Conditional-mean least-squares fitting of Gaussian Markov random fields to Gaussian fields

Noel Cressie
Department of Statistics
The Ohio State University
Columbus, OH, USA
(ncressie@stat.osu.edu)
and
Nicolas Verzelen
Department of Mathematics
Université Paris Sud
Orsay, France

ABSTRACT

This article discusses the following problem, often encountered when analyzing spatial lattice data. How can one construct a Gaussian Markov random field (GMRF), on a lattice, that reflects well the spatial-covariance properties present either in data or in prior knowledge? The Markov property on a spatial lattice implies spatial dependence expressed conditionally, which allows intuitively appealing site-by-site model building. There are also cases, such as in biological network analysis, where the Markov property has a deep scientific significance. Moreover, the model is often important for computational efficiency of Markov chain Monte Carlo algorithms. In this article, we introduce a new criterion to fit a GMRF to a given Gaussian field, where the Gaussian field is characterized by its spatial covariances. We establish that on the one hand this criterion is computationally appealing, and on the other hand it allows nonstationary fields to be fitted efficiently.

1 Introduction

Gaussian Markov random fields (GMRFs) use undirected graphs to represent dependencies between stochastic variables. They are widely used in spatial statistics (e.g., Cressie, 1993, Chs. 6 and 7), image analysis (Geman and Geman, 1984), and stochastic expert systems (Lauritzen, 1996). Moreover, they are becoming promising tools in other fields, such as the analysis of gene interactions (Schäfer and Strimmer, 2005). This success is explained by two main benefits of these models:

- The underlying graphs often provide a biological or a physical meaning that is easily visualized by the scientist. For example, in protein-interaction networks, the edges in the graph represent biochemical interactions or regulatory activities.

- The Markov property makes them appealing when the statistical analysis depends on Markov chain Monte Carlo (MCMC) techniques (see Gilks et al., 1996, for an introduction). Bayesian spatial models (e.g., Rue and Held, 2005) are often analyzed using MCMC, for which the GMRF is well adapted to achieve computational efficiency.

In this article, we address the following question in a general framework: How can we fit a GMRF to a given (or estimated) Gaussian field? This question can be addressed once we define a good "pseudo-distance" between two Gaussian processes.

Besag and Kooperberg (1995) performed a study on small lattices to fit the parameters of a GMRF to a given spatial covariance matrix. They used the Kullback-Leibler divergence as a pseudo-distance and an algorithm introduced by Dempster (1972). Subsequently, Rue and Tjelmeland (2002) focused on homogeneous GMRFs on a torus and proposed an improved pseudo-distance (a weighted Kulback-Leibler divergence). These two studies do not consider irregular lattices, such as one might find in applications to disease mapping. In this article, we introduce a new pseudo-distance that is computationally feasible for any type of lattice and has an intuitive statistical interpretation. It is based on conditional regression and weighted squared error loss.

The plan of the paper is as follows. In Section 2, we define GMRFs, discuss their computational advantages, and define formally our objective in this article. In Section 3, we examine three different pseudo-distances between Gaussian fields, namely the Kullback-Leibler-divergence criterion, the (subsequently called Matched-Correlation) criterion of Rue and Tjelmeland (2002), and finally our new criterion that is based on the conditional properties of Gaussian fields. In Section 4, we compare these criteria on simulated and real datasets, and Section 5 contains discussion and conclusions. Technical results and proofs are given in an appendix.

2 Gaussian Markov random fields

In this section, we define a Gaussian Markov random field (GMRF) and give some properties that are needed later. This is followed by a formal definition of the objective of the article, which we shall see reduces to solving a matrix optimization problem.

2.1 Notation

A Gaussian field X on a finite lattice D is a Gaussian random vector of length $|D| \equiv n$ with mean μ and covariance matrix Σ. If $\mathcal{L}(X)$ denotes the distribution of the variable X, then for a Gaussian field, $\mathcal{L}(X) = N(\mu, \Sigma)$; equivalently, we can write

$$X = (X_1, \ldots, X_n) \sim N(\mu, \Sigma)\,.$$

Let $\mathcal{G} = (V, E)$ be an undirected graph where V is the set of vertices and E is the set of edges. In the spatial context, $V = D$, the spatial domain, and for any site $i \in V$, $\mathcal{N}(i)$ is the set of neighbours of i defined by those sites $j \in V$ connected to i by an edge in E. Let $X_{\mathcal{N}(i)}$ denote the random vector of those X_j, where $j \in \mathcal{N}(i)$. The notation $i \sim j$, for i and j two sites in V means that i and j are neighbours.

One final piece of notation is needed for Section 2.4. Let S_n denote the set of all symmetric $n \times n$ matrices, S_n^+ denote the set of all positive-semidefinite symmetric $n \times n$ matrices, and S_n^{++} denote the set of all positive-definite symmetric $n \times n$ matrices.

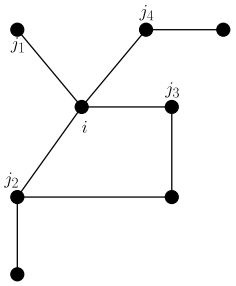

Figure 1: *Graph $\mathcal{G}$ where $\mathcal{N}(i)$ is the set of vertices $\{j_1, j_2, j_3, j_4\}$.*

2.2 Definition and properties of a GMRF

Consider the random vector X indexed by the vertices of an undirected graph $\mathcal{G}$. The Markov property on the graph $\mathcal{G} = (V, E)$ states that if $a, b \in V$ and there is no edge between a and b, then X_a is conditionally independant of X_b, given all remaining variables $X_{V \setminus \{a,b\}}$. A finite Markov random field is a random vector that follows the Markov property with respect to a given graph. More details about Markov random fields can be found, for example, in the texts of Cressie (1993), Ch. 6, or Lauritzen (1996). As a special case, the GMRFs are particularly appealing, since we can obtain their likelihood in closed form (contrary to most Markov random fields). Moreover, they have the appealing property that the underlying graph is easily obtained from inspection of the precision matrix. We now present this formally.

