Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

ESTIMATION WHEN BOTH COVARIANCE AND PRECISION MATRICES ARE SPARSE

Shev MacNamara
Erik Schlögl

School of Mathematical & Physical Sciences
University of Technology Sydney
Broadway, Ultimo NSW 2007
Sydney, AUSTRALIA

Zdravko Botev

School of Mathematics & Statistics
The University of New South Wales
High St. Kensington, NSW 2052
Sydney, AUSTRALIA

ABSTRACT

We offer a method to estimate a covariance matrix in the special case that *both* the covariance matrix and the precision matrix are sparse — a constraint we call double sparsity. The estimation method is maximum likelihood, subject to the double sparsity constraint. In our method, only a particular class of sparsity pattern is allowed: both the matrix and its inverse must be subordinate to the same chordal graph. Compared to a naive enforcement of double sparsity, our chordal graph approach exploits a special algebraic local inverse formula. This local inverse property makes computations that would usually involve an inverse (of either precision matrix or covariance matrix) much faster. In the context of estimation of covariance matrices, our proposal appears to be the first to find such special pairs of covariance and precision matrices.

1 INTRODUCTION

We begin with a quote from Rothman et al. (2010): "*As a rule of thumb in practice, if it is not clear from the problem whether it is preferable to regularize the covariance or the inverse, we would recommend fitting both and choosing the sparser estimate*." The authors are describing methods that can estimate a covariance matrix or a precision matrix when one – but not both – of these matrices is sparse. For example, Bickel and Levina (2008) exploit sparsity patterns when estimating a covariance matrix or its inverse, and Ravikumar et al. (2011) explore related problems for high dimensional estimation.

In this article we demonstrate that it is in fact possible to be greedy, and ask for simultaneous sparsity in both the covariance and precision matrices. Our estimator is a maximum likelihood estimator with the constraint that both the covariance and its inverse be sparse. To the best of our knowledge, this is the first proposal to seek efficient estimation for the doubly sparse case (when both covariance matrix and precision matrix are sparse).

Some of the advantages to imposing sparsity in both the covariance and its inverse are simplicity of interpretation and faster computation via local formulas that we describe below (these formulas allow us to work with both the matrix and its inverse in an efficient way).

Estimating a covariance matrix is an important problem in the subject of statistics. Methods of imposing sparsity in the covariance matrix — or more commonly in the inverse of the covariance matrix (the precision matrix) — have attracted a great deal of attention. A very popular example of such an approach is the *graphical LASSO* of Friedman et al. (2008). One reason these methods are important is that the corresponding model is more interpretable when the precision matrix is sparse, which is important in the subjects of Gaussian Markov random fields (Salemi, Song, Nelson, and Staum 2019) and of graphical models and covariance selection (Lauritzen 1996).

When there is a known relationship between entries in a matrix (covariance matrix, precision matrix, or Cholesky factor, for example) and coefficients in regression, then it is possible to apply many available

978-1-6654-3311-2/21/$31.00 ©2021 IEEE

methods from regression that impose zeros in the regression coefficients, as a way to impose zeros in the matrix. Financial applications of such methods include the shrinkage estimation algorithm of Ledoit and Wolf (2004) and the missing data completion algorithm of Georgescu et al. (2018).

Missing data completion is closely related to the *Dempster completion*, which is closely related to the local inverse formula that we describe later. The Dempster (1972) completion of a covariance matrix in which there are missing entries has an especially satisfying form when the inverse is chordal. Then Georgescu et al. (2018) note the completion brings together a number of attractive properties (Strang and MacNamara 2018; Grone et al. 1984; Dym and Gohberg 1981; Johnson and Lundquist 1998):

- sub-blocks of these matrices away from the main diagonal have low rank, as in The Nullity Theorem (Strang 2010; Strang and Nguyen 2004) ; these matrices are examples of semi-separable matrices (Vandebril et al. 2007);
- the inverse can be found directly, without completing the covariance matrix, via a 'local inverse formula' (as shown in the symmetric positive definite case by Lauritzen 1996; Speed and Kiiveri 1986) that uses only information in the blocks on the main diagonal of the incomplete covariance matrix;
- the completed matrix maximizes the log-determinant amongst the cone of symmetric positive definite matrices consistent with the original incomplete covariance matrix, and it is a maximum entropy estimate.

When an existing method succeeds in imposing sparsity in the precision matrix, then typically the corresponding covariance matrix is not sparse. *Vice-versa*, if the covariance matrix is sparse then the precision matrix is typically not sparse. In summary, all methods that are currently available in the literature are *not* able to impose sparsity simultaneously in *both* the covariance matrix and the precision matrix. Indeed, informally, if a sparse matrix is chosen "at random" then *all* entries of the corresponding inverse matrix are non-zero. A sparse matrix with a sparse inverse is an extremely exceptional case. That exceptional case, applied to the problem of estimating a covariance matrix, is the subject of the subsequent sections.

