times & & & \times & \times & & & & & & \\
\times & \times & \times & & \times & \times & \times & & & & & \\
 & \times & \times & \times & & \times & \times & \times & & & & \\
 & & \times & \times & & & \times & \times & & & & \\
\hline
\times & \times & & & \times & \times & & & \times & \times & & \\
\times & \times & \times & & \times & \times & \times & & \times & \times & \times & \\
 & \times & \times & \times & & \times & \times & \times & & \times & \times & \times \\
 & & \times & \times & & & \times & \times & & & \times & \times \\
\hline
 & & & & \times & \times & & & \times & \times & & \\
 & & & & \times & \times & \times & & \times & \times & \times & \\
 & & & & & \times & \times & \times & & \times & \times & \times \\
 & & & & & & \times & \times & & & \times & \times
\end{array}\right]
\end{array}
$$

Figure 6.1: A matrix that is specified on a multiple band.

approximations to the triangular factors of the inverses of the corresponding submatrices in the completely specified matrix. The approximations are suboptimal — the principal submatrices $\mathbf{A}(l, l+b)$ can be permuted to Hermitian matrices that are specified on a staircase band, and the triangular factors of the inverses of the maximum entropy extensions of these matrices are suboptimal sparse approximations (in the Frobenius norm) to the triangular factors of the inverses of the permuted versions of the completely specified matrices (see Chapter 2). Next, we approximate L^{-*} by a lower triangular matrix that is composed of (the first columns of) the triangular factors of the inverses of the maximum entropy extensions of the principal submatrices $\mathbf{A}(l, l+b)$ — see the following definition.

Definition 6.1 *Suppose that $A = [a_{ij}]$, $i, j = 1, \ldots, n$, is a positive definite matrix with property $\mathcal{M}$, and let $\mathbf{L}_H^{-*}$ be a lower triangular block*

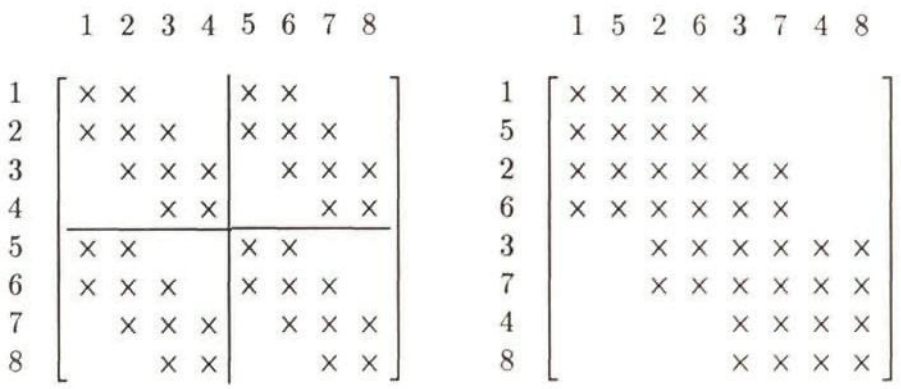


Figure 6.2: The principal submatrix $\mathbf{A}(1,2)$ and the permuted matrix.

matrix whose columns are such that for $l = 1, \ldots, m-b-1$

$$\left(\mathbf{L}_H^{-*}\right)_{kl} = \left(\mathbf{L}^{-*}_{\mathbf{A}(l,l+b)_{ME}}\right)_{(k-l+1)1}, \qquad k = l, \ldots, l+b;$$

$$\left(\mathbf{L}_H^{-*}\right)_{kl} = 0, \qquad k = l+b+1, \ldots, m,$$

and for $l = m-b, \ldots, m$

$$\left(\mathbf{L}_H^{-*}\right)_{kl} = \left(\mathbf{L}^{-*}_{\mathbf{A}(m-b,m)_{ME}}\right)_{(k-(m-b)+1)(l-(m-b)+1)}, \qquad k = l, \ldots, m.$$

Then, the hierarchical approximation of A *is defined as* $A_H = L_H L_H^*$.

The adjective 'hierarchical' refers to the fact that $\mathbf{A}_H$ is the maximum entropy extension of a block matrix that is specified on a block band $\mathcal{B}$ in which the largest principal submatrices are (nearly) equal to the maximum entropy extensions of the corresponding submatrices in $\mathbf{A}$.

The hierarchical approximation of A has two shortcomings. To begin with, its inverse does not vanish in every entry on the complement of $\mathcal{S}$. It does, except in the positions that correspond to unspecified entries in the blocks $\mathbf{a}_{kl}$ with $|k-l| \leq b-1$ and $2 \leq k, l \leq m-1$. For example, if A is

as shown in Figure 6.1 and partitioned as before, then A_H^{-1} does not vanish in the positions that correspond to unspecified entries in $\mathbf{a}_{22}$. Furthermore, it is not possible to determine L_H^{-*} or A_H^{-1} in an efficient way. Although the inverses of the maximum entropy extensions of the principal submatrices $\mathbf{A}(l, l+b)$ can be computed efficiently — see Chapter 4 — this is not the case for their triangular factors (the factors of the inverses are not sparse).

To obtain an approximation to A^{-1} that vanishes on the complement of $\mathcal{S}$, we seek (1) a matrix B with the property that $A_H^{-1} + B$ is zero on the complement of $\mathcal{S}$ and (2) a C that vanishes on the complement of $\mathcal{S}$, and that is such that $\| A_H^{\frac{1}{2}} (B - C) A_H^{\frac{1}{2}} \|$ is small and $B - C$ positive semidefinite. We define the new approximation as $A_H^{-1} + B - C$ — it is zero on the complement of $\mathcal{S}$, close to A_H^{-1}, and positive definite. We take

$$B = \sum_{l=1}^{m-b-1} \square \left[\left(\mathbf{A}(l, l+b)_{ME}(2, b+1) \right)^{-1}; (l+1, l+b) \right], \tag{6.1}$$

because (after some direct calculations) $A_H^{-1} + B$ appears to be equal to

$$\sum_{l=1}^{m-b} \square \left[\mathbf{A}(l, l+b)_{ME}^{-1}; (l, l+b) \right],$$

which vanishes on the complement of $\mathcal{S}$ — the matrices of which it is composed are zero on the complement of $\mathcal{S}$ — and it can be computed efficiently (we obtain its components as described in Chapter 4). We can formulate the problem of finding an optimal choice for C as a nonlinear optimization problem, which can be solved by using standard techniques, but these methods are computationally very expensive.

For this reason, we assume that $\mathbf{A}(l+1, l+b)_{ME}^{-1}$ is a good approximation to $(\mathbf{A}(l, l+b)_{ME}(2, b+1))^{-1}$ (for the sake of clearness: $\mathbf{A}(l+1, l+b)_{ME}$ is the maximum entropy extension of the principal submatrix of $\mathbf{A}$ that lies in the rows and columns indexed by $l+1, \ldots, l+b$ and $\mathbf{A}(l, l+b)_{ME}(2, b+1)$ the principal submatrix of $\mathbf{A}(l, l+b)_{ME}$ that lies in the rows and columns indexed by $2, \ldots, b+1$). The approximation is suboptimal — $\mathbf{A}(l+1, l+b)_{ME}^{-1}$ can be permuted to a Hermitian matrix with support on a staircase band, and

the triangular factors of this matrix are suboptimal sparse approximations (in the Frobenius norm) to the triangular factors of the permuted version of $(\mathbf{A}(l,l+b)_{ME}(2,b+1))^{-1}$. We define

$$C = \sum_{l=1}^{m-b-1} \square \left[\mathbf{A}(l+1,l+b)^{-1}_{ME};(l+1,l+b)\right]. \tag{6.2}$$

It is zero on the complement of $\mathcal{S}$, and its components are suboptimal sparse approximations to the components of B, but we have violated the condition that $B - C$ must be positive semidefinite (it is indefinite). We proceed with the following definition.

Definition 6.2 *Suppose that $A = [a_{ij}]$, $i,j = 1,\ldots,n$, is a positive definite matrix with property $\mathcal{M}$, and let*

$$A^{-1}_{SI} = \sum_{l=1}^{m-b} \square\left[\mathbf{A}(l,l+b)^{-1}_{ME};(l,l+b)\right] - \sum_{l=1}^{m-b-1} \square\left[\mathbf{A}(l+1,l+b)^{-1}_{ME};(l+1,l+b)\right].$$

Then, the sparse-inverse approximation of A is defined as A_{SI}.