Definition 2.1. *A Gaussian field X on a spatial lattice with sites labelled $V \equiv \{1, 2, \ldots, n\}$ is a Gaussian Markov Random Field (GMRF) with respect to the graph $\mathcal{G} = (V, E)$ if its covariance matrix Σ is non-singular and if X satisfies the following Markov property with respect to $\mathcal{G}$: For all $i \in \{1, \ldots, n\}$,*

$$\mathcal{L}(X_i | X_{-i}) = \mathcal{L}(X_i | X_{\mathcal{N}(i)}) \,,$$

where $\mathcal{L}(A|B)$ denotes the conditional distribution of A given B, $X_{-i} \equiv (X_1, \ldots, X_{i-1}, X_i, \ldots, X_n)$, and $\mathcal{N}(i)$ is the neighborhood of i defined by the edges E of $\mathcal{G}$.

Now let $R \equiv \Sigma^{-1}$ denote the precision matrix of X. The following result is straightforward to prove (e.g., Besag and Kooperberg, 1995).

Proposition 2.2. *Let X be a GMRF with respect to $\mathcal{G}$ with mean zero and precision matrix R. Then the (i, j) element of R, $R[i, j]$, is zero unless $j \in \mathcal{N}(i)$, and the conditional means and variances are*

$$\mathbb{E}(X_i | x_{-i}) = -\sum_{j \neq i} \frac{R[i,j]}{R[i,i]} x_j \,; \;\; i = 1, \ldots, n \,, \tag{1}$$

$$\mathrm{var}(X_i | x_{-i}) = 1/R[i,i] \,; \quad i = 1, \ldots, n \,. \tag{2}$$

2.3 GMRFs on a torus

For some of our spatial applications, we shall consider a particular GMRF, namely the zero-mean, stationary, isotropic GMRF on a regular $n_r \times n_c$ toroidal grid Λ (see Figure 2 for an example of

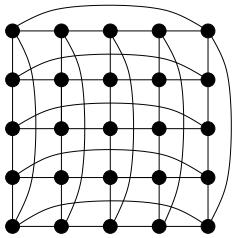

Figure 2: *Example of a graph on a* 5×5 *toroidal grid.*

a 5×5 grid). For the sake of simplicity, we use a double subscript for these fields; that is, X_{ij} refers to the variable corresponding to the site in the ith row and jth column of Λ. Following the same convention, we denote $\Sigma[ij, i'j']$ to be the covariance between X_{ij} and $X_{i'j'}$. The stationarity and isotropy of a GMRF can be expressed in the terms of the covariance matrix Σ, namely for stationarity:

$$\Sigma[ij, i'j'] = f\left((i - i')_r, (j - j')_c\right) ,$$

for all (i, j) and (i', j') on the toroidal grid, where f is a given function and

$$(i - i')_r \equiv sgn(i - i')I(|i - i'| < n_r/2)|i - i'| - sgn(i - i')I(|i - i'| \geq n_r/2)(n_r - |i - i'|),$$

and likewise for $(j - j')_c$; and for isotropy:

$$\Sigma[ij, i'j'] = f^0\left(|(i - i')_r|^2 + |(j - j')_c|^2\right) ,$$

where f^0 is a given function.

2.4 Models of GMRFs

We now formally present the objective of this article. Consider a class (or a model) $\mathcal{M}$ of zero-mean GMRFs with n sites and positive-definite covariance matrix Σ. As these GMRFs in $\mathcal{M}$ have mean zero, they are uniquely defined by their precision matrix $R \equiv \Sigma^{-1}$. Thus, $\mathcal{M}$ is in 1-1 correspondence with a subset of $\mathcal{S}_n^{++}$, the set of all positive-definite symmetric $n \times n$ matrices (Section 2.1). For the sake of simplicity, we also denote $\mathcal{M}$ to be this subset of precision matrices. The goal is to choose $\mathcal{M}$ to be as flexible as possible but such that the precision matrix is as sparse as possible. Thus, we assume that $\mathcal{M}$ is of the following form:

$$\mathcal{M} = S_n^{++} \cap H_m \,, \tag{3}$$

where H_m is a linear subspace of S_n (set of all symmetric $n \times n$ matrices; see Section 2.1) whose matrices are sparse (i.e., very few non-zero entries).

This type of model enables us to formulate our problem as an optimization over convex subsets, as we shall see in the next section. The model $\mathcal{M}$ is quite flexible and includes many interesting classes of GMRFs that one would want to fit (e.g., Besag and Kooperberg, 1995; Rue and Tjelmeland, 2002), including the following two.

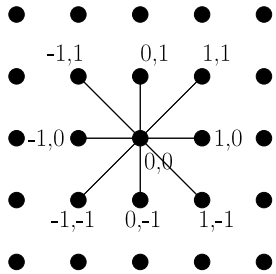

Figure 3: *Neighbourhood of* $(0,0)$*, corresponding to a second-nearest neighborhood structure.*

1. Let H_m be the linear subspace of symmetric matrices R such that $R[i,j] = 0$ for $i \neq j$, except when there is an edge between i and j. Such an R corresponds to a zero-mean GMRF on a graph $\mathcal{G}$.

2. The set H_m is made up of block circulant matrices B such that

$$B[ij, i'j'] = g\left((i-i')_r, (j-j')_c\right) ,$$

where $g(0,0) > 0$, which is the type of model studied in Rue and Tjelmeland (2002). Figure 3 shows a second-nearest neighborhood structure, where neighbors of one of the sites are featured; here $g(x,y) = 0$, for $x, y \geq 2$. This type of model results in a zero-mean, *stationary* GMRF on a toroidal grid. Further, if

$$g(x,y) = g^0\left((x^2+y^2)^{\frac{1}{2}}\right) , \tag{4}$$

then the GMRF is *isotropic*.

It is easy to see that the models given by (3) are not closed in the space of matrices. Then, upon defining $\overline{\mathcal{M}}$ to be the closure of $\mathcal{M}$ in the space of matrices, it is straightforward to show that

$$\overline{\mathcal{M}} = S_n^+ \cap H_m ,$$

where recall from Section 2.1 that S_n^+ is the set of all positive-semidefinite symmetric $n \times n$ matrices. For some criteria, it will be easier to fit the GMRF over $\overline{\mathcal{M}}$ rather than over $\mathcal{M}$. Now, if the resulting precision matrix is singular, the field is not a GMRF but an Intrinsic Gaussian Markov random field (IGMRF). For further details on IGMRFs, see Besag and Kooperberg (1995) and Rue and Held (2005), Ch. 5.

