

Fitting Gaussian Markov Random Fields to Gaussian Fields

HÅVARD RUE and HÅKON TJELMELAND

Norwegian University of Science and Technology

ABSTRACT. This paper discusses the following task often encountered in building Bayesian spatial models: construct a homogeneous Gaussian Markov random field (GMRF) on a lattice with correlation properties either as present in some observed data, or consistent with prior knowledge. The Markov property is essential in designing computationally efficient Markov chain Monte Carlo algorithms to analyse such models. We argue that we can restate both tasks as that of fitting a GMRF to a prescribed stationary Gaussian field on a lattice when both local and global properties are important. We demonstrate that using the Kullback–Leibler discrepancy often fails for this task, giving severely undesirable behaviour of the correlation function for lags outside the neighbourhood. We propose a new criterion that resolves this difficulty, and demonstrate that GMRFs with small neighbourhoods can approximate Gaussian fields surprisingly well even with long correlation lengths. Finally, we discuss implications of our findings for likelihood based inference for general Markov random fields when global properties are also important.

Key words: conditional autoregressive models, empirical Bayes, Gaussian fields, Gaussian Markov random field, Kullback–Leibler discrepancy, Markov chain Monte Carlo, maximum likelihood

1. Introduction

A task often encountered in Bayesian spatial, spatio-temporal and imaging models is to represent a spatial variable x by a stationary Gaussian field for stochastic modelling purposes. We assume that x is defined on an $n_r \times n_c$ lattice \mathcal{A} , where $n_r \leq n_c$. In some cases we have one or more realizations of the variable of interest, and we can then use an empirical Bayes approach in estimating a Gaussian field from the realizations. The usual procedure is to subtract a parametric form of the mean value, estimate the variance and then fit a parametric correlation function to the empirical estimate found from the data. The estimated Gaussian field can then be fed back into the Bayesian model. In those cases where the empirical Bayes approach is not appropriate, perhaps because the spatial variable x is not observed, it is common to specify a parametric correlation function, perhaps using a hierarchical model for its parameters.

Bayesian spatial models are often analysed using Markov chain Monte Carlo (MCMC) techniques [refer to Gilks *et al.* (1996) for an introduction to MCMC]. Such techniques generate (in the limit) dependent realizations from the posterior density by running a Markov chain which has the posterior as its equilibrium distribution. The Markov chain is usually constructed so that only a small set of variables, often single sites, are updated at each iteration. This is done by sampling from the conditional density of the selected variable or group of variables, or from a density close to it, given the rest. For this to be computationally tractable, we often need a Markov property in the posterior, meaning that the conditional density for one variable conditioned on the rest does not depend on all other variables but only those in the Markov neighbourhood. Typically, neighbourhoods are close either in space and/or time. It is common to ensure a Markov property in the posterior by using prior fields with a Markov property; the observations in the posterior enter the Bayesian model through the likelihood and usually do not destroy the Markov properties present in the prior.

The above scenario requires that we provide a Gaussian field on A based on the observed data and that the Gaussian field should have a Markov property. This subclass of Gaussian fields (on A) are called Gaussian Markov random fields (GMRF) or conditional autoregressive models (Cressie, 1993; Besag & Kooperberg, 1995). In the case where an empirical Bayes approach is not appropriate, we might have to specify the GMRF by specifying its neighbourhood and a set of conditional densities uniquely identifying the joint density. Two important questions now arise:

1. How should we fit a GMRF to the observed data, to the estimated Gaussian field or to our prior knowledge about the spatial behaviour? Here, global as well as local properties of the GMRF are important.
2. Is a GMRF with a small neighbourhood capable of approximating the behaviour of a Gaussian field with long correlation length?

This paper is about these questions. We first argue that it suffices to discuss the problem of how to fit a GMRF to a given Gaussian field on a lattice, whether there are observed data or not. This is because the empirical correlations from observed data can be thought of as the correlation function of a Gaussian field. Further, for those cases where we need to specify the field from our prior knowledge only, we find it easier to specify the variance and a correlation function in a Gaussian field, which gives a good (and correct) intuition of the behaviour of the field since the marginal joint density of any two distinct sites is then known. A GMRF specified by conditional properties is much harder to interpret in terms of correlation properties. This is because the marginal joint density of two distinct sites, x_{ij} and x_{kl} , does not depend just on the conditional density for x_{ij} and x_{kl} conditioned on the rest, but also on the marginal density for the rest. We present counter-intuitive full conditionals which support this claim.

Besag & Kooperberg (1995) performed a study where they fitted the parameters of a non-homogeneous GMRF to account for the boundary effects in small lattices. The fitting was done using the Kullback–Leibler discrepancy between a GMRF and a prescribed covariance matrix (i.e. Gaussian field), and solved numerically using an algorithm by Dempster (1972). This numerical algorithm for non-homogeneous GMRFs is only computationally tractable for small lattices. We focus on homogeneous GMRFs on a torus in order to derive a computational feasible algorithm for large lattices, and demonstrate in section 3 that the Kullback–Leibler discrepancy is not suitable for fitting a GMRF to a Gaussian field; correlations outside the neighbourhood can have severely undesirable behaviour which is not supported by the data. Hence, we do not obtain a GMRF which approximates well the correlation function of a Gaussian field using the Kullback–Leibler discrepancy. However, we propose an improved criterion which gives a good overall fit (in terms of correlations and variance) between the GMRF and the Gaussian field.

The plan of this paper is as follows. In section 2 we present GMRFs and discuss their computational advantages over Gaussian fields. In section 3 we discuss how to use the Kullback–Leibler discrepancy between the Gaussian field and the GMRF as a fitting criterion. The Kullback–Leibler discrepancy has close connections to both the principle of maximum-likelihood and maximum entropy, which are basic in statistical inference, and is thus a natural choice. By using properties of covariance matrices for stationary fields on a torus, we are able to evaluate the discrepancy and obtain the best fit using numerical minimization, however, the results obtained are not satisfactory, although still interesting. We propose in section 4 an improved criterion for fitting GMRFs, which includes the Kullback–Leibler discrepancy as a special case. The results demonstrate that GMRFs are indeed capable of fitting Gaussian fields even for long correlation lengths. In section 5 we demonstrate how our approach could be

used to analyse a Bayesian model for a synthetic aperture radar scene. Finally, section 6 discusses implications of our findings for likelihood based inference for GMRFs and general Markov random fields from real data.

2. Preliminaries

A general Gaussian field \mathbf{x} on an $n_r \times n_c$ lattice A , $n_r \leq n_c$, is a multivariate Gaussian variable with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. A Gaussian field on A is a Gaussian Markov random field (GMRF) if it satisfies the Markov property

$$\pi(x_{ij} \mid x_{kl}, (k, l) \in A \setminus \{(i, j)\}) = \pi(x_{ij} \mid x_{kl}, (k, l) \in \partial_{ij}),$$

where ∂_{ij} is the neighbourhood of (i, j) . In most cases, ∂_{ij} is a set of sites close to (i, j) in some norm,

$$\partial_{ij} = \{(k, l) \in A \setminus \{(i, j)\} : \|(k, l) - (i, j)\| \leq c\}. \tag{1}$$

GMRFs have been used in many different areas, typically in situations where the Markov property is needed. Refer to Besag & Kooperberg (1995) for an overview.

The Markov property induced by the neighbourhood system $\{\partial_{ij}, (i, j) \in A\}$ imposes a zero-pattern structure in the precision matrix $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ as follows: $Q_{ij,kl}$ is zero unless (i, j) and (k, l) are neighbours. A more formal statement is that x_{ij} and x_{kl} are conditionally independent given the remaining variables iff $Q_{ij,kl} = 0$. The conditional mean and variance are expressed as follows

$$E(x_{ij} \mid \mathbf{x}_{A \setminus \{(i,j)\}}) = - \sum_{(k,l) \in \partial_{ij}} \frac{Q_{ij,kl}}{Q_{ij,ij}} x_{kl} \quad \text{and} \quad \text{var}(x_{ij} \mid \mathbf{x}_{A \setminus \{(i,j)\}}) = 1/Q_{ij,ij} \tag{2}$$

which depend only on sites in the neighbourhood ∂_{ij} . The coefficients in \mathbf{Q} have nice conditional interpretations: $-Q_{ij,kl}/Q_{ij,ij}$ is the conditional correlation coefficient between x_{ij} and x_{kl} conditioned on the rest, and $1/Q_{ij,ij}$ is the conditional variance of x_{ij} conditioned on the rest.

A GMRF has several benefits compared to a general Gaussian field, amongst which are the following:

Markov property. The valuable Markov property of a GMRF is often necessary for model analysis using Markov chain Monte Carlo techniques as discussed in section 1.

