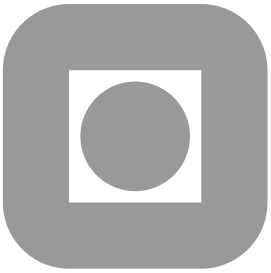
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Specifying a Gaussian Markov random field by a sparse Cholesky triangle

by

Hanne T. Wist & Håvard Rue

PREPRINT
STATISTICS NO. 2/2005



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This preprint has URL http://www.math.ntnu.no/preprint/statistics/2005/S2-2005.pdf
Hanne T. Wist has homepage: http://www.math.ntnu.no/∼hannewis
E-mail: hannewis@math.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

Specifying a Gaussian Markov random field by a sparse Cholesky triangle

Hanne T. Wist & Håvard Rue
Department of Mathematical Sciences
Norwegian University of Science and Technology
Norway

April 5, 2005

Abstract

This note discusses the approach of specifying a Gaussian Markov random field (GMRF) by the Cholesky triangle of the precision matrix. A such representation can be made extremely sparse using numerical techniques for incomplete sparse Cholesky factorisation, and provide very computational efficient representation for simulating from the GMRF. However, we provide theoretical and empirical justification showing that the sparse Cholesky triangle representation is fragile when conditioning a GMRF on a subset of the variables or observed data, meaning that the computational cost increases.

1 Introduction

Gaussian Markov random fields (GMRFs) are frequently used in statistics due to the nice analytical properties of the Gaussian distribution combined with Markov properties. GMRFs have a wide area of applications, including structural time-series analysis, analysis of longitudinal and survival data, graphical models, semiparametric regression and splines, image analysis, spatial statistics and geostatistics. For references and examples, see Rue and Held (2005, Ch. 1). GMRFs are also named as conditional autoregressive models (CARs) due to the seminal work of Besag (1974, 1975) in spatial statistics. Intrinsic versions of GMRFs are also extensively used, see for example Besag and Higdon (1999), Fahrmeir and Lang (2001), Knorr-Held and Rue (2002) and Banerjee et al. (2004).

One major feature of GMRFs, is the computational superiority due to the close link between GMRFs and numerical methods for sparse matrices (Rue, 2001). These algorithms allow for fast algorithms for simulation and evaluation of the log-density and includes those based on the Kalman-filter for Gaussian dynamic models as a special case (Knorr-Held and Rue, 2002, Appendix). These fast algorithms also includes the possibilities for treating various forms for conditioning, like conditioning on subset of variables and linear constraints. See Rue and Held (2005, Ch. 2) for a complete discussion. These fast algorithms can also be extended to construct non-Gaussian approximations to (non-Gaussian) hidden GMRFs which can be sampled exactly with computable normalising constants (Rue et al., 2004), and to compute marginal variances for GMRFs (Rue, 2005).

A GMRF is nearly always specified through its precision matrix or (equivalently) through specifying the set of the full conditionals in the spirit of Besag (1974, 1975). Although a such approach is natural, we investigate in this report the option of specifying a GMRF, implicitly, by the Cholesky triangle of the precision matrix. The main idea is to compute the *incomplete* Cholesky factorisation of the precision matrix for a GMRF, and use this factorisation to specify an extremely sparse representation

of a GMRF, which properties are close to the original one. The extremely sparse representation implies that computing with this model is extremely efficient and computational costs is nearly linear in the dimension. We demonstrate in this report that it is possible to construct such extremely sparse representations, but unfortunately, this construction is fragile when doing conditioning. We provide theoretical results which explain why a such construction is fragile in this sense. The fraglie nature of these representations limits their usefulness in pratical applications.

The outline of this paper is as follows. In section 2 we give some background of GMRFs and interpretation of the Cholesky triangle. Then we describe how to construct extremely sparse precision matrices by using incomplete Cholesky factorisation in section 3. In section 4 we examine how sparse the Cholesky triangle can be and discuss the possibility of using this model in various application. We conclude with a discussion in section 5.

2 Background

2.1 Definition of a GMRF

A Gaussian Markov random field (GMRF) is a Gaussian distributed random vector $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ with additional Markov properties. The Markov properties are commonly represented by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices $\{1, \ldots, n\}$ and $\mathcal{E}$ a set of edges, where there is no edge between node i and node j if and only if $x_i \perp x_j \mid \boldsymbol{x}_{-ij}$. The Markov properties appear in the precision matrix $\boldsymbol{Q}$ (the inverse covariance matrix) of the GMRF, as

$$Q_{ij} = 0 \quad \Longleftrightarrow \quad x_i \perp x_j | \boldsymbol{x}_{-ij}, \quad i \neq j. \tag{1}$$

We then say that $\boldsymbol{x}$ is a GMRF with respect to $\mathcal{G}$.

Since $\boldsymbol{Q}$ is a symmetric positive definite matrix, then there exists a unique Cholesky triangle $\boldsymbol{L}$ such that $\boldsymbol{L}$ is a lower triangular matrix and $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$. To sample $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$, we first sample $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Then the solution of $\boldsymbol{L}^T\boldsymbol{v} = \boldsymbol{z}$ has the correct covariance matrix

$$\mathrm{Cov}(\boldsymbol{v}) = \mathrm{Cov}(\boldsymbol{L}^{-T}\boldsymbol{z}) = (\boldsymbol{L}\boldsymbol{L}^T)^{-1} = \boldsymbol{Q}^{-1}. \tag{2}$$

Finally, $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{v}$. If the mean is only known through the canonical representation $\boldsymbol{\mu} = \boldsymbol{Q}^{-1}\boldsymbol{b}$, then we calculate $\boldsymbol{\mu}$ by solving $\boldsymbol{L}\boldsymbol{w} = \boldsymbol{b}$ and $\boldsymbol{L}^T\boldsymbol{\mu} = \boldsymbol{w}$.

2.2 The sparsity pattern in the Cholesky triangle

The sparsity pattern of $\boldsymbol{L}$ can be decided before computing the Cholesky factorisation. It is no need to compute elements known to be zero, hence the factorisation can be computed rapidly for sparse matrices. The following result (see Rue and Held (2005, Ch. 2)) can be used to decide the sparsity pattern of $\boldsymbol{L}$. Define for $1 \leq i < j \leq n$ the set

$$F(i, j) = \{i + 1, \ldots, j - 1, j + 1, \ldots, n\}, \tag{3}$$

which is the future of i except j. Then

$$x_i \perp x_j \mid \boldsymbol{x}_{F(i,j)} \iff L_{ji} = 0. \tag{4}$$

Using the global Markov property, it follows that

$$F(i, j) \text{ separates } i \text{ and } j \implies x_i \perp x_j \mid \boldsymbol{x}_{F(i,j)}. \tag{5}$$

Thus, if $F(i,j)$ is a separating subset for node i and j, then x_i and x_j are conditionally independent in the marginal distribution, and $L_{ji} = 0$. One should note that if $L_{ji} = 0$ from the statement above, then it is zero for all $\boldsymbol{Q} > 0$ with the same graph $\mathcal{G}$.