If we do not want to include IGMRFs in the set, it also possible to consider the convex closed set $\overline{\mathcal{M}^\alpha}$:

$$\overline{\mathcal{M}^\alpha} = \overline{S_n^\alpha} \cap H_m ,$$

where $\overline{S_n^\alpha}$ denotes the set of symmetric real matrices whose eigenvalues are greater than or equal to α for a given $\alpha > 0$; notice that $\overline{S_n^\alpha}$ is convex and closed. In this case, we need to choose a value for α; in our examples, we chose α equal to half the smallest eigenvalue of the given Gaussian field Σ'^{-1} (Section 4.2). Optimizing over $\overline{\mathcal{M}^\alpha}$ enables us to compare the corresponding covariance matrices since every fitted matrix is invertible, in contrast to optimizing over $\overline{\mathcal{M}}$.

3 Fitting a GMRF using different criteria

In this section, we discuss the use of three criteria for fitting a GMRF. Specifically, we introduce three pseudo-distances between Gaussian distributions, namely the Kullback-Leibler divergence, the Matched-Correlation criterion used by Rue and Tjelmeland (2002), and our new, Conditional-Mean Least-Squares criterion. In what follows, we seek to minimize each of these criteria over either $\mathcal{M}$, $\overline{\mathcal{M}}$, or $\overline{\mathcal{M}^\alpha}$.

3.1 Kullback-Leibler divergence

The Kullback-Leibler divergence measures the entropy between two densities f and g with respect to a common measure μ. It is defined as

$$K(f,g) \equiv \int f(x) \log\left(\frac{f(x)}{g(x)}\right) d\mu(x)\,.$$

Notice that $K(f,g)$ does not formally define a distance because it is not symmetric and does not satisfy the triangle inequality. However, it can be used to motivate maximum likehood estimation and is closely related to Shannon's measure of entropy. Thus, $K(f,g)$, with $d\mu(x) = dx$, was used by Besag and Kooperberg (1995) to fit the parameters of a non-homogeneous GMRF to a Gaussian field.

If f and g are the joint densities of zero-mean Gaussian fields with covariance matrices Σ_1 and Σ_2, respectively, then it can be shown that the Kullback-Leibler divergence between these two distributions, which we write as $K(\Sigma_1, \Sigma_2)$, is

$$K(\Sigma_1,\Sigma_2) = (1/2)\left(\log(|\Sigma_2|/|\Sigma_1|) + tr(\Sigma_1\Sigma_2^{-1}) - n\right)\,, \tag{5}$$

where $|\Sigma|$ is the determinant of the matrix Σ. In our case, we consider a given zero-mean Gaussian field with covariance matrix Σ' and zero-mean GMRFs from the model $\mathcal{M}$. We wish to minimize

$$h(R) \quad \equiv \quad 1/2\left(-\log\left(|R|\,|\Sigma'|\right) + tr(\Sigma' R) - n\right)\,, \tag{6}$$

over $R \in \mathcal{M}$, and we choose the precision matrix R^0 that achieves this minimum. This problem was considered by Besag and Kooperberg (1995) and Rue and Tjelmeland (2002), and both gave iterative algorithms to find an approximate solution.

Besag and Kooperberg (1995) used an algorithm introduced by Dempster (1972); and, for stationary Gaussian fields on a toroidal grid (as in Figure 2), Rue and Tjelmeland (2002) improved the algorithm to handle large regular lattices. For GMRFs on irregular lattices, Dempster (1972) showed existence and uniqueness of the closest GMRF, but his algorithm can fail to converge for large graphs. Furthermore, the algorithm of Rue and Tjelmeland (2002) cannot be applied in the irregular case. Therefore, we need to look for another criterion that can be used in these more general situations.

3.2 Matched-Correlation criterion

Rue and Tjelmeland (2002) considered stationary, isotropic GRMFs on a toroidal grid and pointed out that the use of the criterion based on the Kullback-Leibler divergence tends to fit the covariance function very well within the GMRF's neighborhood, but gives a poor fit beyond it. Thus, they looked for a criterion that would take into account the whole covariance function.

Consider a zero-mean, stationary Gaussian field X with covariance matrix Σ' on a $n_r \times n_c$ toroidal grid Λ. That is, Σ' satisfies the definition for stationarity given in Section 2.3. Hence, the field is uniquely defined by the variance of X_{00} and the correlation function $\rho \equiv \{\rho_{ij}\}$, where $\rho_{ij} \equiv \text{corr}(X_{00}, X_{ij})$. Consider the following pseudo-distance between two correlation functions ρ and ρ', introduced by Rue and Tjelmeland (2002):

$$q_\omega(R) \equiv \|\rho - \rho'\|_\omega^2 \equiv \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} (\rho_{ij} - \rho'_{ij})^2 \omega_{ij} \,, \tag{7}$$

where $\{\omega_{ij}\}$ are nonnegative weights determined by the user. In the case of isotropy, they recommend,

$$\omega_{ij} \propto \begin{cases} 1 + r & , \text{ if } ij = 00 \\ [1 + r/d(ij, 00)]/d(ij, 00) & , \text{ otherwise,} \end{cases}$$

where r represents the range of the neighbourhood and $d(ij, 00)$ denotes the Euclidean toroidal distance between (i, j) and $(0, 0)$ defined in Section 2.3.

The goal is to minimize the Matched-Correlation criterion (7) over $\mathcal{M}$ for a given Σ'; Rue and Tjelmeland (2002) give an algorithm based on discrete Fourier transforms that yields an approximate solution. The result is a correlation matrix that is subsequently converted to a covariance matrix by matching its variance to that of the original Gaussian field.

To generalize (7) beyond stationary Gaussian fields, one could define an L_2-type distance between a GMRF covariance matrix Σ and a Gaussian field covariance matrix Σ' as follows:

$$\sum_{ij} (R^{-1}[i, j] - \Sigma'[i, j])^2 \omega_{ij} \,,$$

where now the sum is over pairs of sites i and j in Λ, and recall that $R = \Sigma^{-1}$. However, for large nonstationary Gaussian fields, the computational cost of its minimization would be even worse than for the Kullback-Leibler divergence criterion.