In all subsequent sections, we refer to a chordal graph $\mathscr{G}$ and the junction tree $\mathscr{J}$ for that graph. A definition of a chordal graph is that all cycles of four or more vertices have a chord. A chord is an edge that is not part of the cycle, but that connects two vertices of the cycle. There are other equivalent characterizations of chordal graphs, such as the graphs that have perfect elimination orderings. (See, e.g. (Vandenberghe and Andersen 2015; Strang and MacNamara 2018; Johnson and Lundquist 1998).) Chordal graphs are also known as decomposable graphs, in the graphical models literature (Lauritzen 1996).

2 DOUBLY SPARSE COVARIANCE AND CONSTRAINED LIKELIHOOD

Block diagonal covariance matrices are the simplest examples for which both the matrix and its inverse are sparse. A diagonal covariance matrix $V = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ has inverse $V^{-1} = \mathrm{diag}(1/\sigma_1^2, \ldots, 1/\sigma_n^2)$. Similarly, a block-diagonal matrix and its inverse have the same sparsity pattern.

Other than such (block) diagonal trivial examples, it is not clear if it is possible to construct nontrivial examples in which both the covariance and its inverse have the same sparsity pattern — a phenomenon we dub *doubly sparse covariance*.

Do such doubly sparse covariance matrices exist? The answer is 'yes', and examples can be constructed, where both the covariance matrix and its inverse (or precision matrix) are assumed chordal:

$$V \equiv \begin{pmatrix} 13 & 8 & 4 & 2 & 0 & 0 \\ 8 & 13 & 2 & 1 & 0 & 0 \\ 4 & 2 & 10 & 6 & 1 & 1 \\ 2 & 1 & 6 & 13 & 10 & 10 \\ 0 & 0 & 1 & 10 & 13 & 8 \\ 0 & 0 & 1 & 10 & 8 & 13 \end{pmatrix} \tag{1}$$

with inverse (or precision matrix) $\Theta \equiv V^{-1}$ given by

$$\Theta = \begin{pmatrix} 2960 & -1690 & -900 & 90 & 0 & 0 \\ -1690 & 2675 & 150 & -15 & 0 & 0 \\ -900 & 150 & 8715 & -12180 & 5385 & 5385 \\ 90 & -15 & -12180 & 23835 & -10770 & -10770 \\ 0 & 0 & 5385 & -10770 & 7539 & 3231 \\ 0 & 0 & 5385 & -10770 & 3231 & 7539 \end{pmatrix} /21540. \tag{2}$$

It can be quickly checked this pair are also positive definite. Both V and Θ are *not* block diagonal, demonstrating that the class of matrices we are considering in this article is richer than simply block diagonal matrices.

While chordal covariance matrices have been studied before, there seem to be no examples in the literature where both the covariance and its inverse are chordal. We will revisit this same matrix example later in (6) to show that it has some local inverse properties.

2.1 Constrained Maximum Likelihood

We want both the matrix M and the inverse M^{-1} to be *subordinate* to the same chordal graph. i.e. if there is no edge between nodes i and j then the (i, j) entry of M is zero, and the (i, j) entry of M^{-1} is zero. We assume that we are given iid sample data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^p$, drawn from $\mathcal{N}(\boldsymbol{\mu}, M)$ for some unknown $\boldsymbol{\mu}$ and M. Let the sample covariance matrix be

$$S = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})^\top, \quad \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i.$$

Under the normality assumption and after eliminating the nuisance parameter $\boldsymbol{\mu}$ (by replacing it with its MLE $\hat{\boldsymbol{\mu}}$), the maximization of the log-likelihood is equivalent to the minimization (with respect to M):

$$\operatorname{trace}(SM^{-1}) + \ln|M|.$$

In our constrained MLE framework, we want to find a symmetric positive definite matrix M that *minimizes the objective function*

$$\operatorname{trace}\left(M^{-1}S\right) + \ln|M| \tag{3}$$

subject to the constraint that

$$M \text{ and } M^{-1} \quad \text{are subordinate to } \mathscr{G}. \tag{4}$$

The main novelty in our constrained optimization is that we will make use of the so-called *local inverse formula* to impose the doubly sparse constraint in (4). We describe that formula in detail in the next section. This is a simple formula that relates the inverse to inverses of sub-blocks of the matrix, and which applies when the graph is subordinate to a chordal graph.

The same idea allows us to compute $\det(M)$ in the objective function by a local formula for the determinant (Johnson and Lundquist 1998). A main advantage is that we can easily define a function $L(M)$ based on the local inverse formula that allows us to parameterize the set of matrices in the constraint using only one subset of the variables. For example, we can use only the subset of nonzero entries of M, and then $L(M)$ can be used in place of M^{-1}. (Or vice-versa: we could parameterize with only the subset of entries of M^{-1}, and then use $L(M^{-1})$ in place of M.)

The optimization also exploits the following well-known observations:

- We need only optimize over a subset of the entries of a Cholesky factor, R, that correspond to the chordal graph. Then form M, if required, as $M = R^\top R$, for example. That is, *we parameterize only with the subset of entries in a Cholesky factor*, R.