For two neighbours $i \sim j$, $F(i,j)$ is not a separating subset, since $Q(i,j) \neq 0$. Hence, we know that $\boldsymbol{L}$ is always more or equally dense than the lower triangular part of $\boldsymbol{Q}$, and for many graphs the number of possible non-zero elements in $\boldsymbol{L}$ ($n_{\boldsymbol{L}}$) is much larger than the number of non-zero elements in the lower triangular part of $\boldsymbol{Q}$ ($n_{\boldsymbol{Q}}$). The difference $n_{\boldsymbol{L}} - n_{\boldsymbol{Q}}$ is called number of fill-ins. Since the positions of the fill-ins can be determined from the graph $\mathcal{G}$, we compute and store only the terms in $\boldsymbol{L}$ that we know is possibly non-zero. This makes the algorithms for sparse matrices computationally efficient.

2.3 Example

We will illustrate the relationship between the Cholesky triangle and the graph by considering a GMRF $\boldsymbol{x}$ with respect to the graph $\mathcal{G}$ in Fig. 1. The precision matrix is given by

$$\boldsymbol{Q} = \begin{pmatrix} \times & \times & \times & \\ \times & \times & & \times \\ \times & & \times & \times \\ & \times & \times & \times \end{pmatrix},$$

where $\times$ denotes possible non-zero terms. The only nodes that are not neighbours are nodes 1 and 4 and nodes 2 and 3. Since $F(1,4) = \{2,3\}$ is a separating subset for nodes 1 and 4, then $L_{41} = 0$. However, $F(2,3) = \{4\}$ is not a separating subset for nodes 2 and 3, hence L_{32} is a possibly non-zero element. The corresponding Cholesky triangle is then given by

$$\boldsymbol{L} = \begin{pmatrix} \times & & & \\ \times & \times & & \\ \times & \times & \times & \\ & \times & \times & \times \end{pmatrix}.$$

The relationship between $\boldsymbol{L}$ and the conditional dependency in $\boldsymbol{x}$ is easily seen if we express the joint pdf in terms of $\boldsymbol{L}$ as

$$\pi(x_1, x_2, x_3, x_4) \propto \exp\left\{ -\frac{1}{2} \sum_{k=1}^{4} L_{kk}^2 \left(x_k + \frac{1}{L_{kk}} \sum_{j=k+1}^{4} L_{jk} x_j \right)^2 \right\}. \tag{6}$$

The density function of x_1 and x_4 given x_2 and x_3 is given by

$$\pi(x_1, x_4 | x_2, x_3) \propto \ \pi(x_1, x_2, x_3, x_4) \propto \exp\left\{ f_1(x_1) + f_4(x_4) - L_{11} L_{41} x_1 x_4 \right\}, \tag{7}$$

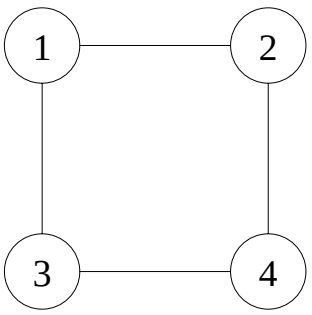


Figure 1: The graph $\mathcal{G}$.

where $f_1(x_1)$ and $f_4(x_4)$ are functions of x_1 and x_4, respectively, and do not contain cross-terms x_1x_4. Since $Q(4,1) = 0$, i.e., $x_1 \perp x_4 \mid x_2, x_3$, then from (7) we must have $L_{41} = 0$, as concluded above.

The density function of x_2 and x_3 given x_1 and x_4 is given by

$$\pi(x_2, x_3|x_1, x_4) \propto \pi(x_1, x_2, x_3, x_4) \propto \exp\{f_2(x_2) + f_3(x_3) - (L_{22}L_{32} + L_{21}L_{31})x_2x_3\}, \quad (8)$$

where $f_2(x_2)$ and $f_3(x_3)$ are functions of x_2 and x_3, respectively, and do not contain cross-terms x_2x_3. Since $Q(3,2) = 0$, then from (8) we must have $L_{22}L_{32} + L_{21}L_{31} = 0$. This means that L_{32} is zero if and only if at least one of L_{21} and L_{31} is zero.

The density function of x_2 and x_3 given x_4 is given by

$$\pi(x_2, x_3|x_4) \propto \pi(x_2, x_3, x_4) \propto \exp\{f_2(x_2) + f_3(x_3) - L_{22}L_{32}x_2x_3\}. \quad (9)$$

If $L_{32} \neq 0$, then x_2 and x_3 are not conditionally independent in the marginal distribution.

We will continue with this example later on.

3 GMRFs specified by an extremely sparse Cholesky triangle

3.1 Construction of extremely sparse Cholesky triangle

The sparsity pattern of the Cholesky triangle can be determined from the graph $\mathcal{G}$. We do not calculate the elements in $\boldsymbol{L}$ that we know is zero. Thus, if we could decrease the number of fill-ins we could save computations. Here we will present an idea using incomplete, instead of complete, Cholesky factorisation, which results in an extremely sparse Cholesky triangle.

A number of iterative solution algorithms are used to solve large sets of sparse linear equations. The incomplete Cholesky factorisation, or LU factorisation for rectangular matrices, has been shown to be a good preconditioner, significantly reducing the number of iterations needed for convergence. An incomplete Cholesky triangle of $\boldsymbol{Q}$ is a lower triangular matrix $\widetilde{\boldsymbol{L}}$ such that

$$\boldsymbol{Q} = \widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^T + \boldsymbol{R} \quad (10)$$

where the number of possible non-zero elements in $\widetilde{\boldsymbol{L}}$ is smaller than $n_{\boldsymbol{L}}$.

There are two main strategies to obtain the incomplete Cholesky triangle: fixed-fill strategies and drop-tolerance strategies. It is also possible to use a combination of these two strategies.

Fixed-fill strategies fix the non-zero pattern of the incomplete Cholesky triangle prior the factorisation. One of the first methods was the ICCG (Incomplete Cholesky and Conjugate Gradient) method proposed by Meijerink and van der Vorst (1977). They considered two choices of non-zero patterns: let the sparsity pattern of $\widetilde{\boldsymbol{L}}$ be equal to the sparsity pattern of the lower triangular of $\boldsymbol{Q}$, or allow fill-ins. A modification of this method was given by Gustavson (1978). This has later been denoted the ILU(p) factorisation, where p is the level of fill. ILU(0) is to use the non-zero pattern of $\boldsymbol{Q}$. For $p = 1$ the sparsity pattern of $\widetilde{\boldsymbol{L}}_1$ is equal to the sparsity pattern of the lower triangular of $\widetilde{\boldsymbol{L}}_0\widetilde{\boldsymbol{L}}_0{}^T$ obtained from ILU(0). See also Saad (1996) for a discussion of level of fill. A disadvantage of the fixed-fill strategies is that the dropped elements in $\widetilde{\boldsymbol{L}}$ do not depend on the value of the elements in $\boldsymbol{Q}$, only the structure of $\boldsymbol{Q}$.