Theorem 6.2 *Suppose that $A = [a_{ij}]$, $i,j = 1,\ldots,n$, is a positive definite matrix with property $\mathcal{M}$ — it is specified on a multiple band $\mathcal{S}$, and can be partitioned as $\mathbf{A} = [\mathbf{a}_{kl}]$, $k,l = 1,\ldots,m$, so that for some block band $\mathcal{B} = \{(k,l) \mid |k-l| \leq b\}$ (1) the principal submatrices $\mathbf{A}(l,l+b)$ can be permuted to Hermitian matrices that are specified on a staircase band and (2) the blocks on the complement of $\mathcal{B}$ are unspecified — and let A_{SI} be as defined in Definition 6.2. Then, A^{-1}_{SI}*

1. *is zero on the complement of $\mathcal{S}$;*
2. *can be computed in $O(nc^2)$ operations and with $O(nc)$ storage, where c is the average number of elements in $\mathcal{S}$ per row of A.*

Proof Because A^{-1}_{SI} is equal to $A^{-1}_{H} + B - C$, where B and C are as defined in Equations 6.1 and 6.2, and $A^{-1}_{H} + B$ and C vanish on the complement of $\mathcal{S}$, the first property holds. Furthermore, $\mathbf{A}(l,l+b)^{-1}_{ME}$ and $\mathbf{A}(l+1,l+b)^{-1}_{ME}$ can be computed as described in Chapter 4, and the second assertion is readily verified. ■

Because by assumption $\|A_H^{\frac{1}{2}}(B-C)A_H^{\frac{1}{2}}\|$ is small, we conjecture that A_{SI}^{-1} (which is equal to $A_H^{-1}+B-C$) may fail to be positive definite only when A (and hence A_H) is ill conditioned. This conjecture has been confirmed by numerical experiments — see [Gen90, Mei90], the next chapter, and Figures 6.3 and 6.4, where A is 21×21 (it is the covariance matrix of 2 two dimensional sinusoids in white noise), $\mathcal{S}$ is a triple band, and the width of the bands in $\mathcal{S}$ is 2. The first figure shows $\|I - L_{ME}^{-1}L\|$, $\|I - L_{SI}^{-1}L\|$, and 10 times the smallest singular value of A_{SI}^{-1} (which gives an idea of the positivity of A_{SI}^{-1}) as a function of the condition number $\kappa(A)$; the second gives the condition numbers of A_{ME}^{-1} and A_{SI}^{-1}, also as a function of $\kappa(A)$. If $\kappa(A)$ is less than

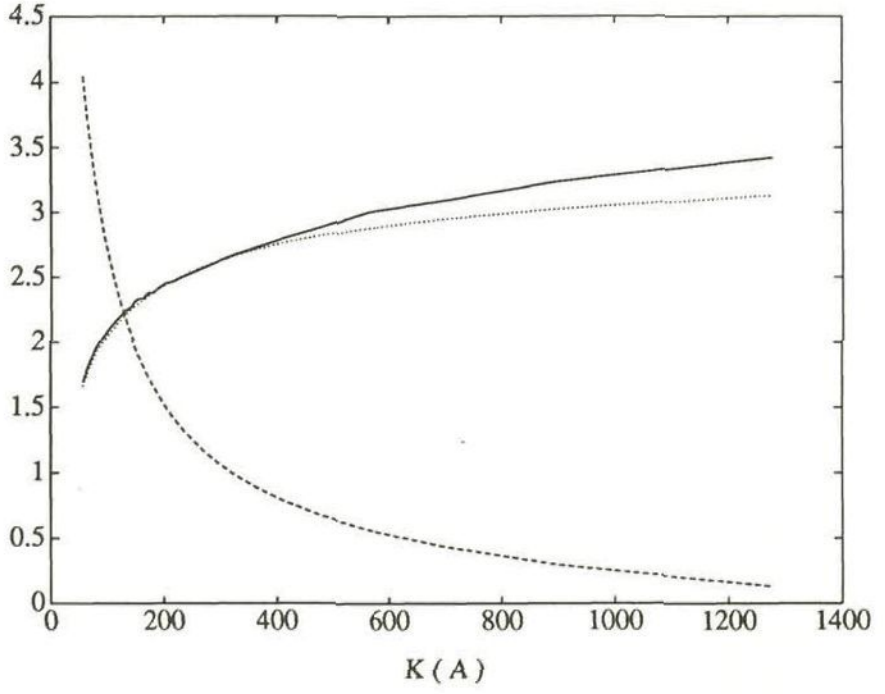


Figure 6.3: The norms $\|I - L_{ME}^{-1}L\|$ (solid) and $\|I - L_{SI}^{-1}L\|$ (dotted) and 10 times the smallest singular value of A_{SI}^{-1} (dashed) as a function of $\kappa(A)$.

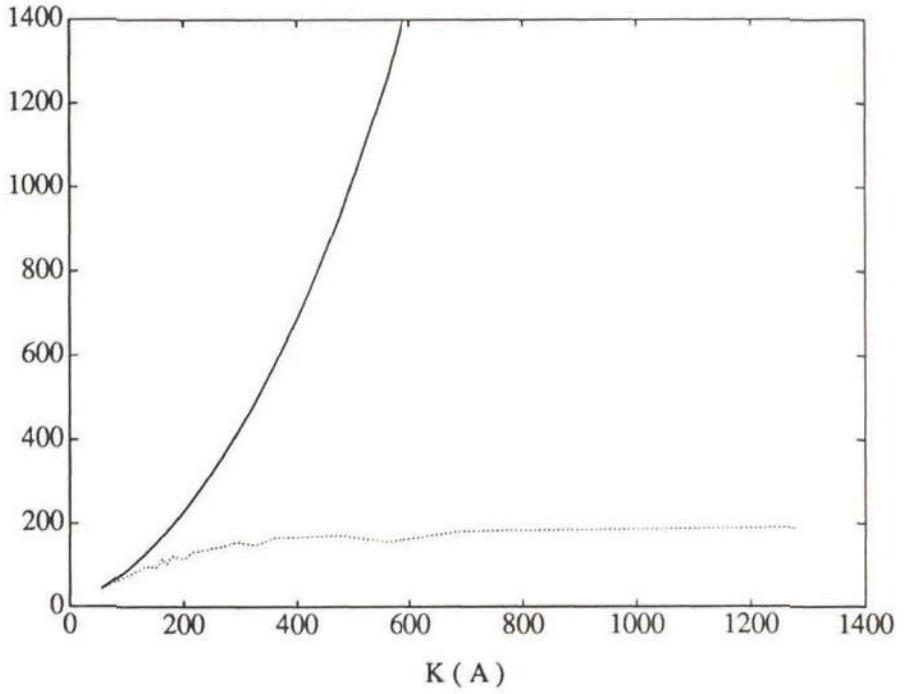


Figure 6.4: The condition numbers $\kappa(A_{SI}^{-1})$ (solid) and $\kappa(A_{ME}^{-1})$ (dotted) as a function of $\kappa(A)$.

400, then $\|I - L_{ME}^{-1}L\|$ and $\|I - L_{SI}^{-1}L\|$ are close and A_{SI}^{-1} is well conditioned and positive definite. When $\kappa(A)$ increases, the condition of A_{SI}^{-1} gets worse, and the sparse-inverse approximation loses its meaning.

6.2 Error Analysis

We now derive bounds on the distance between A_{SI} and A, and similarly for A_{SI}^{-1} and A^{-1}. We start out with estimating $\|I - L_H^{-1}L\|$.

Theorem 6.3 *Suppose that $A = [a_{ij}]$, $i, j = 1, \ldots, n$, is a positive definite matrix with property $\mathcal{M}$ — it is specified on a multiple band $\mathcal{S}$, and*

can be partitioned as $\mathbf{A} = [\mathbf{a}_{kl}]$, $k,l = 1,\ldots,m$, *(assume that the blocks* $\mathbf{a}_{kl}$ *are of size* $n_k \times n_l$*) so that for some block band* $\mathcal{B} = \{(k,l) \mid |k-l| \leq b\}$ *(1) the principal submatrices* $\mathbf{A}(l,l+b)$ *can be permuted to Hermitian matrices that are specified on a staircase band and (2) the blocks on the complement of* $\mathcal{B}$ *are unspecified — and let* A_H *be as defined in Definition 6.1. Let* ρ_{ij} *be the* ij*th reflection coefficient of* A*, and let* $\rho_{l_{ij}}$ *and* $\mathcal{S}_l$ *be the* ij*th reflection coefficient and the support of* $P\mathbf{A}(l,l+b)P^*$ *respectively. Furthermore, assume that for* $l = 1,\ldots,m-b-1$

1. $|\rho_{ij}| < \epsilon$ *for* $i = e_l+1,\ldots,n_l$ *and* $j = f_l+1,\ldots,n$,

and for $l = 1,\ldots,m-b$

2. $|\rho_{l_{ij}}| < \epsilon$ *for* $i = 1,\ldots,g_l$ *and* $j = b_{l_i}+1,\ldots,g_l$,

where $\epsilon < 1$, $e_l = \sum_{r=1}^{l-1} n_r$, $f_l = \sum_{r=1}^{l+b} n_r$, $g_l = \sum_{r=l}^{l+b} n_r$, *and* $b_{l_i} = \max\{j \mid (i,j) \in \mathcal{S}_l\}$. *Then,*

$$\|I - L_H^{-1}L\| < \sqrt{\sum_{l=1}^{m-b-1} 2(1-\gamma_l)g_l + \sum_{l=1}^{m-b} 4\kappa(\mathbf{L}^{-1}_{P\mathbf{A}(l,l+b)P^*})^2\delta_l},$$

where $\gamma_l = 1-(n-f_l)\frac{1}{2}\epsilon^2 + O(\epsilon^4)$, $\delta_l = t_l\epsilon^2 + O(\epsilon^4)$, *and* t_l *is the number of elements in the upper triangular part of the complement of* $\mathcal{S}_l$.

The first condition in Theorem 6.3 corresponds to the assumption that L_{ME}^{-1} is a close approximation to L^{-1} ($\|I - L_{ME}^{-1}L\|$ is small); the second corresponds to the assumption that the triangular factors of the inverses of the maximum entropy extensions of the permuted versions of the principal submatrices $\mathbf{A}(l,l+b)$ are good approximations to the triangular factors of the inverses of the permuted versions of the corresponding submatrices in the completely specified matrix.

To prove the theorem, we need two more results. The first one is due to Stewart [Ste77], and gives perturbation bounds for the QR factorization of a matrix.