3.3 Conditional-Mean Least-Squares (CMLS) criterion

In this section, we define a pseudo-distance between Gaussian fields that has a nice probabilistic interpretation and allows a computationally straightforward way of fitting GMRFs to a Gaussian field. We use the fact that a GMRF can be expressed as a conditional autoregressive (CAR) model, and hence we have the following straightforward result (e.g., Besag, 1974):

Lemma 3.1. *Suppose that $X = (X_1, \ldots, X_n)$ is a zero-mean Gaussian random vector with positive-definite covariance matrix Σ. Then the conditional expectation and conditional variance of the individual component X_i given all the other components can be obtained from the Gaussian conditional distribution,*

$$\mathcal{L}(X_i | x_{-i}) = N\left(\sum_{j \neq i} -\Sigma^{-1}[i, j](\Sigma^{-1}[i, i])^{-1} x_j, (\Sigma^{-1}[i, i])^{-1} \right) . \tag{8}$$

Notice that this matches the relations (1) and (2) given earlier.

The idea behind our criterion is to focus on the conditional distributions of the GMRF, a suggestion made by Guyon (1995) in a more specific context. Then for $i \in \{1, \ldots, n\}$, (8) implies that under the distribution of the GMRF with precision matrix $R = \Sigma^{-1}$, we have:

$$\begin{cases} \mathbb{E}_R(X_i|x_{-i}) = \sum_{j \neq i} -R[i,j](R[i,i])^{-1} x_j \\ \mathrm{var}_R(X_i|x_j) = (R[i,i])^{-1}\,, \end{cases} \tag{9}$$

where the subscript 'R' is used to emphasize dependence of the moments on the GMRF's precision matrix R.

To measure how well a zero-mean GMRF with precision matrix R fits to the (conditional) distribution of a given zero-mean Gaussian field with covariance matrix Σ', consider the following criterion:

$$c(R) \equiv \left(\sum_{i=1}^{n} \frac{1}{\mathrm{var}_R(X_i|x_{-i})} \right)^{-1} \mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{\mathrm{var}_R(X_i|x_{-i})} \left(x_i - \mathbb{E}_R(X_i|x_{-i})\right)^2 \right], \tag{10}$$

where the expectation $\mathbb{E}(.)$ is taken with respect to the distribution of the given Gaussian field with covariance matrix Σ'. We refer to (10) as the Conditional-Mean Least-Squares (CMLS) criterion.

Notice that we use weights inversely proportional to the conditional variance. These are "Gauss-Markov" type weights used in generalized-least-squares estimation. Roughly speaking, we wish to obtain the GMRF that gives the "best" prediction of an individual component based on the other components. The weights are used to scale the components so that they have comparable variability.

If T is a $n \times n$ matrix, define D_T to be the $n \times n$ diagonal matrix whose diagonal is the same as that of T. Then from (9), it is straightforward to show that (10) can be written as

$$c(R) \quad \equiv \quad \frac{tr(D_R^{-1} R \Sigma' R)}{tr(R)}\,. \tag{11}$$

Our goal is to minimize the criterion (11) over a model $\mathcal{M}$ of GMRFs, for a given Gaussian field with covariance matrix Σ'. Clearly, from (11), the minimization can be performed without loss of generality over the set $\mathcal{M}_1 \subset \mathcal{M}$, further restricted to have $\mathrm{tr}(R) \equiv 1$. Minimizing (11) over $\overline{\mathcal{M}}_1$ would yield R_1^0, from which the optimal solution in $\overline{\mathcal{M}}$ can be obtained:

$$R^0 \quad = \quad tr(\Sigma'^{-1}) R_1^0\,. \tag{12}$$

It is always possible to express uniquely a given precision matrix R as follows:

$$R = \tau^2 \Phi^{-1}(I - C)\,,$$

where Φ is a diagonal matrix with $\mathrm{tr}(\Phi) = 1$, and C is a matrix whose diagonal elements are all zero (e.g., Cressie, 1993, p. 434). The matrix $\tau^{-2}\Phi$ is diagonal, corresponding to the vector of conditional variances. Let us consider the case where we *know* the diagonal matrix Φ (e.g., a stationary GMRF on a regular toroidal grid, or the spatial-rates model of Cressie et al., 2005). Then minimizing $c(R)$ given by (11) is equivalent to minimizing

$$d(C) \equiv tr\left(\Phi(I - C)\Sigma'(I - C)\right)\,, \tag{13}$$

over the (closure of the) convex set of matrices C such that C has zero diagonal elements and $\Phi^{-1}(I - C)$ is symmetric and positive-definite. Minimizing (13) yields an optimal C^0, from which the optimal solution in $\overline{\mathcal{M}}$ can be obtained:

$$R^0 = tr(\Sigma'^{-1}) \Phi^{-1}(I - C^0)\,, \tag{14}$$

where recall that Φ is assumed known. In Section 4.1, we show how to minimize the CMLS criterion $c(R)$ very quickly by using the Fast Fourier Transform (as in Rue and Tjelmeland, 2002), when the Gaussian field is stationary on a torus. Finally, we note that $c(R)$ can also be minimized over the set $\overline{\mathcal{M}^\alpha}$, for a given $\alpha > 0$. The computational burden is similar to that for minimizing $c(R)$ over the set $\overline{\mathcal{M}}$.

3.4 A variant of the CMLS criterion

The minimization of (11) is complicated by the presence of the term D_R^{-1}; if it is suppressed, then a computationally more amenable criterion results. Suppose instead we minimize

$$u(R) \equiv tr(R\Sigma' R)\,, \tag{15}$$

over R in $\overline{\mathcal{M}}_1$. In terms of conditional regressions, minimizing $u(R)$ over $\overline{\mathcal{M}}_1$ is the same as minimizing:

$$\left(\sum_{i=1}^n \frac{1}{\mathrm{var}_R(X_i|x_{-i})}\right)^{-2} \mathbb{E}\left[\sum_{i=1}^n \frac{1}{\mathrm{var}_R(X_i|x_{-i})^2}\left(x_i - \mathbb{E}_R(X_i|x_{-i})\right)^2\right],$$

over $\overline{\mathcal{M}}$. Thus, with this variant of the CMLS criterion, we are putting more weight on the terms with small conditional variances and less on those with large conditional variances. However, if the conditional variances are not too different from each other, $u(R)$ will be a good approximation of the CMLS criterion $c(R)$.

From the appendix, we see that

$$u(R) \quad = \quad tr\left[\left(R - \frac{\Sigma'^{-1}}{tr(\Sigma'^{-1})}\right)\Sigma'\left(R - \frac{\Sigma'^{-1}}{tr(\Sigma'^{-1})}\right)\right] + \frac{1}{tr(\Sigma'^{-1})}\,. \tag{16}$$

Thus, minimizing $u(R)$ over $\overline{\mathcal{M}}_1$ is equivalent to computing the orthogonal projection of the matrix $\Sigma'^{-1}/tr(\Sigma'^{-1})$, onto $\overline{\mathcal{M}}_1$, equipped with the inner product

$$\langle A|B\rangle \equiv tr(A\Sigma' B^*)\,,$$

where B^* is the complex transpose of the matrix B. The norm of the matrix A is $\|A\|$, where $\|A\|^2 \equiv \langle A|A\rangle$.