Drop-tolerance strategies includes only elements in $\widetilde{\boldsymbol{L}}$ if they are larger than a given threshold parameter. A possible choice of the threshold parameter is given by Munksgaard (1980): if the size of the element relative to the diagonal elements of its row and column is smaller than a given drop tolerance, it is not included in $\widetilde{\boldsymbol{L}}$. A disadvantage of the drop tolerance strategies is that the non-zero pattern of $\widetilde{\boldsymbol{L}}$ is determined dynamically. Thus, all the elements in $\widetilde{\boldsymbol{L}}$ must be calculated to see if they can be neglected.

3.2 Defining GMRF models using extremely sparse Cholesky triangles

Incomplete sparse Cholesky triangles have mostly been used as preconditioners as mentioned above. Here the idea is to use incomplete Cholesky factorisation to construct a sparse $\widetilde{\boldsymbol{L}}$-matrix for $\boldsymbol{x}$. The new model is then defined by the precision matrix

$$\widetilde{\boldsymbol{Q}} = \widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^T \tag{11}$$

One should note that this will only be an approximate model. However, if we can construct very fast models that reduce the CPU time required for sampling from the model, then it can be preferable to have an approximate model.

Using a different precision matrix means that the graph of the original model have changed. We will come back to the difference between the original and the new precision matrix and see what influence this has on the graph of the model. But first we will demonstrate how we can construct a very sparse $\widetilde{\boldsymbol{L}}$ representation of $\boldsymbol{x}$ in an approximate model of a stationary Gaussian field.

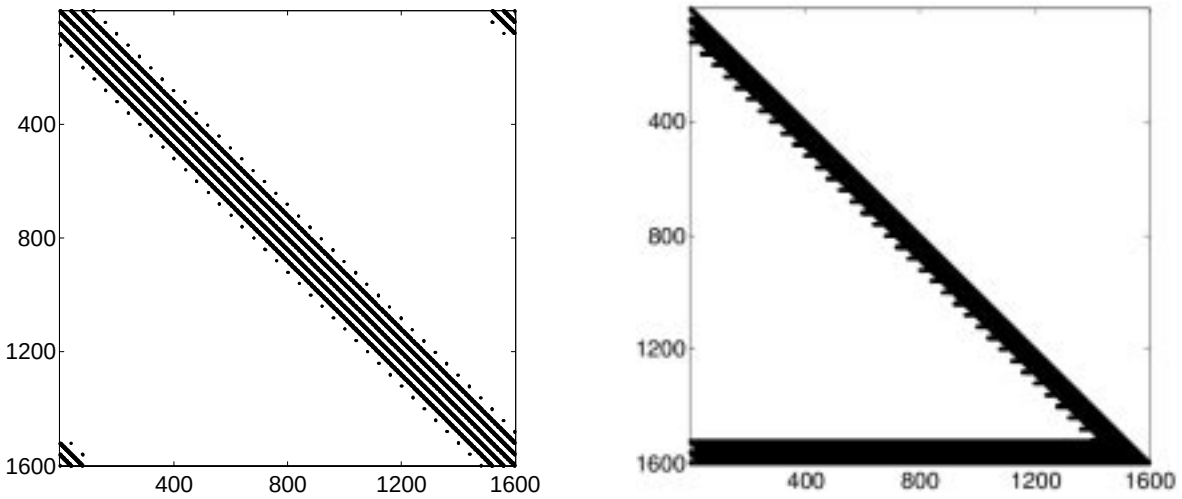
Stationary Gaussian fields are commonly used in spatial modelling, e.g., in geostatistical applications. However, a Gaussian field with long correlation length will be computationally demanding. Rue and Tjelmeland (2002) approximates stationary Gaussian fields by GMRFs with small neighbourhoods. This results in a quite sparse $\boldsymbol{L}$ matrix, and we will show that we can make this even sparser.

Consider a Gaussian field on a 40×40 torus. We use an exponential correlation function with range = 16, and precision = 1.0. Using the approximative technique as described in Rue and Tjelmeland (2002) and Rue and Held (2005) a GMRF with a 5×5 neighbourhood is fitted to the Gaussian field. Figure 2 (a) shows the precision matrix $\boldsymbol{Q}$. Note that the precision matrix is cyclic due to the boundary conditions on a torus. The corresponding Cholesky triangle $\boldsymbol{L}$ is shown in Fig 2 (b). We see that the band structure is preserved in the factorisation, but due to the cyclic precision matrix $\boldsymbol{L}$ has non-zero elements in the last rows. The properties of band matrices and matrices where only a part of the matrix has a band structure will be discussed later.

In general, we would use a permutation matrix to reorder the elements in $\boldsymbol{Q}$ to obtain a low number of fill-ins. We will not do this here, since the difference between complete and incomplete Cholesky factorisation in this case will be most visible with the original $\boldsymbol{Q}$ matrix.

Using incomplete Cholesky factorisation we can reduce the number of non-zero elements in $\boldsymbol{L}$. The incomplete Cholesky triangle is calculated column-wise, and for each column j, the elements that are smaller in magnitude than the local drop tolerance ($droptol \times (\sum_{i=1}^{n} |Q_{ij}|^2)^{1/2}$) are removed from $\widetilde{\boldsymbol{L}}$. Then the rows are scaled by the square root of the diagonal entries in that column. Setting $droptol = 0$ produces the complete Cholesky factorisation. Figures 3 (a) and (b) show the incomplete Cholesky triangle for $droptol = 0.0005$ and 0.005, respectively. It is clear that $n_{\tilde{L}}$ decreases with increasing value of $droptol$.