- Parameterizing via a Cholesky factor automatically ensures that we optimise over positive definite matrices. Also, parameterizing via a Cholesky factor exploits our knowledge of numerical analysis and our knowledge of chordal graphs, that chordal graphs also correspond to 'perfect eliminators.' That is, the sparsity pattern is preserved (Blair and Peyton 1993).
- Sometimes it may be better to parameterize by the entries of the precision matrix, or by the entries of the Cholesky factor of the precision matrix (Pourahmadi 2013), but we have made no attempt to compare the relative merits of those two approaches.

Before proceeding with the details of the local inverse formula, we make two remarks.

Imposing the double sparse constraint in a naive way. A naive approach to (3)+(4) is to apply off-the-shelf optimization software, that only optimizes over the subset of the entries in M that are allowed to be possibly nonzero (so that the first part of the constraint, M subordinate to $\mathscr{G}$, is automatically fulfilled), and then impose some form of penalty based on terms in M^{-1} that are nonzero and that are not allowed to be nonzero according to the graph $\mathscr{G}$. There are other naive ways that one could imagine to impose the doubly sparse constraint in (4). An issue with such naive approaches is that they do not make use of the underlying algebraic structure of such special pairs of matrices.

Separable nonlinear least squares problems. Finding a pair of matrices M and M^{-1} that are both subordinate to the same graph $\mathscr{G}$, as in (4), is related to so-called separable nonlinear least squares problems (Golub and Pereyra 2003). Roughly speaking, that class of problems corresponds to a nonlinear optimization problem where the variables can be partitioned into two subsets, and where knowing the values of the variables for one of the subsets, leads to a linear problem to be solved in order to find the unknown values of the remaining variables in the other subset. For our application at hand, the two subsets are the unknown entries of M and of M^{-1}. If we knew the true nonzero entries of M, then it is a simple linear problem to find the nonzero entries of M^{-1} (and vice-versa). In the context of separable nonlinear least squares problems, one approach is the so called *variable projection method*. Such an algorithm alternates between updating the two subsets of the variables. At first glance, the algorithm only involves linear optimization in each iteration. Unfortunately, such an alternating approach often has disappointing performance, and there are challenges with parameterization.

2.2 Local Inverse Formula

If the inverse matrix, V^{-1}, is subordinate to a chordal graph, then this local inverse formula (5) tells us how to find the inverse *using only sub-blocks* of the matrix V. The formula reads

$$\Theta \equiv V^{-1} = \sum_{[c]\in\mathscr{C}} \left(V_{[c]}\right)^{-1} - \sum_{[j]\in\mathscr{J}} \left(V_{[j]}\right)^{-1}, \tag{5}$$

where $V_{[c]}$ denotes a square sub-block of the matrix V, which corresponds to a *maximal clique* c in the set of maximal cliques $\mathscr{C}$ that are the nodes of the *clique tree* associated with the chordal graph of the matrix V; and for each element in the set of edges $\mathscr{J}$ of the clique tree, $V_{[j]}$ is a square sub-block of the matrix V that corresponds to the *intersection* (or 'separator') j of two maximal cliques that are connected in the clique tree. Proofs of (5) can be found in the references, under mild assumptions (Johnson and Lundquist 1998). There are also other terminologies for the same thing, see, e.g., Bartlett's lecture notes on Undirected graphical models: Chordal graphs, decomposable graphs, junction trees, and factorizations https://people.eecs.berkeley.edu/~bartlett/courses/2009fall-cs281a/, or (Lauritzen 1996; Speed and Kiiveri 1986).

For our special class of matrices, that satisfy the doubly sparse constraint in (4), both are subordinate to the same chordal graph, so we are allowed to swap the roles of V and V^{-1} and the local inverse formula (5) still holds. We can exploit this local property in whichever way is most convenient. This is what distinguishes the class of covariance matrices we study here from the rest of the literature.

Example: A 5×5 Local Inverse Formula. While the formula (5) scales to large matrices in a computer code, here we only give small examples that can be displayed on a page. The numerical conditioning depends on the conditioning of the cliques and the separators. How well the computations scale depends on the clique tree, and especially the size of the maximal cliques. In other words, the speed depends more on the graph, rather than the size of the matrix.

Let us revisit the same example in (1). We will demonstrate the Local Formula (5) holds. In plain English, formula (5) roughly states that *"the inverse is the sum of the inverses of the blocks, minus the inverse of the overlaps,"* as in this example:

$$\Theta = V^{-1} = \left(\left(\begin{pmatrix} 13 & 8 & 4 & 2 \\ 8 & 13 & 2 & 1 \\ 4 & 2 & 10 & 6 \\ 2 & 1 & 6 & 13 \end{pmatrix}^{-1} \right) + \left(\left(\begin{pmatrix} 10 & 6 & 1 & 1 \\ 6 & 13 & 10 & 10 \\ 1 & 10 & 13 & 8 \\ 1 & 10 & 8 & 13 \end{pmatrix}^{-1} \right) \right. \right. \tag{6}$$

$$\left. \left. - \left(\left(\begin{pmatrix} 10 & 6 \\ 6 & 13 \end{pmatrix}^{-1} \right) \right) \right) .$$

We are also seeing examples of the sub-matrices that correspond to the two maximal cliques (corresponding to the two sets of indices $\{1,2,3,4\}$ and $\{3,4,5,6\}$) and to their intersection (corresponding to the set of indices $\{3,4\}$) in the clique tree coming from the chordal graph associated with this matrix example.