Lemma 6.1 ([Ste77]) *Suppose that A is a square and nonsingular matrix, and let $A = QR$, where Q is unitary and R upper triangular with positive diagonal entries. Furthermore, assume that E is such that $A + E$ is nonsingular, and that*

$$A + E = (Q + W)(R + F),$$

where $Q + W$ is unitary and $R + F$ upper triangular with positive diagonal entries. Then, for $\|E\|$ sufficiently small,

$$\|FR^{-1}\| \leq 2\kappa(A)\frac{\|E\|}{\|A\|}.$$

With the help of this result, and denoting $A_P = PAP^*$ and $B_P = PBP^*$, where P is a permutation matrix, we next show that if $\|I - L_{B_P}^{-1}L_{A_P}\|$ is small, and $L_{A_P}^{-1}$ is well conditioned, then $\|I - L_B^{-1}L_A\|$ is small as well.

Lemma 6.2 *Suppose that A and B are positive definite, and let $A_P = PAP^*$ and $B_P = PBP^*$, where P is a permutation matrix. Furthermore, let*

$$\|I - L_{B_P}^{-1}L_{A_P}\| < \epsilon.$$

Then,

$$\|I - L_B^{-1}L_A\| < 2\kappa(L_{A_P}^{-1})\epsilon.$$

Proof Because the QR factorizations of $L_{A_P}^{-1}P$ and $L_{B_P}^{-1}P$ are unique, and

$$L_A^{-*}L_A^{-1} = P^*L_{A_P}^{-*}L_{A_P}^{-1}P$$

and

$$L_B^{-*}L_B^{-1} = P^*L_{B_P}^{-*}L_{B_P}^{-1}P,$$

we have

$$L_{A_P}^{-1}P = Q_A L_A^{-1}$$

and

$$L_{B_P}^{-1}P = Q_B L_B^{-1},$$

where Q_A and Q_B are unitary. Identifying $L_{A_P}^{-1}P$ with A in Lemma 6.1, $L_{B_P}^{-1}P$ with $A+E$, L_A^{-1} with R, and L_B^{-1} with $R+F$, we obtain

$$\begin{aligned}\|L_{B_P}^{-1}P - L_{A_P}^{-1}P\| &= \|(L_{B_P}^{-1}L_{A_P} - I)L_{A_P}^{-1}P\| \\ &< \epsilon\|L_{A_P}^{-1}P\|\end{aligned}$$

($\|E\| < \epsilon\|A\|$) and hence

$$\|I - L_B^{-1}L_A\| < 2\kappa(L_{A_P}^{-1}P)\epsilon$$

($\|FR^{-1}\| < 2\kappa(A)\|E\|/\|A\|$). Because the Frobenius norm is invariant under unitary transformations, we have $\kappa(L_{A_P}^{-1}P) = \kappa(L_{A_P}^{-1})$. ■

We proceed with the proof of Theorem 6.3.

Proof Let $\mathbf{A}(l)$ be shorthand for $\mathbf{A}(l, l+b)$. Adding and subtracting a number of terms, we obtain

$$\begin{aligned}\mathrm{tr}(\mathbf{L}_H^{-1}\mathbf{A}\mathbf{L}_H^{-*}) &= \sum_{l=1}^{m-b} \mathrm{tr}(\mathbf{L}_{\mathbf{A}(l)_{ME}}^{-1}\mathbf{A}(l)\mathbf{L}_{\mathbf{A}(l)_{ME}}^{-*}) - \\ &\quad \sum_{l=1}^{m-b-1} \mathrm{tr}(\mathbf{L}_{\mathbf{A}(l)_{ME}(2,b+1)}^{-1}\mathbf{A}(l)(2,b+1)\mathbf{L}_{\mathbf{A}(l)_{ME}(2,b+1)}^{-*}).\end{aligned}$$

Because by Lemma 2.5

$$\mathrm{tr}(\mathbf{L}_{\mathbf{A}(l)_{ME}}^{-1}\mathbf{A}(l)\mathbf{L}_{\mathbf{A}(l)_{ME}}^{-*}) = \mathrm{tr}\mathbf{I},$$

and similar matrices have equal traces it follows that

$$\mathrm{tr}(\mathbf{L}_H^{-1}\mathbf{A}\mathbf{L}_H^{-*}) = \mathrm{tr}\mathbf{I} + \sum_{l=1}^{m-b-1} \mathrm{tr}\left(\mathbf{I} - \mathbf{A}(l)(2,b+1)\big(\mathbf{A}(l)_{ME}(2,b+1)\big)^{-1}\right)$$

and hence, as

$$\|\mathbf{I} - \mathbf{L}_H^{-1}\mathbf{L}\|^2 = \mathrm{tr}\mathbf{I} - 2\mathrm{tr}(\mathbf{L}_H^{-1}\mathbf{L}) + \mathrm{tr}(\mathbf{L}_H^{-1}\mathbf{A}\mathbf{L}_H^{-*}),$$

that

$$\|\mathbf{I} - \mathbf{L}_H^{-1}\mathbf{L}\|^2 = \tag{6.3}$$

$$2\mathrm{tr}\mathbf{I} - 2\mathrm{tr}(\mathbf{L}_H^{-1}\mathbf{L}) + \sum_{l=1}^{m-b-1} \mathrm{tr}\left(\mathbf{I} - \mathbf{A}(l)(2,b+1)\left(\mathbf{A}(l)_{ME}(2,b+1)\right)^{-1}\right).$$

We now derive a lower bound on $\mathrm{tr}(\mathbf{L}_H^{-1}\mathbf{L})$. It is readily checked that for $l = 1, \ldots, m-b-1$

$$\mathrm{tr}(\mathbf{L}_H^{-1}\mathbf{L})_{ll} = \mathrm{tr}\left((\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)})_{11} W_l\right),$$

where

$$W_l = \bigoplus_{i=e_l+1}^{n_l} \prod_{j=f_l+1}^{n} \sqrt{1-|\rho_{ij}|^2}.$$

Because $|\rho_{ij}| < \epsilon$ for $i = e_l + 1, \ldots, n_l$ and $j = f_l + 1, \ldots, n$, it follows that

$$\mathrm{tr}(\mathbf{L}_H^{-1}\mathbf{L})_{ll} > \gamma_l \mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)})_{11}$$

and hence

$$\mathrm{tr}(\mathbf{L}_H^{-1}\mathbf{L}) > \sum_{l=1}^{m-b-1} \gamma_l \mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)})_{11} + \mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(m-b)_{ME}}\mathbf{L}_{\mathbf{A}(m-b)}). \tag{6.4}$$

We next need an upper bound on $\mathrm{tr}(\mathbf{I} - \mathbf{A}(l)(2,b+1)(\mathbf{A}(l)_{ME}(2,b+1))^{-1})$. Because

$$\|\mathbf{I} - \mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)}\|^2 = \mathrm{tr}\mathbf{I} - 2\mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)}) + \mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{A}(l)\mathbf{L}^{-*}_{\mathbf{A}(l)_{ME}}),$$

and

$$\mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{A}(l)\mathbf{L}^{-*}_{\mathbf{A}(l)_{ME}}) = \mathrm{tr}\mathbf{I},$$

we have

$$\mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)}) = \mathrm{tr}\mathbf{I} - \frac{1}{2}\|\mathbf{I} - \mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)}\|^2 \tag{6.5}$$

and hence

$$\mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}(2,b+1)}\mathbf{L}_{\mathbf{A}(l)(2,b+1)}) =$$

$$\mathrm{tr}\mathbf{I} - \frac{1}{2}\|\mathbf{I} - \mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)}\|^2 - \mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)})_{11}.$$

Since

$$\|\mathbf{I} - \mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}(2,b+1)}\mathbf{L}_{\mathbf{A}(l)(2,b+1)}\|^2 =$$

$$\mathrm{tr}\mathbf{I} - 2\mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}(2,b+1)}\mathbf{L}_{\mathbf{A}(l)(2,b+1)}) + \mathrm{tr}\Big(\mathbf{A}(l)(2,b+1)\Big(\mathbf{A}(l)_{ME}(2,b+1)\Big)^{-1}\Big),$$

this implies that

$$\mathrm{tr}\left(\mathbf{I} - \mathbf{A}(l)(2,b+1)\Big(\mathbf{A}(l)_{ME}(2,b+1)\Big)^{-1}\right) =$$

$$2\mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)})_{11} - 2\mathrm{tr}\mathbf{I} + \|\mathbf{I} - \mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)}\|^2 - \|\mathbf{I} - \mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}(2,b+1)}\mathbf{L}_{\mathbf{A}(l)(2,b+1)}\|^2,$$

so that

$$\mathrm{tr}\left(\mathbf{I} - \mathbf{A}(l)(2,b+1)\Big(\mathbf{A}(l)_{ME}(2,b+1)\Big)^{-1}\right) < \tag{6.6}$$

$$2\mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)})_{11} - 2\mathrm{tr}\mathbf{I} + \|\mathbf{I} - \mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)}\|^2,$$

and it follows from Equations 6.3–6.6 that

$$\|\mathbf{I} - \mathbf{L}_H^{-1}\mathbf{L}\|^2 < \sum_{l=1}^{m-b-1} 2(1-\gamma_l)\mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)})_{11} + \sum_{l=1}^{m-b} \|\mathbf{I} - \mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)}\|^2.$$

Denoting $P\mathbf{A}(l)P^*$ by $\mathbf{C}(l)$, we obtain from Theorem 4.3 that

$$\|\mathbf{I} - \mathbf{L}^{-1}_{\mathbf{C}(l)_{ME}}\mathbf{L}_{\mathbf{C}(l)}\| = \sqrt{2g_l - 2\sum_{i=1}^{g_l}\prod_{j=b_{l_i}+1}^{g_l}\sqrt{1-|\rho_{l_{ij}}|^2}}.$$