The original covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$ is shown in Fig. 4 (a). The resulting covariance matrices $\widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{Q}}^{-1}$ using $droptol = 0.0005$ and 0.005 are shown in Figs. 4 (b) and (c), respectively. It is possible to reduce the number of non-zero elements in $\boldsymbol{L}$ to less than one third and still have almost the same covariance matrix when using $droptol = 0.0005$. However, if we use too high value of

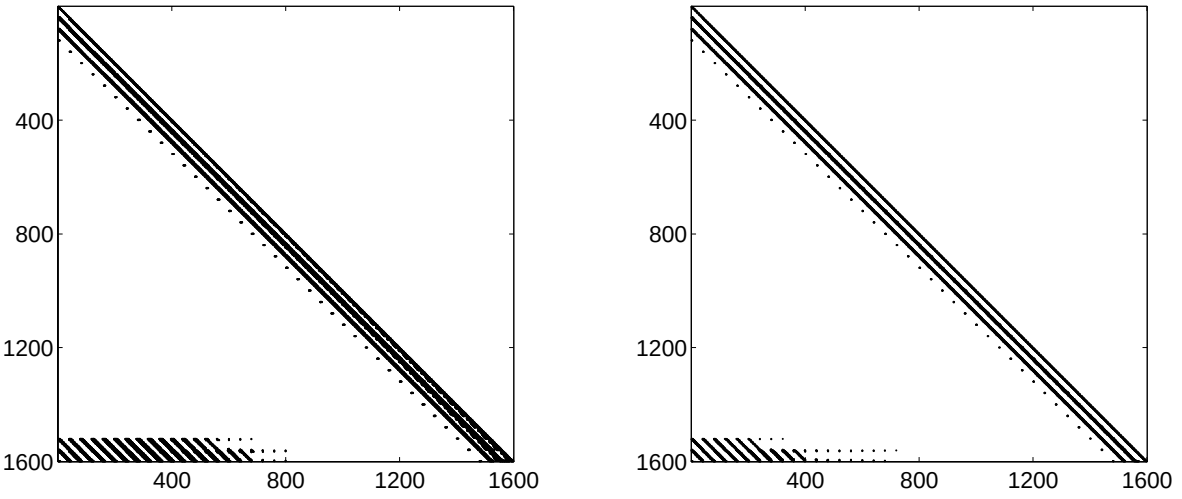


(a) (b)

Figure 2: (a) $\boldsymbol{Q}$ with $n_{\boldsymbol{Q}} = 20\,800$. (b) $\boldsymbol{L}$ with $n_{\boldsymbol{L}} = 247\,600$.

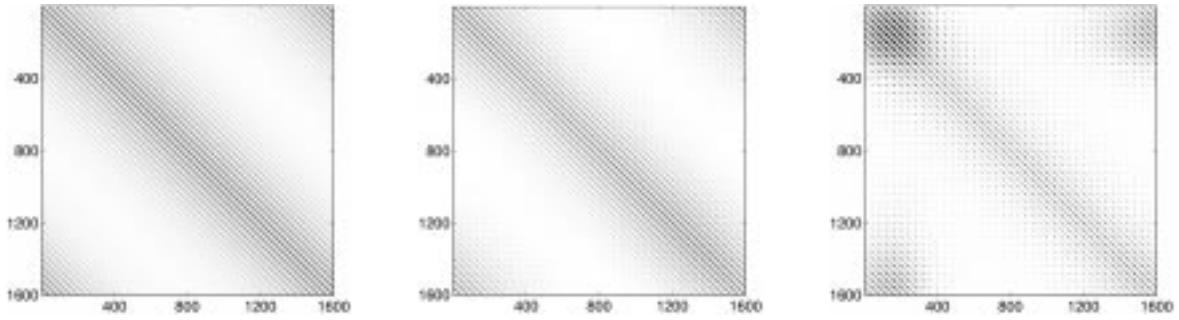
droptol, e.g. 0.005, then the resulting covariance matrix will be quite different from the original. One should also note that the covariance matrices in (b) and (c) are not stationary, due to some of the elements in the Cholesky triangle have been removed. The non-stationarity is most significant in (c).

The spatial covariance function for the node in the center of the graph is shown in Fig. 5 for the different cases. The incomplete Cholesky factorisation with $droptol = 0.0005$ produces a very similar covariance function compared with the original model. If we use $droptol = 0.005$, then the peak is much lower. Thus, this is too large value of *droptol*.



(a) (b)

Figure 3: (a) $\widetilde{\boldsymbol{L}}$ with $droptol = 0.0005$ and $n_{\widetilde{L}} = 69\,811$. (b) $\widetilde{\boldsymbol{L}}$ with $droptol = 0.005$ and $n_{\widetilde{L}} = 33\,839$.



(a) (b) (c)

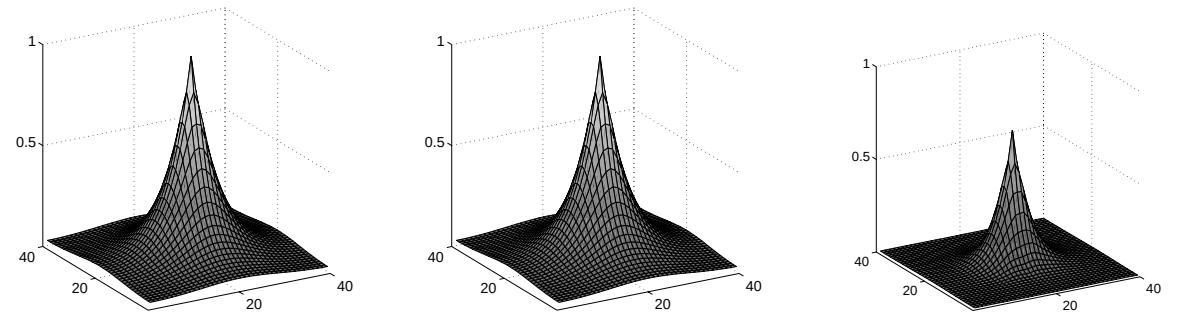
Figure 4: Covariance matrices. (a) $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$. (b) $\widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{Q}}^{-1}$, $droptol = 0.0005$. (c) $\widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{Q}}^{-1}$, $droptol = 0.005$. $\widetilde{\boldsymbol{\Sigma}}$ is not stationary.

Figure 6 (a) shows a sample from the GMRF model using complete Cholesky factorisation. A sample using incomplete Cholesky factorisation with $droptol = 0.0005$ is shown in Fig. 6 (b), where we have used the same random vector $\boldsymbol{z}$ as in (a).

4 Effect on precision matrix when using incomplete Cholesky factorisation

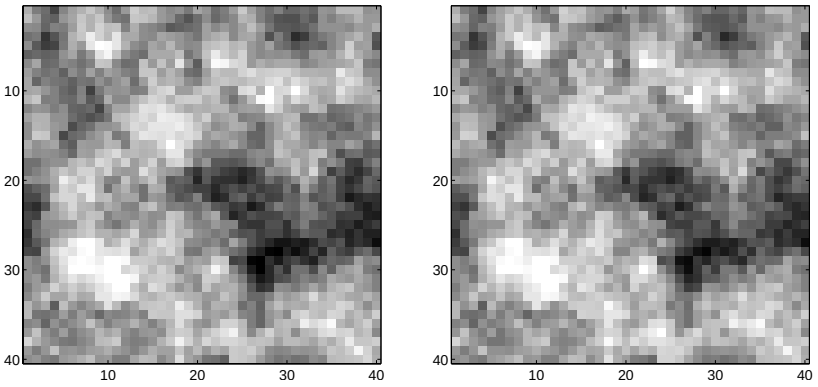
4.1 Difference between the original and new precision matrix

Replacing the Cholesky triangle $\boldsymbol{L}$ with a sparser matrix $\widetilde{\boldsymbol{L}}$ will reduce the computational cost. In the example we could see that it was possible to reduce the number of non-zero elements to less than one third without getting any significant change in the covariance matrix. However, even though the covariance matrix is still almost the same, the new precision matrix $\widetilde{\boldsymbol{Q}}$ will be different from the original precision matrix $\boldsymbol{Q}$. Thus, the graph in the new model will also be different from the original graph.