Recall that what is novel about the general class of matrices that we consider in this article is that they satisfy that local inverse formula, (5), in *both* directions. That is, for these special examples, we can swap the roles of V and V^{-1}, and the Local Formula (5) remains true! In this example we obtain:

$$V = 21540 \left(\left(\begin{pmatrix} 2960 & -1690 & -900 & 90 \\ -1690 & 2675 & 150 & -15 \\ -900 & 150 & 8715 & -12180 \\ 90 & -15 & -12180 & 23835 \end{pmatrix}^{-1} \right) \right. \tag{7}$$

$$+21540 \left(\left(\begin{pmatrix} 8715 & -12180 & 5385 & 5385 \\ -12180 & 23835 & -10770 & -10770 \\ 5385 & -10770 & 7539 & 3231 \\ 5385 & -10770 & 3231 & 7539 \end{pmatrix}^{-1} \right) \right.$$

$$\left. \left. -21540 \left(\left(\begin{pmatrix} 8715 & -12180 \\ -12180 & 23835 \end{pmatrix}^{-1} \right) \right) \right) .$$

Having simultaneously both (6) and (7) hold is an example of the local algebraic property that we exploit in this article for covariance matrix estimation, and is our main contribution.

Example: Local inversion for Block Matrices. Consider the block matrix

$$M = \begin{pmatrix} M_{11} & M_{12} & * \\ M_{21} & M_{22} & M_{23} \\ * & M_{32} & M_{33} \end{pmatrix} . \tag{8}$$

Suppose we know that the matrix is invertible and that the inverse has the sparsity pattern

$$M^{-1} = \begin{pmatrix} \times & \times & \mathbf{0} \\ \times & \times & \times \\ \mathbf{0} & \times & \times \end{pmatrix}. \tag{9}$$

We are not specifying the entry in the top right, nor in the bottom left, of M, because it can be shown that the sparsity pattern of M^{-1} implies that those two blocks must be

$$M_{13} = M_{12}M_{22}^{-1}M_{23},$$

and

$$M_{31} = M_{32}M_{22}^{-1}M_{21}.$$

There are some mild assumptions, such as that M_{22} is invertible. Then the local inverse formula tells us that:

$$M^{-1} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}^{-1} + \begin{pmatrix} M_{22} & M_{23} \\ M_{32} & M_{33} \end{pmatrix}^{-1} - \begin{pmatrix} M_{22} \end{pmatrix}^{-1}. \tag{10}$$

Direct matrix multiplication of M with this claimed form of M^{-1} to arrive at the (block) identity matrix is one way to prove this. Equation (10) is an example of the Local Inverse Formula (5). Although this example is only 3×3, in fact the Local Inverse Formula (5) can be applied to arbitrarily large matrices (subject to mild assumptions and of course being subordinate to a chordal graph as previously stated), and the proof for larger matrices essentially boils down to this 3×3 non-commutative block matrix algebra, applied in a recursive way (and the proof is thus by induction), and combined with the key property of chordal graphs that they always have a junction tree (Johnson and Lundquist 1998).

2.3 The Local Function

Let a chordal graph $\mathscr{C}$ and its junction tree $\mathscr{J}$ be given. Then we define a function $L : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^{n\times n}$ by the right hand side of (5)

$$L(M) \equiv \sum_{[c]\in\mathscr{C}} \left(M_{[c]}\right)^{-1} - \sum_{[j]\in\mathscr{J}} \left(M_{[j]}\right)^{-1}. \tag{11}$$

We now make the following observations:

1. *L depends on a chordal graph*, but that dependence is not explicit in the notation on the left side of (11), i.e., we could plausibly use notation $L_{\mathscr{G}}$ to indicate the dependence on the chordal graph $\mathscr{G}$.
2. The dependence on the chordal graph is seen on the right side of (11) with the first sum being over the maximal cliques $\mathscr{C}$ of the graph, and the second sum being over the separators $\mathscr{J}$. This pair $(\mathscr{C}, \mathscr{J})$, must correspond to a clique tree (sometimes called a junction tree) of the graph, which has a running intersection property. It is an equivalence characterisation of chordal graphs that a chordal graph always has at least one clique tree with this property.
3. At first glance, it also looks like L depends on the choice of junction tree, but it can be shown that in fact all junction trees would lead to the same final sum on the right of (11).
4. The domain of L is a strict subset of $\mathbb{R}^{n\times n}$. The input does not need to be an example from our special class of matrices in order to apply L. The only requirement on the input matrix is that the submatrices on the right hand side of (11) are indeed invertible.
5. The output matrix $L(M)$ is subordinate to the chordal graph, by definition in (11).
6. For an arbitrary matrix M, we usually expect $L(M) \neq M^{-1}$. However, for matrices for which the inverse is known to be subordinate to a chordal graph, then we have by the results of the local formula that $L(M) = M^{-1}$ is indeed the inverse.

