Comparing composite likelihood methods based on pairs for spatial Gaussian random fields

Moreno Bevilacqua
DEUV, Universidad de Valparaiso - Valparaiso, Chile
and
Carlo Gaetan
DAIS, Università Ca' Foscari - Venezia, Italy.

March 11, 2014

Abstract

In the last years there has been a growing interest in proposing methods for estimating covariance functions for geostatistical data. Among these, maximum likelihood estimators have nice features when we deal with a Gaussian model. However maximum likelihood becomes impractical when the number of observations is very large. In this work we review some solutions and we contrast them in terms of loss of statistical efficiency and computational burden. Specifically we focus on three types of weighted composite likelihood functions based on pairs and we compare them with the method of covariance tapering. Asymptotic properties of the three estimation methods are derived. We illustrate the effectiveness of the methods through theoretical examples, simulation experiments and by analyzing a data set on yearly total precipitation anomalies at weather stations in the United States.

Keywords: Covariance estimation, Geostatistics, Large datasets, Tapering.

1 Introduction

Geostatistical data come from a limited number of monitoring stations for which we suppose that the observations are a partial realization from a random field defined on the continuum space (Cressie, 1993) and the Gaussian random field plays a central role in providing a building block model in many fields like atmospheric, environmental and geological sciences.

For a Gaussian random field with a given parametric covariance function, exact computation of the likelihood requires calculation of the inverse and determinant of the covariance matrix and this evaluation is slow when the number of observations is large.

More precisely, in the sequel we consider a Gaussian random field $\{Z(s), s \in \mathbb{R}^d\}$ such that the mean function is $E(Z(s)) = x(s)^\intercal \beta$, where $x(s)$ is a vector of q spatially-referenced explanatory variables and $\beta \in \mathbb{R}^q$ is an unknown parameter vector to be estimated. We suppose also that the covariance function $Cov(Z(s), Z(s'))$, for $s, s' \in \mathbb{R}^d$, is unknown up to a small number of parameters that we collect in the vector $\psi \in \mathbb{R}^p$. Without loss of generality, the covariance function is stationary and can be written as $Cov(Z(s), Z(s')) = \sigma^2 \rho(s - s'; \phi)$, with $\sigma^2 > 0$ and $\phi \in \mathbb{R}^{p-1}$.

The unknown parameters $\theta = (\beta^\intercal, \psi^\intercal)^\intercal$ must be estimated on the basis of a finite number of n observations $Z = (Z(s_1), \dots, Z(s_n))^\intercal$. Then $Z \sim \mathcal{N}_n(X\beta, \Sigma(\psi))$, where the i-th row of $n \times q$ matrix X contains the explanatory variables $x(s_i)^\intercal$ and $[\Sigma(\psi)]_{ij} = \sigma^2 \rho(s_i - s_j; \phi)$. The log-likelihood up to an additive constant can be written as

$$l(\theta) = -\frac{1}{2} \log |\Sigma(\psi)| - \frac{1}{2} (Z - X\beta)^\intercal [\Sigma(\psi)]^{-1} (Z - X\beta). \tag{1}$$

The most time-consuming part when calculating (1) is to evaluate the determinant and inverse of $\Sigma(\psi)$. This evaluation could be theoretically carried out in $O(n^{2.81})$ steps (see, e.g., Aho et al. (1974), although the most widely used algorithms such as Cholesky decomposition require up to $O(n^3)$ steps. This can be prohibitive if n is large. This motivated to look for either approximations to the likelihood function or different minimum-contrast-type methods that require less than $O(n^3)$ steps to evaluate (Whittle, 1954; Vecchia, 1988; Curriero and Lele, 1999; Stein et al., 2004; Caragea and Smith, 2006; Fuentes, 2007; Kaufman et al., 2008; Cressie and Johannesson, 2008; Stein, 2008; Lindgren et al., 2011).

Significant computational gain is achieved when the sampling locations form a regular lattice. In this case, the covariance matrix has a special structure (Whittle, 1954) that can be exploited by using spectral methods, reducing the computational burden. For irregularly spaced data, Fuentes (2007) extended the Whittle's idea and suggested to integrate the spatial process over grid cells, obtaining an approximation to the likelihood on a lattice structure. The method requires $O(n \log_2 n)$ operations and does not involve calculating determinants.

Rue and Tjelmeland (2002) approximated the inverse of the covariance matrix to be the precision matrix of a Gaussian Markov random field wrapped on a torus. In this case the numerical factorization of the precision matrix can be done at a cost of $O(n^{3/2})$ for a two-dimensional Gaussian Markov random field. Recently Lindgren et al. (2011) exploited the representations of certain Gaussian random fields with Matérn covariance structure by the solution of a stochastic partial differential equation and derived an approximation based on a Markov Gaussian random field with sparse precision matrix. One drawback of this approach is that we can only find the explicit form for those Gaussian random fields that have a Matèrn covariance structure at certain degree of smoothness (see the discussion to Lindgren et al., 2011). Other drawbacks are the distortion due to edge effects and the necessity of placing nodes at all data locations, both observed and predictive.

Another idea (Banerjee et al., 2008; Cressie and Johannesson, 2008; Stein, 2008) is putting a low rank structure on the covariance matrix. This allows to calculate the inverse and the determinant of a large covariance by inverting and calculating the determinant of a matrix of lower dimension.

All these methods have their relative strengths but they can lead to making unnatural assumptions about the random fields giving a less appropriate model. Instead in the sequel we will concentrate on two estimation methods that preserve the starting model, and, with some adjustments, allow us to perform standard inference as in the case of classical likelihood estimation.

In the tapering approach (Kaufman et al., 2008) certain elements of the covariance matrix that correspond to pairs with large distance are set to zero. This is done, see Section 2, in a way to preserve the property of being positive definite in the resulting 'tapered' matrix. Then sparse matrix algorithms can be used to evaluate efficiently an approximate likelihood where the original covariance has been replaced by the ‘tapered’ matrix. The intuition behind this approach is that correlations between pairs of distant sampling locations are often nearly zero, so little information is lost in taking them to be independent.

With composite likelihood (CL) we will indicate a general class of objective functions based on the likelihood of marginal or conditional events (Lindsay, 1988; Varin et al., 2011). This kind of estimation method can be helpful when it is difficult to evaluate or to specify the full likelihood. In our case the evaluation of the likelihood of the whole set of the observations is too expensive and composing likelihoods with a smaller number of observations is computationally appealing.

Different types of CL functions have been proposed in the literature for estimating the covariance model of spatial and spatio-temporal Gaussian random fields. For instance Stein et al. (2004) proposed a CL based on conditional events improving a previous proposal of Vecchia (1988). More recently, Bevilacqua et al. (2012) considered a weighted CL based on the difference of Gaussian pairs in the space time context and Eidsvik et al. (2013) developed a pairwise Gaussian block composite likelihood in the similar vein of Caragea and Smith (2006).

As outlined in Lindsay et al. (2011), for a given estimation problem the choice of a suitable CL function should be driven by statistical and computational considerations. In particular, for Gaussian random fields, there is a clear computational advantage when we consider only CL based on pairs of observations.

Therefore in this paper we contrast CL functions based on the marginal distribution of a pair or the distribution of an observation conditionally to another observation or the distribution of the difference between two observations. Since the three CL functions are equivalent from a computational point of view, the main purpose of the paper is to compare them based on statistical efficiency. Moreover we establish the asymptotic properties of the associated estimators. Lastly we argue that the CL approach based on pairs is a valuable competitor of the tapering approach with respect to the efficiency when the computational burden is heavy. This is done through theoretical examples and simulations.

The paper is organized as follows. In Section 2 we present in more detail the tapering method while Section 3 describes the three CL estimating methods based on Gaussian pairs. In Section 4 we compare the methods described in Section 2 and 3 through theoretical examples and numerical results. As a real data example, in Section 5 we apply CL and tapering methods on a real data set of yearly total precipitation anomalies already analyzed in Kaufman et al. (2008). Finally, in Section 6 we give some conclusions.

2 Tapered likelihood

In the tapering approach, proposed by Kaufman et al. (2008), certain elements of the covariance matrix $\Sigma(\psi)$ are set to zero multiplying $\Sigma(\psi)$ element by element by a correlation matrix coming from a compactly supported isotropic correlation function. More precisely, we consider a correlation function $r(s-s';d)$ that is identically 0 whenever $\|s-s'\| > d > 0$. The 'tapered' matrix $\Sigma_T(\psi) = \Sigma(\psi) \circ R(d)$, where $[R(d)]_{ij} = r(s_i - s_j; d)$ and $\circ$ is the Schur product, is still positive definite and sparse matrix algorithms can be used to evaluate an approximated log-likelihood efficiently (Furrer and Sain, 2010). There are several ways to construct compactly supported correlation functions (Gneiting, 2002). An example is given by a specific type of Wendland function (Wendland, 1995):

$$r(h;d) = \begin{cases} (1-r)^4 (1+4r) & 0 \le r \le 1 \\ 0 & r > 1 \end{cases} \tag{2}$$

where $r = \|h\|/d$. Kaufman et al. (2008) and Du et al. (2009) support the choice of this function by results under infill asymptotic framework. Another possible choice is the Bohman taper function:

$$r(h;d) = \begin{cases} (1-r)\left(\frac{\sin(2\pi r)}{2\pi r}\right) + \left(\frac{1-\cos(2\pi r)}{2\pi^2 r}\right) & 0 \le r \le 1 \\ 0 & r > 1 \end{cases} \tag{3}$$

Stein (2013) in a simulation study finds that the taper function (3) generally performs better than (2) in terms of statistical efficiency in the estimation procedure. For this reason we will consider in the sequel only the taper function (3).

Kaufman et al. (2008) proposed two approximations of the log-likelihood (1), namely

$$l_{T,1}(\theta,d) = -\frac{1}{2}\log|\Sigma_T(\psi)| - \frac{1}{2}(Z - X\beta)^\intercal[\Sigma_T(\psi)]^{-1}(Z - X\beta), \tag{4}$$

and

$$l_{T,2}(\theta, d) = -\frac{1}{2}\log|\Sigma_T(\psi)| - \frac{1}{2}(Z - X\beta)^{\intercal}([\Sigma_T(\psi)]^{-1} \circ R(d))(Z - X\beta). \tag{5}$$

In (4) the covariance matrix $\Sigma(\psi)$ is tapered, instead in (5) the $\Sigma(\psi)$ as well as the empirical covariance matrix $ZZ^{\intercal}$ are tapered. So the first approximation is computationally more efficient nevertheless the derivative of (5) leads to an unbiased estimating equation. For this reason the recent literature (Shaby and Ruppert, 2012; Stein, 2013) has been focused on (5) and henceforth we consider only the second approximation setting $l_{T,2} = l_T$.