(a) (b) (c)

Figure 5: Spatial covariance function for node 820 (in the centre of the graph). (a) Original values, (b) $droptol = 0.0005$, (c) $droptol = 0.005$.



(a) (b)

Figure 6: Samples from the GMRF model with zero mean and precision = 1.0 using (a) complete Cholesky factorisation and (b) incomplete Cholesky factorisation with *droptol* = 0.0005. The random vector $\boldsymbol{z}$ is the same.

In order to see the difference between the new model with the extremely sparse precision matrix and the original model we have to examine the difference between $\boldsymbol{Q}$ and $\widetilde{\boldsymbol{Q}}$. Then we will discuss if the new model can be useful in GMRFs. Consider replacing only one non-zero element in $\boldsymbol{L}$ with 0.

Theorem 1 *Let $\boldsymbol{Q} > 0$ and $\boldsymbol{L}$ its lower triangular Cholesky triangle so that $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$. Fix $i^* > j^*$ such that $L_{i^*j^*} \neq 0$ and let*

$$\widetilde{L}_{ij} = \begin{cases} L_{ij} & \text{if} \quad ij \neq i^*j^* \\ 0 & \text{if} \quad ij = i^*j^* \end{cases},$$

where $\widetilde{\boldsymbol{Q}} = \widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^T > 0$. Then

$$Q_{ij} \neq \widetilde{Q}_{ij} \quad \Longleftrightarrow \quad i = i^*,\ j \geq j^* \ \text{ and } \ L_{jj^*} \neq 0. \tag{12}$$

Proof. From $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$ we obtain for $i \geq j$ that $Q_{ij} = \sum_{k=1}^{j} L_{ik}L_{jk}$ and similarly $\widetilde{Q}_{ij} = \sum_{k=1}^{j} \widetilde{L}_{ik}\widetilde{L}_{jk}$. Since $\boldsymbol{L}$ differ from $\widetilde{\boldsymbol{L}}$ only in the i^*j^*th term, only those $\widetilde{Q}_{ij}$'s which include $\widetilde{L}_{i^*j^*}$ in the sum will differ from Q_{ij}, i.e. for $j \geq j^*$,

$$\begin{aligned} \widetilde{Q}_{i^*j} - Q_{i^*j} &= \widetilde{L}_{i^*j^*}L_{j^*j} - L_{i^*j^*}L_{jj^*} \\ &= -L_{i^*j^*}L_{jj^*}. \end{aligned} \tag{13}$$

Hence, $\widetilde{Q}_{i^*j} \neq Q_{i^*j}$ iff $L_{jj^*} \neq 0$ since $L_{i^*j^*} \neq 0$ by assumption. ■

The difference between $\boldsymbol{Q}$ and $\widetilde{\boldsymbol{Q}}$ in the general case can then be found by recursion.

Choosing $j = j^*$ in (13) the following Corollary follows, since $L_{j^*j^*} > 0$.

Corollary 1 *If $Q_{i^*j^*} = 0$ then $\widetilde{Q}_{i^*j^*} \neq 0$.*

So, incomplete Cholesky factorisation of a sparse matrix $\boldsymbol{Q}$ will generally lead to a less sparse matrix $\widetilde{\boldsymbol{Q}}$. Including more elements in $\widetilde{\boldsymbol{Q}}$ corresponds to including more edges in $\mathcal{G}$, i.e., including conditional dependency.

One should note that if originally $L_{ij} \neq 0$, it means that the nodes j and i are either neighbours, or $F(j,i)$ is not a separating subset for j and i. Including more edges in $\mathcal{G}$ will not change this fact. Hence, we can not use (4) and (5) to check that $\widetilde{L}_{ij} = 0$. This means that for other $\widetilde{\boldsymbol{Q}} > 0$ with the same graph, $\widetilde{L}_{ij}$ is not necessarily zero. We will come back to this problem later.

To illustrate Theorem 1 consider the given $\boldsymbol{Q}$ matrix and its lower Cholesky triangle $\boldsymbol{L}$ in Figs. 7 (a) and (b), respectively. The non-zero elements are black, while the zero elements are white. Select a non-zero element in $\boldsymbol{L}$, e.g., element (8,3) and let

$$\widetilde{L}_{ij} = \begin{cases} L_{ij} & \text{if} \quad i,j \neq 8,3 \\ 0 & \text{if} \quad i,j = 8,3 \end{cases}.$$

$\widetilde{\boldsymbol{Q}}$ and $\widetilde{\boldsymbol{L}}$ are shown in Figs. 7 (c) and (d). Using Theorem 1 we can predict which elements $\widetilde{Q}_{ij}$ that