Because $|\rho_{l_{ij}}| < \epsilon$ for $i = 1, \ldots, g_l$ and $j = b_{l_i} + 1, \ldots, g_l$, it follows from Lemma 6.2 that

$$\|\mathbf{I} - \mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)}\| < 2\kappa(\mathbf{L}^{-1}_{\mathbf{C}(l)})\sqrt{\delta_l},$$

and as by Equation 6.5

$$\mathrm{tr}(\mathbf{L}^{-1}_{\mathbf{A}(l)_{ME}}\mathbf{L}_{\mathbf{A}(l)})_{11} < g_l,$$

the proof of the theorem follows readily. ■

Theorem 4.4 implies that $\|A^{-\frac{1}{2}}(A - A_H)A^{-\frac{1}{2}}\|$ and $\|A^{\frac{1}{2}}(A^{-1} - A_H^{-1})A^{\frac{1}{2}}\|$ are essentially twice $\|I - L_H^{-1}L\|$ (of which an estimate is given in Theorem 6.3). In a way similar to Theorem 6.3 we can estimate $\|A_H^{\frac{1}{2}}(A_H^{-1} - A_{SI}^{-1})A_H^{\frac{1}{2}}\|$ — it is equal to $\|A_H^{\frac{1}{2}}(B - C)A_H^{\frac{1}{2}}\|$, where B and C are as defined in Equations 6.1 and 6.2. From standard perturbation theory (see, e.g., [HJ85]) we obtain that if $\|A_H^{\frac{1}{2}}(A_H^{-1} - A_{SI}^{-1})A_H^{\frac{1}{2}}\| < 1$, then

$$\|A_H^{-\frac{1}{2}}(A_H - A_{SI})A_H^{-\frac{1}{2}}\| < \frac{\|A_H^{\frac{1}{2}}(A_H^{-1} - A_{SI}^{-1})A_H^{\frac{1}{2}}\|}{1 - \|A_H^{\frac{1}{2}}(A_H^{-1} - A_{SI}^{-1})A_H^{\frac{1}{2}}\|},$$

and it follows that both A^{-1}, A_H^{-1}, and A_{SI}^{-1} and A, A_H, and A_{SI} are close.

6.3 Concluding Remarks

To illustrate the above, we use the method of steepest descent described in Chapter 5 to minimize the function $g(C) = \mathrm{tr}(AC) - \log\det C$ subject to $C \in \mathcal{C}$ — the minimum occurs for $C = A_{ME}^{-1}$ — where A is 21×21 (it is the covariance matrix of 2 two dimensional sinusoids in white noise), $\kappa(A) \approx 98$, $\mathcal{S}$ is a triple band, the width of the bands in $\mathcal{S}$ is 2, and $\mathcal{C}$ is the set of positive definite matrices that vanish on the complement of $\mathcal{S}$. Figure 6.5 shows the course of the Frobenius norm of the gradient of $g(C)$ for initial guesses $C_0 = I$ and $C_0 = A_{SI}^{-1}$. We see that A_{SI}^{-1} and A_{ME}^{-1} are close — the gradient of $g(C)$ for $C = A_{SI}^{-1}$, a measure for the departure of A_{SI} from being an extension of A, is small.

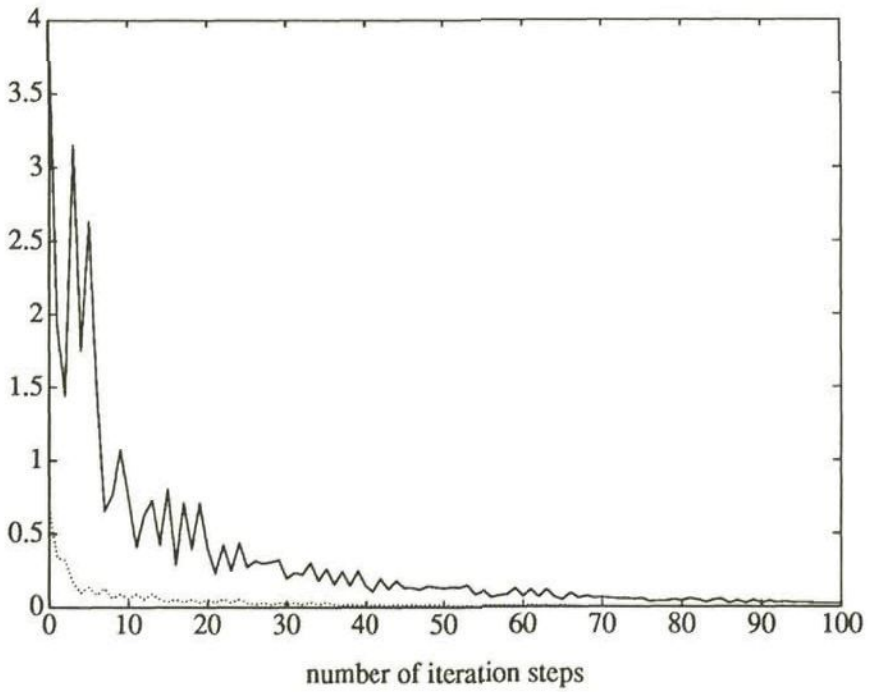


Figure 6.5: The Frobenius norm of the gradient of $g(C)$ as a function of the number of iteration steps for initial guesses $C_0 = I$ (solid) and $C_0 = A_{SI}^{-1}$ (dotted).

Chapter 7

A Model Reduction Example

IN THIS CHAPTER we apply the techniques developed in the preceding sections to a modeling problem in electrical engineering: modeling the parasitic capacitance of interconnections in a VLSI (Very Large Scale Integration) circuit. This capacitance largely determines the performance of a chip, and has a global character — interconnecting lines may influence each other over long distances, in a way that is dependent on their relative position in the (three dimensional) space. Since modern chips have hundreds of interconnects that crisscross their surface — see Figure 7.1 for a moderate example — it is nearly impossible to model the scene with reasonable accuracy. By using the methods described in the previous chapters, we are able, however, to obtain a reduced, yet accurate model for this complex situation.

7.1 The Problem

To determine the capacitance of interconnections in a VLSI circuit, we have to obtain a relationship between the potential of the conductors and the charge on their surfaces. For the three conductor problem shown in Figure 7.2 we have the following expression:

$$\begin{aligned} q_1 &= c_{11}v_1 + c_{12}(v_1 - v_2) + c_{13}(v_1 - v_3) \\ q_2 &= c_{21}(v_2 - v_1) + c_{22}v_2 + c_{23}(v_2 - v_3) \end{aligned}$$

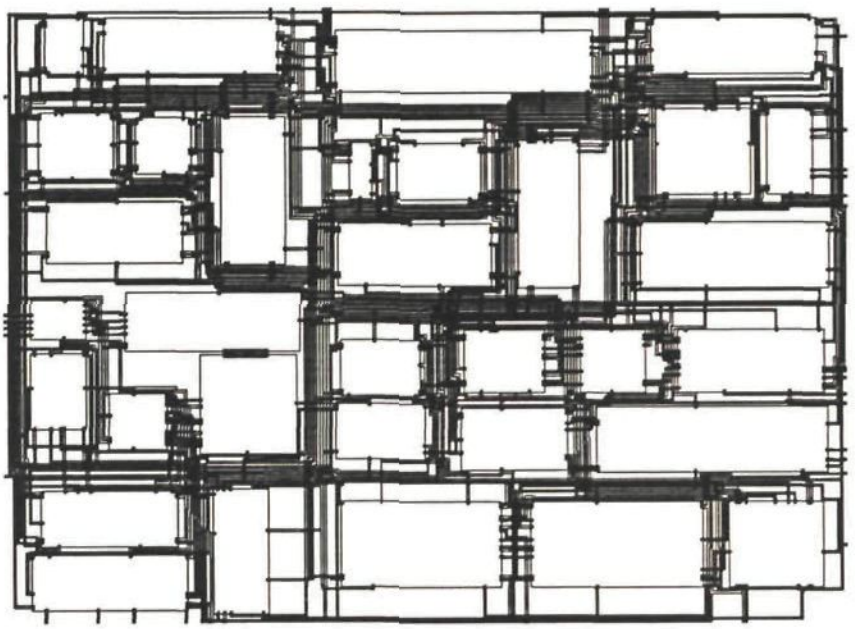


Figure 7.1: Interconnects in the layout of a chip.

$$q_3 = c_{31}(v_3 - v_1) + c_{32}(v_3 - v_2) + c_{33}v_3,$$

where q_i and v_i are the charge on and the potential of the ith conductor and c_{ij} $(i \neq j)$ is the coupling capacitance between conductor i and j (c_{ii} is the capacitance between conductor i and the ground).