Fast and exact simulation. A GMRF can be quickly and exactly simulated using $\mathcal{O}(n_r^3 n_c)$ flops by the algorithm of Rue (2001). This algorithm is also available for a GMRF defined on a graph. Repeated i.i.d. samples are available in only $\mathcal{O}(n_r^2 n_c)$ flops. The constant of proportionality in the flops-count depends on the size of the neighbourhood. Rue (2001) discusses various strategies for sampling a GMRF on a large lattice (above 256×256 with a 5×5 neighbourhood), involving blocking and a divide-and-conquer step. The algorithm is valid for general coefficients in the precision matrix, and no assumptions about homogeneous coefficients are needed unlike the simulation algorithm based on the discrete Fourier transform (Wood & Chan, 1994; Dietrich & Newsam, 1997). Thus we have full generality which is important in applications. The example we discuss in section 5 illuminates this point. In comparison, simulation algorithms for general Gaussian fields require $\mathcal{O}(n_r^3 n_c^3)$ flops. Furthermore, the memory requirement is often lower for GMRFs and this is important for large fields.

Likelihood. The likelihood of a GMRF can be computed in $\mathcal{O}(n_r^3 n_c)$ flops (Rue, 2001), which is important if the normalization constant is needed in estimation or in an MCMC algorithm. Examples of the latter are where the coefficients in \mathbf{Q} depend on hyperparameters, or the region of interest varies. For a general Gaussian field, we need $\mathcal{O}(n_r^3 n_c^3)$ flops to compute the normalization constant.

For the neighbourhood of the GMRF, we assume a maximum norm is used in (1), so that ∂_{ij} is a window of size $(2m + 1) \times (2m + 1)$, $m = 1, 2, \dots$ centred at (i, j) . The motivation for using a square and not for example a discrete circle, is that the computational cost of the algorithm of Rue (2001), including the computation of the likelihood, is (almost) identical for a square of size $(2m + 1) \times (2m + 1)$ and a discrete circle with diameter $(2m + 1)$. Hence, our choice gives us some extra parameters for little computational cost.

The Gaussian fields we use for simulation experiments in later sections are all stationary and isotropic. We use the four commonly used correlation functions, Gaussian, exponential, spherical and Matérn (Cressie, 1993):

$$\begin{aligned}
 \text{Exponential: } \rho(k, l) &= \exp(-3d(k, l)/r) \\
 \text{Gaussian: } \rho(k, l) &= \exp(-3d(k, l)^2/r^2) \\
 \text{Spherical: } \rho(k, l) &= \begin{cases} 1 - \frac{2}{\pi} \left(\frac{d(k, l)}{r} \sqrt{1 - \left(\frac{d(k, l)}{r} \right)^2} + \arcsin \frac{d(k, l)}{r} \right), & d(k, l) \leq r \\ 0, & d(k, l) > r \end{cases} \quad (3) \\
 \text{Matérn: } \rho(k, l) &= \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(s_\nu \frac{d(k, l)}{r} \right)^\nu K_\nu \left(s_\nu \frac{d(k, l)}{r} \right)
 \end{aligned}$$

where $d(k, l)$ denotes the Euclidean distance from $(0, 0)$ to (k, l) , $K_\nu(\cdot)$ is the modified Bessel function of order ν , and the scaling factor s_ν in the Matérn correlation function is chosen so that $\rho(k, l) = 0.05$ for $d(k, l) = r$ for comparison. For the spherical correlation function r is the correlation length, whereas in the exponential, Gaussian and Matérn cases r defines an effective range of correlation, i.e. the distance where the correlation drops below 0.05. With a slight loss of preciseness, in the remainder of this paper we use the term correlation length or range for r in all four cases.

We denote by $\gamma(k, l)$ the covariance function for the Gaussian field defined as the covariance between x_{00} and x_{kl} . Thus, $\rho(k, l) = \gamma(k, l)/\gamma(0, 0)$ where $\gamma(0, 0)$ denotes the variance. The corresponding quantities for the GMRF are $\tilde{\rho}$ and $\tilde{\gamma}$. The number of parameters in a $(2m + 1) \times (2m + 1)$ neighbourhood is $(m(m + 1)/2) + m + 1$ when we impose rotation and reflection invariance. We denote by θ_{kl} the element $Q_{00,kl}$. Figure 1 displays the parameters taking into account rotation and reflection invariance for $m = 2$.

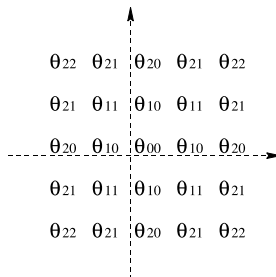


Fig. 1. The six parameters for a GMRF with a 5×5 neighbourhood, labelled as $\theta_{00}, \theta_{10}, \theta_{20}, \theta_{11}, \theta_{21}$ and θ_{22} .

3. Fitting a GMRF using Kullback–Leibler discrepancy

In this section we discuss the use of Kullback–Leibler discrepancy between two densities as a criterion for fitting a GMRF to a Gaussian field. We then show how to evaluate the discrepancy using only $\mathcal{O}(n_r n_c \log(n_r n_c))$ flops, which makes numerical minimization of the Kullback–Leibler discrepancy feasible in practice for large fields.

The Kullback–Leibler discrepancy measures the discrepancy when the density $f(\mathbf{x})$ is approximated by $\tilde{f}(\mathbf{x})$, and is defined as

$$\text{KL}(f, \tilde{f}) = \int_{\mathbb{R}^{n_r n_c}} f(\mathbf{x}) \log \left(\frac{f(\mathbf{x})}{\tilde{f}(\mathbf{x})} \right) d\mathbf{x}. \tag{4}$$

$\text{KL}(f, \tilde{f})$ is not symmetric in its arguments and defines a pseudo-distance as the triangle inequality does not hold. $\text{KL}(f, \tilde{f})$ has close connections to both the principle of maximum-likelihood and maximum entropy, which are basic in statistical inference. $\text{KL}(f, \tilde{f})$ is a natural candidate for a pseudo-metric between a Gaussian field and the approximating GMRF, and was also used by Besag & Kooperberg (1995) as discussed in section 1.

For a stationary Gaussian field f and a GMRF \tilde{f} on a torus, the computation of $\text{KL}(f, \tilde{f})$ is especially efficient.

Lemma 1

Let f be a zero mean stationary Gaussian field on a $n_r \times n_c$ torus with covariance function $\gamma(k, l)$ and covariance matrix Σ , and \tilde{f} a zero mean stationary GMRF on the same torus with precision matrix \mathbf{Q} parametrized with θ . Then

$$\text{KL}(f, \tilde{f}) = -\frac{1}{2} \sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} (\log(\lambda_{ij} q_{ij}(\theta)) - \lambda_{ij} q_{ij}(\theta) + 1) \tag{5}$$

where

$$\lambda_{ij} = \sum_{k=0}^{n_r-1} \sum_{l=0}^{n_c-1} \gamma(k, l) \exp(-2\pi i(ki/n_r + lj/n_c)), \tag{6}$$

$$q_{ij}(\theta) = \sum_{(k,l) \in \partial_{00} \cup \{00\}} \mathbf{Q}_{00,kl}(\theta) \exp(-2\pi i(ki/n_r + lj/n_c)) \tag{7}$$

and $i = \sqrt{-1}$.

The proof is a simple consequence of a property of Toeplitz (block) circulant matrices and is given in the appendix.

The result shows that we can express the Kullback–Leibler discrepancy between a Gaussian field and a GMRF in terms of the eigenvalues of the respective covariance and precision matrices. The eigenvalues are known explicitly as the two-dimensional discrete Fourier transform of the first row of Σ and $\mathbf{Q}(\theta)$, respectively, interpreted as $n_r \times n_c$ arrays. All eigenvalues λ_{ij} and $q_{ij}(\theta)$ are real since Σ and $\mathbf{Q}(\theta)$ are symmetric. We can use a fast two-dimensional discrete Fourier transform algorithm and evaluate $\text{KL}(f, \tilde{f})$ in $\mathcal{O}(n_r n_c \log(n_r n_c))$ flops, with successive iterations being even faster if some extra coding effort is made. We can then solve numerically the minimization problem to obtain the best fit, θ_{KL} ,

$$\theta_{\text{KL}} = \arg \min_{\theta} \sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} [\lambda_{ij} q_{ij}(\theta) - \log(q_{ij}(\theta))], \quad q_{ij}(\theta) > 0, \quad \forall i, j. \tag{8}$$