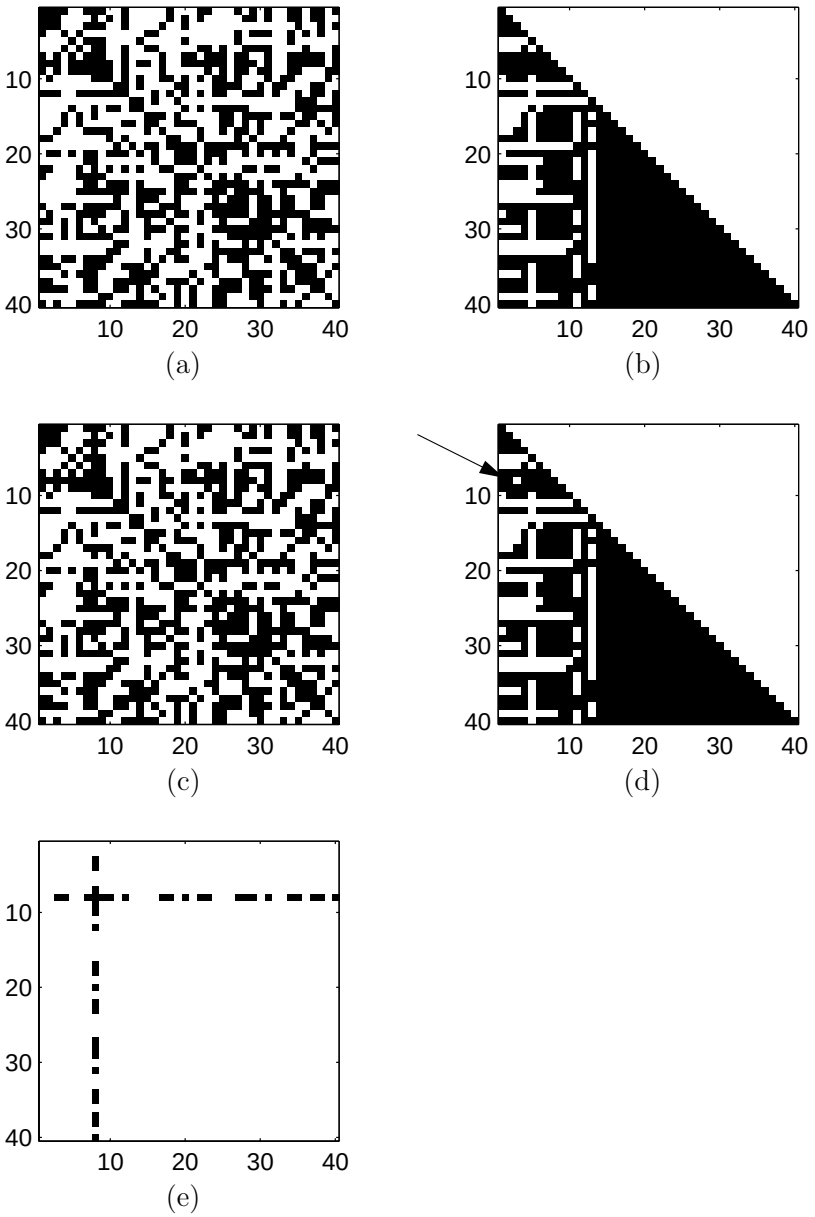


Figure 7: (a) Precision matrix $\boldsymbol{Q}$, (b) Cholesky triangle $\boldsymbol{L}$, (c) new precision matrix $\widetilde{\boldsymbol{Q}} = \widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^T$, (d) modified Cholesky triangle $\widetilde{\boldsymbol{L}}$ with $\widetilde{L}_{8,3} = 0$, (e) predicted $\boldsymbol{Q} - \widetilde{\boldsymbol{Q}}$.

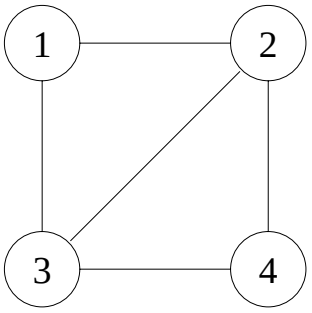


Figure 8: The graph when $\widetilde{Q}_{3,2} \neq Q_{3,2}$.

will differ from Q_{ij}. The result is shown in Fig. 7 (e), where the black elements indicate $\widetilde{Q}_{ij} \neq Q_{ij}$. The predicted result is equal to the difference $\boldsymbol{Q} - \widetilde{\boldsymbol{Q}}$.

4.2 Example

We now go back to the example in Fig. 1, and use this simple graph to demonstrate the results in Theorem 1. Let

$$\widetilde{L}_{ij} = \begin{cases} L_{ij} & \text{if} \quad i,j \neq 3,2 \\ 0 & \text{if} \quad i,j = 3,2 \end{cases}.$$

From (12) we have that $\widetilde{Q}_{ij} \neq Q_{ij}$; $i = 3, j = 2, 3, 4$. Since $Q_{32} = 0$, then $\widetilde{Q}_{32} \neq 0$. The result is

$$\widetilde{\boldsymbol{L}} = \begin{pmatrix} \times & & & \\ \times & \times & & \\ \times & & \times & \\ & \times & \times & \times \end{pmatrix} \qquad \widetilde{\boldsymbol{Q}} = \begin{pmatrix} \times & \times & \times & \\ \times & \times & \surd & \times \\ \times & \surd & \surd & \surd \\ & \times & \surd & \times \end{pmatrix},$$

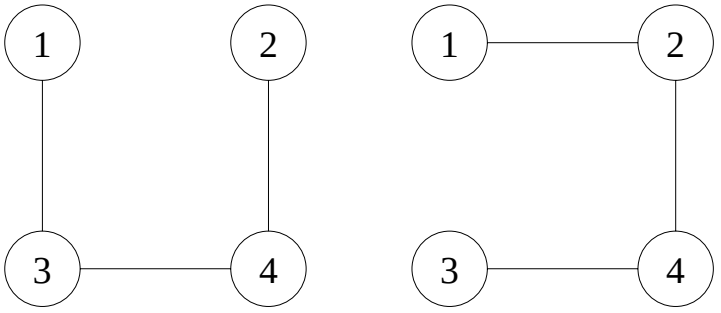
where $\surd$ denotes $\widetilde{Q}_{ij} \neq Q_{ij}$. This graph is shown in Fig. 8. Since $\widetilde{Q}_{32} \neq 0$, then x_2 and x_3 are no longer conditionally independent in the joint distribution. This can also be seen from (8), where now $\widetilde{L}_{3,2} = 0$ while $\widetilde{L}_{21}$ and $\widetilde{L}_{31}$ are possibly non-zero elements.

However, since $\widetilde{L}_{3,2} = 0$, then x_2 and x_3 are conditionally independent in the marginal distribution (from (9)). Also note that $F(2,3) = \{4\}$ is not a separating subset for 2 and 3. If we add small elements on the diagonal, or just change $\widetilde{Q}(1,1)$, then the corresponding $\widetilde{\boldsymbol{L}}$ matrix will not have $\widetilde{L}_{32} = 0$.

In order to obtain $\widetilde{L}_{ij} = 0$ for all $\widetilde{\boldsymbol{Q}} > 0$ with the same graph, we can assure $F(j,i)$ to be a separating subset for j and i by removing edges in $\mathcal{G}$, i.e., remove conditional dependency. Then (4) and (5) are satisfied.

If we want $L_{32} = 0$ for all $\widetilde{\boldsymbol{Q}} > 0$ for the same graph, we must assure that $F(2,3) = \{4\}$ is a separating subset for 2 and 3. This is obtained by removing the edge between node 1 and 2 or between node 1 and 3. These graphs are shown in Figs. 9 (a) and (b), respectively. Consider the graph in Fig. 9 (a) with precision matrix $\widetilde{\boldsymbol{Q}}$ and Cholesky triangle $\widetilde{\boldsymbol{L}}$

$$\widetilde{\boldsymbol{L}} = \begin{pmatrix} \times & & & \\ & \times & & \\ \times & & \times & \\ & \times & \times & \times \end{pmatrix} \qquad \widetilde{\boldsymbol{Q}} = \begin{pmatrix} \times & & \times & \\ & \times & & \times \\ \times & & \times & \times \\ & \times & \times & \times \end{pmatrix}.$$



(a) (b)

Figure 9: The graph when putting (a) $\widetilde{Q}_{1,2} = \widetilde{Q}_{2,1} = 0$ and (b) $\widetilde{Q}_{1,3} = \widetilde{Q}_{3,1} = 0$.