7. Recall the doubly sparse constraint in (4). We typically fulfill one of the conditions in that constraint (4) automatically by parameterizing by the allowed nonzero entries of, say, M. If the matrix M was truly in the special class that satisfied the constraint, then we would have that both $L(M) = M^{-1}$ and $L(M^{-1}) = M$. We can attempt to exploit this when imposing the constraint, and to take advantage of the property that $L(M)$ or $L(M^{-1})$ will usually be much better to compute than an inverse, because of the local formula.

2.4 Optimization Summary and Theoretical Justification

The following two key points make it clear that the constraints in (4) can be enforced via the local formula, in the way that we describe in our algorithm below.

Theorem 1 (Blair and Peyton 1993, Vandenberghe and Andersen 2015.) Let $\mathscr{G}$ be a chordal graph, and let $(\mathscr{C}, \mathscr{J})$ be a corresponding list of cliques and separators, which we call a *perfect elimination ordering*, in terms of a clique tree. Then M and a Cholesky factor R based on this ordering have the same sparsity pattern, in the sense that, if we consider only entries in the triangular part, i.e. $(i < j)$,

$$M_{ij} = 0 \qquad \Leftrightarrow \qquad R_{ij} = 0, \qquad (i < j).$$

One direction of the above result is the 'no fill-in' property for perfect eliminators, which has been known a long time in the numerical linear algebra literature (see, e.g., references by Rose and by Rose, Tarjan, & Lueker in the exposition of Blair and Peyton 1993, or Theorem 9.1 of Vandenberghe and Andersen 2015). The other direction, is discussed, in relation to *monotone transitivity* properties of chordal graphs, in Figure 4.1, and in equation (9.7) of Vandenberghe and Andersen 2015.

Theorem 2 (Johnson and Lundquist 1998) Let $\mathscr{G}$ be a chordal graph, and let L be defined as in (11) for this graph $\mathscr{G}$. Let M and M^{-1} be a given pair of matrices. Assume that $L(M)$, as defined by the right side of equation (11), is defined. The following two statements are equivalent. We have

$$M^{-1} \text{ subordinate to } \mathscr{G}$$

is equivalent to

$$M^{-1} = L(M).$$

One direction of this result is trivial, by definition of our function L. The other direction is a result in Johnson and Lundquist 1998 where it is more general, allowing non-symmetric matrices. More examples, exposition, and references can be found in Strang and MacNamara 2018.

As a consequence of these theoretical results, the optimization (3) and (4), is equivalent to the following.

Let $\boldsymbol{x} \in \mathbb{R}^p$ where p is the number of edges in the given chordal graph $\mathscr{G}$. Let $R(\boldsymbol{x})$ be a triangular matrix with a sparsity pattern corresponding to the chordal graph $\mathscr{G}$, i.e., each entry of $\boldsymbol{x}$ corresponds to a particular entry of R. We will let

$$M(\boldsymbol{x}) = R(\boldsymbol{x})^\top R(\boldsymbol{x}).$$

Let S be the sample covariance matrix from n observations. Then we find the numbers in $\boldsymbol{x}$ that *maximize the objective function*

$$\operatorname{tr}\Big(L(R(\boldsymbol{x}))L(R(\boldsymbol{x}))^\top S\Big) + 2\ln\det(R(\boldsymbol{x})). \tag{12}$$

subject to the constraint that

$$C = 0 \tag{13}$$

where the matrix C is defined to be

$$C \equiv ML(M) - I.$$

The fact that we can meet the first constraint of (4) in the original optimization problem simply by optimizing over entries with the same sparsity pattern in a Cholesky factor, $R(\boldsymbol{x})$, depends on Theorem 1. The fact that we can meet the second constraint of (4) in our original problem by requiring $C = 0$ depends on Theorem 2. (There is some redundancy in requiring $C = 0$, since in our application C is symmetric.)

3 NUMERICAL EXPERIMENTS

We provide two experiments, one with simulated data, and the other with financial data obtained from Yahoo Finance.

3.1 Simulated Data

In this experiment, we start with the 'true' matrix V_{true} that is given in equation (1). We reproduce that example matrix here for convenience:

$$V_{\text{true}} = \begin{pmatrix} 13 & 8 & 4 & 2 & 0 & 0 \\ 8 & 13 & 2 & 1 & 0 & 0 \\ 4 & 2 & 10 & 6 & 1 & 1 \\ 2 & 1 & 6 & 13 & 10 & 10 \\ 0 & 0 & 1 & 10 & 13 & 8 \\ 0 & 0 & 1 & 10 & 8 & 13 \end{pmatrix}.$$

We draw n random samples from the multivariate normal distribution with zero mean and with this covariance matrix V_{true}, and we form the sample covariance matrix S. We then optimize with Matlab's `fmincon.m`. We supply `fmincon.m` a function handle that it can call to compute the objective function (12) at a given $\boldsymbol{x}$ using the local inverse formula. The only constraint is that $C = 0$ as in (13), which is also supplied to `fmincon.m` as a function handle. We now make the following observations.