For a system with m conductors we have

$$q_i = c_{ii}v_i + \sum_{j=1}^{m} c_{ij}(v_i - v_j), \quad i = 1, \ldots, m$$

or, equivalently,

$$q_i = \sum_{j=1}^{m} c_{sij}v_j, \quad i = 1, \ldots, m,$$

where

$$c_{sij} = -c_{ij}, \quad i \neq j$$

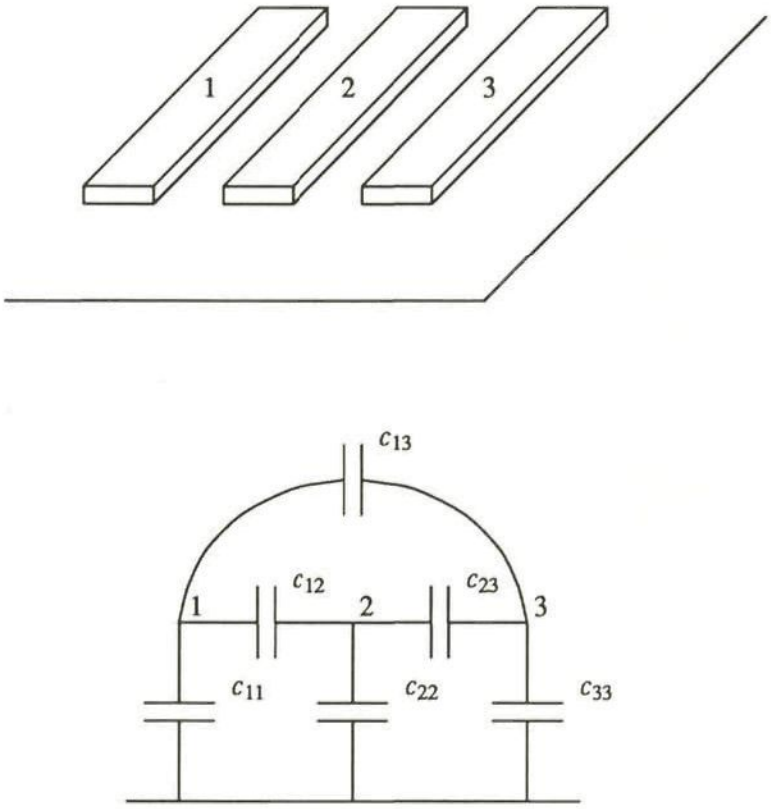


Figure 7.2: Three conductors above the ground and the equivalent circuit.

and

$$c_{sii} = \sum_{j=1}^{m} c_{ij}.$$

Assembling the charges and potentials in the vectors

$$Q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_m \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}$$

and the c_{sij}'s in the matrix $C_s = [c_{sij}]$, we obtain

$$Q = C_s V. \tag{7.1}$$

Because the system is conservative, C_s is positive definite. The c_{ij}'s and c_{sij}'s are called the *network* and *short circuit* capacitances respectively, and the modeling problem consists in determining the c_{ij}'s.

The silicon composite in which the interconnects lie can be viewed as a sandwich of essentially dielectric silicon, silicon dioxide, and silicon nitride, bounded at the bottom by a conducting layer of higher doped silicon and at the top by a coating and air. We assume that the composite consists of stratified, homogeneous, dielectric layers — see Figure 7.3 — and that the ground plane is ideally conducting.

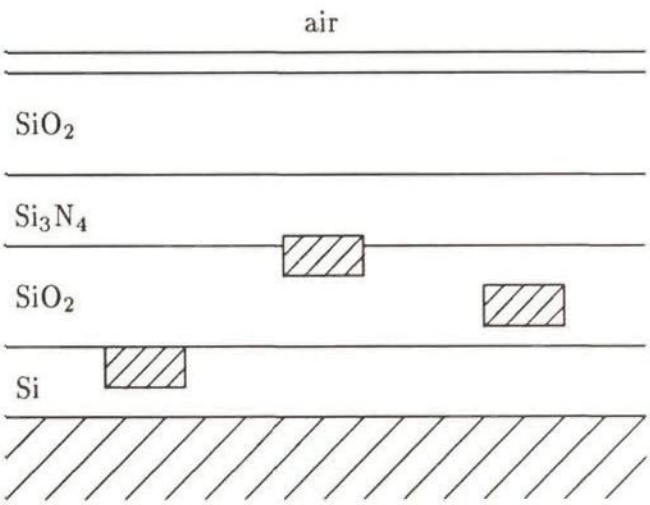


Figure 7.3: The silicon composite as a stratified medium. Conductors are hatched.

The potential $v(X)$ at a point X in the medium satisfies the Laplace equation, viz.,

$$\nabla^2 v(X) = 0,$$

with as boundary condition that $v(X)$ vanishes for X on the ground or at infinity. The solution to this equation is

$$v(X) = \int g(X,Y)\sigma(Y)dY, \tag{7.2}$$

where $g(X,Y)$ is the appropriate Green's function, which describes the potential induced at X by a unit point charge at Y, $\sigma(Y)$ is the charge density at Y, and the integration extends over the whole space (in fact, we integrate only where the charge density is nonzero, i.e., on the surfaces of the conductors). The Green's function for a medium with permittivity ϵ is given by

$$g(X,Y) = \frac{1}{4\pi\epsilon}\frac{1}{\|X-Y\|}.$$

For the stratified medium shown in Figure 7.3 it is a superposition of such functions.

By discretizing the charge on the surfaces of the conductors, we transform the Green's integral equation into a matrix equation. We concentrate the charge on a web of edges that spans the conductor surfaces — see Figure 7.4 — and attach to every node in the mesh a finite element (FE) of the *spider* type (see [Nin89]). Spider i consists of the edges that are in-

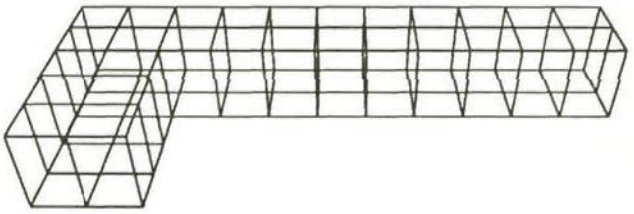


Figure 7.4: FE mesh for a conductor.

cident to node i and an elementary spline function $f_i(s)$ with the following properties:

- it slopes down linearly from its value at node i;

- it is zero at adjacent nodes and on the complement of the spider;
- $\int f_i(s)ds = 1$, where the integration extends over the edges of the spider.

We next approximate Equation 7.2 by

$$v(X) = \sum_{j=1}^{n} \left(\int g(X,s) f_j(s) ds \right) \bar{q}_j, \tag{7.3}$$

where n is the total number of FE's, the integration extends over the edges of the jth spider, and $\bar{q}_j$ is the charge on that element.

By evaluating Equation 7.3 at every node, we find

$$\bar{V} = \bar{G}\bar{Q}, \tag{7.4}$$

where

$$\bar{V} = \begin{bmatrix} \bar{v}_1 \\ \bar{v}_2 \\ \vdots \\ \bar{v}_n \end{bmatrix}, \quad \bar{Q} = \begin{bmatrix} \bar{q}_1 \\ \bar{q}_2 \\ \vdots \\ \bar{q}_n \end{bmatrix},$$

$\bar{v}_i$ is the potential of FE i, $\bar{G} = [\bar{g}_{ij}]$,

$$\bar{g}_{ij} = \int g(X_i, s) f_j(s) ds,$$

and X_i is the position of node i.

Denoting the FE-conductor incidence matrix by $F = [f_{ij}]$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, that is,

$$f_{ij} = \begin{cases} 1 & \text{if FE } i \text{ is on conductor } j; \\ 0 & \text{otherwise,} \end{cases}$$

we obtain from Equations 7.1 and 7.4 that

$$Q = F^* \bar{Q} = F^* \bar{G}^{-1} \bar{V} = F^* \bar{G}^{-1} F V,$$

and hence that

$$C_s = F^* \bar{G}^{-1} F,$$

and the network capacitances are calculated as

$$c_{ij} = -c_{sij}, \quad i \neq j$$

and

$$c_{ii} = \sum_{j=1}^{m} c_{sij}.$$

The number of components in the model can be very large: for a system with 500 interconnects we have a circuit with 125,250 capacitances. Such a model provides too much detail, and requires too much storage. Moreover, the computation of C_s is very expensive (for the above example and with 50 spiders per conductor we have to invert a $25,000 \times 25,000$ matrix, which takes $O(10^{13})$ operations and $O(10^9)$ storage) — we have to compute a reduced model.

7.2 Results

Many entries in C_s are so small that the corresponding network capacitances have almost no effect on the signals that propagate along the interconnections. For this reason, we can approximate C_s by a matrix that vanishes in the positions where C_s is small. Because of the special structure of the incidence matrix F, we can do so by approximating $\bar{G}^{-1}$ by a matrix that vanishes where $\bar{G}^{-1}$ is small (C_s is equal to $F^* \bar{G}^{-1} F$). As we have seen in the previous chapters, when the desired support $\mathcal{S}$ of the approximation is staircase, the inverse of the maximum entropy extension of the part of $\bar{G}$ with support on $\mathcal{S}$ is a suboptimal sparse approximation (in the Frobenius norm) to $\bar{G}^{-1}$, and we can compute it efficiently: we approximate $\bar{G}^{-1}$ by $\bar{G}_{ME}^{-1}$. When $\mathcal{S}$ is a multiple band, we determine $\bar{G}_{SI}^{-1}$ — it is close to $\bar{G}_{ME}^{-1}$, which is an optimal approximation (in the Kullback-Leibler measure) to $\bar{G}^{-1}$, and it can be obtained in an efficient way.

To illustrate the above, we approximate the circuit shown in Figure 7.5, which consists of four conductors with one FE each (in this case C_s is

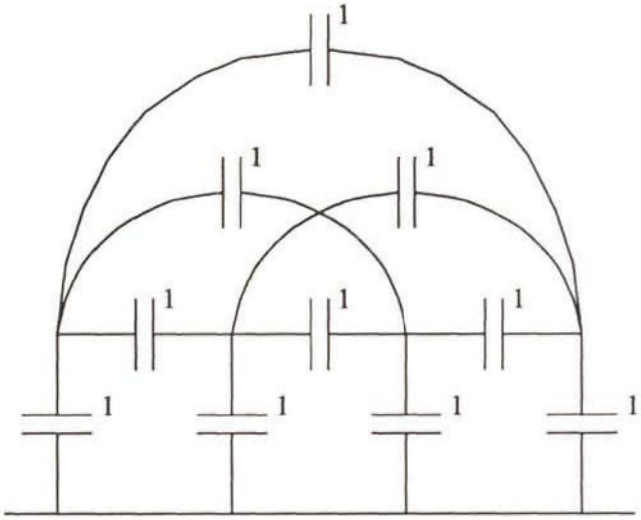


Figure 7.5: Original network.

equal to $\bar{G}^{-1}$). The first order approximation — see Figure 7.6 — is obtained by computing the inverse of the maximum entropy extension of the diagonal of $\bar{G}$. The second order approximation results when the first upper and lower diagonals are also taken into account, etc. The fourth order approximation is equal to the original circuit. The circuits have the following properties: the first order circuit is such that the driving point capacitances of the nodes are the same as in the original circuit; the second order circuit is such that the 2×2 capacitance matrices for the ports $(1,2)$, $(2,3)$, and $(3,4)$ (assume that the nodes are numbered from left to right) are equal to those of the original circuit when the other nodes are floating, etc.