From (8) we can see that x_2 and x_3 are conditionally independent in the joint distribution, Here $F(2,3) = \{4\}$ will be a separating subset for 2 and 3. Thus, $\widetilde{L}_{32} = 0$ for all $\widetilde{\boldsymbol{Q}} > 0$. The result from the graph in Fig. 9 (b) will be similar. However, this approach is not advisable, since we don't want to remove some of the original edges and thereby getting a significantly different covariance matrix.

4.3 Results for band matrices

The ordering of the nodes is important for the sparsity pattern of $\boldsymbol{L}$. If $\boldsymbol{Q}$ is a band matrix, then the computational cost in the Cholesky factorisation can be reduced (see Rue and Held (2005)) due to the following well known result. If $\boldsymbol{Q} > 0$ is a band matrix with bandwidth b_w, i.e., $Q_{i+k,i} = 0$ for $k > b_w$, then $F(i, i+k)$ separates i and $i+k$ when $k > b_w$. Hence, $L_{i+k,i} = 0$ for $k > b_w$, and $\boldsymbol{L}$ has bandwidth b_w.

This result can be generalised to the case where $\boldsymbol{Q}$ is not a complete band matrix, as in Fig. 2. If $\boldsymbol{Q}$ has a band structure in parts of the matrix, this will be preserved in $\boldsymbol{L}$.

Theorem 2 *Let $\boldsymbol{Q} > 0$ and $\boldsymbol{L}$ its lower triangular Cholesky triangle so that $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$. Then*

$$Q_{i,1:j} = 0 \quad \Longleftrightarrow \quad L_{i,1:j} = 0. \tag{14}$$

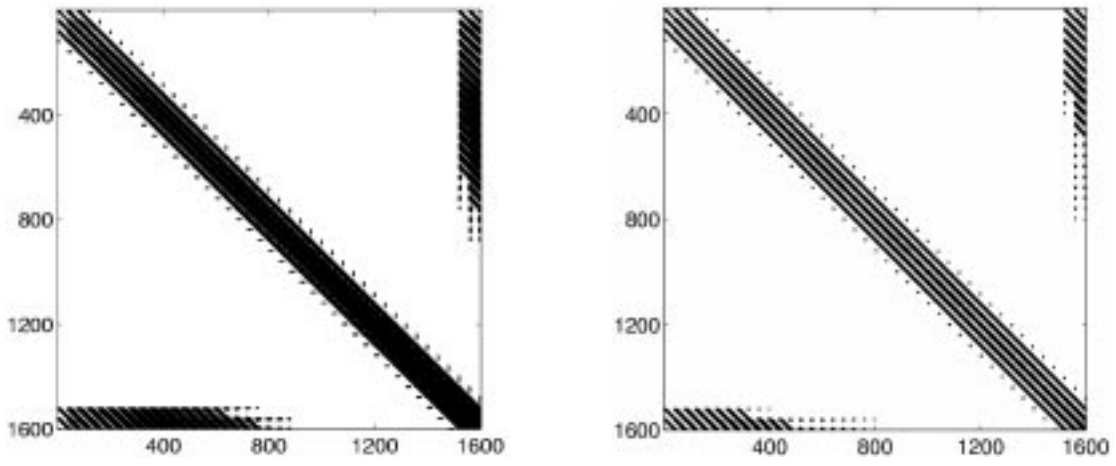
Proof. If $Q_{i,1:j} = 0$, then $F(k,i)$ is a separating subset for node k and i, $k = 1,..,j$. From (4) and (5) it follows that $L_{i,1:j} = 0$. If $L_{i,1:j} = 0$, then $Q_{i,1:j} = 0$ since $Q_{ij} = \sum_{k=1}^{j} L_{ik}L_{jk}$. ■

If $\boldsymbol{Q}$ is a band matrix, then the corresponding $\widetilde{\boldsymbol{Q}}$ resulted from incomplete Cholesky factorisation will have the same bandwidth.

Theorem 3 *Let $\boldsymbol{Q} > 0$ be a band matrix with bandwidth $b_w > 0$, and $\boldsymbol{L}$ its lower triangular Cholesky triangle so that $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^T$ and $\boldsymbol{L}$ has bandwidth b_w. Fix $i^* > j^*$ such that $L_{i^*j^*} \neq 0$ and let*

$$\widetilde{L}_{ij} = \begin{cases} L_{ij} & \text{if} \quad ij \neq i^*j^* \\ 0 & \text{if} \quad ij = i^*j^* \end{cases}.$$

Then $\widetilde{\boldsymbol{Q}} = \widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{L}}^T$ has bandwidth b_w.



(a) (b)

Figure 10: $\widetilde{\boldsymbol{Q}}$ matrices corresponding to the $\widetilde{\boldsymbol{L}}$ matrices in Fig. 3. (a) $\widetilde{\boldsymbol{Q}}$ with *droptol* = 0.0005 and $n_{\widetilde{\boldsymbol{Q}}} = 129\,563$. (b) $\widetilde{\boldsymbol{Q}}$ with *droptol* = 0.005 and $n_{\widetilde{\boldsymbol{Q}}} = 62\,260$.

Proof.

$$\widetilde{Q}_{i+k,i} = \sum_{j=1}^{i} \widetilde{L}_{i+k,j}\widetilde{L}_{ij} = 0 \quad \text{for } k > b_w,$$

since $\widetilde{L}_{i+k,j} = 0$ for $k > b_w$ and $j \leq i$. ■

If $\boldsymbol{Q}$ is not a band matrix, but only part of the matrix has a band structure, then the band structure is preserved during the incomplete Cholesky factorisation. When $L_{i,1:j} = 0$, then $\widetilde{L}_{i,1:j} = 0$, which results in $\widetilde{Q}_{i,1:j} = 0$ since $\widetilde{Q}_{ij} = \sum_{k=1}^{j} \widetilde{L}_{ik}\widetilde{L}_{jk}$. An example of a matrix where only part of the matrix has a band structure is in Fig. 2 (a), where the cyclic boundary conditions prevents this from being a band matrix. The corresponding precision matrices $\widetilde{\boldsymbol{Q}}$ are shown in Fig. 10. We see that the matrices keep their band structure apart from the rows with the cyclic boundary conditions. They are also more dense than the original $\boldsymbol{Q}$ with $n_{\boldsymbol{Q}} = 20\,800$.