If n is very large then the optimization returns an estimate that is close to the true covariance matrix. Also, the likelihood at the estimate is very close to the likelihood at the true matrix.

However, if n is not large then the estimate can be very noticeably different to the true matrix, and the likelihood at the estimate is higher than the likelihood at the true matrix. For example, with $n = 100$ samples, in one experiment, the sample covariance matrix is

$$S = \begin{pmatrix} 16.703 & 8.774 & 4.113 & 2.629 & -0.25 & 1.16 \\ 8.774 & 11.559 & 1.92 & 0.01 & -1.605 & -0.854 \\ 4.113 & 1.92 & 10.07 & 5.813 & 1.245 & 0.947 \\ 2.629 & 0.01 & 5.813 & 12.424 & 10.227 & 9.68 \\ -0.25 & -1.605 & 1.245 & 10.227 & 13.958 & 7.88 \\ 1.16 & -0.854 & 0.947 & 9.68 & 7.88 & 13.345 \end{pmatrix},$$

and the estimate from the optimization procedure is

$$\begin{pmatrix} 37.126 & 16.09 & 2.384 & 1.175 & 0 & 0 \\ 16.09 & 26.676 & 1.477 & 0.728 & 0 & 0 \\ 2.384 & 1.477 & 10.933 & 2.507 & -2.974 & -2.811 \\ 1.175 & 0.728 & 2.507 & 6.07 & 4.989 & 4.715 \\ 0 & 0 & -2.974 & 4.989 & 18.091 & 5.179 \\ 0 & 0 & -2.811 & 4.715 & 5.179 & 18.607 \end{pmatrix}.$$

3.2 S&P100 Financial Data

To further showcase the proposed method of estimation, in Figure 1, we considered data based on daily prices of the S&P 100 stocks from 1 January 2019 to 26 March 2021, downloaded from Yahoo Finance. In financial markets, portfolio theory suggests nearly all assets have some correlation with common market factors, so perhaps insisting on zeros in the covariance matrix is only justified if we are looking at 'residual' covariance matrices, after conditioning on market factors. Therefore, we calculated the residuals, after

regressing the log returns of all stocks in the S&P100 on the S&P100 index. Then we naively estimated the sample covariance matrix of those residuals. For the purpose of demonstration, we chose the subset of the S&P100 index that corresponds to 'Consumer Discretionary' (13 stocks) and 'Information Technology' (10 stocks), and the corresponding 23×23 sample covariance matrix. To apply our proposed method of estimation, we also need a sparsity pattern that corresponds to a chordal graph, so we specified the junction tree to have two cliques, corresponding to the 'Consumer Discretionary' stocks, and to the 'Information Technology' stocks together with Amazon and Tesla, and specified the separator to correspond to the two stocks Amazon and Tesla.

The result of the procedure is displayed in Figure 1. The examples we used earlier to illustrate this special class of matrices, in (1), were in exact arithmetic, so everything was pleasingly exact and algebraic. Here in this application to real data, we obtain our estimates by an optimization procedure, and in finite precision. The norm of the constraint, $\|ML(M) - I\|$ evaluated at the final estimate M gives a sense of the accuracy, and in this numerical example the optimization terminated with $\|ML(M) - I\| \approx 10^{-4}$.

Note that in the S&P100 constituents classification, Amazon and Tesla are classified as 'Consumer Discretionary,' but arguably these two stocks have more to do with the information technology sector, so visualizing the covariance matrix that we estimate with this sparsity pattern (as in Figure 1), could be useful in exploring the coupling of these two bigger industry groups, via Amazon and Tesla. In this case the right panel of Figure 1 shows the magnitude of the entries in rows corresponding to Amazon and Tesla, and which indicate the strength of the coupling.

Note also that we could not examine such couplings between groups if we were to only allow simply block diagonal matrices as our estimates, so it is important that the methods are more general, and that is one benefit of allowing the class of chordal graphs. All the examples considered here are suggestive of the potential applications of the method we propose.