When the FE's lie in the two or three dimensional space, we cannot number them such that the significant entries in the capacitance matrix form a staircase band. With a lexicographic ordering, and when only capacitances between nearest neighbor FE's are considered to be important, they form a multiple band — see Figure 7.7 (when we draw an arc between FE's with a direct coupling, we obtain Figure 2.1).

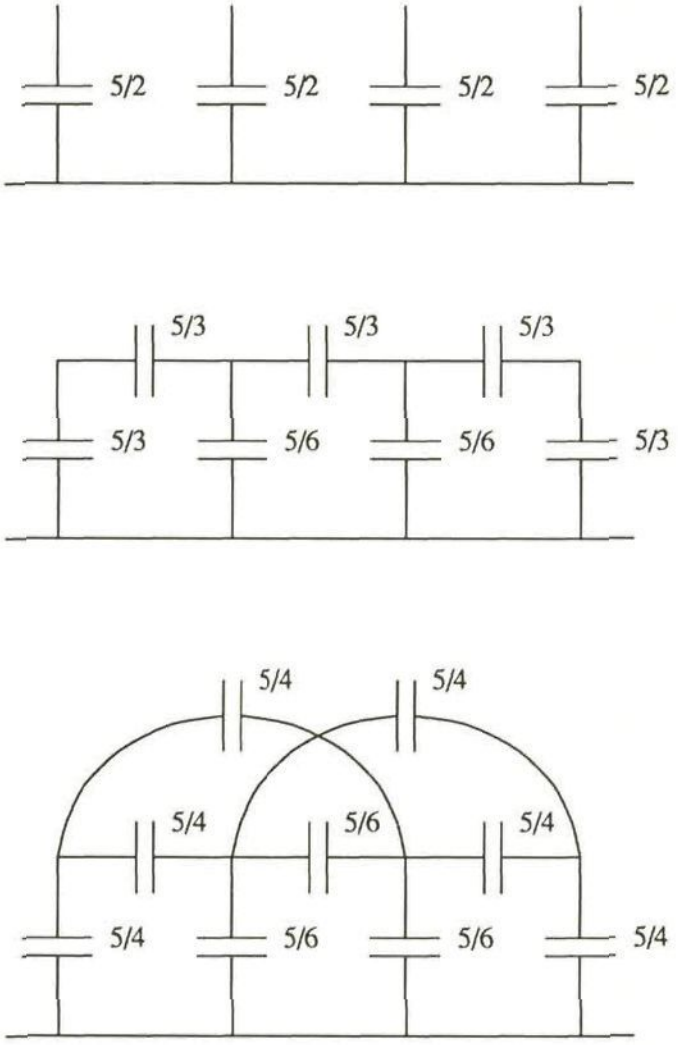
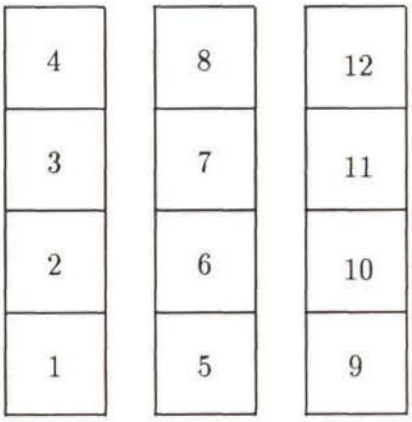


Figure 7.6: First, second, and third order approximation.



$$\bar{G}^{-1} =$$

	1	2	3	4	5	6	7	8	9	10	11	12
1	×	×			×	×						
2	×	×	×		×	×	×					
3		×	×	×		×	×	×				
4			×	×			×	×				
5	×	×			×	×			×	×		
6	×	×	×		×	×	×		×	×	×	
7		×	×	×		×	×	×		×	×	×
8			×	×			×	×			×	×
9					×	×			×	×		
10					×	×	×		×	×	×	
11						×	×	×		×	×	×
12							×	×			×	×

Figure 7.7: Two dimensional numbering scheme and matrix. Nonzero entries are marked with an '×'; vanishing entries are blank

The techniques described above have been integrated in a layout-to-circuit extractor called SPACE (see [Gen90, Mei90]). The input is a layout specification. The output is a description of the circuit. The extractor defines a window of size $w \times w$ that is swept over the layout. Within the window a mesh of spiders is created, the entries in $\bar{G}$ on $\mathcal{S}$ are determined, and $\bar{G}_{SI}^{-1}$ is computed on the fly.

We close this section with some experimental results that give an impression of the accuracy and efficiency of the method. We first consider a situation with 5 parallel conductors that lie 1μm above the substrate and 1.25μm apart. They are 1μm high and wide, and have a length of 10μm. The mesh for this geometry consists of 220 spiders, located 1μm apart on top and bottom of the conductors — see Figure 7.8. The capacitance values and the

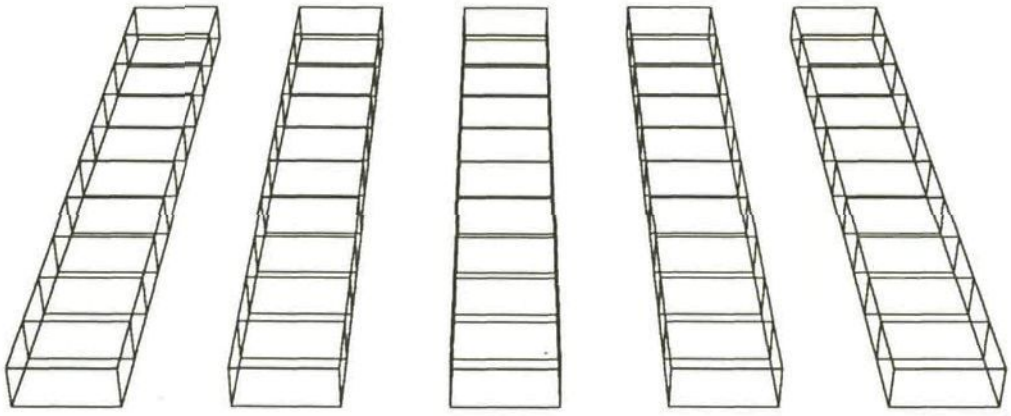


Figure 7.8: FE mesh for 5 parallel conductors.

cpu time and memory usage of SPACE are shown in Table 7.1. The last two figures refer to the matrix inversion only. The experiments were done on a HP 9000-840 computer with 8 Mbyte of physical memory. The case with the window of 10μm $\times$ 10μm corresponds to an exact inversion of $\bar{G}$. The c_{sii}'s largely determine the timing of the circuit, and are already very accurate for small window sizes.

window (μm $\times$ μm)	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}	c_{s11}	cpu time (min:sec)	storage (Kbyte)
			(10^{-18} F)					
10 x 10	1128	557.0	37.59	14.93	9.132	1746	2:06	2441
8 x 8	1134	556.9	37.70	17.21		1746	1:30	1196
6 x 6	1148	556.7	42.79			1748	0:46	526
4 x 4	1184	568.4				1752	0:15	169
2 x 2	1578					1578	0:02	27

Table 7.1: Capacitance values, cpu time, and memory usage for 5 parallel conductors.

Tables 7.2 and 7.3 give the cpu time and memory usage for situations with an increasing number of parallel conductors of increasing length. The former shows the results for an exact extraction. When the number of conductors is 14 or more, we cannot compute $\tilde{G}^{-1}$ — it requires too much storage. The latter corresponds to an approximate extraction (using a window of 4μm $\times$ 4μm).

# conductors	length (μm)	# spiders	# Green's evaluations	cpu time (min:sec)	storage (Kbyte)
5	10	220	24310	2:11	1152
5	20	420	88410	16:08	4200
10	20	840	353220	176:14	16800
14	29	1680	-	-	-
20	41	3360	-	-	-

Table 7.2: Cpu time and memory usage for an increasing number of parallel conductors of increasing length (exact extraction).

Table 7.4 gives the results for a practical example: the extraction of a static RAM cell in a 1 μm CMOS technology (using a window of 6μm $\times$ 3μm). The FE mesh for this case is shown in Figure 7.9.

# conductors	length (μm)	# spiders	# Green's evaluations	cpu time (min:sec)	storage (Kbyte)
5	10	220	12250	0:15	169
5	20	420	26150	0:33	322
10	20	840	55848	1:06	322
14	29	1680	122340	2:26	461
20	41	3360	253872	5:14	645

Table 7.3: Cpu time and memory usage for an increasing number of parallel conductors of increasing length (4μm $\times$ 4μm window).

cpu time (min:sec)	10:41
time for matrix inversion	5:25
time for Green's evaluation	4:34
memory (Mbyte)	6.9

Table 7.4: Cpu time and memory usage for a CMOS static RAM cell.