Shaby and Ruppert (2012) show that, under increasing domain asymptotic framework (Cressie, 1993), the maximizer of (5) has an asymptotic Gaussian distribution and the asymptotic variance is given by the inverse of the Godambe information matrix

$$G_{TAP}(\theta, d) = H_{TAP}(\theta, d)J_{TAP}(\theta, d)^{-1}H_{TAP}(\theta, d)^{\intercal}, \tag{6}$$

where

$$H_T(\theta, d) = -\mathrm{E}[\nabla^2 l_T(\theta, d)], \qquad J_T(\theta, d) = \mathrm{E}[\nabla l_T(\theta, d)\nabla l_T(\theta, d)^{\intercal}]. \tag{7}$$

The generic entries of the $H_T(\theta, d)$ and $J_T(\theta, d)$ matrices are

$$[H_T(\theta, d)]_{ij} = \left(\frac{\partial X\beta}{\partial\theta_i}\right)^T \left([\Sigma_T(\psi)]^{-1} \circ R(d)\right) \frac{\partial X\beta}{\partial\theta_j} + \frac{1}{2}\mathrm{tr}\left\{B_i\left(\frac{\partial\Sigma(\psi)}{\partial\theta_j} \circ R(d)\right)\right\}$$

and

$$\begin{aligned}[J_T(\theta, d)]_{ij} \;\; = \;\; & \left(([\Sigma_T(\psi)]^{-1} \circ T)\frac{\partial X\beta}{\partial\theta_i}\right)^T \Sigma\left(([\Sigma_T(\psi)]^{-1} \circ T)\frac{\partial X\beta}{\partial\theta_j}\right) \\ & + \frac{1}{2}\mathrm{tr}\left\{[B_i \circ R(d)]\,\Sigma(\psi)\,[B_j \circ R(d)]\,\Sigma(\psi)\right\},\end{aligned}$$

where $B_i = [\Sigma_T(\psi)]^{-1}\left(\frac{\partial\Sigma(\psi)}{\partial\theta_i} \circ R(d)\right)[\Sigma_T(\psi)]^{-1}$.

Note that $\lim_{d\to\infty} l_T^2(\theta, d) = l(\theta)$, that is when increasing the taper range an improvement of the statistical efficiency is expected. At the limit the asymptotic variance is given by the Fisher information matrix (Mardia and Marshall, 1984)) whose generic entries are

$$[I_{ML}(\theta)]_{ij} = \left(\frac{\partial X\beta}{\partial\theta_i}\right)^T [\Sigma(\theta)]^{-1}\frac{\partial X\beta}{\partial\theta_j} + \frac{1}{2}\mathrm{tr}\left([\Sigma(\psi)]^{-1}\frac{\partial\Sigma}{\partial\theta_i}[\Sigma(\psi)]^{-1}\frac{\partial\Sigma}{\partial\theta_j}\right). \tag{8}$$

3 Composite likelihood estimation based on pairs

Let A_k be a marginal or conditional set of the data, the composite likelihood (CL) (Lindsay, 1988) is an objective function defined as a product of K sub-likelihoods

$$CL(\theta) = \prod_{k=1}^{K} L(\theta; A_k)^{w_k}, \tag{9}$$

where $L(\theta; A_k)$ is a likelihood calculated by considering only the random variables in A_k and w_k are suitable non negative weights that do not depend on θ. The maximum CL estimate is given by $\hat{\theta} = \operatorname{argmax}_{\theta} CL(\theta)$.

The choice of which and how many factors to include in (9) can be related to the computational and statistical efficiency (Lindsay et al., 2011). For instance joint densities of blocks of observations (Caragea and Smith, 2006) or joint densities of pairs of blocks (Eidsvik et al., 2013) have been considered for Gaussian random field estimation.

In the sequel we will consider more simple instances of A_k. Setting $A_k = (Z(s_i), Z(s_j))^\intercal$, we obtain the pairwise marginal Gaussian likelihood L_{ij}. If we let $A_k = Z(s_i)|Z(s_j)$ we obtain the pairwise conditional Gaussian likelihood $L_{i|j}$ and finally setting $A_k = Z(s_i) - Z(s_j)$ we obtain the pairwise difference Gaussian likelihood L_{i-j}. The computational cost for considering all possible pairs is of order $O(n^2)$ while it is of order $O(n^3)$ in considering all possible triplets *i.e.* the same order of the evaluation of the likelihood for Gaussian random fields. Thus from a computational point of view only the pairwise CL is opportune.

The expression for the logarithm of the sub-likelihoods are

$$\begin{aligned} l_{ij}(\theta) &= -\frac{1}{2}\left\{2\log\sigma^2 + \log(1-\rho_{ij}^2) + \frac{B_{ij}}{\sigma^2(1-\rho_{ij}^2)}\right\} && (10)\\ l_{i|j}(\theta) &= -\frac{1}{2}\left\{\log\sigma^2 + \log(1-\rho_{ij}^2) + \frac{G_{ij}^2}{\sigma^2(1-\rho_{ij}^2)}\right\} && (11)\\ l_{i-j}(\theta) &= -\frac{1}{2}\left\{\log\sigma^2 + \log(1-\rho_{ij}) + \frac{U_{ij}^2}{2\sigma^2(1-\rho_{ij})}\right\} && (12)\end{aligned}$$

where $\rho_{ij} = \rho(s_i - s_j; \phi)$, $B_{ij} = (Z(s_i) - \mu_i)^2 + (Z(s_j) - \mu_j)^2 - 2\rho_{ij}(Z(s_i) - \mu_i)(Z(s_j) - \mu_j)$, $G_{ij} = (Z(s_i) - \mu_i) - 2\rho_{ij}(Z(s_j) - \mu_j)$, $U_{ij} = (Z(s_i) - \mu_i) - (Z(s_j) - \mu_j)$ and $\mu_i = x_i^\intercal\beta$. The corresponding weighted composite log-likelihoods are:

$$pl_M(\theta) = \sum_{i=1}^{n}\sum_{j>i}^{n} l_{ij}(\theta)w_{ij}, \tag{13}$$

$$pl_C(\theta) = \sum_{i=1}^{n}\sum_{j\neq i}^{n} l_{i|j}(\theta)w_{ij} = \sum_{i=1}^{n}\sum_{j>i}^{n}(2l_{ij}(\theta) - l_i(\sigma^2,\beta) - l_j(\sigma^2,\beta))w_{ij}, \tag{14}$$

$$pl_D(\theta) = \sum_{i=1}^{n}\sum_{j>i}^{n} l_{i-j}(\theta)w_{ij}, \tag{15}$$

where

$$l_i(\sigma^2,\beta) = -\frac{\log\sigma^2}{2} - \frac{(Z(s_i)-\mu_i)^2}{2\sigma^2}$$

is the marginal likelihood. Observe that $l_{ij} = l_{ji}$, $l_{i-j} = l_{j-i}$ but $l_{i|j} \neq l_{j|i}$ thus the definition of pl_C involves the sum of all the possible pairs. Note that equation (14) is true assuming symmetric weights and that when $w_{ij} = 1$ (Lindsay et al., 2011):

$$pl_C(\theta) = \sum_{i=1}^{n}\sum_{j>i}^{n} 2l_{ij}(\theta) - (n-1)\sum_{i=1}^{n} l_i(\sigma^2,\beta).$$

When the marginal parameters σ^2 and β are known, then marginal and the conditional pairwise likelihood have the same efficiency. Otherwise it is not obvious which kind of estimation is more efficient.

A distinctive feature of pl_a, $a = M, C, D$, is that the associated estimating function, $\nabla pl_a(\theta)$, is unbiased, irrespective of the distributional assumptions on the pairs. In Appendix A we will show that the maximum CL estimators are consistent and asymptotically normal under increasing domain asymptotic framework. Note that the aforementioned results have been derived in a more general settings than those in Bevilacqua et al. (2012) for strictly increasing sequence on evenly-spaced lattices. In contrast here we do not impose any particular restrictions on the geometry and growth behavior of the lattice, allowing unevenly spaced locations. This framework is more suited for real data analysis as for instance the precipitation data in Section 5.

Under these results, again the inverse of the Godambe information matrix

$$G_a(\theta) = H_a(\theta)J_a(\theta)^{-1}H_a(\theta)^{\intercal}, \qquad a = M, C, D \tag{16}$$

is the asymptotic variance of the CL estimator with

$$H_a(\theta) = -\mathrm{E}[\nabla^2 pl_a(\theta)], \qquad J_a(\theta) = \mathrm{E}[\nabla pl_a(\theta)\nabla pl_a(\theta)^{\intercal}]. \tag{17}$$

In the Appendix B we can find closed form expressions for the Godambe information assuming, for notational simplicity, a constant mean function $E(Z(s)) = \mu$. Note that the J_a is a block diagonal matrix so that expressions when $E(Z(s)) = x(s)^{\intercal}\beta$, can be easily derived.

The role of the weights in CL function is to save computational time and improve the statistical efficiency. A compactly supported weight function, i.e. $w_{ij}(d) > 0$ if $\|s_i - s_j\| \leq d$, and 0 otherwise, has evident computational advantages. Moreover even a simple cut-off weight function, $w_{ij}(d) = 1$ if $\|s_i - s_j\| \leq d$, and 0 otherwise, can improve the efficiency as it has been shown in Joe and Lee (2009), Davis and Yau (2011) and Bevilacqua et al. (2012). The intuition behind this approach is that the correlations between pairs of distant sampling locations are often nearly zero. Therefore the use of the whole pairs may lose efficiency since too many redundant pairs of observations can skew the information confined in pairs of near observations. Hereafter we use $pl_a(\theta, d)$ to denote CL function based on pairs using simple cut-off weight function and $G_a(\theta, d)$ the associated Godambe information matrix.

The evaluation of the standard error requires consistent estimation of the inverse of the matrix $G_a(\theta)$. It can be obtained through the plug-in estimates $H_a(\widehat{\theta}_a)$ and $J_a(\widehat{\theta}_a)$ where $\widehat{\theta}_a$, $a = M, C, D$ is the maximizer of (13), (14) and (15) respectively. Nevertheless the latter becomes computationally unfeasible for large data sets since it is of order $O(n^4)$. In order to estimate $J_a(\widehat{\theta}_a)$ we use a subsampling method as described in Heagerty and Lumley (2000). Provided that $W^{-1}J_a(\widehat{\theta}_a)$ converges to a matrix J_a^* as $n \longrightarrow \infty$, where $W = \sum_{(i,j)} w_{ij}$, we use the subsampling method in order to obtain an estimate $\widehat{J}_a^*$ of J_a^* and then estimate $J_a(\widehat{\theta}_a)$ by $W\widehat{J}_a^*$. Given $S_1, \ldots, S_m$ subsets of the observation points $\{s_1, s_2, \ldots, s_n\}$, the estimator is

$$\widehat{J}_a^* = \frac{1}{m}\sum_{k=1}^{m}\frac{1}{W^{(k)}}\sum_{\substack{(i,j)\in S_k\\(i',j')\in S_k}}[\nabla pl_a(\widehat{\theta}_a)]_{ij}[\nabla pl_a(\widehat{\theta}_a)^{\intercal}]_{i'j'}w_{ij}w_{i'j'}, \tag{18}$$

where $W^{(k)} = \sum_{(i,j)\in S_k} w_{ij}$. The subsets are derived gathering the points that fall in a collection of overlapping sub-regions of the same shape of the region of observations but of smaller volume (Lee and Lahiri, 2002). Finally, the asymptotic covariance matrix of $\widehat{\theta}_a$

can then be estimated using the subsampling approximation

$$\widehat{G_a^{-1}}(\widehat{\theta}_a) = W H_a^{-1}(\widehat{\theta}_a) \widehat{J}_a^* H_a^{-1}(\widehat{\theta}_a) \tag{19}$$

and standard error estimation of each parameter is computed taking the square root of the diagonal elements of $\widehat{G_a^{-1}}(\widehat{\theta}_a)$.