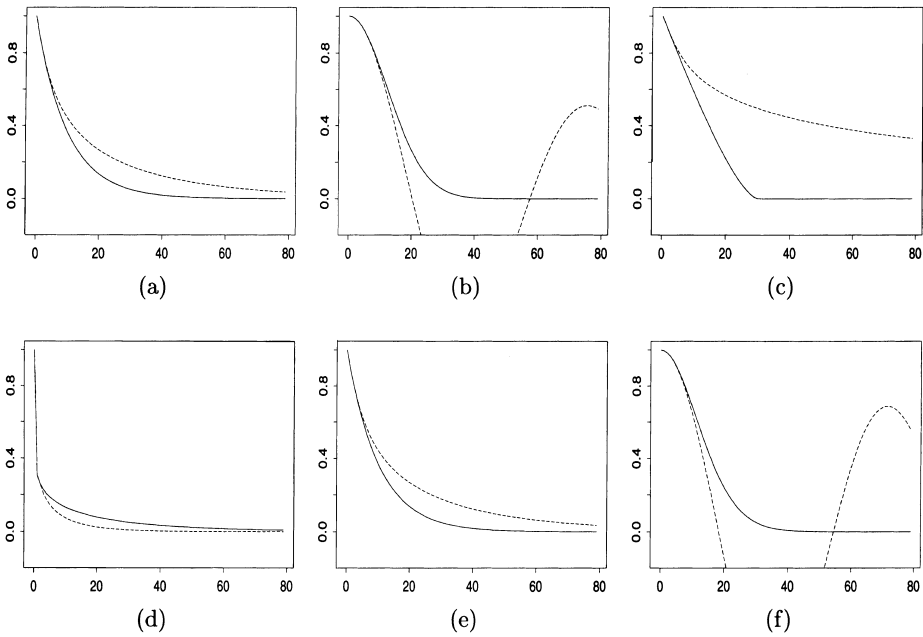


Fig. 2. Target (solid) and fitted (dashed) correlation functions for 5×5 neighbourhood using $KL(f, \tilde{f})$ for the Gaussian field with (a) exponential, (b) Gaussian, (c) spherical and Matérn ((d) $\nu = 0.05$, (e) $\nu = 0.50$, (f) $\nu = 10.0$) correlation functions with correlation length $r = 30$.

The constraint that all eigenvalues are positive is equivalent to positive definiteness. Even if this constraint is linear in θ the dimensions involved are usually so large that we can not treat it as a linear inequality constraint. We approach the constrained minimization problem as an unconstrained one, using a simplex-based iterative algorithm and adding a penalty if some of the eigenvalues are negative. This approach works satisfactorily in practice.

Figure 2 shows the fitted correlation functions $\tilde{\rho}_{KL}(k, l)$, along the main axis, for a GMRF with a 5×5 neighbourhood for the Gaussian fields in section 2 (exponential, Gaussian, spherical and Matérn with $\nu = 0.05, 0.5$ and 10.0) along the main axis with correlation length $r = 30$. The fit along the diagonal axis is slightly better due to the square neighbourhood. Table 1 shows the upper right part (compare with Fig. 1) of the coefficients in the fitted

Table 1. Coefficients in the fitted GMRF using the $KL(f, \tilde{f})$ criterion and a 5×5 neighbourhood for exponential, Gaussian, spherical and Matérn correlation functions with range $r = 30$. The table shows the upper right part of Fig. 1 and displays the conditional standard deviation and conditional correlations

Exponential			Gaussian			Spherical		
-0.026	-0.011	-0.013	-0.028	-0.006	0.017	-0.008	-0.021	-0.011
0.260	0.050	-0.011	0.510	-0.237	-0.006	0.268	0.043	-0.021
0.277	0.260	-0.026	0.003	0.510	-0.028	0.174	0.268	-0.008
Matérn ($\nu = 0.05$)			Matérn ($\nu = 0.5$)			Matérn ($\nu = 10.0$)		
0.033	0.032	0.033	-0.027	-0.010	-0.014	-0.023	-0.014	0.016
0.072	0.048	0.032	0.259	0.051	-0.010	0.467	-0.182	-0.014
0.855	0.072	0.033	0.278	0.259	-0.027	0.004	0.467	-0.023

GMRF for the four correlation functions converted to the interpretable quantities $1/\sqrt{\theta_{00}}$ and $-\theta_{kl}/\theta_{00}$, $(k, l) \in \partial_{00}$. The match for the exponential correlation function is reasonably good, which is natural as the exponential correlation function has the Markov property on the line. The Gaussian, spherical and Matérn (with $\nu = 10.0$) correlation functions are more difficult to fit, and the matches are not at all satisfactory. The fitted GMRFs do not have similar global behaviour to the corresponding Gaussian fields. Even though larger neighbourhoods improve the fit somewhat, the fundamental behaviour of $\tilde{\rho}_{\text{KL}}$ outside ∂_{00} is still present for all neighbourhoods of reasonable size. The question is whether the inaccurate match is due to the fitting procedure or is a feature of the GMRF itself.

To gain more insight into the fitting procedure based on KL discrepancy, assume the GMRF is parametrized with $(2m + 1)^2$ free parameters indexed as θ_{kl} , $(k, l) \in \partial_{00} \cup \{00\}$. The (k', l') component of the gradient of $\text{KL}(f, \tilde{f})$, evaluated at the optimal fit, becomes

$$\sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} \left(\lambda_{ij} - \frac{1}{q_{ij}(\boldsymbol{\theta}_{\text{KL}})} \right) \frac{\partial}{\partial \theta_{k'l'}} q_{ij}(\boldsymbol{\theta}_{\text{KL}}) = 0, \tag{9}$$

where

$$\frac{\partial}{\partial \theta_{k'l'}} q_{ij}(\boldsymbol{\theta}_{\text{KL}}) = \exp(-2\pi i(k'i/n_r + l'j/n_c)). \tag{10}$$

Note that (9) equals the inverse discrete Fourier transform of $\{\lambda_{ij}\}$ and $\{1/q_{ij}(\boldsymbol{\theta}_{\text{KL}})\}$ at (k', l') . Since $\{1/q_{ij}(\boldsymbol{\theta}_{\text{KL}})\}$ are the eigenvalues of the inverse precision matrix $\boldsymbol{Q}(\boldsymbol{\theta}_{\text{KL}})^{-1}$, (9) gives that

$$\gamma(k', l') = \tilde{\gamma}_{\text{KL}}(k', l'), \quad (k', l') \in \partial_{00} \cup \{00\}. \tag{11}$$

This result gives an alternative interpretation of the optimal fit using the KL discrepancy. All covariances within the neighbourhood $\partial_{00} \cup \{00\}$ are fitted exactly and all the remaining ones are determined by inversion of the fitted precision matrix. Dempster (1972) shows this result for maximum-likelihood estimators of the precision matrix in general, see also Whittaker (1990) and Cressie (1993, eq. 7.2.31). Using (11) we can now understand better the results in Fig. 2; only correlations within the 5×5 neighbourhood are fitted and these determine the behaviour of the fitted correlation function for larger lags. All Gaussian fields with similar correlations within the neighbourhood will be fitted with the same GMRF. Hence, we cannot expect to fit the correlation function for a Gaussian field by a GMRF for lags outside $\partial_{00} \cup \{00\}$ using the KL discrepancy. This fact has implications for likelihood based inference for GMRF and general Markov random fields from real data, which we discuss further in section 6.

4. Fitting a GMRF using matched correlation function

In this section we introduce our criterion for fitting a GMRF to a stationary Gaussian field on a torus. The criterion takes all lags of the correlation function into account and is therefore able to fit both local and global properties.

The KL discrepancy fits the variance and the correlation function $\tilde{\rho}(k, l)$ for the GMRF to $\rho(k, l)$ using only lags in ∂_{00} . We extend this idea to take lags outside ∂_{00} into account. We define

$$\mathcal{D}(f, \tilde{f}) = \sum_{(k,l) \in \mathcal{A} \setminus \{00\}} (\rho(k, l) - \tilde{\rho}(k, l))^2 W(k, l, \rho(k, l)), \tag{12}$$

where $W(k, l, \rho(k, l)) \geq 0$ is a user-defined weight function. Our criterion does not depend on the variances of the two fields, but the variance is trivial to fit exactly via a scaling.

$\mathcal{D}(f, \tilde{f})$ defines a proper metric between the correlation functions ρ and $\tilde{\rho}$ if we restrict the weight function $W(k, l, \rho(k, l))$ not to depend on $\rho(k, l)$ and to be strictly positive for all (k, l) . Since the correlation function uniquely identifies the Gaussian density scaled to have unit variance, $\mathcal{D}(f, \tilde{f})$ is also a proper metric between the densities scaled to have unit variance. In other cases, $\mathcal{D}(f, \tilde{f})$ has the interpretation of a discrepancy between f and \tilde{f} .

We can compute the correlation function $\tilde{\rho}$ of our GMRF via the covariance function $\tilde{\gamma}$ which is given as the inverse discrete two-dimensional Fourier transform of $\{1/q_{ij}(\boldsymbol{\theta})\}$,

$$\tilde{\gamma}(k, l) = \frac{1}{n_r n_c} \sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} \frac{1}{q_{ij}(\boldsymbol{\theta})} \exp(2\pi i(ki/n_r + lj/n_c)). \tag{13}$$

Hence, for reasonable choices of n_r and n_c we can compute (12) using $\mathcal{O}(n_r n_c \log(n_r n_c))$ flops.