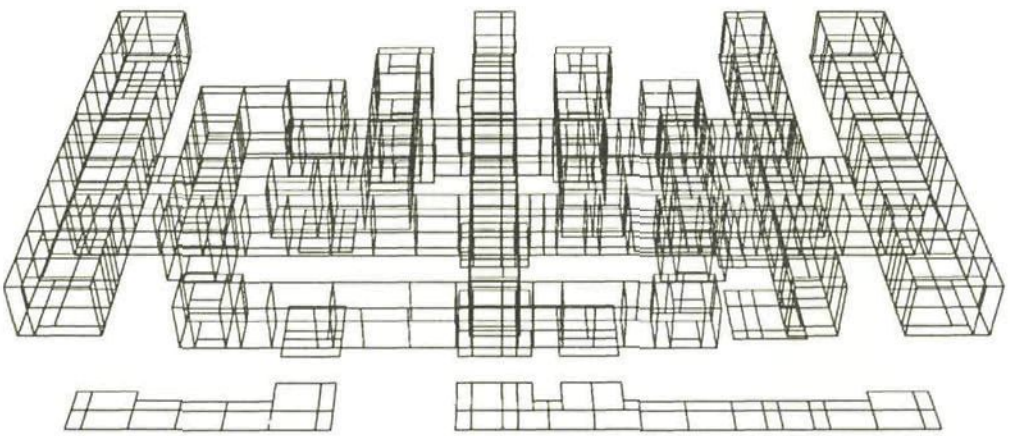


Figure 7.9: FE mesh for a CMOS static RAM cell.

Chapter 8

Concluding Remarks

IN THIS DISSERTATION we have studied the problem of finding an optimal sparse approximation to the inverse of a positive definite matrix. It has been viewed as a model reduction problem. We have followed an approach that is based on matrix extension theory and inverse scattering theory, and we have devised algorithms for determining an optimal or suboptimal approximation. The method can be applied, for example, to obtain reduced models for problems that are governed by the Laplace equation. For the case of modeling parasitic capacitances in VLSI circuits very good results have been obtained. We conclude this thesis by indicating some of the problems that have not been solved yet.

In many problems we have an underlying stationarity or homogeneity (invariance under displacements in time or space) property that often leads to the matrix A being Toeplitz. It is known that a Toeplitz matrix can be inverted in $O(n^2)$ operations and with $O(n)$ storage, where n is the size of the matrix. For instance, see the first example in Chapter 4, where because of the Toeplitz structure of A many of the computations are identical. In [KKM79] Kailath et al. studied the more general, so-called α-stationary case. The integer α, $1 \leq \alpha \leq n$, is called the displacement rank of the matrix, and provides a measure of how close a matrix is to being Toeplitz. They showed that a matrix with index α can be inverted with (about) α times as much computations and storage as required for a Toeplitz matrix. This also

has impact on the computation of the inverse of the sparse-inverse approximation defined in Chapter 6 — see Definition 6.2. For example, consider the case where A is as shown in Figure 8.1 (matrices of this type arise in, e.g., two dimensional spectral estimation). The inverse of the sparse-inverse

$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \end{array} \begin{array}{c} \begin{array}{cccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \end{array} \\ \left[\begin{array}{cccc|cccc|cccc} 1 & a & & & c & d & & & & & & \\ a & 1 & a & & b & c & d & & & & & \\ & a & 1 & a & & b & c & d & & & & \\ & & a & 1 & & & b & c & & & & \\ \hline c & b & & & 1 & a & & & c & d & & \\ d & c & b & & a & 1 & a & & b & c & d & \\ & d & c & b & & a & 1 & a & & b & c & d \\ & & d & c & & & a & 1 & & & b & c \\ \hline & & & & c & b & & & 1 & a & & \\ & & & & d & c & b & & a & 1 & a & \\ & & & & & d & c & b & & a & 1 & a \\ & & & & & & d & c & & & a & 1 \end{array}\right] \end{array}$$

Figure 8.1: A Toeplitz-block Toeplitz matrix that is specified on a multiple band.

approximation of A is given by

$$A_{SI}^{-1} = \square\left[\mathbf{A}(1,2)_{ME}^{-1};(1,2)\right] + \square\left[\mathbf{A}(2,3)_{ME}^{-1};(2,3)\right] - \square\left[\mathbf{A}(2,2)_{ME}^{-1};(2,2)\right],$$

and we only need to compute $\mathbf{A}(1,2)_{ME}^{-1}$ and $\mathbf{A}(2,2)_{ME}^{-1}$ — $\mathbf{A}(2,3)_{ME}^{-1}$ is equal to $\mathbf{A}(1,2)_{ME}^{-1}$. To determine $\mathbf{A}(1,2)_{ME}^{-1}$, we permute $\mathbf{A}(1,2)$ to a matrix that is specified on a staircase band — see Figure 8.2. This destroys the Toeplitz structure, but the permuted matrix has displacement rank $\alpha = 4$ — we can determine the inverse of its maximum entropy extension with twice as much computations and storage as required for a Toeplitz matrix.

Often there is structure in the physical problem, but the resulting matrix has a high displacement rank. For example, consider the problem of

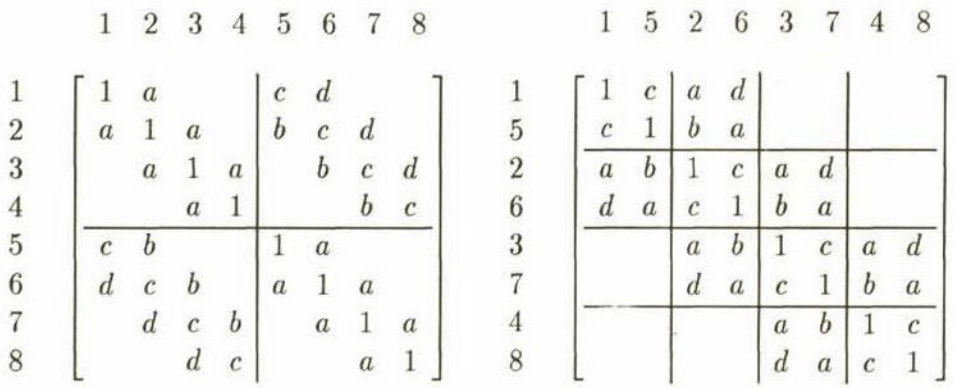


Figure 8.2: The principal submatrix $\mathbf{A}(1,2)$ and the permuted matrix.

modeling parasitic capacitances in a VLSI circuit that has been described in Chapter 7. A memory chip has a very regular structure — it consists of thousands of identical cells — which would lead to a matrix with a low displacement rank, but irregularities at the boundaries and interconnections that crisscross the chip spoil the game. A technique for obtaining a model with a reduced complexity is proposed in [Gen90]. It is hierarchical — it computes a model for a cell and a model for the boundaries of the chip and the interconnects, and combines them to a model for the complete system — and based on heuristics. It can be shown, however, that assumptions similar to those made in Chapter 6 (replace 'maximum entropy' by 'sparse-inverse') and an error analysis similar to the one in that chapter provide a justification for the method.

Another area for further investigation is the derivation of other types of reduced models. For example, it is possible to represent the inverse of the hierarchical approximation defined in Chapter 6 — see Definition 6.1 — in a mixed form that consists of matrices with support on a staircase band and inverses of such matrices. Some computations with representations of this kind are efficient: for instance, the product of a vector and the matrix can be computed in an efficient way.

We hope that we have convinced the reader of the power of our approach, and that this booklet provides an impetus for further research in this fascinating field.

Bibliography

[Bur75] J. Burg. *Maximum-Entropy Spectral Analysis.* PhD thesis, Dept. of Geophysics, Stanford University, Stanford, California, 1975.

[DD87] P. Dewilde and E. Deprettere. Approximate inversion of positive matrices with applications to modelling. In *Modelling, Robustness and Sensitivity Reduction in Control Systems*, Springer-Verlag, 1987.

[DD88] P. Dewilde and E. Deprettere. The generalized Schur algorithm: approximation and hierarchy. *Operator Theory: Advances and Applications*, 29, 1988.

[Dep81] E. Deprettere. Mixed form time-variant lattice recursions. In *Outils et Modèles Mathématiques pour l'Automatique, l'Analyse des Systèmes et le Traitement du Signal*, CNRS France, 1981.

[DG79] H. Dym and I. Gohberg. Extensions of matrix valued functions with rational polynomial inverses. *Integral Equations and Operator Theory*, 2, 1979.

[DG81] H. Dym and I. Gohberg. Extensions of band matrices with band inverses. *Linear Algebra and its Applications*, 36, 1981.

[DG88] H. Dym and I. Gohberg. A new class of contractive interpolants and maximum entropy principles. *Operator Theory: Advances and Applications*, 29, 1988.

[DGK80] Ph. Delsarte, Y. Genin, and Y. Kamp. A method of matrix inverse triangular decomposition, based on contiguous principal submatrices. *Linear Algebra and its Applications*, 31, 1980.

[DVK78] P. Dewilde, A. Vieira, and T. Kailath. On a generalized Szegö-Levinson realization algorithm for optimal linear prediction based on a network synthesis approach. *IEEE Transactions on Circuits and Systems*, 25, 1978.

[Gen90] A. van Genderen. *Verification of Complex VLSI Circuits.* PhD thesis, Dept. of Electrical Engineering, Delft University of Technology, Delft, The Netherlands, to appear Januari 1990.

[GJSW84] R. Grone, C. Johnson, E. Sá, and H. Wolkowicz. Positive definite completions of partial Hermitian matrices. *Linear Algebra and its Applications*, 58, 1984.