Computational reasons can drive the choice of the number of subsets m. Suppose the n observations are divided into subsets of roughly q observations per subset, so that $n \simeq mq$. Evaluation of H_a and $\widehat{J}_a^*$ are of order $O(n^2) = O(m^2q^2)$ and $O(n^4/m^3) = O(mq^4)$, respectively. So, if m grows at a rate $O(n^{2/3})$, both evaluations have the same computational order of the composite likelihood.

4 Numerical examples

We have considered three models:

1. the exponential covariance function

$$C(h;\theta) = \sigma^2 \exp(-3\|h\|/\phi) \tag{20}$$

2. the Cauchy covariance function

$$C(h;\theta) = \frac{\sigma^2}{1 + (\sqrt{19}\|h\|/\phi)^2} \tag{21}$$

3. the wave or cardinal sine covariance function

$$C(h;\theta) = \sigma^2(20.371\,\|h\|/\phi)^{-1} \sin(20.371\,\|h\|/\phi). \tag{22}$$

Figure 1 illustrates the behavior of the covariance function and the tapered correlation functions using the Bohman function (3), with $\sigma^2 = 1$ and $d = \phi = 0.1$. The covariance models (20), (21) and (22) are parametrized in terms of practical range that is the correlation is lower than 0.05 when $\|h\| \geq \phi$. The aforementioned models cover a wide spectrum of situations that can arise in geostatistics. The first model probably is the most commonly used model in geostatistics and it is a special case of the Matérn model when $\nu = 1/2$.

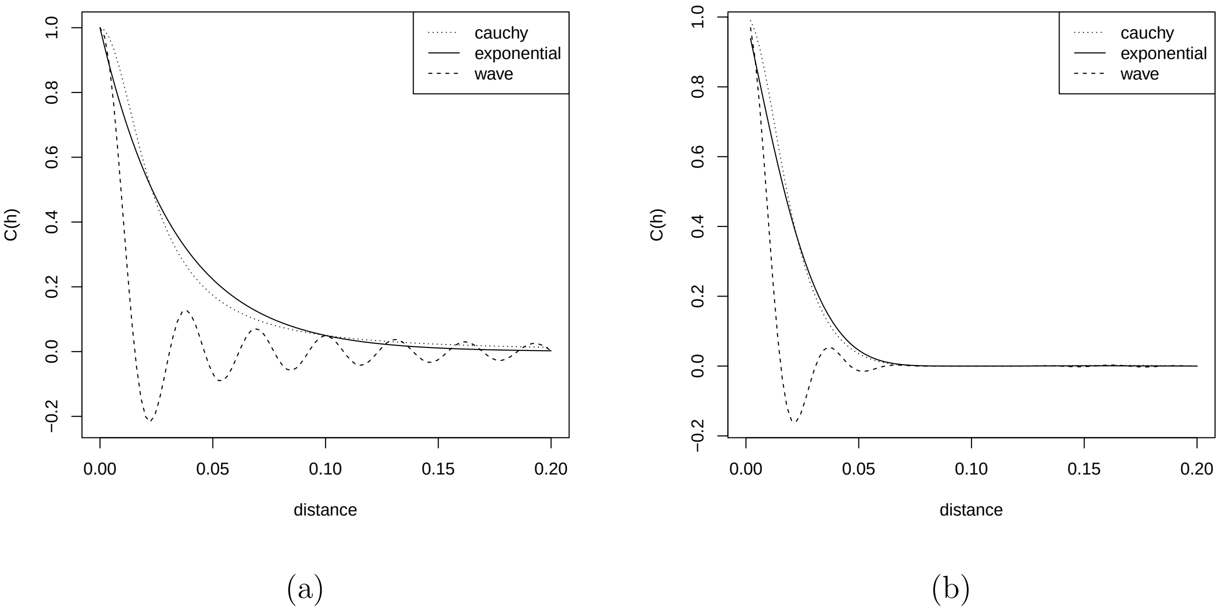


Figure 1: (a) Covariance functions with equivalent practical range, where $\sigma^2 = 1$, $\phi = 0.1$, (b) the tapered covariance functions using (3)), with $d = \phi$.

Model (21) is polynomially decreasing and hence more suitable than the exponential model for modeling of a slowly decaying covariance. Model (22) allows for negative correlations which is used, for instance, in meteorology, with high and low pressure zones.

First of all we compare the computational time required for one evaluation of the likelihood (1), the tapered likelihood (5), the weighted marginal pairwise likelihood (13) and its unweighted version using the exponential covariance function with $\phi = 0.1$. As taper function we consider the Bohman function (3).

For evaluating (1) and (5) we follow the implementation of Kaufman et al. (2008) (available at `www.image.ucar.edu/Data/precip_tapering/`) and we use the sparse matrix implementation in the R package `spam` (Furrer and Sain, 2010). As an anonymous referee suggested, the `spam` package allows users to separate structural and numerical computations needed for Cholesky factorization. The result is that for a given sparsity structure, the full factorization needs only to be done once. In subsequent factorizations, one can pass in the structure and have spam only compute the numerical part. This can save a lot of time when the tapered likelihood function is evaluated repeatedly. In the sequel the sparsity structure is given, and we have recorded the total time for computing the Cholesky factor, the log determinant of the covariance matrix and the quadratic form in (5). For the implementation of the $pl_M(.,d)$, the vectors of the distances $\|s_i - s_j\|$, $i \neq j$ and the statistics for l_{ij} are calculated using C code, then the evaluation of the pairwise likelihood is made by R code. Only this evaluation has been considered in the experiment, making the comparison appropriate with the code for the likelihood and its tapered version.

For the data locations we have followed Kaufman et al. (2008). We have considered a regular grid with increments 0.03 over W_k where $W_k = [0, 2^{k/2}] \times [0, 2^{k/2}]$, $k = 0, \ldots, 5$. The grid points have been perturbed adding a uniform random value on $[-0.01, 0.01]$ to each coordinates and, finally, we have randomly chosen $n_k = 500 \cdot 2^k$ points without replacement. As taper range and cut-off distance for the weighted composite likelihood estimator we have set $d = \phi$, i.e. the practical range. Because we keep d as fixed, increasing k and consequently the number n of observations, the fraction of nonzero elements in the resulting tapered covariance matrix decreases. Finally, in carrying out the experiment, we have used a 2.4 GHz processor with 16 GB of memory and the reported time statistics are means over 10 evaluations of each function.

Table 1 depicts the saving in terms of computational burden for large datasets for the marginal pairwise likelihood estimates. In this setup the tapering method takes advantage of the sparsity of the matrix for overtaking the unweighted version of the CL. However the saving is quite remarkable when we consider the weighted version of the CL.

Now we compare the asymptotic relative efficiency (ARE) of the estimates $\widehat{\sigma}^2$, $\widehat{\phi}$ and $\widehat{\mu}$ under the covariance models (20), (21), and (22). We have considered the case $k = 0$, that is 500 locations over $[0, 1] \times [0, 1]$, and a sequence of increasing values of the taper range d, corresponding to increasing percentages 0.1%, 0.2%, ..., 2% of non zero values in the tapered covariance matrix. As practical range for the models we have chosen $\phi = 0.1$ because this value of ϕ over this spatial domain is consistent to increasing domain framework.

As overall measure of the ARE for the multi-parameter case we consider:

$$ARE_a(d) = \left(\frac{|G_a(\theta; d)|}{|I_{ML}(\theta)|} \right)^{1/p}, \quad a = C, M, T \tag{23}$$

n	$l(\theta)$	$l_T(\theta, d)$	$pl_M(\theta, \infty)$	$pl_M(\theta, d)$	%
500	0.201	0.036	0.048	0.001	0.00436
1000	0.884	0.127	0.266	0.002	0.00221
2000	3.809	0.470	1.076	0.002	0.00113
4000	20.616	1.722	4.050	0.002	0.00056
8000	154.031	7.271	14.067	0.002	0.00029
16000	1624.439	29.524	56.065	0.004	0.00014

Table 1: Time in seconds for evaluating $l(\theta)$, $l_T(\theta, d)$, $pl_M(\theta, \infty)$ and $pl_M(\theta, d)$ functions with $d = 0.1$ under increasing domain setup. The column (%) indicates the associated percentages of non zero values in the tapered covariance matrix.

where $I_{ML}(\theta)$ is the Fisher information matrix (8) and $p = 3$ is the number of unknown components in θ. Note that in the overall measure, the case $a = D$ is not considered since composite likelihood based on differences does not involve the mean estimation when the mean is constant. Beyond the overall measure we have evaluated the relative efficiences of each single parameter for the cases $a = C, D, M, T$.

In Figures 2, 3, 4 we depict the ARE of the estimates as a function of the percentages of non zero values obtained in varying the value d. As general remark for the tapering method the asymptotic relative efficiency is a monotonic increasing function of the percentages of non zero values as expected. The tapering method is more efficient in estimating the marginal parameters μ and σ^2. On the other hand, for small percentages of non zero values, where the maximum tapered likelihood estimates takes advantage from the sparsity of the covariance matrix, the maximum marginal and conditional pairwise likelihood estimates outperform the maximum tapered likelihood and the maximum difference pairwise likelihood estimates. This is true for the overall measure (23) and this performance is owed to the gain in estimating more efficiently the practical range ϕ. Moreover there is no practical difference in considering marginal and conditional likelihood estimates, so that a preference should be given to the first one because it requires less computation. Finally we note that the estimates based on the marginal or conditional pairwise likelihood are better of the ones based on the difference pairwise likelihood. Note also that asymptotic efficiency of the maximum CL estimates is not a increasing function of the distance considered in the weight function with the exception of the wave model. These examples suggest that a proper choice of the distance d can improve significantly the statistical efficiency of maximum CL estimates under specific models. Our findings add more evidence to similar results reported in the literature (Joe and Lee, 2009; Davis and Yau, 2011; Bevilacqua et al., 2012). Furthermore such distance, i.e. the number of pairs, in the marginal and conditional pairwise CL should be different with respect to the distance of the difference CL.