The variance of a GMRF can always be fitted exactly by a scaling of the precision matrix. Therefore we minimize $\mathcal{D}(f, \tilde{f})$ with respect to $\boldsymbol{\theta}$ for fixed $\theta_{00} = 1$, and then scale the solution to fit the variance. We approach the constrained minimization problem in the same manner as in section 3, adding a penalty if $\boldsymbol{\theta}$ does not define a positive definite precision matrix, i.e. gives at least one non-positive eigenvalue.

The KL discrepancy is a special case of (12) (in the sense that it leads to equivalent estimating equations) choosing $W(k, l, \rho(k, l))$ positive for $(k, l) \in \partial_{00}$ and zero otherwise. We choose

$$W(k, l, \rho(k, l)) = \frac{1}{2\pi d(k, l)}, \tag{14}$$

where $d(k, l)$ is the Euclidean distance on the torus between (k, l) and $(0, 0)$. With this choice we obtain a fitted GMRF with good average properties, since the fitted correlation function should be close to $\rho(k, l)$ for all lags. Further, $\mathcal{D}(f, \tilde{f})$ is a proper metric. With (14) the contribution along each radii has roughly the same weight. Other choices are of course possible and the weight function should be chosen by considering the application at hand.

We have fitted the exponential, Gaussian, spherical and Matérn (for $\nu = 0.05, 0.25, 0.5, 1.0$ and 10.0) correlation functions with ranges 10, 30, and 50, to GMRFs with neighbourhoods from 3×3 up to 9×9 using the $\mathcal{D}(f, \tilde{f})$ criterion. The results presented are for a torus size of 512×512 , but we also tried other array sizes and the results were almost identical. Figures 3–8 and Table 2 summarize the results. Figures 3–8 show the differences between fitted and target correlation functions along the horizontal axis, one figure for each of exponential, Gaussian,

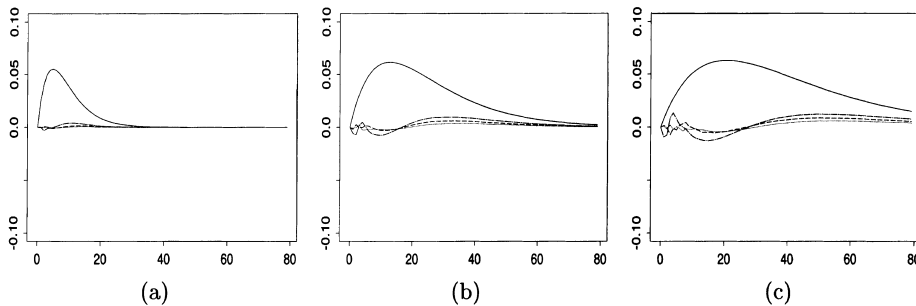


Fig. 3. Difference between fitted and target correlation functions for a Gaussian field with exponential correlation function. Figure (a) is for correlation length $r=10$ and with neighbourhood sizes 3×3 (solid), 5×5 (dash-dot), 7×7 (dashed) and 9×9 (dotted). Figures (b) and (c) show similar results for range $r=30$ and $r=50$, respectively.

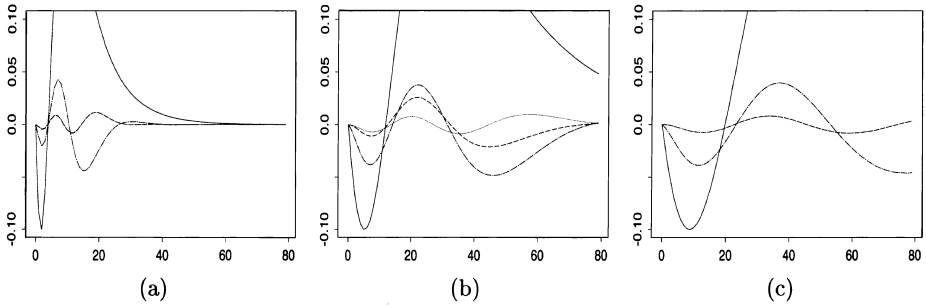


Fig. 4. Difference between fitted and target correlation functions for a Gaussian field with Gaussian correlation function. Figure (a) is for correlation length $r=10$ and with neighbourhood sizes 3×3 (solid), 5×5 (dash-dot), 7×7 (dashed) and 9×9 (dotted). Figures (b) and (c) show similar results for range $r=30$ and $r=50$, respectively.

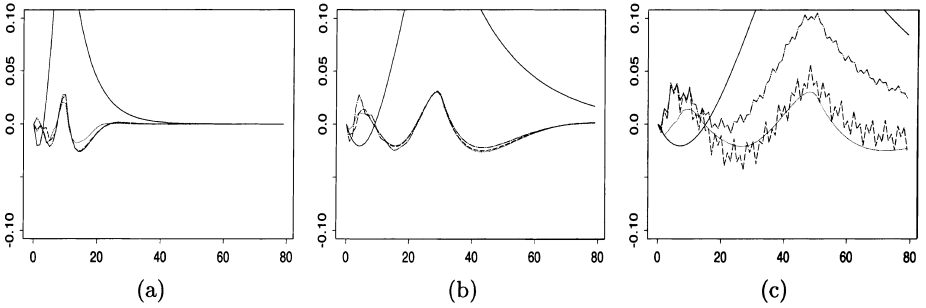


Fig. 5. Difference between fitted and target correlation functions for a Gaussian field with spherical correlation function. Figure (a) is for correlation length $r=10$ and with neighbourhood sizes 3×3 (solid), 5×5 (dash-dot), 7×7 (dashed) and 9×9 (dotted). Figures (b) and (c) show similar results for range $r=30$ and $r=50$, respectively. We believe that the wiggly curves in (c) are because we were not able to find a “good” optimum for this case.

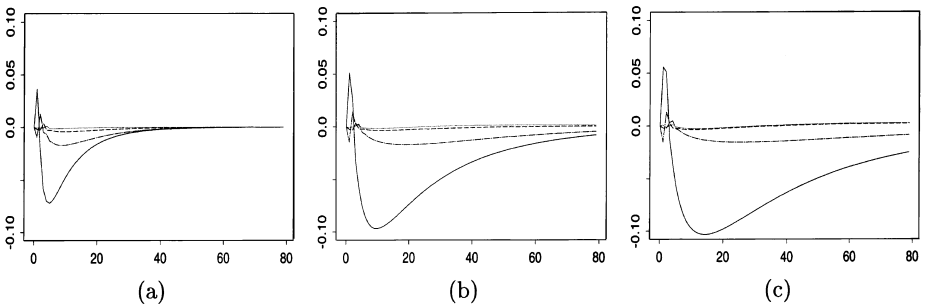


Fig. 6. Difference between fitted and target correlation functions for a Gaussian field with Matérn correlation function with $\nu=0.05$. Figure (a) is for correlation length $r=10$ and with neighbourhood sizes 3×3 (solid), 5×5 (dash-dot), 7×7 (dashed) and 9×9 (dotted). Figures (b) and (c) show similar results for range $r=30$ and $r=50$, respectively.

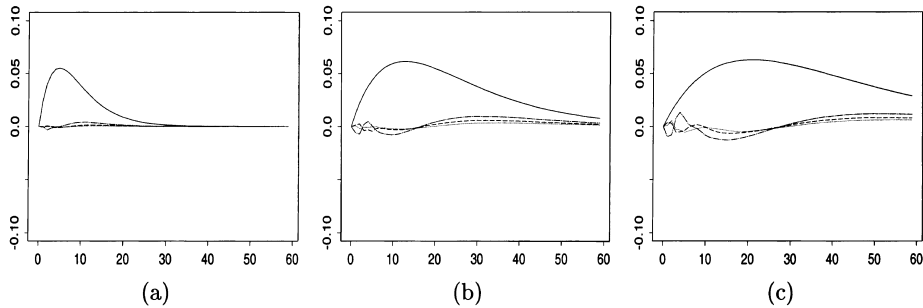


Fig. 7. Difference between fitted and target correlation functions for a Gaussian field with Matérn correlation function with $\nu=0.5$. Figure (a) is for correlation length $r=10$ and with neighbourhood sizes 3×3 (solid), 5×5 (dash-dot), 7×7 (dashed) and 9×9 (dotted). Figures (b) and (c) show similar results for range $r=30$ and $r=50$, respectively.