Let R_1^0 be the minimizer of (16), and recall that the minimization is restricted to be over matrices with unit trace. Therefore, to obtain the best GMRF in $\overline{\mathcal{M}}$, we have to find a scalar λ to be substituted into the expression, $R^0 = \lambda R_1^0$. Suppose λ is chosen to minimize $\|\lambda R_1^0 - \Sigma'^{-1}\|$. Then, from the appendix, we see that $\lambda^0 = 1/tr(R_1^0\Sigma' R_1^0)$, and hence the "best" R is:

$$R^0 \quad = \quad \lambda^0 R_1^0\,. \tag{17}$$

Again from the appendix, we see that, in fact, R^0 in (17) minimizes

$$v(R) \equiv tr(R\Sigma' R) - 2tr(R)\,, \tag{18}$$

over R in $\overline{\mathcal{M}}$, which from the appendix is equivalent to minimizing $\|R - \Sigma'^{-1}\|$ over R in $\overline{\mathcal{M}}$.

This criterion v is particularly appealing from a computational perspective, as it is fast to compute even for nonstationary models and does not need any inversions. Notice that optimizing

v is equivalent to projecting Σ'^{-1} onto $\overline{\mathcal{M}}$ with respect to the norm $\|.\|$. Recall from Section 2.4, that $\mathcal{M} = S_n^{++} \cap H_m$. For models with a lot of parameters, the projection of Σ'^{-1} onto the space H_m (w.r.t. $\|.\|$) should be close to Σ'^{-1}. As the set of positive-definite symmetric matrices S_n^{++} is open, it follows that for models with a lot of parameters, the projection of Σ'^{-1} onto H_m should belong to $\mathcal{M}$. Thus, for such models, one has only to compute the minimizer of v over H_m and then check that it belongs to $\overline{\mathcal{M}}$.

Minimizing $v(R)$ over H_m is fast and easy. Let the symmetric matrices $A_1, \ldots, A_k$ be a basis for H_m of dimension k and write $R = \sum_{j=1}^{k} x_j A_j$. Then minimizing $v(R)$ given by (18) is equivalent to minimizing:

$$\sum_{i=1}^{k} \sum_{j=1}^{k} x_i x_j tr(A_i \Sigma A_j) - 2 \sum_{i=1}^{k} x_i tr(A_i),$$

with respect to $x_1, \ldots, x_k$. Upon taking partial derivatives with respect to $x_1, \ldots, x_k$ and setting the result equal to 0, we obtain a $k \times k$ linear system of equations that can be solved directly. Even if, after checking, the solution does not belong to $\overline{\mathcal{M}}$, it can be used as a starting value for an iterative optimization algorithm.

As in the previous subsection, we can also apply the criterion $v(R)$ over the set $\overline{\mathcal{M}^\alpha}$ for a given $\alpha > 0$. Again, for GMRF models with a lot of parameters, the projection of Σ'^{-1} onto H_m should belong to $\overline{\mathcal{M}^\alpha}$ if α is smaller than the smallest eigenvalue of Σ'^{-1}. Hence the minimization of $v(R)$ over $\overline{\mathcal{M}^\alpha}$ proceeds in the same manner as that over $\overline{\mathcal{M}}$.

4 Computation and simulations

4.1 Gaussian fields on a torus

In this section, we compare results from using $q_\omega(R)$, the Matched-Correlation criterion (7), and $c(R)$, the CMLS criterion (10). We shall develop evaluation functions to make this comparison. Further, we make the comparison when fitting to stationary Gaussian fields, since $q_\omega(R)$ is based on this.

For a stationary Gaussian field Σ' and a stationary GMRF R on a torus, the computation of the criteria under consideration can be carried out efficiently. Following Rue and Tjelmeland (2002), consider the following lemma, specifically for the criterion $c(R)$.

Lemma 4.1. *Let Σ' be the covariance matrix of a zero-mean stationary Gaussian field on a $n_r \times n_c$ torus, and let R be the precision matrix in $\overline{\mathcal{M}_1}$ of a zero-mean stationary GMRF on the same torus. Then,*

$$c(R) = n_r n_c \sum_{i=0}^{n_r-1} \sum_{j=0}^{n_c-1} q_{ij}(R)^2 \lambda_{ij}, \tag{19}$$

where the eigenvalues of Σ' are:

$$\lambda_{ij} = \sum_{k=0}^{n_r-1} \sum_{l=0}^{n_c-1} \Sigma[00, kl] \exp(-2\iota\pi(ki/n_r + lj/n_c)),$$

the eigenvalues of R are:

$$q_{ij}(R) = \sum_{k=0}^{n_r-1} \sum_{l=0}^{n_c-1} R[00, kl] \exp(-2\iota\pi(ki/n_r + lj/n_c)),$$

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

The proof is straightforward: Because Σ' and R are defined on a torus, they are Toeplitz circulant matrices and hence they are both diagonalizable using the same basis and their eigenvalues are given above. Hence (19) follows from (11).

The other criteria under consideration can be computed likewise using $\{\lambda_{ij}\}$ and $\{q_{ij}(R)\}$. Again following Rue and Tjelmeland (2002), we use a two-dimensional Fast Fourier Transform algorithm and evaluate $c(R)$ in $\mathcal{O}(n_r n_c \log(n_c n_c))$ flops. We can then solve the minimization problem numerically to obtain the best fit R in $\overline{\mathcal{M}_1}$.

The simulation was conducted using the same design as Rue and Tjelmeland (2002). That is, we consider a Gaussian field on a 64×64 torodial grid, with the following exponential correlation function. Let (i, j) and (i', j') be two sites on the grid; then the correlation between X_{ij} and $X_{i'j'}$ is given by,

$$\text{corr}(d(ij, i'j')) \equiv \exp(-3d(ij, i'j')/r)\,, \tag{20}$$

where the range parameter r is chosen here to be equal to 7. We then fitted stationary, isotropic GMRFs with neigborhood sizes of 3×3, 5×5, and 7×7 to the Gaussian field with Σ' defined by (20), using the criteria $q_\omega(R)$ and $c(R)$.