Looking at the behavior for the different models, we see that the the maximum tapered likelihood estimate performs reasonably well under the exponential and the Cauchy model, but probably a larger taper range is required for outperforming the maximum pairwise likelihood estimates, vanishing the computational advantage of the sparsity of the covariance matrix.

Finally we have simulated $1,000$ random samples drawn from a Gaussian random fields under the same spatial configuration as before but with different values of the range parameter, namely $\phi = 0.05, 0.1, 0.2$. All the estimates have been carried out using the version 1.0.3 of R package `CompRandFld` (Padoan and Bevilacqua, 2013), avalaible on CRAN (`http://cran.r-project.org/`). This package offer a full implementation of all the estimation methods described here, including the evaluation of the standard errors of the estimates.

The numerical results collected in Tables 2, 3, 4, are consistent with the theoretical results. Each estimation methods appear unbiased and the increment of the spatial dependence, i.e. increasing the taper range, leads to an increment of the variability of the estimates. As general comment the tapering approach outperforms the CL methods in estimating the marginal parameters, μ and σ^2, of the random field. This dominance is not attained when we consider the estimation of the range parameter. In particular, when we consider small spatial dependence ($\phi = 0.05$), pl_C and pl_M provide a better efficiency than the tapering approach in the estimates of ϕ (see Table 2). Moreover the tapering approach is less efficient of the marginal and conditional pairwise likelihood when we consider the Wave model. Finally we remark that the pl_d yields to estimates with large variability except

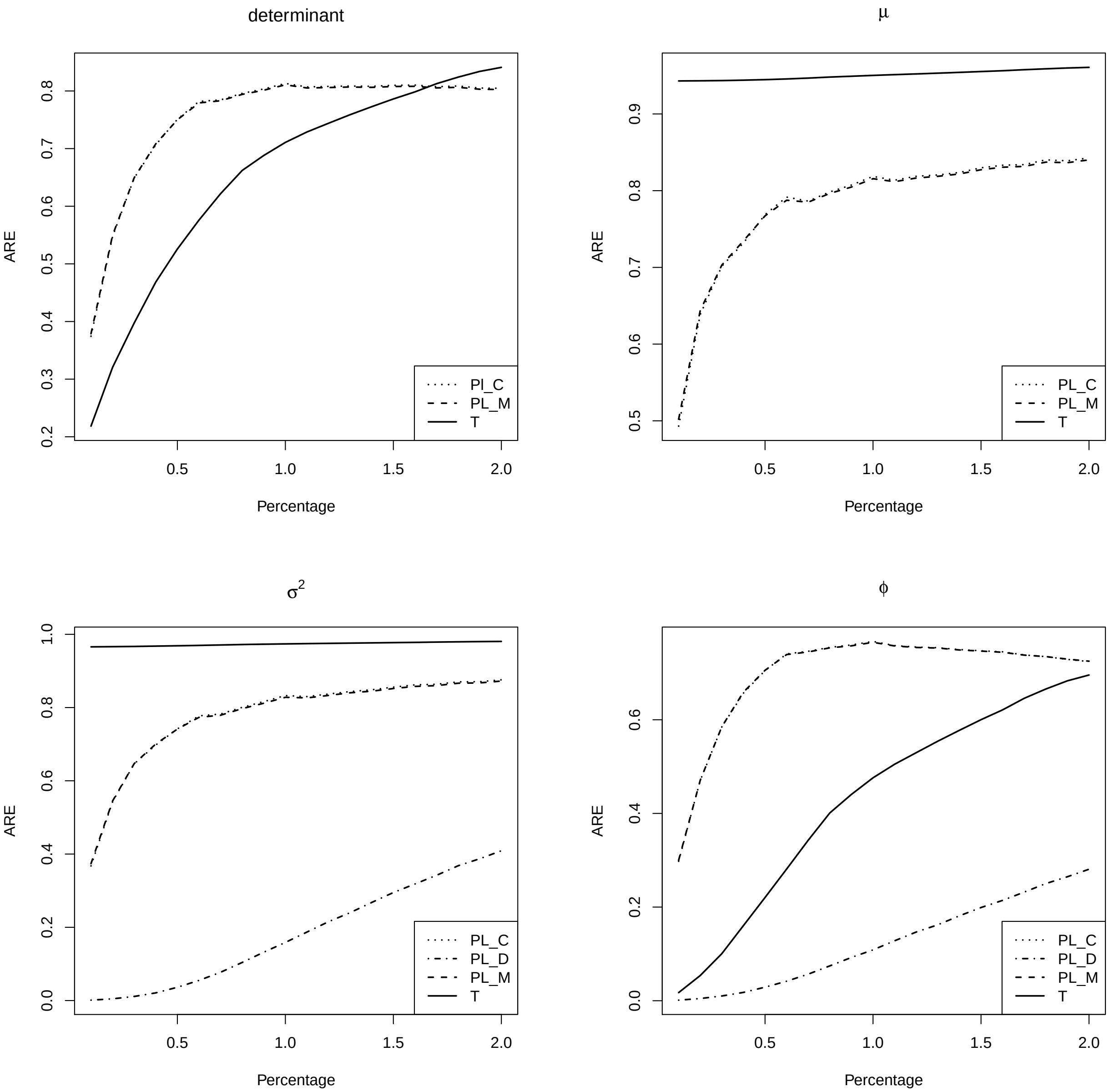


Figure 2: AREs of the tapering and CL estimators with respect of the percentages of non zeros values in the tapering matrix, for the exponential model (20) with $\phi = 0.1$, $\sigma^2 = 1$.

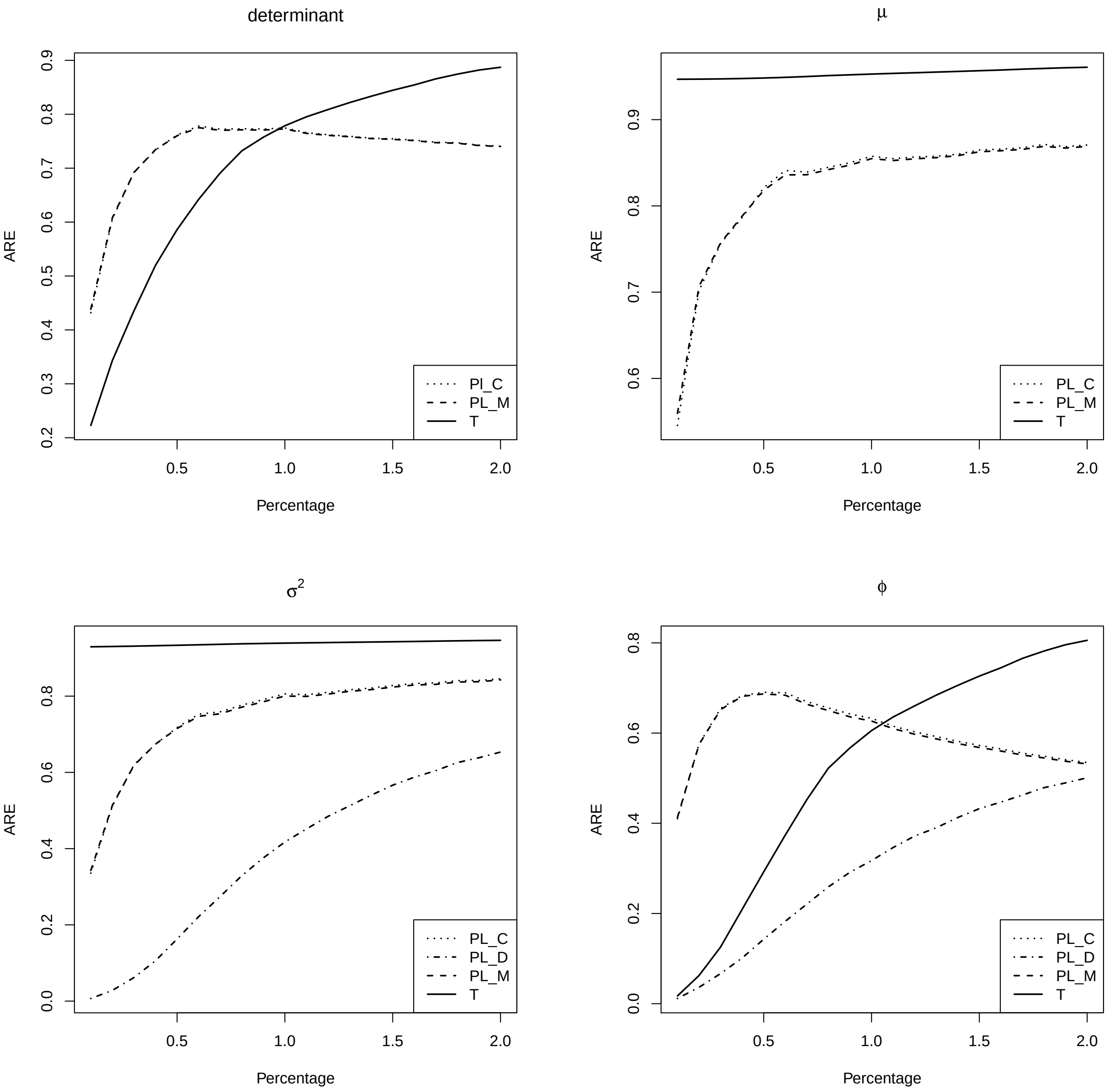


Figure 3: AREs of the tapering and CL estimators with respect of the percentages of non zeros values in the tapering matrix, for the Cauchy model (21) with $\phi = 0.1$, $\sigma^2 = 1$.