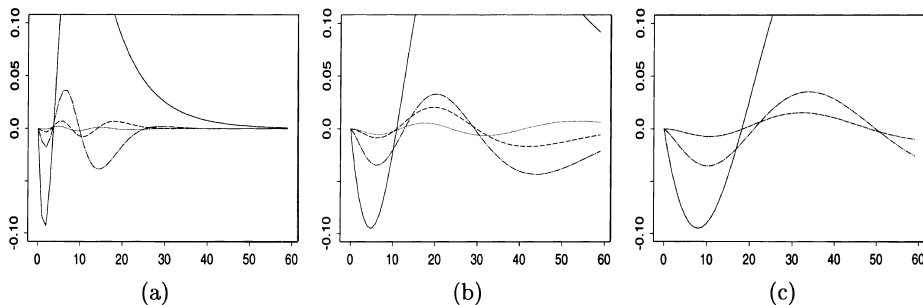


Fig. 8. Difference between fitted and target correlation functions for a Gaussian field with Matérn correlation function with $\nu=10.0$. Figure (a) is for correlation length $r=10$ and with neighbourhood sizes 3×3 (solid), 5×5 (dash-dot), 7×7 (dashed) and 9×9 (dotted). Figures (b) and (c) show similar results for range $r=30$ and $r=50$, respectively.

spherical and Matérn (with $\nu = 0.05, 0.5$ and 10.0) correlation functions. Again the fit along the diagonal axis is better than that along the horizontal axis due to the square neighbourhood. The tables give the maximum absolute difference between $\rho(k, l)$ and $\tilde{\rho}(k, l)$ (now also for $\nu = 0.25$ and 1.0 for the Matérn function).

In a similar way to the GMRF on the line, the auto-regressive process, the correlation function of a GMRF on a torus consists of exponentials and exponentially decaying sin functions, see for example Besag & Kooperberg (1995). The exponential and Matérn (with $\nu = 0.05$ and 0.5) correlation functions are the easiest case to fit, providing near perfect matches for a neighbourhood larger than 3×3 . The Gaussian and spherical correlation functions represent difficult cases, as they consist of exponential decaying sin functions of all orders. Despite this, the fit is surprisingly accurate even for a 5×5 neighbourhood. The accuracy of the fit increases as the neighbourhood increases, but so too does the computational cost related to the GMRF. The optimal size of the neighbourhood depends on the application. If we consider the (lack of) accuracy commonly observed when fitting correlation functions to real data, then our fitted GMRF could equally be used instead of the fitted Gaussian field. It is of course not hard to find correlation functions which is not possible

Table 2. Maximum absolute values of the differences between $\rho(k, l)$ and $\tilde{\rho}(k, l)$ when fitting parameter values by minimizing $\mathcal{D}(f, \tilde{f})$. Values are given for the exponential, Gaussian, spherical and Matérn correlation functions and for neighbourhood sizes $3 \times 3, 5 \times 5, 7 \times 7$ and 9×9

		3×3	5×5	7×7	9×9
Exponential	$r = 10$	0.0551	0.0043	0.0014	0.0006
	$r = 30$	0.0613	0.0112	0.0073	0.0044
	$r = 50$	0.0628	0.0136	0.0120	0.0083
Gaussian	$r = 10$	0.2538	0.0442	0.0117	0.0116
	$r = 30$	0.2597	0.0484	0.0259	0.0097
	$r = 50$	0.2601	0.0463	0.0095	0.0095
Spherical	$r = 10$	0.2045	0.0308	0.0301	0.0243
	$r = 30$	0.2117	0.0318	0.0303	0.0302
	$r = 50$	0.2120	0.1062	0.0566	0.0303
Matérn, $\nu = 0.05$	$r = 10$	0.0718	0.0274	0.0080	0.0017
	$r = 30$	0.1199	0.0369	0.0085	0.0017
	$r = 50$	0.1405	0.0380	0.0081	0.0036
Matérn, $\nu = 0.25$	$r = 10$	0.0200	0.0127	0.0041	0.0015
	$r = 30$	0.0426	0.0203	0.0107	0.0069
	$r = 50$	0.0483	0.0284	0.0161	0.0122
Matérn, $\nu = 0.50$	$r = 10$	0.0551	0.0043	0.0014	0.0006
	$r = 30$	0.0613	0.0112	0.0069	0.0043
	$r = 50$	0.0629	0.0137	0.0108	0.0079
Matérn, $\nu = 1.00$	$r = 10$	0.1354	0.0070	0.0010	0.0002
	$r = 30$	0.1475	0.0016	0.0009	0.0002
	$r = 50$	0.1484	0.0012	0.0001	0.0001
Matérn, $\nu = 10.0$	$r = 10$	0.2417	0.0385	0.0077	0.0028
	$r = 30$	0.2468	0.0433	0.0205	0.0072
	$r = 50$	0.2471	0.0410	0.0155	0.0155

to fit well with a local GMRF, but those we have studied here are well fitted and they are also the most commonly used.

Table 3 shows the upper right part (compare with Fig. 1) of the coefficients in the fitted GMRF for the correlation functions with range 30 and a 5×5 neighbourhood, again converted to the interpretable quantities $1/\sqrt{\theta_{00}}$ and $-\theta_{kl}/\theta_{00}, (k, l) \in \partial_{00}$. We find the full conditionals counter-intuitive when we take into account that the correlation functions we are trying to match are all positive. The full conditionals contain both positive and negative conditional correlations and the different patterns are no easy to interpret. The coefficients are

Table 3. Coefficients in the fitted GMRF using the $\mathcal{D}(f, \tilde{f})$ criterion and a 5×5 neighbourhood for exponential, Gaussian, spherical and Matérn correlation functions with range $r = 30$. The table shows the upper right part of Fig. 1 and displays the conditional standard deviation and conditional correlations

Exponential		Gaussian		Spherical				
-0.220	0.128	-0.087	-0.054	0.008	0.003	-0.360	0.220	-0.134
0.287	0.013	0.128	0.457	-0.171	0.008	0.299	0.005	0.220
0.262	0.287	-0.220	0.040	0.457	-0.054	0.120	0.299	-0.360
Matérn ($\nu = 0.05$)			Matérn ($\nu = 0.5$)			Matérn ($\nu = 10.0$)		
0.033	0.046	0.055	-0.220	0.128	-0.087	-0.054	0.007	0.003
0.044	0.025	0.046	0.286	0.014	0.128	0.457	-0.171	0.007
0.852	0.044	0.033	0.262	0.286	-0.220	0.042	0.457	-0.054

not easier to interpret for larger neighbourhoods. Thus, our intuition of the full conditionals can not be trusted. Therefore, we suggest that modelling (from prior knowledge only) of a Gaussian spatially varying variable with the Markov property, should be done through a correlation function and then fitted by a GMRF with a prescribed neighbourhood using the $\mathcal{D}(f, \tilde{f})$ criterion and numerical minimization. Specifying a joint density implicitly via the full conditionals is common in modern statistical modelling, and sometimes there is no viable alternative. However, our examples show that we should be careful about trusting too much in our intuition about the global properties of conditionally specified models. To obtain an accurate fit of the Gaussian field, we need more accurate coefficients than those displayed in Table 3.

Note that the \mathbf{Q} matrix derived from the coefficients in Table 3 is far from being diagonally dominant, $Q_{ij,ij} > \sum_{kl \neq ij} |Q_{ij,kl}|$. Diagonal dominance is a local property and is a sufficient but not necessary criterion for a matrix to be positive definite (which is a global property). Hence, if we only consider diagonal dominant parametrizations we do not seem to get interesting models: none of our fits are diagonally dominant.

5. An example

Our example is a Bayesian model for estimating the underlying true signal in a synthetic aperture radar image, and demonstrates the importance and computational gain we obtain approximating Gaussian fields by a GMRF. The savings follow from the Markov property of the GMRF and the efficient simulation algorithms that exist for GMRF (Rue, 2001). We further demonstrate how the use of GMRFs allows construction of efficient block updates in MCMC algorithms.

5.1. Data and model

Figure 9 shows a 128×128 region of a 512×512 image acquired by synthetic aperture radar (SAR). The full image is of size 5000×6300 with a 20×15.9 metre resolution. The scene is taken offshore from the Lofoten area in northern Norway and the ocean waves can be clearly seen. SAR is an active imaging remote sensing technique using coherent microwave radiation. The imaging technique used in SAR produces multiplicative noise, called speckle. For an introduction to the physics of SAR, see Kovaly (1976). To reduce the effect of speckle it is common to take several “looks”, i.e. independent images. Our image is based on three looks.