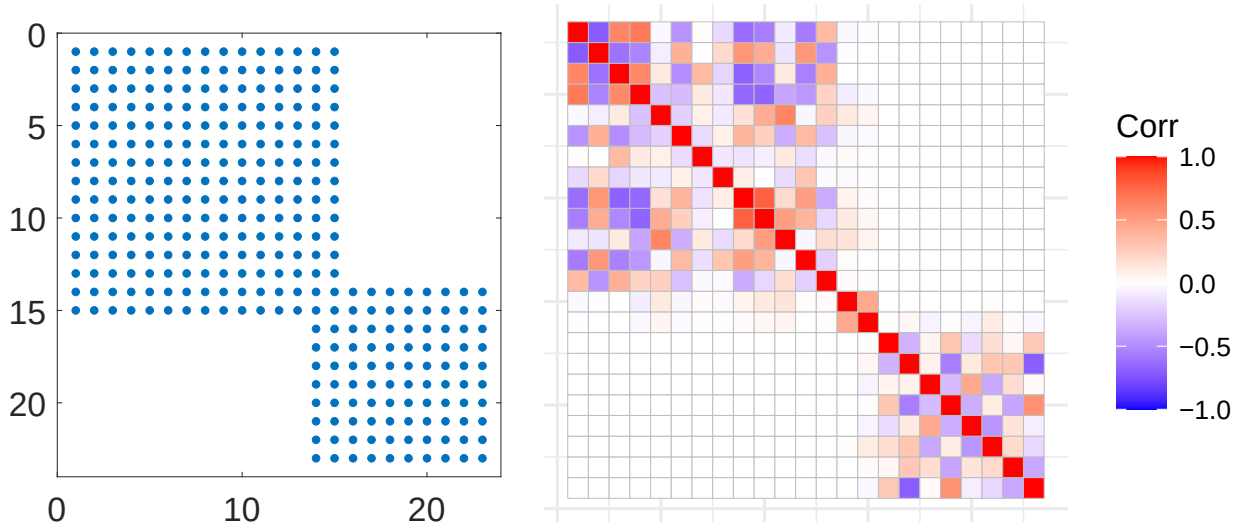

Figure 1: Left: sparsity pattern of the covariance matrix that is estimated by the method proposed in this article. Note this is NOT simply block diagonal. Right: Visualisation of the corresponding correlation matrix. Rows 14 and 15 correspond to Amazon and Tesla, while rows 1-13 correspond to 'Consumer Discretionary' stocks, and the remaining rows correspond to 'Information Technology' stocks, from the S&P100 data described in the main text.

3.3 Discussion

We now highlight a number of issues:

- We have chosen a maximum likelihood framework here merely because it is the simplest way to illustrate our ideas that estimation of doubly sparse matrices is indeed possible. However, note that the algebraic structures and methods for local computations that we describe do not depend on that maximum likelihood framework – so, for example, it should be possible to also use these computational approaches in other frameworks that do not make assumptions about the distribution.
- The choice of norm is likely important (for example, the $1-$norm may be preferable to the $2-$norm), but that issue is not explored here.
- Instead of prescribing the sparsity pattern, it would be better to estimate the sparsity graph from the data. (For instance, in the finance example, we would prefer not to require any hunch about the significance of Amazon or Tesla.) A crude first approach could be in three separate steps: first use existing methods such as the graphical LASSO to estimate a graph, and then second force the resulting graph to be chordal, and then third and finally apply the suggested method of this paper. But instead of such a crude approach of consecutive but separate estimation problems, it seems more natural to simultaneously estimate the graph together with the matrix estimation problem.
- We have required the matrix and its inverse be subordinate to the same chordal graph, because that it is the simplest first idea to explore. Note that – although we have not displayed such an example pair of matrices in this article – if the (i, j) edge is present in the graph, then it is possible that the (i, j) entry of M is nonzero and the (i, j) entry of M^{-1} is zero, and that both matrices are subordinate to the same chordal graph. It is natural to also consider the generalisation of our problem to the case where the matrix and the inverse are subordinate to different chordal graphs, but we did not explore that generalisation here.
- If the matrix M was truly in the special doubly sparse class, then the matrix would be a fixed point of the special Local Function we defined in (11), composed with itself, i.e. $L(L(M)) = M$. We have not explored fixed-point iteration algorithms. Nor have we explored the algebraic structure of this special class of matrices from this point of view of the properties of the local function $L(\cdot)$, although The Nullity Theorem gives constraints on ranks of subblocks (Strang and Nguyen 2004).
- Chordal graphs can be considerably more complicated than only the simplest examples we have illustrated here – see many varied examples of matrix sparsity patterns and corresponding graphs in the references, e.g., Vandenberghe and Andersen 2015.
- Any given graph can be 'forced' to become a chordal graph by a known process of adding edges, thus allowing the methods of this paper to always be used. However, whether or not such a process is computationally worthwhile depends on the example.

4 CONCLUSION

We proposed a method to estimate a covariance matrix when both the covariance and the precision matrix are sparse (which we called double sparsity). This is a maximum likelihood approach, subject to the double sparsity constraint. This appears to be the first work to estimate such special pairs of covariance and precision matrices. The sparsity patterns we consider are restricted to the class of chordal graphs (also known as decomposable graphs in the graphical models literature). This class includes the banded matrices with banded inverse. Restricting to this class of sparsity pattern allows us to exploit a special algebraic structure – the local inverse formula, as we described – that can make computations faster (and that is computationally more attractive than simply naively imposing the double sparsity constraint during any optimization). For future work, it should be possible to extend these approaches to simultaneously estimate both the sparsity pattern, and the corresponding special pair of covariance matrix and precision matrix.