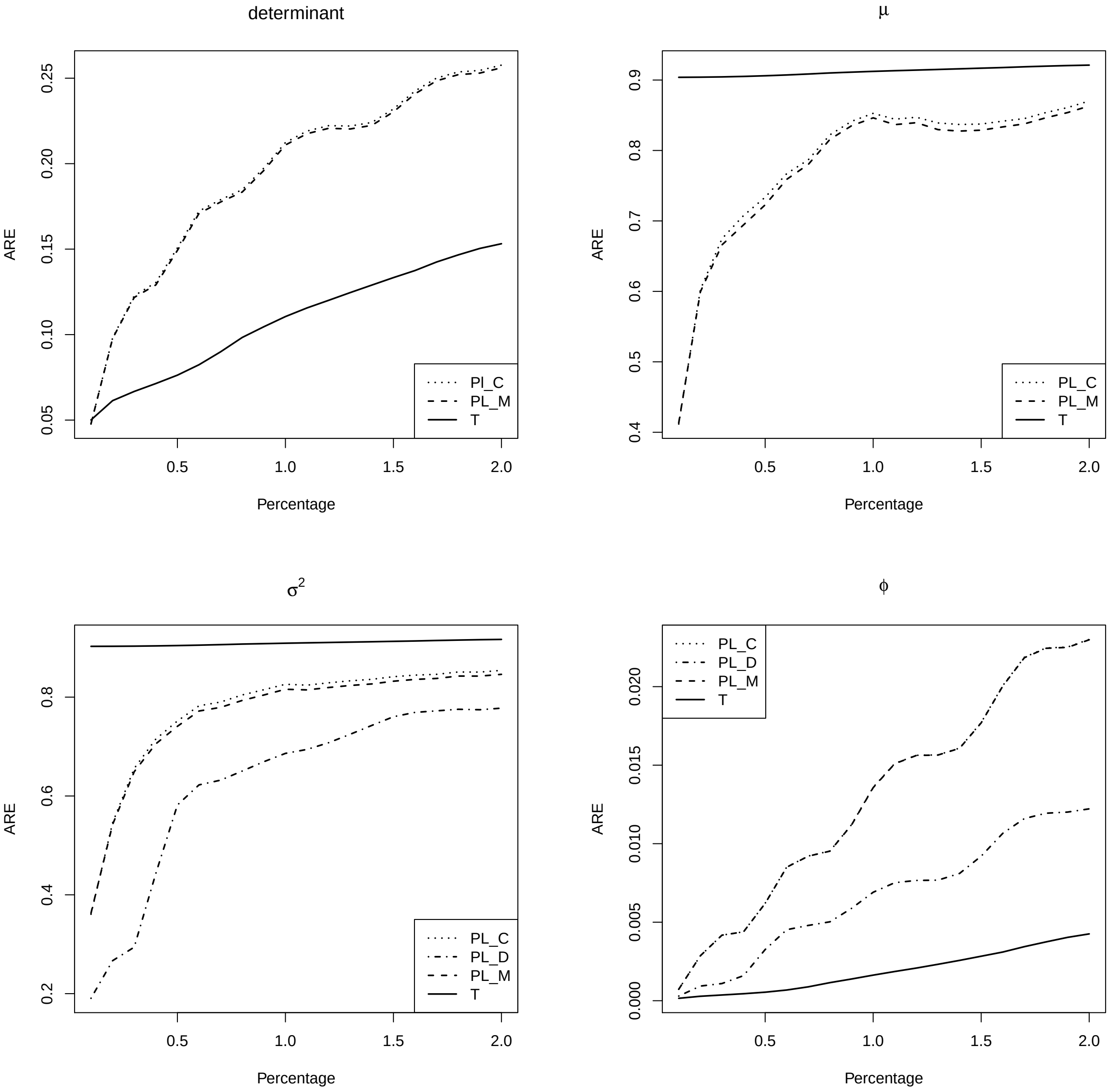


Figure 4: AREs of the tapering and CL estimators with respect of the percentages of non zeros values in the tapering matrix, for the wave model (22) with $\phi = 0.1$, $\sigma^2 = 1$.

in the case of $\phi = 0.2$ for the Cauchy and Wave models (see Table 4).

		Exponential model			Cauchy model			Wave model		
		μ	ϕ	σ^2	μ	ϕ	σ^2	μ	ϕ	σ^2
ML	bias	0.0017	-0.0009	-0.0023	0.0022	-0.0016	-0.0021	0.0011	-0.0017	-0.0028
	rmse	0.0552	0.0075	0.0672	0.0643	0.0098	0.0669	0.0443	0.0091	0.0651
TAP	bias	0.0016	0.0002	-0.0020	0.0023	-0.0008	-0.0030	0.0007	0.0199	0.0002
	rmse	0.0553	0.0138	0.0674	0.0647	0.0143	0.0668	0.0444	0.0425	0.0652
pl_C	bias	0.0008	-0.0010	-0.0022	0.0015	-0.0017	-0.0030	0.0002	0.0000	-0.0010
	rmse	0.0614	0.0078	0.0725	0.0693	0.0099	0.0719	0.0477	0.0165	0.0701
pl_M	bias	0.0008	-0.0010	-0.0021	0.0015	-0.0017	-0.0029	0.0001	-0.0006	-0.0001
	rmse	0.0613	0.0078	0.0726	0.0693	0.0099	0.0719	0.0478	0.0178	0.0702
pl_D	bias		0.0054	0.0571		-0.0010	0.0149		0.0065	-0.0007
	rmse		0.0310	0.2330		0.0201	0.1189		0.0263	0.0787

Table 2: Bias and root mean square error (rmse) of the estimates when $\mu = 0$, $\phi = 0.05$, $\sigma^2 = 1$.

5 A real data example

As data example we consider the data-set in Kaufman et al. (2008) that can be retrieved from `www.image.ucar.edu/Data/precip_tapering/`. We consider yearly total precipitation anomalies registered at $7,352$ location sites in the USA from 1895 to 1997.

The yearly totals have been standardized by the long-run mean and standard deviation for each station from 1962. The data-set can be considered of medium size allowing ML estimation although it is very slow to compute.

Kaufman et al. (2008) adapted a zero mean Gaussian random field with an exponential covariance model using the maximum likelihood and the tapering method. Here we choose an exponential covariance model plus a nugget effect , i.e.

$$C(h;\theta) = \tau^2 I(||h|| = 0) + \sigma^2 \exp\{-||h||/\phi\}, \qquad (24)$$

		Exponential model			Cauchy model			Wave model		
		μ	ϕ	σ^2	μ	ϕ	σ^2	μ	ϕ	σ^2
ML	bias	0.0037	-0.0015	-0.0067	0.0044	-0.0010	-0.0041	0.0012	-0.0001	-0.0022
	rmse	0.0843	0.0114	0.0792	0.0973	0.0089	0.0758	0.0409	0.0027	0.0648
TAP	bias	0.0039	-0.0013	-0.0074	0.0048	-0.0015	-0.0096	0.0011	-0.0026	-0.0022
	rmse	0.0851	0.0129	0.0791	0.0989	0.0094	0.0767	0.0420	0.0155	0.0675
pl_C	bias	0.0041	-0.0028	-0.0089	0.0050	-0.0034	-0.0110	0.0013	-0.0010	-0.0020
	rmse	0.0900	0.0133	0.0826	0.1033	0.0120	0.0800	0.0433	0.0086	0.0691
pl_M	bias	0.0041	-0.0028	-0.0087	0.0049	-0.0034	-0.0109	0.0013	-0.0005	-0.0017
	rmse	0.0901	0.0133	0.0827	0.1033	0.0120	0.0800	0.0435	0.0069	0.0694
pl_D	bias		0.0011	0.0088		-0.0004	0.0020		-0.0044	-0.0001
	rmse		0.0201	0.1142		0.0120	0.0891		0.0167	0.0717

Table 3: Bias and root mean square errors (rmse) of the estimates when $\mu = 0$, $\phi = 0.1$, $\sigma^2 = 1$.

as suggested by inspecting the empirical semi-variogram in Figure 5.

The parameter $\theta = (\tau^2, \sigma^2, \phi)^\intercal$ is estimated with maximum likelihood, tapered likelihood and $pl_a(\theta; d)$, $a = C, D, M$ methods. The distance between two sites are measured using the great-circle distance and the exponential covariance function is still positive definite for this distance (Huang et al., 2011). As taper function we use the Bohman taper with $d = 112.654$ Km, as in Kaufman et al. (2008). This leads to 0.0063% of non zero values in the tapered covariance matrix. The same value d has been adopted for the weighted version of the pairwise likelihood. However the estimates that we obtained using the pairwise likelihood based on difference using d were unrealistic. Note that the difference pairwise likelihood estimates can be calculated by nonlinear weighted least squares in the model

$$(Z(s_i) - Z(s_j))^2 = 2\gamma(s_i - s_j; \theta) + \epsilon_{i,j}, \qquad \epsilon_{i,j} \sim \mathcal{N}(0, 8\gamma(s_i - s_j; \theta)^2)$$

where $\gamma(s_i - s_j; \theta)$ is the semi-variogram model. Since pl_D is basically based on semi-variogram the selection of the distance d requires some care after considering Figure 5. The distance $d = 112.654$ Km seems too limiting for catching the actual behavior of the variogram so we fixed a different distance, namely $d = 3 \times 112.654$ in $pl_D(\theta, d)$.

		Exponential model			Cauchy model			Wave model		
		μ	ϕ	σ^2	μ	ϕ	σ^2	μ	ϕ	σ^2
ML	bias	0.0060	-0.0046	-0.0170	0.0069	-0.0009	-0.0052	0.0005	0.0000	-0.0026
	rmse	0.1474	0.0283	0.1178	0.1591	0.0104	0.0972	0.0375	0.0009	0.0647
TAP	bias	0.0073	-0.0058	-0.0219	0.0082	-0.0031	-0.0223	0.0010	0.0001	-0.0008
	rmse	0.1514	0.0286	0.1183	0.1632	0.0118	0.1057	0.0408	0.0048	0.0727
pl_C	bias	0.0077	-0.0139	-0.0270	0.0085	-0.0124	-0.0298	0.0014	0.0002	-0.0010
	rmse	0.1624	0.0391	0.1251	0.1731	0.0297	0.1195	0.0442	0.0029	0.0780
pl_M	bias	0.0077	-0.0138	-0.0264	0.0084	-0.0124	-0.0296	0.0014	0.0002	-0.0010
	rmse	0.1621	0.0390	0.1250	0.1728	0.0297	0.1195	0.0444	0.0029	0.0783
pl_D	bias		0.0031	0.0125		-0.0004	0.0020		0.0001	0.0011
	rmse		0.0420	0.1587		0.0193	0.1206		0.0037	0.0789

Table 4: Bias and root mean square errors (rmse) of the estimates when $\mu = 0$, $\phi = 0.2$, $\sigma^2 = 1$.

Table 5 reports the estimates of maximum likelihood, tapering and pl_a, $a = C, D, M$ methods and the associate standard errors. For maximum likelihood and tapering methods standard errors are computed using the square root of the diagonal elements of the inverse of the Fisher and Godambe information matrices in (8) and (6). The Godambe information matrix for the composite likelihood methods is estimated using the subsampling method as explained in Section 3 with overlapping rectangular subregions of length and width respectively 76.15 Km and 52.17 Km. The number of sub-regions involved in the subsampling estimation is 116. Note that the standard errors of the maximum likelihood estimates are not necessarily smaller than the standard errors of the other estimation methods. What we expect is that the difference between the Fisher information matrix and the Godambe information matrix of the other estimation methods is a non-negative definite matrix.

Figure 5 gives a rough evaluation of the goodness of fit of the proposed model and shows that the estimates with different estimation methods look similar enough.

We gain a better insight if we compare the prediction performance. In doing this, we have used a leave-one-out cross-validation, i.e. we have set aside one observation and we have predicted it using the other observations. As overall criteria we have considered three

	ML	$TAP(d)$	$pl_C(d)$	$pl_M(d)$	$pl_D(d)$
τ^2	0.1033 (0.0042)	0.06381 (0.0091)	0.1069 (0.0072)	0.1070 (0.0024)	0.1144 (0.02842)
ϕ	168.1174 (12.2329)	121.36855 (9.6355)	186.2457 (18.7000)	185.7594 (18.9975)	161.6937 (27.3767)
σ^2	0.6693 (0.0632)	0.7418 (0.0449)	0.5890 (0.0251)	0.5866 (0.0616)	0.4909 (0.03658)

Table 5: Maximum likelihood, tapering and composite likelihood estimates for the exponential covariance model with nugget effect (estimated standard errors are reported between parentheses).

predictive scores described in Gneiting and Raftery (2007) and Zhang and Wang (2010).