The first evaluation involves a visual comparison of the fitted correlation functions compared to the true correlation function. The second evaluation is based on the function,

$$\zeta(R) = \mathbb{E}\left(\mathbb{E}(X_0|x_{-0}) - \mathbb{E}_R(X_0|x_{-0})\right)^2 / \text{var}(X_0|x_{-0})\,, \tag{21}$$

where '0' denotes the site $(0, 0)$, $\mathbb{E}_R$ is defined by (9) and $\mathbb{E}(.)$ and var(.) are moments under the given Gaussian field with covariance matrix Σ'. This evaluation function is based on how well the GMRF predicts a value from its neighbors, which is appropriate if we wish to approximate a Gaussian field in a Gibbs sampler. Values of $\zeta(R)$ are presented in Table 1.

Neighborhood	CMLS Criterion, $c(R)$	Matched-Correlation Criterion, $q_w(R)$
3×3	1.15×10^{-2}	0.22
5×5	2.32×10^{-5}	1.32×10^{-3}
7×7	7.22×10^{-7}	1.55×10^{-5}

Table 1: *Evaluation function $\zeta(R)$ for each fitted GMRF*

From Table 1, the CMLS fits perform at least an order of magnitude better than the Matched-Correlation fits, when using the evaluation function $\zeta(R)$. From Figures 4-7, we see visually that the correlation functions are somewhat better approximated when one uses the Matched-Correlation criterion. We conclude that when one wants to predict a missing value given other components

(e.g., in image processing or for kriging), the CMLS criterion considerably outperforms the Matched-Correlation criterion. Moreover, these algorithms share the same computational efficiency as they are all based on the Fast Fourier Transform and iterative minimization.

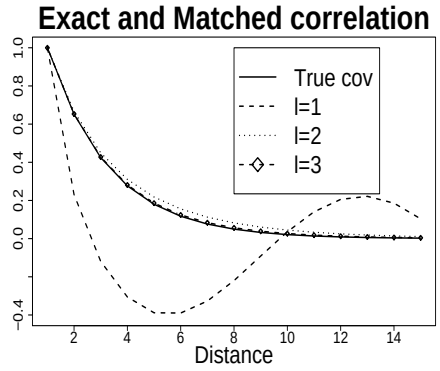

Figure 4: *Correlation function of the original Gaussian field and of the three fitted GMRFs (fitted using the CMLS criterion $c(R)$).*

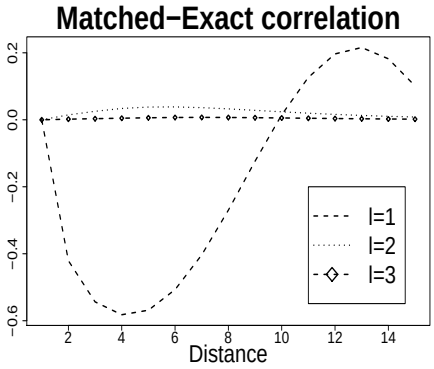

Figure 5: *Difference between the given Gaussian-field correlation function and the three fitted GMRF correlation functions (fitted using the CMLS criterion $c(R)$).*

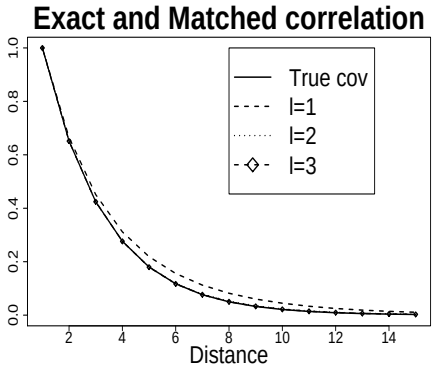

Figure 6: *Correlation function of the original Gaussian field and of the three fitted GMRFs (fitted using the Matched-Correlation criterion $q_\omega(R)$).*

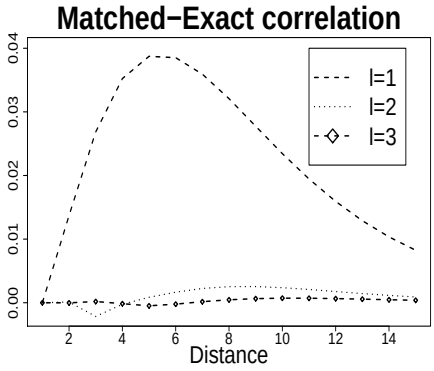

Figure 7: *Difference between the given Gaussian-field correlation function and the three fitted GMRF correlation functions (fitted using the Matched-Correlation criterion $q_\omega(R)$).*

4.2 Radionuclide concentrations on Rongelap Island

In this section, we consider the data set of Diggle et al. (1997) on radionuclide concentrations on Rongelap Island, a small island in the Pacific Ocean. These data were analyzed in several papers, including Diggle et al. (1997), Diggle et al. (1998), and Hrafnkelsson and Cressie (2003). See Diggle et al. (1997) for details of the study.

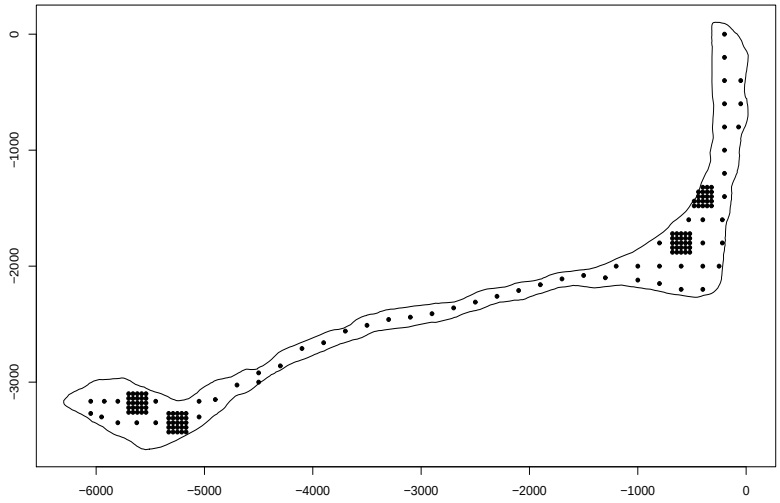

Figure 8: *Map of Rongelap Island. The points show the locations of the 157 sampling sites. The more intensively sampled locations show a* 5×5 *array at 40m spacing. The distance scale is in meters relative to an arbitrary origin. [Subject to permission from the Royal Statistical Society.]*

The levels of ^{137}Cs contamination were measured by *in situ* γ-ray counting at $n = 157$ locations on the island. Notate the data as $\{Y_i : i = 1, \ldots, n\}$, which are the counts at each location; associated with the i-th count is its duration t_i. In this illustration, we focus on the values $Z_i = Y_i/t_i$, the counts per unit time. Figure 8 shows a map of Rongelap Island and the locations of the 157 sites where measurements were taken. Most of the locations are on a grid with a spacing of 200m. The sampling grid was supplemented by four 5×5 arrays of sites at 40m spacing. Hrafnkelsson and Cressie (2003) laid down a fine-scale grid of 40m over the island to predict the ^{137}Cs over the whole island. The total number of points on this fine-scale grid is $m = 1323$.