[GKW89a] I. Gohberg, M. Kaashoek, and H. Woerdeman. The band method for positive and contractive extension problems. *Journal of Operator Theory*, to appear 1989.

[GKW89b] I. Gohberg, M. Kaashoek, and H. Woerdeman. The band method for positive and contractive extension problems: an alternative version and new applications. *Integral Equations and Operator Theory*, 12, 1989.

[GV83] G. Golub and C. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, 1983.

[HJ85] R. Horn and C. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

[KKM79] T. Kailath, S.-Y. Kung, and M. Morf. Displacement ranks of matrices and linear equations. *Journal of Mathematical Analysis and Applications*, 68, 1979.

[Lev47] N. Levinson. The Wiener rms (root mean square) error criterion in filter design and prediction. *Journal of Mathematical Physics*, 25, 1947.

[Lev85] H. Lev-Ari. Multidimensional maximum-entropy covariance extension. In *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, 1985.

[LHDB88] A. de Lange, A. van der Hoeven, E. Deprettere, and J. Bu. An optimal floating-point pipeline CMOS CORDIC processor: algorithm, automated design, layout, and performance. In *Proceedings International Symposium on Circuits and Systems*, 1988.

[LK81] H. Lev-Ari and T. Kailath. Schur and Levinson algorithms for nonstationary processes. In *Proceedings International Conference on Acoustics, Speech, and Signal Processing*, 1981.

[LM81] J. Lim and N. Malik. A new algorithm for two-dimensional maximum entropy power spectrum estimation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29, 1981.

[LM82] S. Lang and J. McClellan. Multidimensional MEM spectral estimation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 30, 1982.

[LPK89] H. Lev-Ari, S. Parker, and T. Kailath. Multi-dimensional maximum-entropy covariance extension. *IEEE Transactions on Information Theory*, 35, 1989.

[Lue73] G. Luenberger. *Introduction to Linear and Nonlinear Programming.* Addison-Wesley Publishing Company, Inc., 1973.

[MD81] M. Morf and J.-M. Delosme. Matrix decompositions and inversions via elementary signature-orthogonal transformations. In *Proceedings International Symposium on Mini- and Micro computers in Control and Measurement*, 1981.

[Mei90] N. van der Meijs. *Accurate and Efficient Layout Extraction.* PhD thesis, Dept. of Electrical Engineering, Delft University of Technology, Delft, The Netherlands, to appear Januari 1990.

[Nin89] Z.-Q. Ning. *Accurate and Efficient Modeling of Global Circuit Behaviour in VLSI Layouts.* PhD thesis, Dept. of Electrical Engineering, Delft University of Technology, Delft, The Netherlands, 1989.

[Ros72] D. Rose. A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations. In *Graph Theory and Computing*, Academic Press, 1972.

[RP87] N. Rozario and A. Papoulis. Spectral estimation from nonconsecutive data. *IEEE Transactions on Information Theory*, 33, 1987.

[RR85] M. Rosenblum and J. Rovnyak. *Hardy Classes and Operator Theory.* Oxford University Press, 1985.

[Sch17] I. Schur. Über Potentzreihen die im Innnern des Einheitskreises beschrankt sind. *Journal für die Reine und Angewandte Mathematik*, 147, 1917.

[Ste77] G. Stewart. Perturbation bounds for the QR factorization of a matrix. *SIAM Journal of Numerical Analysis*, 14, 1977.

[Vol59] J. Volder. The Cordic trigonometric computing technique. *IRE Transactions on Electronic Computers*, 8, 1959.

[Wal71] J. Walther. A unified algorithm for elementary functions. In *Proceedings Spring Joint Computer Conference*, 1971.

[Woe89] H. Woerdeman. *Matrix and Operator Extensions.* PhD thesis, Dept. of Mathematics and Computer Science, Free University, Amsterdam, The Netherlands, 1989.

Samenvatting

HET OPLOSSEN van een groot aantal problemen in de natuurwetenschappen en de techniek leidt uiteindelijk tot het inverteren van een positief definiete matrix. Wanneer deze matrix groot is (bijv. 10.000 × 10.000), zoals het geval kan zijn in modelleringsproblemen, is dit vrijwel onbegonnen werk. Om het aantal berekeningen te beperken, probeert men vaak eigenschappen van het oorspronkelijke, fysische probleem in rekening te brengen. In dit proefschrift leggen we structuur op aan de inverse van de matrix — we nemen aan dat hij benaderd kan worden door een ijle matrix (d.w.z. een matrix waarvan veel elementen gelijk aan nul zijn).

We laten zien hoe technieken uit de inverse verstrooiingstheorie zoals de Wiener-Hopf ontbinding en het Schur algoritme gebruikt kunnen worden om een optimale of suboptimale ijle benadering van de inverse van een positief definiete matrix te bepalen. We gebruiken alleen elementen uit de oorspronkelijke matrix die overeenkomen met niet-nul elementen in de benadering. De algoritmen die voorgesteld worden hebben een complexiteit die evenredig is met het aantal niet-nul elementen.

Hoofdstuk 2 gaat over het bepalen van een optimale ijle benadering van de inverse van een positief definiete matrix. Eerst bekijken we het geval waar het ijlheidspatroon monotoon transitief is. We tonen aan dat de driehoeksfactoren van de inverse van de zogenaamde maximum entropie uitbreiding — we gebruiken de elementen uit de oorspronkelijke matrix die overeenkomen met niet-nul elementen in de benadering, en schatten de andere — optimale ijle benaderingen in de Frobenius norm zijn van de driehoeksfactoren van de inverse van de oorspronkelijke matrix. Wanneer het

ijlheidspatroon willekeurig is gaat dit niet op, maar van alle matrices die een ijle inverse hebben ligt de maximum entropie uitbreiding het dichtst bij de oorspronkelijke matrix in de Kullback-Leibler maat.

Hoodstuk 3 beschrijft een veralgemening van de Wiener-Hopf ontbindingstheorie naar het geval van algemene, positief definiete matrices met eindige dimensies die gespecificeerd zijn op een blok band. Deze theorie geeft het verband tussen de klassieke verstrooiingstheorie en de uitbreidingstheorie. Zij stelt ons in staat een globale oplossing van een veralgemeend invers verstrooiingsprobleem te construeren, dat gelijkwaardig blijkt te zijn aan het maximum entropie uitbreidingsprobleem. We bepalen de oplossing en daarmee de driehoeksfactoren van de inverse van de maximum entropie uitbreiding door een stelsel lineaire vergelijkingen op te lossen.

Hoofdstukken 4, 5 en 6 gaan over algoritmen voor het berekenen van de inverse van de maximum entropie uitbreiding. Wanneer het ijlheidspatroon een trapvorm heeft, gebruiken we het Schur algoritme om de driehoeksfactoren van deze matrix uit te rekenen. Het algoritme is bij uitstek geschikt om uitgevoerd te worden op een array processor van het systolische of golffront type. Voor algemene ijlheidspatronen zijn we aangewezen op iteratieve algoritmen. Omdat ze veel berekeningen en geheugen vergen, leiden we voor het belangrijke geval waar het ijlheidspatroon een meervoudige band is een algoritme af om een benadering van de inverse van de maximum entropie uitbreiding te bepalen. Het algoritme is gebaseerd op het Schur algoritme, en berekent de inverse van de maximum entropie uitbreiding van een matrix die dicht bij de oorspronkelijke ligt.

Hoofdstuk 7 behandelt een toepassing: het modelleren van parasitaire capaciteiten in grote geintegreerde circuits. Hier blijkt de kracht van de methodes — we zijn in staat om een nauwkeurig model voor een systeem met een groot aantal geleiders te berekenen, terwijl men in de literatuur het modelleren van een systeem met slechts een paar geleiders al als een enorm probleem beschouwt.

About the Author

Hendrik Willem Nelis was born in Haarlem, The Netherlands, on February 19, 1964. In May 1982 he received the Gymnasium diploma from the 'Christelijk College Marnix van St. Aldegonde' in Haarlem. In September of that year he started his study of electrical engineering at the Delft University of Technology in Delft, The Netherlands. Four years later he received the Ingenieur degree (the equivalent of an M.Sc.). In September 1986 he joined the Laboratory for Network Theory of the Department of Electrical Engineering at the Delft University of Technology, where he embarked on his work towards the Ph.D. degree. From June 1987 until September of that year he was a visiting researcher at the Information Systems Laboratory of Stanford University, Ca., U.S.A. Most of his work has been published in international journals. His main interest is in parallel algorithms and architectures.

Stellingen

behorende bij het proefschrift van

Harry Nelis

1. Het verschil tussen het schatten van één- en twee-dimensionale spectra is meer dan een dimensie.

2. Een gereduceerd model geeft meer inzicht dan een gedetailleerd model, en het is makkelijker te berekenen.

3. De maximum entropie uitbreiding van een Toeplitz-blok Toeplitz matrix die gespecificeerd is op een meervoudige band is niet meer Toeplitz-blok Toeplitz.

4. Een alternatieve probleemstelling maakt het gebruik van parallel rekentuig vaak overbodig.

5. Het keep-your-lane systeem op snelwegen drukt de maximum snelheid, en het geeft een rustiger verkeersbeeld.

6. Met een geavanceerde tekstverwerker gaat het vervaardigen van een document niet noodzakelijk sneller dan met een eenvoudige.

7. The purpose of computing is insight, not numbers. — R. Hamming

8. Het verkrijgen van goede resultaten is moeilijk. Ze op een duidelijke manier presenteren is zo mogelijk nog lastiger.

9. Een promotie heeft veel weg van een (mini) triathlon: het is pas echt leuk als het voorbij is.