REFERENCES

Bickel, P. J., and E. Levina. 2008. "Regularized estimation of large covariance matrices". *The Annals of Statistics* 36:199–227.

Blair, J., and B. Peyton. 1993. "Introduction to Chordal Graphs and Clique Trees". In *Graph Theory and Sparse Matrix Computation*. Springer.

Dempster, A. 1972. "Covariance selection". *Biometrics* 28(1):157–175.

Dym, H., and I. Gohberg. 1981. "Extensions of band matrices with band inverses". *Linear algebra and its applications* 36:1–24.

Friedman, J., T. Hastie, and R. Tibshirani. 2008. "Sparse inverse covariance estimation with the graphical lasso". *Biostatistics* 9(3):432–41.

Georgescu, D. I., N. J. Higham, and G. W. Peters. 2018. "Explicit solutions to correlation matrix completion problems, with an application to risk management and insurance". *Royal Society open science* 5(3):172348.

Golub, G., and V. Pereyra. 2003. "Separable nonlinear least squares: the variable projection method and its applications". *Inverse problems* 19(2):R1.

Grone, R., C. R. Johnson, E. M. Sá, and H. Wolkowicz. 1984. "Positive definite completions of partial Hermitian matrices". *Linear algebra and its applications* 58:109–124.

Johnson, C., and M. Lundquist. 1998. "Local inversion of matrices with sparse inverses". *Linear Algebra and Its Applications* 277:33–39.

Lauritzen, S. 1996. *Graphical Models*. Oxford University Press.

Ledoit, O., and M. Wolf. 2004. "Honey, I shrunk the sample covariance matrix". *The Journal of Portfolio Management* 30(4):110–119.

Pourahmadi, M. 2013. *High-dimensional covariance estimation*. Wiley.

Ravikumar, P., M. J. Wainwright, G. Raskutti, and B. Yu. 2011. "High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence". *Electronic Journal of Statistics* 5:935–980.

Rothman, A. J., E. Levina, and J. Zhu. 2010. "A new approach to Cholesky-based covariance regularization in high dimensions". *Biometrika* 97:539–550.

Salemi, P., E. Song, B. Nelson, and J. Staum. 2019. "Gaussian Markov random fields for discrete optimization via simulation: Framework and algorithms". *Operations Research* 67(1):250–266.

Speed, T., and H. Kiiveri. 1986. "Gaussian Markov distributions over finite graphs". *The Annals of Statistics* 14:138–150.

Strang, G. 2010. "Fast transforms: Banded matrices with banded inverses". *Proceedings of the National Academy of Sciences* 107(28):12413–12416.

Strang, G., and S. MacNamara. 2018. "A local inverse formula and a factorization". In *Contemporary Computational Mathematics - A Celebration of the 80th Birthday of Ian Sloan*, edited by W. H. Dick J., Kuo F., 1109–1126. Springer.

Strang, G., and T. Nguyen. 2004. "The interplay of ranks of submatrices". *SIAM review* 46(4):637–646.

Vandebril, R., M. van Barel, and N. Mastronardi. 2007. *Matrix Computations and Semiseparable Matrices*. Johns Hopkins.

Vandenberghe, L., and M. S. Andersen. 2015. "Chordal graphs and semidefinite optimization". *Foundations and Trends in Optimization* 1(4):241–433.

AUTHOR BIOGRAPHIES

Shev MacNamara is a Senior Lecturer in the School of Mathematical and Physical Sciences at the University of Technology Sydney. His research interests include advanced stochastic modeling and simulation. Previously, he was a postdoctoral scholar in the Department of Mathematics at MIT, and a postdoctoral scholar in the Mathematical Institute at The University of Oxford. He has been a Fulbright Scholar, and has held a David G. Crighton Fellowship at The University of Cambridge. His webpage is https://www.uts.edu.au/staff/shev.macnamara, and his email is shev.macnamara@uts.edu.au.

Erik Schlögl is Professor and Director of the Quantitative Finance Research Centre, University of Technology Sydney, Broadway, NSW 2007, Australia. He also holds an honorary Professorship at the African Institute for Financial Markets and Risk Management (AIFMRM), University of Cape Town, Rondebosch 7701, South Africa; and an honorary affiliation with the Department of Statistics, Faculty of Science, University of Johannesburg, Auckland Park 2006, South Africa. His email address is Erik.Schlogl@uts.edu.au. His website is https://profiles.uts.edu.au/Erik.Schlogl.

Zdravko Botev is a Lecturer of Statistics at UNSW Sydney. His research interest include: 1) Monte Carlo variance reduction methods, especially for rare-event probability estimation; 2) nonparametric kernel density estimation, and more recently 3) fast model-selection algorithms for large-scale statistical learning. He is well-known as the inventor of the widely-used method of *kernel density estimation via diffusion*, as well as the *generalized splitting algorithm* for Monte Carlo rare-event simulation and combinatorial counting. His email address is botev@unsw.edu.au. His website is https://web.maths.unsw.edu.au/~zdravkobotev/.