Our purpose here is not to give a new method of prediction but to evaluate the efficiency of GMRF approximation methods on real-world data. As a consequence, we simply fit a Gaussian geostatistical model to the data $\{Z_i : i = 1, \ldots, n\}$, use it to define a Gaussian field on the fine grid, and then use our theory to approximate it by a well chosen GMRF. Initial analysis of the $\{Z_i\}$ showed them to be approximately Gaussian and hence no transformation of the data was deemed necessary. We consider the process on the fine grid to be a stationary, isotropic Gaussian field with constant mean β. The covariance structure of the process is described by the Matérn covariance function (Matérn, 1986; Handcock and Stein, 1993),

$$\begin{aligned} C_\theta(h) &= \frac{\sigma^2}{2^{\theta_2-1}\Gamma(\theta_2)} \left(\frac{2h\sqrt{\theta_2}}{\theta_1}\right)^{\theta_2} \mathcal{K}_{\theta_2}\left(\frac{2h\sqrt{\theta_2}}{\theta_1}\right); h \neq 0, \\ C_\theta(0) &= \sigma^2 + \tau^2, \end{aligned}$$

where $\theta = (\theta_1, \theta_2, \sigma^2, \tau^2)$, $\mathcal{K}_{\theta_2}$ is the modified Bessel function of order θ_2 (e.g., Abramowitz and Stegun, 1994, Sect. 9), and τ^2 is the nugget effect. The estimates of β, σ^2, τ^2, θ_1, and θ_2 are obtained

straightforwardly using weighted least squares (Cressie, 1993, Ch. 2); they are, respectively, 7.60, 2.92, 3.67, 161.7, and 52.42.

We now address the question of the approximation of this estimated Gaussian field by a GMRF on the fine-scale prediction grid. As this grid is not a torus and is quite large, for computational reasons we chose the CMLS criterion $v(R)$ given by (18).

It is obvious from Figure 8 that the shape of Rongelap Island is thin, and hence its boundary is large for its size. Therefore, it will be important to take boundary effects into account. We do this in a way that respects the neighborhood size. We expect that an entry of R could be of several different types, according to its distance from the boundary, relative to the neighborhood size. That is, we assume that H_m (defined in (3)) is the result of having sites that belong to r possible categories (see below). For any site (i, j), let c_{ij} denote its category. We use the following structure for R:

$$R(ij, i'j') = h\left(|i - i'|^2 + |j - j'|^2, c_{ij} \vee c_{i'j'}\right)$$

where $\vee$ denotes the maximum and the function h is 0 for $|i-i'|^2+|j-j'|^2$ sufficiently large. These models are similar to the models used in the toroidal case, except that the function h now depends on the categories of the sites. It follows that for a GMRF model whose R is given just above, a neighborhood of size $k \times k$ and r categories result in $r(k+1)(k+3)/8$ parameters in R.

The remaining issue is the choice of the category for each site, which will be ordered according to its distance to the boundary. A site falls in category 1 if it is near or on the boundary, resulting in very few neighbors; conversely, a site in the center of the island would fall in category r (see below). This defines H_m, $\mathcal{M}$, $\overline{\mathcal{M}}$, and $\overline{\mathcal{M}^\alpha}$.

In the simulations that follows, we chose $r = 1, 2, 3, 4$, for which we need the notion of l-th nearest neighbor. Let δ denote the largest distance (on the fine-scale grid) from a site to its neighbors. Then the first-nearest neighbors are defined by $\delta = 1$, the second-nearest neighbors by $\delta = \sqrt{2}$, and the third-nearest neighbors by $\delta = 2$. Suppose that $r = 4$. Category 1 consists of any site whose number of first-nearest neighbors is < 4. Category 2 consists of any site whose number of second-nearest neighbors is < 8, which is not in category 1. Category 3 consists of any site whose number of third nearest-neighbors is < 12, which is not in category 1 or 2. Category 4 consists of all remaining sites. For $r = 3, 2, 1$, the definitions are similar taking care that the last categories consist of all sites not contained in the previous categories.

Recall that Σ' denotes the covariance matrix of the Gaussian field. The GMRF approximation is performed for 16 models: we specify neighborhoods of sizes 3×3, 5×5, 7×7, and 9×9 and the number of categories r ranging from 1 to 4. We used the modified CMLS criterion $v(R)$ given by (18). If we minimize $v(R)$ over $\overline{\mathcal{M}}$ to obtain a fitted matrix R, the matrix R could be singular (as noted in Section 3.3). As a consequence, it is not always possible to compute R^{-1} and to compare it with Σ'. Consequently, we minimize v over $\overline{\mathcal{M}^\alpha}$, where we choose α to be half the smallest eigenvalue of Σ'^{-1}. Among these 16 optima, six are optimal over the whole vector space H_m. Hence, from the result at the end of Section 3.4, these optima are easily computable. The other 10 optima are obtained by iterative minimization.

To evaluate the quality of the approximation R, we use two evaluation criteria. First, for each model we compute the difference, $c(R) - c(\Sigma'^{-1})$; this function is related to the mean squared prediction error at each site. The second evaluation criterion is the Kullback-Leibler divergence $K(R^{-1}, \Sigma')$ given by (5).

No. of categories Neighborhood	1	2	3	4
3×3	23.02	18.45	18.19	18.15
5×5	13.40	6.79	5.06	4.18
7×7	10.84	5.17	4.12	3.27
9×9	8.73	2.89	2.24	1.79

Table 2: *Evaluation criterion* $c(R) - c(\Sigma'^{-1})$*, for each fitted GMRF, in units of* 10^{-2}.