Let $\hat{Z}(s_i)$ be the best linear prediction on the location site s_i based on all data except $Z(s_i)$, i.e. its kriging prediction (Cressie, 1993). We will consider

1. the root mean square error (RMSE)

$$\text{RMSE} = \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ Z(s_i) - \hat{Z}(s_i) \right\}^2 \right]^{1/2}, \tag{25}$$

2. the logarithmic score (LSCORE)

$$\log S = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{2} \log\{2\pi\sigma(s_i)\} + \frac{1}{2}\{Y(s_i)\}^2 \right], \tag{26}$$

where $Y(s_i) = (Z(s_i) - \hat{Z}(s_i))/\sigma(s_i)$ and $\{\sigma(s_i)\}^2$ is the prediction variance associated with $\hat{Z}(s_i)$,

3. the continuous ranked probability score (CPRS)

$$\text{CPRS} = \frac{1}{n} \sum_{i=1}^{n} \sigma(s_i) \left(Y(s_i) \left(2F(Y(s_i)) - 1\right) + 2F(Y(s_i)) - \frac{1}{\sqrt{\pi}} \right), \tag{27}$$

where F is the cumulative distribution of the Gaussian distribution.

The prediction scores required the computation of $\hat{Z}(s_i)$ and its associated variance $\{\sigma(s_i)\}^2$ that can be evaluated one by one. Zhang and Wang (2010) describe how to

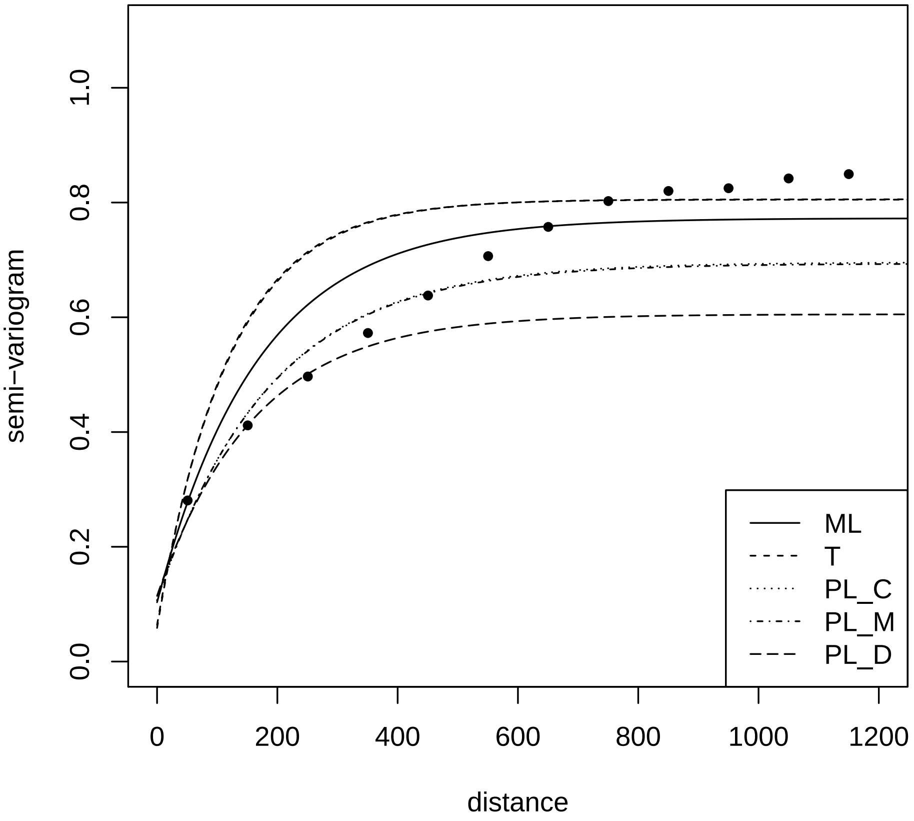


Figure 5: The empirical semi-variogram for yearly total precipitation anomalies. The lines represent the fitted semi-variograms.

compute them in one step. For instance, assuming a zero mean function mean, RMSE can be evaluated as

$$RMSE = \left(\frac{Z^{\intercal}\Sigma(\psi)^{-1}D^{-1}D^{-1}\Sigma(\psi)^{-1}Z}{n}\right)^{1/2}$$

where $D = diag(\Sigma(\psi)^{-1})$. Since $\Sigma(\psi)$ is unknown, we can use a plug-in estimate of Σ, $\Sigma(\hat{\psi})$.

In Table 6, 7 and 8 we report RMSE, LSCORE and CRPS for the exponential model and we contrast them with an exponential model without nugget effect, as proposed by Kaufman et al. (2008). Our findings highlight how we have an effective improvement in the RMSE and LSCORE criteria when we consider an additional nugget effect. Moreover pl_a,

Model	ML	$TAP(d)$	$pl_C(d)$	$pl_M(d)$	$pl_D(d)$
Exponential with nugget	0.467	0.470	0.469	0.467	0.467
Exponential without nugget	0.479	0.479	0.481	0.481	0.480

Table 6: Prediction performance in terms of MSPE for exponential covariance model with and without nugget effect estimated with maximum likelihood, tapering and composite likelihood methods.

Model	ML	$TAP(d)$	$pl_C(d)$	$pl_M(d)$	$pl_D(d)$
Exponential with nugget	0.638	0.639	0.642	0.642	0.642
Exponential without nugget	0.677	0.670	0.868	0.869	0.834

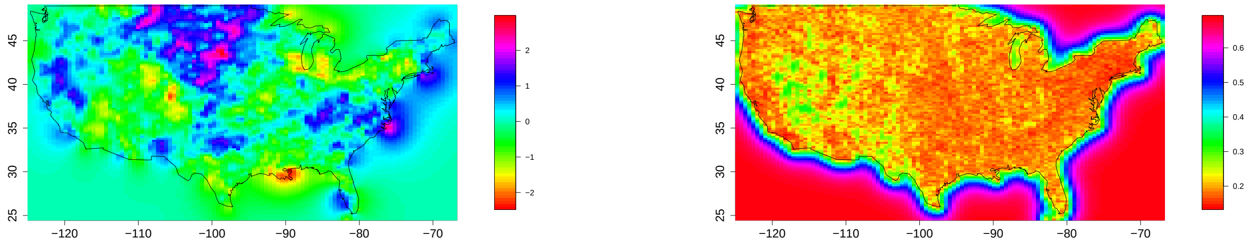
Table 7: Prediction performance in terms of LSCORE for exponential covariance model with and without nugget effect estimated with maximum likelihood, tapering and composite likelihood methods.

$a = C, M, D$ estimates provides comparable results with respect to the tapering method.

Finally Figure 6 shows prediction map of the precipitation anomalies over USA and the associate map of standard error prediction where the covariance matrix has been estimated using the pl_M estimates. It can be appreciated that precipitation anomalies are mainly concentrated in the north of the country and in the east/west coast. Note that, as expected, standard errors tend to be higher in the west of the country where there are few location sites.

Model	ML	$TAP(d)$	$pl_C(d)$	$pl_M(d)$	$pl_D(d)$
Exponential with nugget	0.446	0.447	0.444	0.444	0.445
Exponential without nugget	0.456	0.460	0.440	0.439	0.452

Table 8: Prediction performance in terms of CRPS for exponential covariance model with and without nugget effect estimated with maximum likelihood, tapering and composite likelihood methods.



(a)

(b)

Figure 6: (a) Prediction map of the precipitation anomalies data; (b) the associated standard error prediction map.

6 Concluding remarks

The class of CL functions is very large and for a given estimation problem it is not clear how to choose in this class. In the Gaussian case, if the choice of the CL is driven by computational concerns then the CL based on pairs have clear computational advantages with respect to other type of CL.

In this paper through theoretical and numerical examples we have compared three versions of the weighted pairwise likelihood (marginal, conditional and difference), using the covariance tapering approach as benchmark. All approaches rely on the choice of a distance d. As pointed out by an anonymous referee, the distance does not play the same role in the considered approaches. In the tapering approach we pretend that pairs of observations that are far apart a certain distance d are independent. The role of the distance d in the weight function of the CL is different since it allows to keep out the (marginal, conditional or difference) likelihood of pairs of observations.

One advantage of the tapering approach is that the balance between the statistical and computational efficiency is clear. Instead for the CL approach the gain in statistical efficiency is less undimmed when we increase the distance and preliminary evaluation of this gain as in Bevilacqua et al. (2012) could be computationally hard in particular for large data-set.

In this paper the theoretical and numerical examples highlight a better performance of the weighted version of the conditional and marginal pairwise likelihood with respect to the one of the difference pairwise likelihood. Moreover the weighted marginal pairwise likelihoods are computationally preferable with respect to the tapering approach while the tapering approach shows better statistical efficiency when increasing the taper range. Our suggestion for the practitioners is to consider both the approaches when they are computationally feasible, as in the real data example. For data sets of large dimension the pairwise likelihood approach is preferable since a little loss of statistical efficiency is offset by good computational performances. Our findings are consistent with those of Stein (2013) who compares the covariance tapering with a specific type of composite likelihood based on independent blocks.

Acknowledgement

Research work of Moreno Bevilacqua was partially supported by grant FONDECYT 11121408 from the Chilean government. We thank the anonymous referees and the Associate Editor for their constructive comments that have improved a lot the first version of this work.

Appendix A

Asymptotic results can be been proved for spatial processes which are observed at finitely many locations in the sampling region. In this case we deal with an increasing domain setup where the sampling region is unbounded. In the sequel we suppose that the mean function for the random field is known and without loss of generality this is zero. More precisely, we consider a weakly dependent random field $\{Z(s), s \in S\}$ defined over an arbitrary lattice S in $\mathbb{R}^d$ that is not necessarily regular. The lattice S is equipped with the metric $\delta(s_k, s_l) = \max_{1\leq l\leq d} |s_{i,l} - s_{j,l}|$ and the distance between any subsets $A, B \subset S$ is defined as $\delta(A, B) = \inf\{\delta(s_k, s_l) : s_k \in A \text{ and } s_l \in B\}$. We denote

$$\alpha(U, V) = \sup_{A,B}\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}(U), B \in \mathcal{F}(V)\},$$

where $\mathcal{F}(E)$ is the σ-algebra generated by the random variables $\{Z(s),\, s \in E\}$. The α-mixing coefficient (Doukhan, 1994) for the random field $\{Z(s), s \in S\}$ is defined as

$$\alpha(a, b, m) = \sup_{U,V}\{\alpha(U, V),\, |U| < a,\, |V| < b,\, \delta(U, V) \geq m\}.$$

where $|C|$ is the cardinality of the set C.