Another example where only part of the precision matrix has a band structure is the following case. Let $\mu \sim \mathcal{N}(0, 1)$ and

$$\boldsymbol{x} \mid \mu \ \sim \ \mathcal{N}(\mu\mathbf{1}, \boldsymbol{Q}), \tag{15}$$

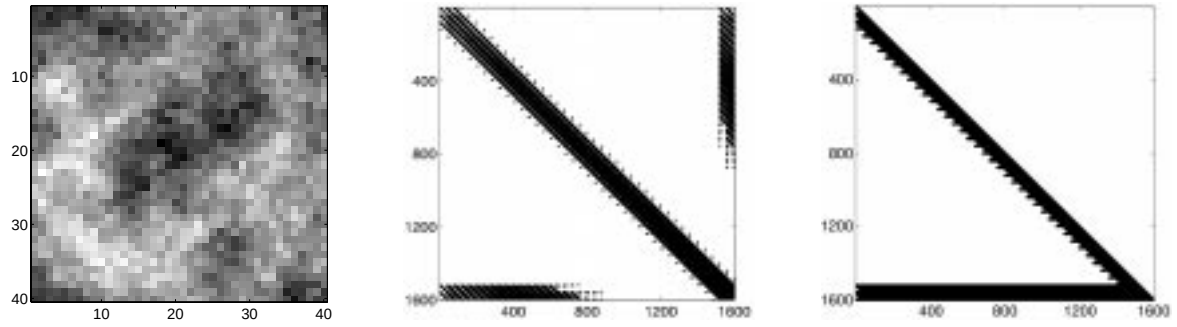
where $\boldsymbol{x}$ has dimension n and $Q_{ij} = 0$ for $|i - j| > p$ and non-zero otherwise. This corresponds to an AR(p) model for $\boldsymbol{x} \mid \mu$. The joint distribution for $(\boldsymbol{x}, \mu)$ is given by

$$\pi(\boldsymbol{x}, \mu) = \pi(\mu)\pi(\boldsymbol{x}|\mu) \propto \exp\left\{-\frac{1}{2}[\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} - 2\mu\boldsymbol{x}^T\boldsymbol{Q}\mathbf{1}^T + \mu^2(1 + \mathbf{1}^T\boldsymbol{Q}\mathbf{1})]\right\}. \tag{16}$$

Thus, the full precision matrix will be

$$\begin{bmatrix} \boldsymbol{Q} & \vdots & \boldsymbol{Q}\mathbf{1} \\ \cdots & & \cdots \\ \mathbf{1}^T\boldsymbol{Q} & \vdots & 1 + \mathbf{1}^T\boldsymbol{Q}\mathbf{1} \end{bmatrix}, \tag{17}$$

where the upper left part of the matrix has bandwidth p.



(a) (b) (c)

Figure 11: (a) Sample from (18). (b) $\widetilde{\boldsymbol{Q}} + \boldsymbol{I}$ with $n_{\boldsymbol{Q}} = 129\,563$. (c) $\boldsymbol{L}$ with $n_{\boldsymbol{L}} = 247\,600$.

4.4 Use of new model in applications

Recall that when we use incomplete Cholesky factorisation, the model is only valid for the corresponding precision matrix $\widetilde{\boldsymbol{Q}}$. If we condition on some of the nodes or on data, then the values in the precision matrix will change, and the corresponding Cholesky triangle will be different from $\widetilde{\boldsymbol{L}}$. The elements $\widetilde{L}_{ij} \neq L_{ij}$ are zero only for the given $\widetilde{\boldsymbol{Q}}$, and they will be different from zero when we change the values in $\widetilde{\boldsymbol{Q}}$. Thus, the Cholesky triangle will again have the same number of fill-ins as the original $\boldsymbol{L}$, and we will not save any computation time.

To illustrate this problem consider the Gaussian field model presented earlier, and use the $\widetilde{\boldsymbol{Q}}$ and $\widetilde{\boldsymbol{L}}$ with *droptol* $= 0.0005$ as the new model. Let $\boldsymbol{y}$ be the observed data, where $\boldsymbol{y} \mid \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}, \boldsymbol{I})$. Then, the density of $\boldsymbol{x}$ conditioned on the data is

$$\begin{aligned} \pi(\boldsymbol{x}|\boldsymbol{y}) &\propto \pi(\boldsymbol{x})\pi(\boldsymbol{y}|\boldsymbol{x}) \\ &\propto \exp\left\{-\frac{1}{2}\boldsymbol{x}^T(\widetilde{\boldsymbol{Q}} + \boldsymbol{I})\boldsymbol{x} + \boldsymbol{y}^T\boldsymbol{x}\right\}. \end{aligned} \tag{18}$$

Figure 11 (a) shows a sample from the density in (18), where we have used the sample from the full model in Fig. 6 (a) as the data $\boldsymbol{y}$. The new precision matrix $(\widetilde{\boldsymbol{Q}} + \boldsymbol{I})$ is shown in Fig. 11 (b) and the corresponding Cholesky triangle is shown in Fig. 11 (c). We see that the sparsity pattern of this matrix is equal to the original $\boldsymbol{L}$ in Fig. 2 (b).

The same problem will occur in MCMC algorithms. The Cholesky triangle will depend on the values of the variables in the current state and the proposed state, and thus it will change for each iteration. Hence, we have to compute each $\widetilde{L}_{ij}$ to check if it can be neglected, and there will be no gain in computation time.

5 Discussion

In this note we have investigated the option of specifying a GMRF by an extremely sparse Cholesky triangle $\widetilde{\boldsymbol{L}}$ obtained from incomplete Cholesky factorisation of the precision matrix of the original model. We include only those elements that are larger than a given threshold parameter, and we are able to reduce the number of non-zero elements in $\boldsymbol{L}$ considerably. The resulting model has propoerties close to the original one. Computing with this model is extremely efficient, and the results are quite good compared with the original model. However, if we select too high threshold

parameter, then the resulting covariance matrix will be quite different from the original, and the results are not good.

The problem with incomplete Cholesky factorisation, is that the resulting Cholesky triangle is only valid for a specific precision matrix. In the case where we condition on some of the nodes or on data, the values in the precision matrix will change, and the elements in $\widetilde{\boldsymbol{L}}$ that were removed because they were smaller than a threshold parameter are no longer zero. Thus, for applications of interest we have to compute each $\widetilde{L}_{ij}$ to check if it can be neglected, and we will not save any computation time.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, volume 101 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36(2):192–225.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195.

Besag, J. and Higdon, D. (1999). Bayesian analysis of agricultural field experiments (with discussion). *Journal of the Royal Statistical Society, Series B*, 61(4):691–746.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society, Series C*, 50(2):201–220.

Gustavson, I. (1978). A class of first order factorization methods. *BIT: Numerical mathematics*, 18:142–156.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614.

Meijerink, J. A. and van der Vorst, H. A. (1977). An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. *Mathematics of Computation*, 31(137):148–162.

Munksgaard, N. (1980). Solving sparse symmetric sets of linear equations by preconditioned conjugate gradients. *ACM Transactions on Mathematical Software*, 6(2):206–219.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B*, 63(2):325–338.

Rue, H. (2005). Marginal variances for Gaussian Markov random fields. Statistics Report No. 1, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London. (To appear).

Rue, H., Steinsland, I., and Erland, S. (2004). Approximating hidden Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B*, 66(4):877–892.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29(1):31–49.

Saad, Y. (1996). *Iterative Methods for Sparse Linear Systems*. The PWS Series in Computational Science. PWS Publishing Company, Boston.