We number the pixels from 1 to $n = 512^2$, and let x_i denote the unobserved true value in pixel i and set $\mathbf{x} = (x_1, \dots, x_n)^T$. The basic observational model for SAR-images is as follows; the speckle gives independent multiplicative noise in each pixel, i.e. with $\mathbf{y} = (y_1, \dots, y_n)^T$ being the observed image one has $y_i = x_i u_i$ where u_1, \dots, u_n are i.i.d. errors for each pixel. For a 1-look image u_i is Rayleigh distributed (density $\propto u \exp(-u^2)$). Our image has been generated as a sum of three one-look images. Thus, the observed value in pixel i is given by

$$y_i = x_i(u'_i + u''_i + u'''_i), \quad (15)$$

where u'_i, u''_i, u'''_i are three independent Rayleigh distributed random variables. Other SAR-images may have combined the looks by squaring the y_i s before summing them, hence the speckle distribution will change accordingly.

The data in Fig. 9 are not the true observations. They are related to the true data by a linear transformation on the log-scale and then rounded to the nearest unsigned two byte integer. Hence, the log of our image-data $\tilde{\mathbf{y}}$ relates linearly to the real data \mathbf{y} , $\ln \tilde{y}_i = a \ln y_i + b$ for some

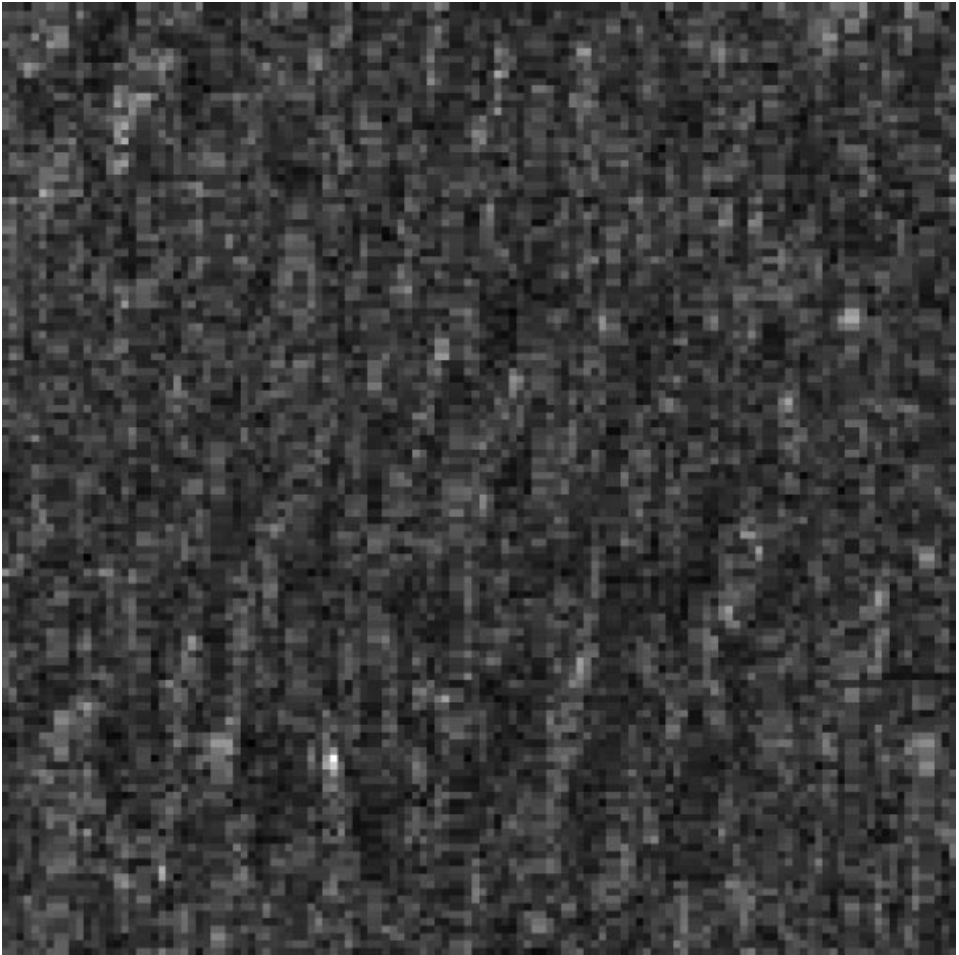


Fig. 9. A 128×128 region of the 512×512 SAR-image, taken over the sea outside the Lofoten area in northern Norway.

unknown scaling parameters a and b . We will ignore the effect of rounding. On a logarithmic scale, (15) gives that

$$\begin{aligned} \check{y}_i &= \check{x}_i + \ln(u'_i + u''_i + u'''_i) \\ &= \check{x}_i + \check{u}_i \end{aligned} \quad (16)$$

where $\check{y}_i = a \ln y_i + b$ and $\check{x}_i = \ln(x_i)$. The density $f_{\check{u}}(\check{u}_i)$ of the additive noise \check{u}_i follows from the distribution for u_i . We have no simple closed form expression for this density, which is displayed in Fig. 10.

For the data described above we adopt a Bayesian model with a stationary Gaussian prior for \check{x} . From the image itself this seems reasonable. We use an anisotropic prior to account for the difference in resolution, 15.9 m along the satellite direction (vertically in the image) and 20.0 m perpendicular to it. We choose an exponential covariance function for the log true signal, with horizontal range r (and vertical range $r/0.795$). This is a reasonable choice for two reasons: (1) the derivative of the covariance is negative at zero meaning that we avoid possible

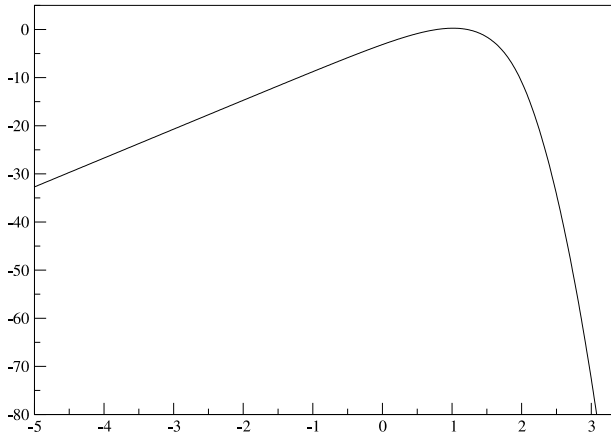


Fig. 10. The log-density of $\check{u} = \log(u' + u'' + u''')$, where u', u'', u''' are three independent Rayleigh distributed variables.

over-smoothing, and (2) the Gaussian field will not be too informative so *a-posteriori* it will adapt to the slowly varying non-stationarity in the data.

We focus on computational issues related to the representation of the Gaussian prior. We adopt the following simple procedure to decide on the unknown scaling-parameters a and b . We assume (without loss of generality) that \check{x} has zero mean and variance σ^2 , and then fix a and b so the two first observed moments of \check{y} equal the theoretical two first moments of $\check{x}_i + \check{u}_i$.

Section 5.2 discusses our MCMC algorithm for sampling from the posterior $\check{x} \mid \check{y}$, including the computational savings of using a GMRF representation of the Gaussian field. Using samples from the MCMC algorithm, we use the relation $x_i = \exp(\check{x}_i)$ to estimate the posterior mean of x which is our estimate of the true signal.

5.2. The MCMC algorithm

Assuming a stationary Gaussian prior with covariance matrix Σ_r (depending on the range r) for the log true signal, \check{x} , the posterior is

$$\pi(\check{x} \mid \mathbf{y}) \propto \exp\left(-\frac{1}{2}\check{x}^T \Sigma_r^{-1} \check{x} + \sum_i \log f_{\check{u}}(\check{y}_i - \check{x}_i)\right). \tag{17}$$

Common MCMC algorithms for posteriors such as (17) use single-site schemes, see for example the recent discussion article by Diggle *et al.* (1998). This means updating each \check{x}_i in some order. For this step we need to sample from or to evaluate the density

$$\pi(\check{x}_i \mid \check{x}_{-i}, \mathbf{y}) \propto \exp\left(-\frac{1}{2}\check{x}^T \Sigma^{-1} \check{x} + \log f_{\check{u}}(\check{y}_i - \check{x}_i)\right). \tag{18}$$

Note that the dimension of Σ is $512^2 \times 512^2$. Due to the lack of (local) Markov property in the prior, we see that the full conditional density for \check{x}_i given the rest, depends on all \check{x}_{-i} . Additionally, due to the large dimension, it is not feasible to compute Σ^{-1} in general. For a stationary Gaussian field however, we could embed Σ into a larger Toeplitz circulant matrix and invert this matrix using the discrete Fourier transform. (We then also need to augment the model to lie on a torus following a rather standard procedure.) But we would still need $\mathcal{O}(512^2)$ operations to evaluate (18), in comparison to $\mathcal{O}(1)$ using a GMRF as shown below.

The same argument applies if we study the full conditional density for \check{x}_w where w is the set of pixels in a small window of the image. Before we are accused of constructing a worst-case scenario, we should simply conclude that a practical statistician would reject the idea of using a Gaussian prior for this problem and chose a more convenient Markov random field prior from the imaging literature [see Winkler (1995) for examples] instead. Although these models are computationally convenient, they lack the nice interpretation and (global) understanding provided by a Gaussian field through the covariance function.