We make the following assumptions:

C1: S is infinite, locally finite: for all $s \in S$ and $r > 0$, $|\mathcal{B}(s, r) \cap S| = O(r^d)$, with $\mathcal{B}(s, r)$ d-dimensional ball of center s and radius r; moreover there exists a set of neighbourhoods, $V_s \subset S$, such that $|V_s|$ is uniformly bounded.

C2: D_n is an increasing sequence of finite subsets of S: $d_n = |D_n| \to \infty$ as $n \to \infty$.

C3: Z is a Gaussian random field with covariance function $C(h; \theta)$, with $\theta \in \Theta$. Θ is a compact set of $\mathbb{R}^p$. The function $\theta \mapsto C(h; \theta)$ has continuous second order partial derivatives with respect to $\theta \in \Theta$, and these functions are continuous with respect to h and $\inf_{\theta\in\Theta} C(h; \theta) > 0$;

C4: The true unknown value of the parameter θ, namely θ^*, is an interior point of Θ.

C5: The Gaussian random field is α-mixing with mixing coefficient $\alpha(m) = \alpha(\infty, \infty, m)$ satisfying:

(C5a) $\exists\, \eta > 0$ s.t. $\sum_{k,l \in D_n} \alpha(\delta(k,l))^{\frac{\eta}{2+\eta}} = O(d_n)$,

(C5b) $\sum_{m \geq 0} m^{d-1} \alpha(m) < \infty$;

C6: Let

$$g_k(Y(k);\theta) = -(1/2) \sum_{l \in V_k, l \neq k} l_{(k,l)}(\theta)$$

with $Y(k) = (Z(s), s \in V_k)$, $l_{(k,l)} = l_{kl}$, $l_{k|l}$ or l_{k-l}. The functions l_{kl}, $l_{k|l}$, and l_{k-l} are defined as in (10),(11) and (12), respectively.

The composite likelihood is defined as

$$Q_n(\theta) = \frac{1}{d_n} \sum_{k \in D_n} g_k(Y(k);\theta); \tag{28}$$

and the composite likelihood estimator is given by

$$\widehat{\theta}_n = \mathrm{argmin}_{\theta \in \Theta}\, Q_n(\theta).$$

C7: the function $\overline{Q}_n(\theta) = \mathrm{E}_{\theta^*}[Q_n(\theta)]$ has a unique global minimum over Θ at θ^*, the true value.

Remarks

1. Assumptions C1-2 are quite general. For instance we can consider a rectangular lattice, as in Shaby and Ruppert (2012), $D_n \subset \Delta\mathbb{Z}^d$, for a fixed $\Delta > 0$, and $D_n \subset D_{n+1}$ for all n.

2. The $\alpha-$mixing assumption C5 are a bit hard to check in general. It is satisfied when we consider compactly supported correlation functions, like the taper functions (2) and (3). When we consider a rectangular lattice the condition is satisfied for a stationary Gaussian random field with correlation function $C(h;\theta) = O(\|h\|^{-c})$, for some $c > d$ and its spectral density bounded below (Doukhan, 1994, Corollary 2, p. 59). In our examples this condition is satisfied by the exponential model.

3. The assumption C7 is an identifiability condition. For each s, the function $\mathrm{E}_{\theta^*}[g_s(Y_s;\theta)]$ has a global minimum at θ^* according the Kullback-Leibler inequality but in the multidimensional case ($p > 1$) θ^* fails, in general, to be the unique minimizer. Assumption

$C7$ is not satisfied, for instance, when we consider a rectangular lattice and a compactly supported correlation functions, see (2) and (3), for instance, and $\Delta > d$. In the case of the covariance models (20), (21) and (22) the condition is clearly satisfied.

4. The assumption C6 is satisfied if we suppose a cut-off weight function for w_{kl}.

5. Any individual log-likelihood $l_{(i,j)}$ can be written as

$$l_{(k,l)} = c_1(\theta, k-l) + c_2(\theta, k-l)Z_k^2 + c_3(\theta, k-l)Z_k^2 + c_4(\theta, k-l)Z_kZ_l,$$

where the functions c_i, $i = 1, \ldots, 4$ are $\mathcal{C}^2$ functions with respect to θ.

Consistency

Given the previous assumptions C1-C7, $\widehat{\theta}_n$ is a consistent estimator for θ_0 provided that $\sup_{\theta\in\Theta} |Q_n(\theta) - \overline{Q}_n(\theta)| \to 0$ in probability, as $n \to \infty$. According Corollary 2.2 in Newey (1991), we have to prove that

1: for each $\theta \in \Theta$, $Q_n(\theta) - \overline{Q}_n(\theta) \to 0$ in probability, as $n \to \infty$;

2: for $M_n = O_p(1)$,

$$|\overline{Q}_n(\theta') - \overline{Q}_n(\theta)| \leq M_n\|\theta' - \theta\|.$$

We sketch the proof for $l_{(k,l)} = l_{kl}$, the same arguments apply for the other sub-likelihoods, using the fourth remark.

1: We prove that $\sup_{k\in D_n} \mathrm{E}[(\sup_{\theta\in\Theta} g_k(Y(k);\theta))^{2+\eta}] < \infty$, for $\eta > 0$. In fact, we have

$$\begin{aligned} g_k(Y(k);\theta) &= \frac{1}{2}\sum_{l\in V_k, l\neq k} \left\{ 2\log\sigma^2 + \log(1-\rho_{kl}^2) + \frac{Z_k^2 + Z_l^2 - 2\rho_{kl}Z_kZ_l}{\sigma^2(1-\rho_{kl}^2)} \right\} \\ &\leq \sum_{l\in V_k, l\neq k} \log\sigma^2 + \frac{1}{2}\log(1-\rho_{kl}^2) + \frac{Z_k^2 + Z_l^2}{\sigma^2(1-\rho_{kl}^2)} \\ &\leq c_1|V_k|\log\sigma^2 + c_2|V_k|Z_k^2 + c_2\sum_{l\in V_k, l\neq k} Z_l^2 \end{aligned}$$

and $|V_k|$ is uniformly bounded according the assumption C6. The uniform bounded moments $g_k(Y(k);\theta)$ entail uniform L^1 integrability of g_k and with the assumption

C5 we obtain (Jenish and Prucha, 2009, Theorem 3)

$$Q_n(\theta) - \overline{Q}_n(\theta) = d_n^{-1} \sum_{k \in D_n} \{g_k(Y(k), \theta) - \mathrm{E}_\theta[g_k(Y(k), \theta)]\} \to 0, \text{ in probability}$$

2: We have

$$\begin{aligned} |g_k(Y(k); \theta') - g_k(Y(k); \theta)| &= \frac{1}{2} \sum_{l \in V_k, l \neq k} \Bigg| 2 \log \frac{\sigma^{'2}}{\sigma^2} + \log \frac{1 - \rho_{kl}^{'2}}{1 - \rho_{kl}^2} \\ &\quad + (Z_k^2 + Z_l^2) \left[\frac{1}{\sigma^{'2}(1 - \rho_{kl}^{'2})} - \frac{1}{\sigma^2(1 - \rho_{kl}^2)} \right] \\ &\quad - 2 Z_k Z_l \left[\frac{\rho_{kl}^{'}}{\sigma^{'2}(1 - \rho_{kl}^{'2})} - \frac{\rho_{kl}}{\sigma^2(1 - \rho_{kl}^2)} \right] \Bigg| \\ &\leq c_1 |V_k| \|\theta' - \theta\| + c_2(|V_k| Z_k^2 + \sum_{l \in V_k, l \neq k} Z_l^2) \|\theta' - \theta\| \end{aligned}$$

$$\begin{aligned} |\overline{Q}_n(\theta') - \overline{Q}_n(\theta)| &\leq d_n^{-1} \sum_{k \in D_n} |q_k(\theta') - q_k(\theta)| \\ &\leq c_3 d_n^{-1} \sum_{k \in D_n} (1 + Z_k^2 + \sum_{l \in V_k, l \neq k} Z_l^2) \|\theta' - \theta\| \\ &= M_n \|\theta' - \theta\| \end{aligned}$$

for some positive constants c_1, c_2 and c_3 and $M_n = c_3 d_n^{-1} \sum_{k \in D_n} (1 + Z_k^2 + \sum_{l \in V_k, l \neq k} Z_l^2)$.

Since $\mathrm{E}_\theta[M_n] < \infty$, we obtain the desired result.

Asymptotic normality

We make the additional assumption:

N1: there exists two symmetric nonnegative definite matrices H and J such that for large n:

$$J_n = \mathrm{var}_{\theta^*}(\sqrt{d_n} \nabla Q_n(\theta^*)) \geq J \quad \text{and} \quad H_n = \mathrm{E}_{\theta^*}(\nabla^2 Q_n(\theta^*)) \geq I.$$

where if A and B are two symmetric matrices, $A \geq B$ means that $A - B$ is a semipositive definite matrix.

We note that because g_s is a $\mathcal{C}^2$ and Θ is a compact space there exists a random variable $h(Y(s))$, $\mathrm{E}_\theta(h(Y(s))) < \infty$ satisfying:

$$\left| \frac{\partial^2}{\partial \theta_k \theta_l} g_s(Y(s), \theta) \right|^2 \leq h(Y(s)).$$

Moreover for all $s \in S$, $\mathrm{E}_\theta[\frac{\partial}{\partial \theta_k} g_s(\theta)] = 0$, because g_s is a sum of log-likelihoods, and it is easy to show that we have that $\sup_{s\in S,\, \theta\in\Theta} \mathrm{E}_\theta \left[\left| \frac{\partial}{\partial \theta_k} g_s(\theta) \right|^{2+\eta} \right] < \infty$ and $\sup_{s\in S,\, \theta\in\Theta} \mathrm{E}_\theta \left[\left| \frac{\partial^2}{\partial \theta_k \theta_l} g_s(\theta) \right|^{2+\eta} \right] < \infty$, for all $\eta > 0$.

Under the condition C1-C7 and N1, conditions (H1-H2-H3) of Theorem 3.4.5 in Guyon (1995) are satisfied and

$$\sqrt{d_n} J_n^{-1/2} H_n(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, I_p),$$

where I_p is the $p \times p$ identity matrix.

Appendix B

In this Section we provide the formulas for the Godambe information matrix associated to $pl_a(\theta)$, $a = M, C, D$ for a Gaussian random field $\{Z(s)\}$ with mean function $E(Z(s)) = \mu$, and covariance function $Cov(Z(s_i), Z(s_j)) = \sigma^2\rho(s_i - s_j; \phi) = \sigma^2\rho_{ij}$. We denote

$$H_a(\theta) = -\mathrm{E}[\nabla^2 pl_a(\theta)], \qquad J_a(\theta) = \mathrm{E}[\nabla pl_a(\theta)\nabla pl_a(\theta)^\intercal] \qquad a = M, C, D.$$

where θ is the vector of the unknown parameters.