No. of categories Neighborhood	1	2	3	4
3×3	34.05	27.90	27.61	27.63
5×5	30.39	16.11	10.71	10.47
7×7	21.48	12.28	9.49	7.19
9×9	19.76	6.35	4.09	3.79

Table 3: *Evaluation criterion* $K(R^{-1}, \Sigma')$*, for each fitted GMRF.*

According to the first evaluation criterion (Table 2), the fitted GMRF performs better in terms of prediction when the neighborhood size and the number of categories increase. When the neighborhood size is small (3×3), there seems to be no benefit from using more than 2 categories. This makes sense, since the boundary effect will be smaller when the neighborhood size is smaller. On the other hand, when the neighborhood size is larger, the boundary effect increases, which explains why models with large neighborhood size but different numbers of categories perform very differently. The second evaluation criterion (Table 3) shows the same type of result, indicating that the minimization of the criterion v given by (18) achieves comparable fits with respect to the CMLS and Kullback-Leibler divergence evaluations. We recall that minimizing the Kullback-Leibler divergence (Section 3.1) or the Matched-Correlation criterion (Section 3.2) for a non-toroidal field is computationally very expensive. The CMLS criterion v allows one to obtain a good fit in terms of both prediction and Kullback-Leibler divergence. Most of all, it allows one to consider large fields, even if they are not toroidal.

5 Discussion and Conclusions

We have shown that a new criterion, the conditional-mean least squares (CMLS) criterion, for fitting Gaussian Markov random fields (GMRFs) to Gaussian fields has attractive properties when the goal is to predict missing values. Because CMLS is defined in terms of GMRF properties, the criterion is immediately relevant to the approximating GMRF. When approximating a regular Gaussian field on a torus with a GMRF, CMLS is computationally competitive with other criteria. Moreover, it can easily handle non-regular GMRFs.

Our new criterion should allow the difficult problem of choice of GMRF neighborhood structure to be addressed directly. Hrafnkelsson and Cressie (2003) performed a study where they estimated the neighborhood and the parameters of a GMRF at the same time. However, they used only a few

parameters on a regular lattice and their results were specific to the lattice chosen. In future work, we would like to estimate the neighborhood and the parameters of not-necessarily-regular GMRFs using an empirical version of the CMLS criterion.

Acknowledgements

This research was supported by the Office of Naval Research under grant number N0014-05-1-0133.

Appendix

We now prove the statements made in Section 3.4.

First, recall $u(R)$ defined by (15). Since $tr(R) = 1$, we obtain:

$$\begin{aligned} u(R) &= tr(R\Sigma' R) - \frac{2}{tr(\Sigma'^{-1})} + \frac{1}{tr(\Sigma'^{-1})} + \frac{1}{tr(\Sigma'^{-1})} \\ &= tr(R\Sigma' R) - \frac{2tr(R\Sigma'\Sigma'^{-1})}{tr(\Sigma'^{-1})} + \frac{tr(\Sigma'^{-1}\Sigma'\Sigma'^{-1})}{tr(\Sigma'^{-1})^2} + \frac{1}{tr(\Sigma'^{-1})} \\ &= tr\left[\left(R - \frac{\Sigma'^{-1}}{tr(\Sigma'^{-1})}\right)\Sigma'\left(R - \frac{\Sigma'^{-1}}{tr(\Sigma'^{-1})}\right)\right] + \frac{1}{tr(\Sigma'^{-1})}\,, \end{aligned}$$

which is (16).

To obtain λ^0 used in (17), we expand the expression,

$$\|\lambda R_1^0 - \Sigma'^{-1}\|^2 = \lambda^2\|R_1^0\|^2 - 2\lambda\langle R_1^0|\Sigma'^{-1}\rangle + \|\Sigma'^{-1}\|^2\,.$$

This is quadratic in λ, so minimizing it is straightforward. Furthermore, $\langle R_1^0|\Sigma'^{-1}\rangle = 1$, and hence

$$\lambda^0 = 1/\|R_1^0\|^2 = 1/tr(R_1^0\Sigma' R_1^0)\,,$$

which is (17).

Finally, consider the criterion (18):

$$v(R) = tr(R\Sigma' R) - 2tr(R)\,,$$

which is to be minimized over R in $\overline{\mathcal{M}}$. This is equivalent to minimizing

$$v(R) + tr(\Sigma'^{-1}) = \|R - \Sigma'^{-1}\|^2\,,$$

over R in $\overline{\mathcal{M}}$.

References

Abramowitz, M., Stegun, I., 1964. *Handbook of Mathematical Functions.* Dover, New York.

Besag, J., Kooperberg, C., 1995. On conditional and intrinsic autoregressions. *Biometika* 82, 733-746.

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

Cressie, N., 1993. *Statistics for Spatial Data*, revised edition. Wiley, New York.

Cressie, N., Perrin, O., Thomas-Agnan, C., 2005. Likelihood-based estimation for Gaussian MRFs. *Statistical Methodology* 2, 1-16.

Dempster, A. P., 1972. Covariance selection. *Biometrics* 28, 157-175.

Diggle, P., Harper, L., Simon, S., 1997. A geostatistical analysis of residual contamination from nuclear weapon testing. In: Barnett, V., Turkman, K. (Eds.), *Statistics for the Environment 3*. Wiley, Chichester, pp. 89-107.

Diggle, P., Tawn, J., Moyeed, R., 1998. Model-based geostatistics (with discussion). *Applied Statistics* 47, 299-350.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6, 721-741.

Gilks, W., Richardson, S., Spiegelhalter, D., 1996. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, London.

Guyon, X., 1995. *Random Fields on a Network*. Springer-Verlag, Berlin.

Handcock, M., Stein, M., 1993. A Bayesian analysis of kriging. *Technometrics* 35, 403-410.

Hrafnkelsson, B., Cressie, N., 2003. Hierarchical modeling of count data with application to nuclear fall-out. *Environmental and Ecological Statistics* 10, 197-200.

Lauritzen, S., 1996. *Graphical Models*. Oxford Science Publications, Oxford, U.K.

Matérn, B., 1986. *Spatial Variation*, 2nd edition. Springer-Verlag, Berlin.

Rue, H., Held, L., 2005. *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC, London.

Rue, H., Tjelmeland, H., 2002. Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29, 31-49.

Schäfer, J., Strimmer, K., 2005. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* 21, 754-764.