We now demonstrate how our approach of approximating a Gaussian field with a GMRF makes it possible to derive a model with nice interpretation and (global) understanding, which additionally has good computational properties. Since the range is unknown, our first step is to fit GMRFs to the Gaussian field for a set of ranges. Observing the fits shown in Fig. 3, we chose a 5×5 neighbourhood for the GMRF. We fit GMRFs for a discrete set of ranges r on a 512×512 torus. This step is needed once only, as they can and will be reused for later applications. Although the coefficients for the GMRF are computed for a 512×512 grid, they are also valid for other grid-sizes.

Although we could approach our sampling problem using single-site MCMC, instead we discuss how to derive block-updates from the posterior. This illuminates an additional strength of GMRFs: the efficient general simulation algorithm of Rue (2001) which allows us to construct computationally efficient block-updates. Consider now the model in (17) where Σ_r^{-1} is replaced with the (matched) precision matrix \mathbf{Q}_r . The full conditional for \mathbf{x}_w where w is a $m \times m$ window of the lattice, is then

$$\pi(\check{\mathbf{x}}_w \mid \check{\mathbf{x}}_{-w}, \mathbf{y}) \propto \exp\left(-\frac{1}{2} \check{\mathbf{x}}^T \mathbf{Q}_r \check{\mathbf{x}} + \sum_{i \in w} \log f_{\check{u}}(\check{y}_i - \check{x}_i)\right). \tag{19}$$

The computationally big savings using a GMRF prior compared to a Gaussian field prior, come from two facts: (1) we do not need to invert Σ_r to compute the quadratic term, (2) we only need a few (of order $\mathcal{O}(|w|)$) terms from $\check{\mathbf{x}}^T \mathbf{Q}_r \check{\mathbf{x}}$, namely only those involving $\{\check{x}_i, i \in w\}$. (This is all x_j in w' : the dilation of w by the 5×5 mask. When w is a $k \times k$ window, then w' is a $(k + 4) \times (k + 4)$ window.)

Following Rue (2001), we propose to use an GMRF approximation to (19) as our proposal density going from $\check{\mathbf{x}}_w$ to $\check{\mathbf{x}}'_w$, where the non-quadratic terms $\log f_{\check{u}}(\check{y}_i - \check{x}_i)$ are approximated by a quadratic with reasonable local fit around the previous values. We use the quadratic passing through the log density for \check{x}_i and $\check{x}_i \pm L$, where we chose $L = 0.25$ (see Rue (2001) for further discussion). Hence, we compute

$$q(\check{\mathbf{x}}'_w \mid \check{\mathbf{x}}_w) \propto \exp\left(-\frac{1}{2} \check{\mathbf{x}}^T \mathbf{Q}_r \check{\mathbf{x}} + \sum_{i \in w} \left(B_{\check{x}_i, \check{x}'_i} - \frac{1}{2} C_{\check{x}_i} (\check{x}'_i)^2\right)\right) \tag{20}$$

where the B s and C s depends on the quadratic fit around the previous value, and we sample a proposal value $\check{\mathbf{x}}'_w$ from this GMRF using the algorithm in Rue (2001). Note that (20) has the same neighbourhood structure as the prior, but is non-homogeneous and has non-zero mean. We do fast sampling from (20) including rapid computation of the normalizing constant and the conditional mean. We accept the proposal with probability

$$\alpha(\check{\mathbf{x}}_w \rightarrow \check{\mathbf{x}}'_w) = \min\left\{1, \frac{\pi(\check{\mathbf{x}}'_w \mid \check{\mathbf{x}}_{-w}, \mathbf{y}) q(\check{\mathbf{x}}_w \mid \check{\mathbf{x}}'_w)}{\pi(\check{\mathbf{x}}_w \mid \check{\mathbf{x}}_{-w}, \mathbf{y}) q(\check{\mathbf{x}}'_w \mid \check{\mathbf{x}}_w)}\right\}. \tag{21}$$

Note that the normalized densities are required in the ratio of the proposals, since the normalizing constant for $q_{\check{\mathbf{x}}'_w}(\check{\mathbf{x}}_w)$ depends on $\check{\mathbf{x}}'_w$ through the constants $C_{\check{x}_i}$ s. The ratio of the prior does not need the normalizing constants as the range is fixed.

Table 4. *Coefficients in the anisotropic GMRF (for $r = 30$) used in the SAR-example in section 5. The table shows the upper right part of Fig. 1 and displays the conditional standard deviation and conditional correlations*

-0.225	0.116	-0.097
0.474	-0.070	0.215
0.222	0.237	-0.315

5.3. Results

We have estimated the true signal for various fixed values of the range. Naturally, the model should allow the range to vary, but we have not yet been able to sample the true signal and the range for large images simultaneously. This is known to be a hard problem. We ran the model for smaller images and succeeded in getting some, but weak, evidence for a range around 30 pixels to be supported reasonably by the data. We therefore present results for a fixed range of 30 pixels and variance $\sigma^2 = 1$.

The parameters in the GMRF for $r = 30$ is shown in Table 4. Note that the GMRF is anisotropic and demonstrates that our approach can also extend to anisotropic covariance functions although we have less experience in this direction.

We ran our model using blocks of 5×5 , 7×7 and 10×10 . We used two sets of blocks, the first set starts in the upper left corner and divides the image into neighbouring blocks, the second one equals to the first but is shifted by half a block length. This choice gave 21 013, 10 660 and 5204 different blocks for the block-sizes 5×5 , 7×7 and 10×10 , respectively. We define ‘‘one iteration’’ as one sweep over both sets of blocks, and for each block propose a new value using (20) which is accepted with probability given in (21). On our Sun Ultra 4 with a 296 MHz SUNW UltraSPARC-II processor, one update of all blocks took about 32 seconds for all three cases. Although this is quite fast, we can improve the speed with more work on the code. The mean acceptance rate was (after burn-in) about 44.8%, 25.6% and 11.0% for the three block-sizes. A block-size of 7×7 seems to be a good compromise between block-size and acceptance rate.

Figure 11 shows a 128×128 region of the estimated posterior mean of $\mathbf{x} | \mathbf{y}$. Here, we ran 100 full iterations for burn-in and then computed the posterior mean from the 1000 following iterations using a block-size of 7×7 . The posterior mean estimate seems quite reasonable.

6. A comment on likelihood based inference for GMRF and Markov random fields

Combining our findings in sections 3 and 4, we now discuss likelihood based inference from real data in GMRFs and in more general Markov random field models.

The first observation is that our fitted GMRF using the Kullback–Leibler discrepancy is the limiting maximum-likelihood estimator for the parameters in the GMRF when the number of replicates of the Gaussian field tends to infinity. Unless the observed data are realizations of a GMRF, we cannot expect the estimated correlation function outside ∂_{00} to follow the empirical correlation function found from the data, see Fig. 2.

The second observation is that a GMRF can fit a Gaussian field surprisingly well using the proposed criterion (12) rather than the Kullback–Leibler discrepancy. As the correlation function identifies a zero mean and standardized Gaussian density, we have two densities close in norm although the limiting maximum-likelihood estimators of the parameters in the GMRF with data from these two densities are quite different (compare Table 1 with Table 3). Hence, the maximum-likelihood estimators for GMRF seem not to be robust for model error.

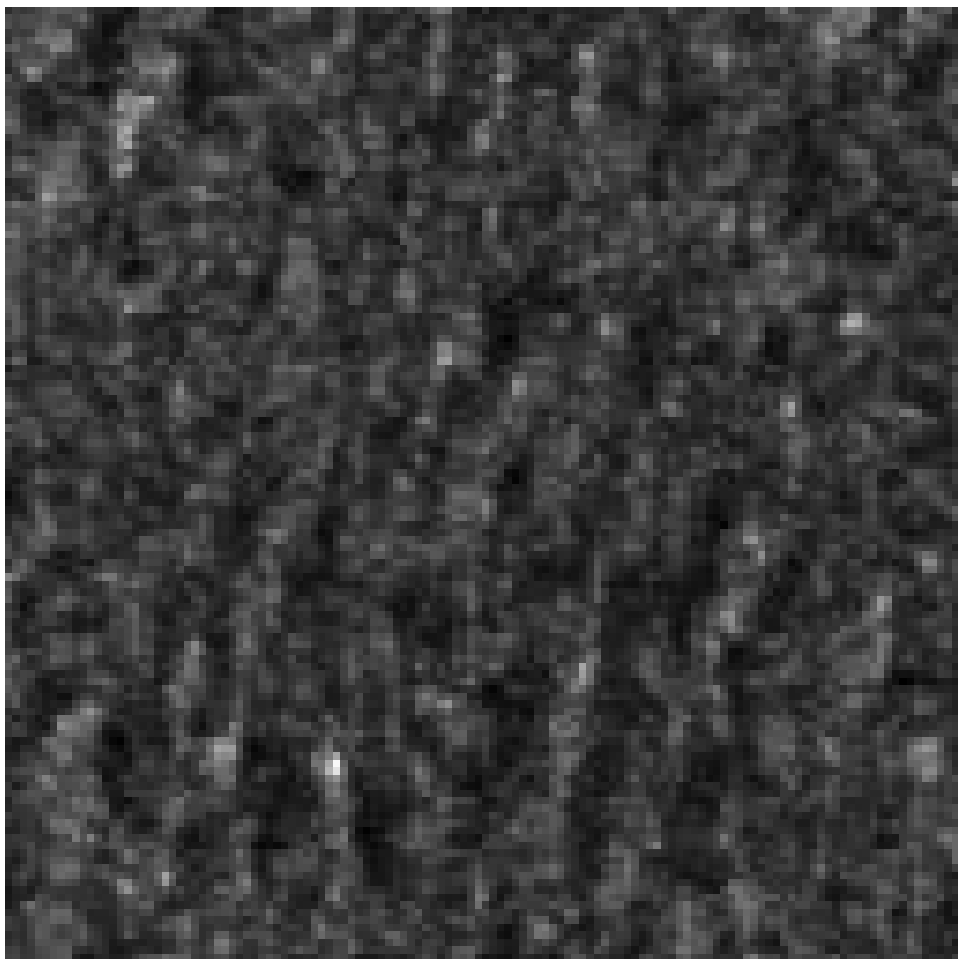


Fig. 11. A 128×128 region of the posterior mean estimate of the true signal \mathbf{x} from the 512×512 image partially shown in Fig. 9.

From our two observations we conclude that likelihood based inference for parameters in GMRFs are questionable for real data, if the global properties of the fitted GMRFs are important. We have investigated empirically the use of (12) as a criterion for fitting the parameters in a GMRF. Although we lose some efficiency if the GMRF is actually the true model, we gain a lot if the data is a Gaussian field close in norm to the GMRF. The estimated correlation function using (12) has reasonable behaviour compared to the empirical one, but the parameters in the GMRF have a larger variance if the GMRF is the wrong model.

We believe that a similar behaviour is also present in many other Markov random fields when estimating parameters from real data using maximum-likelihood. (Although this is computationally difficult, Tjelmeland & Besag (1998) show that it is feasible.) Our argument goes as follows: consider a Markov random field on an exponential family canonical form

$$\pi(\mathbf{x}) \propto \exp\left(\sum_{k=1}^K \beta_k \phi_k(\mathbf{x})\right), \quad (22)$$

where the complete sufficient statistics $\{\phi_k(\mathbf{x})\}$ are known functions and $\boldsymbol{\beta} = \{\beta_k\}$ are unknown parameters. The equations for maximum likelihood can then be formulated as

$$E_{\boldsymbol{\beta}}(\phi_k(\mathbf{x})) = \frac{1}{M} \sum_{m=1}^M \phi_k(\mathbf{x}^{(m)}), \quad k = 1, \dots, K \quad (23)$$

where $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}$ denote observed scenes. As an example, Tjelmeland & Besag (1998) used $\phi_k(\mathbf{x})$ equal to the number of clique configurations within specified sets to model binary Markov random fields. For a zero mean, homogeneous GMRF, (23) becomes

$$E_{\boldsymbol{\beta}} \left(\sum_{(i,j) \in A} \sum_{(k,l) \in A} x_{ij} x_{i+k,j+l} \right) = \frac{1}{M} \sum_{m=1}^M \left[\sum_{(i,j) \in A} \sum_{(k,l) \in A} x_{ij}^{(m)} x_{i+k,j+l}^{(m)} \right] \quad (24)$$

for all $(k, l) \in \partial_{00} \cup \{00\}$. Note that (24) corresponds to the expression in (11) as $M \rightarrow \infty$, and again, the GMRF uses only second order moments within the neighbourhood to estimate the second order properties for all lags. Similarly, (23) shows that only the empirical averages of the local statistics $\phi_k(\mathbf{x})$ are used to estimate $\boldsymbol{\beta}$, and hence these statistics also implicitly define the global properties of the Markov random field.

We expect that local Markov random fields are also able to fit global properties to some extent, but by using likelihood based inference one completely ignores such properties in the estimation process. Hence, we believe that criteria similar to (12) will improve on likelihood based approaches to fit both local and global properties with local Markov models. For the estimation problem of Tjelmeland & Besag (1998), for example, one could in addition to terms corresponding to the sufficient statistics also include pair-wise (or higher order) interaction terms for lags outside the neighbourhood. However, more research along these lines is needed.

Acknowledgements

We thank Arnaldo Frigessi and Gudmund Høst for stimulating discussions, and Harald E. Krogstad for the SAR-image and background information about the SAR-imaging process. HR thanks the EU-TMR project on Spatial Statistics (ERB-FMRX-CT960095) for support and inspiration.

References

- Besag, J. & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82**, 733–746.
- Brockwell, P. J. & Davis, R. A. (1987). *Time series: theory and methods*. Springer Verlag, Berlin.
- Cressie, N. A. C. (1993). *Statistics for spatial data*, 2nd edn. Wiley, New York.
- Dempster, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–175.
- Dietrich, C. R. & Newsam, G. N. (1997). Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM J. Sci. Comput.* **18**, 1088–1107.
- Diggle, P. J., Tawn, J. A. & Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *J. Roy. Statist. Soc. Ser. C* **47**, 299–350.
- Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall, London.
- Kovaly, J. J. (1976). *Synthetic aperture radar*. Artech House, Norwood, MA.
- Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *J. Roy. Statist. Soc. Ser. B*, **63**, 325–338.
- Tjelmeland, H. & Besag, J. (1998). Markov random fields with higher order interactions. *Scand. J. Statist.* **25**, 415–433.
- Whittaker, J. (1990). *Graphical models in applied multivariate statistics*, Wiley series in probability and mathematical statistics. Wiley, Chichester.
- Winkler, G. (1995). *Image analysis, random fields and dynamic Monte Carlo methods*. Springer Verlag, Berlin.

Wood, A. T. A. & Chan, G. (1994). Simulation of stationary Gaussian processes in $[0, 1]^d$. *J. Comput. Graph. Statist.* **3**, 409–432.

Received May 1999, in final form May 2001

H. Rue, Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.
E-mail: havard.rue@math.ntnu.no

Appendix. Proof of lemma 1

The proof uses standard results about Toeplitz circulant matrices, see Brockwell & Davis (1987). An $n \times n$ matrix C is Toeplitz circulant iff $C_{ij} = c_{i-j \bmod n}$ for some vector c . The eigenvectors V and eigenvalues D for Toeplitz circulant matrices, $C = VDV^H$ (H denotes conjugate transpose), are explicitly known: $D = \text{diag}(d_0, \dots, d_{n-1})$, where $d_i = \sum_{j=0}^{n-1} c_j \exp(-2\pi i i j/n)$, and V is the unitary matrix with entries $V_{ij} = n^{-1/2} \exp(-2\pi i i j/n)$. Hence, all Toeplitz circulant matrices have the same set of eigenvectors, and the eigenvalues are found by discrete Fourier transform of the first row of C . Since $C^{-1} = VD^{-1}V^H$, the inverse of a Toeplitz circulant matrix is also Toeplitz circulant. In our two-dimensional case, the covariance matrix of the Gaussian field Σ is a Toeplitz block circulant matrix as is the precision matrix of the GMRF Q . This result about Toeplitz circulant matrices generalizes to Toeplitz circulant block matrices: the eigenvalues are the two-dimensional discrete Fourier transform of the first line interpreted as an $n_r \times n_c$ matrix and all Toeplitz circulant block matrices have the same set of eigenvectors. Denote by λ_{ij} the eigenvalues of Σ and by $q_{ij}(\theta)$ the eigenvalues of Q which depend on the parametrization θ . The expressions for λ_{ij} and $q_{ij}(\theta)$ are given in (6) and (7). Hence we have that

$$\begin{aligned} \text{KL}(f, \tilde{f}) &= \int_{\mathbb{R}^{n_r n_c}} f(x) \log\left(\frac{f(x)}{\tilde{f}(x)}\right) dx \\ &= -\log(|\Sigma|^{1/2} |Q|^{1/2}) - \frac{1}{2} \int_{\mathbb{R}^{n_r n_c}} f(x) x^T (\Sigma^{-1} - Q) x dx \\ &= -\frac{1}{2} \sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} \log(\lambda_{ij} q_{ij}(\theta)) - \frac{1}{2} \sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} \left(\frac{1}{\lambda_{ij}} - q_{ij}(\theta)\right) \hat{\lambda}_{ij} \end{aligned} \tag{25}$$

and the result follows.