The Godambe information matrix is given by

$$G_a(\theta) = H_a(\theta)J_a(\theta)^{-1}H_a(\theta)^\intercal, \qquad a = M, C, D.$$

Let B_{ij}, G_{ij} and U_{ij} be defined in section 3. Moreover we define

$$\begin{aligned} F_{ij} &= \rho_{ij}(Z(s_i) - \mu)^2 + \rho_{ij}(Z(s_j) - \mu)^2 - (Z(s_i) - \mu)(Z(s_j) - \mu)(1 + \rho_{ij}^2) \\ &= \rho_{ij}B_{ij} - (Z(s_i) - \mu)(Z(s_j) - \mu)(1 - \rho_{ij}^2) \\ Q_{ij} &= Z(s_i) + Z(s_j) \end{aligned}$$

For $pl_M(\theta)$ and $pl_C(\theta)$, $\theta = (\phi, \sigma^2, \mu)^\intercal$, the pairwise score functions are given by

$$\begin{aligned} \nabla pl_M(\theta) &= \sum_{i=1,j>i}^{n} w_{ij} \begin{bmatrix} \kappa_{ij}\frac{\rho_{ij}}{(1+\rho_{ij})}\left(1 - \frac{F_{ij}}{\sigma^2\rho_{ij}(1-\rho_{ij}^2)}\right) \\ -\frac{1}{\sigma^2}\left(1 - \frac{B_{ij}}{2\sigma^2(1-\rho_{ij}^2)}\right) \\ \frac{2\mu}{\sigma^2(1+\rho_{ij})}\left(1 - \frac{Q_{ij}}{2\mu}\right) \end{bmatrix} \\ \nabla pl_C(\theta) &= \sum_{i=1,j>i}^{n} w_{ij} \begin{bmatrix} 2\kappa_{ij}\frac{\rho_{ij}}{(1+\rho_{ij})}\left(1 - \frac{F_{ij}}{\sigma^2\rho_{ij}(1-\rho_{ij}^2)}\right) \\ -\frac{1}{\sigma^2}\left(1 - \frac{G_{ij}^2+G_{ji}^2}{2\sigma^2(1-\rho_{ij}^2)}\right) \\ \frac{2\mu(1-\rho_{ij})}{\sigma^2(1+\rho_{ij})}\left(1 - \frac{Q_{ij}}{2\mu}\right) \end{bmatrix} \end{aligned}$$

and for $pl_D(\theta)$, $\theta = (\phi, \sigma^2)^\intercal$, we have

$$\nabla pl_D(\theta) = \sum_{i=1,j>i}^{n} w_{ij} \begin{bmatrix} \kappa_{ij}\left(1 - \frac{U_{ij}^2}{2\gamma_{ij}}\right) \\ -\frac{1}{\sigma^2}\left(1 - \frac{U_{ij}^2}{2\gamma_{ij}}\right) \end{bmatrix}$$

where $\alpha_{ij} = (1+\rho_{ij})^{-1}\sqrt{1+\rho_{ij}^2}$, and $\kappa_{ij} = (1-\rho_{ij})^{-1}\nabla\rho_{ij}$. Moreover we have

$$
\begin{aligned}
H_M(\theta) &= \sum_{i=1,j>i}^{n} w_{ij} \begin{bmatrix} \alpha_{ij}^2\kappa_{ij}\kappa_{ij}^\intercal & -\frac{\rho_{ij}}{\sigma^2(1+\rho_{ij})}\kappa_{ij} & \mathbf{0} \\ - & \sigma^{-4} & 0 \\ - & - & \frac{2}{\sigma^2(\rho_{ij}+1)} \end{bmatrix} \\
J_M(\theta) &= \sum_{i,k=1,j>i,l>k}^{n} w_{ij}w_{lk} \begin{bmatrix} \alpha_{ij}\alpha_{kl}Cor(F_{ij},F_{kl})\kappa_{ij}\kappa_{kl}^\intercal & -\sigma^{-2}\alpha_{ij}Cor(F_{ij},B_{kl})\kappa_{ij} & \mathbf{0} \\ - & \sigma^{-4}Cor(B_{ij},B_{kl}) & 0 \\ - & - & \frac{2Cor(Q_{ij},Q_{kl})}{\sigma^2\sqrt{1+\rho_{kl}}\sqrt{1+\rho_{ij}}} \end{bmatrix} \\
H_C(\theta) &= \sum_{i=1,j>i}^{n} w_{ij} \begin{bmatrix} 2\alpha_{ij}^2\kappa_{ij}\kappa_{ij}^\intercal & -\frac{2\sigma^{-2}\rho_{ij}}{(1+\rho_{ij})}\kappa_{ij}^\intercal & \mathbf{0} \\ - & \sigma^{-4} & 0 \\ - & - & \frac{2(1-\rho_{ij})}{\sigma^2(1+\rho_{ij})} \end{bmatrix} \\
J_C(\theta) &= \sum_{i,k=1,j>i,l>k}^{n} w_{ij}w_{lk} \begin{bmatrix} 4\alpha_{ij}\alpha_{kl}Cor(F_{ij},F_{kl})\kappa_{ij}\kappa_{kl}^\intercal & -\sqrt{2}\sigma^{-2}\alpha_{ij}Cor(F_{ij},G_{kl}^2+G_{lk}^2)\kappa_{ij} & \mathbf{0} \\ - & 2^{-1}\sigma^{-4}Cor(G_{ij}^2+G_{ji}^2,G_{kl}^2+G_{lk}^2) & 0 \\ - & - & \frac{2Cor(Q_{ij},Q_{kl})(1-\rho_{ij})(1-\rho_{kl})}{\sigma^2\sqrt{1+\rho_{kl}}\sqrt{1+\rho_{ij}}} \end{bmatrix} \\
H_D(\theta) &= \frac{1}{2}\sum_{i=1,j>i}^{n} w_{ij} \begin{bmatrix} \kappa_{ij}\kappa_{ij}^\intercal & -\sigma^{-2}\kappa_{ij} \\ - & \sigma^{-4} \end{bmatrix} \\
J_D(\theta) &= \frac{1}{2}\sum_{i,k=1,j>i,l>k}^{n} w_{ij}w_{lk}Cor(U_{ij}^2,U_{kl}^2) \begin{bmatrix} \kappa_{ij}\kappa_{kl}^\intercal & -\sigma^{-2}\kappa_{ij} \\ - & \sigma^{-4} \end{bmatrix}
\end{aligned}
$$

In order to derive the correlations in the $J_a(\theta)$, $a = M, C, D$ matrices, we can exploit the following formula that holds for a stationary Gaussian random field:

$$Cov(Z(s_i)Z(s_j), Z(s_k)Z(s_l)) = \sigma^4(\rho_{ik}\rho_{jl} + \rho_{il}\rho_{jk})$$

After some algebra, we have:

- $Cov(Q_{ij}, Q_{lk}) = \sigma^4 \frac{\rho_{il}+\rho_{ik}+\rho_{jl}+\rho_{jk}}{\sqrt{1+\rho_{ij}}\sqrt{1+\rho_{lk}}}\rho_{ij}\rho_{lk}$
- $Cov(G_{ij}, G_{kl}) = \sigma^2(\rho_{ik} - \rho_{il}\rho_{kl} - \rho_{ij}\rho_{jk} + \rho_{ij}\rho_{kl}\rho_{jk})$

- $Cov(G_{ij}^2, G_{kl}^2) = 2[Cov(G_{ij}, G_{kl})]^2$

- $Cov(U_{ij}, U_{kl}) = \sigma^2(\rho_{ik} - \rho_{il} - \rho_{jk} + \rho_{jl})$

- $Cov(U_{ij}^2, U_{kl}^2) = 2[Cov(U_{ij}, U_{kl})]^2$

- $Cov(B_{ij}, B_{kl}) = 2\sigma^4[\rho_{ik}^2 + \rho_{il}^2 - 2\rho_{kl}\rho_{ik}\rho_{il} + \rho_{jk}^2 + \rho_{jl}^2 - 2\rho_{kl}\rho_{jk}\rho_{jl} - 2\rho_{ij}\rho_{ik}\rho_{jk} - 2\rho_{ij}\rho_{il}\rho_{jl}$

- $Cov(F_{ij}, F_{kl}) = \rho_{ij}\rho_{kl}Cov(B_{ij}, B_{kl}) + \sigma^4(1 - \rho_{ij}^2)(1 - \rho_{kl}^2)(\rho_{ik}\rho_{jl} + \rho_{il}\rho_{jk})$
  $-2\sigma^4\rho_{ij}(1 - \rho_{kl}^2)(\rho_{ik} - \rho_{jk})(\rho_{il} - \rho_{jl}) - 2\sigma^4\rho_{kl}(1 - \rho_{ij}^2)(\rho_{ik} - \rho_{il})(\rho_{jk} - \rho_{jl})$

- $Cov(F_{ij}, B_{kl}) = \rho_{ij}Cov(B_{ij}, B_{kl}) - 2\sigma^4(1 - \rho_{ij}^2)(\rho_{ik} - \rho_{il})(\rho_{jk} - \rho_{jl})$

- $Cov(F_{ij}, G_{kl}^2) = 2\sigma^4[\rho_{ij}\rho_{ik}^2 + \rho_{ij}\rho_{kl}^2\rho_{il}^2 - 2\rho_{ij}\rho_{kl}\rho_{ik}\rho_{il} + \rho_{ij}\rho_{jk}^2 + \rho_{ij}\rho_{kl}^2\rho_{jl}^2 - 2\rho_{ij}\rho_{kl}\rho_{jk}\rho_{jl}$
  $-(1 + \rho_{ij}^2)\rho_{ik}\rho_{jk} - (1 + \rho_{ij}^2)\rho_{kl}^2\rho_{il}\rho_{jl} + (1 + \rho_{ij}^2)\rho_{kl}(\rho_{ik}\rho_{jl} + \rho_{il}\rho_{jk})]$

References

Aho, A. V., Hopcroft, J. E., and Ullman, J. D. (1974), *The Design and Analysis of Computer Algorithms*, Reading, Massachusset: Addison-Wesley.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society: Series B*, 70, 825–848.

Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2012), "Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach," *Journal of the American Statistical Association*, 107, 268–280.

Caragea, P., and Smith, R. (2006), Approximate likelihoods for spatial processes,, Technical report, Department of Statistics, Iowa State University.

Cressie, N. (1993), *Statistics for Spatial Data*, revised edn, New York: Wiley.

Cressie, N., and Johannesson, G. (2008), "Fixed rank kriging for very large spatial data sets," *Journal of the Royal Statistical Society, Series B*, 70, 209–226.

Curriero, F., and Lele, S. (1999), "A composite likelihood approach to semivariogram estimation," *Journal of Agricultural, Biological and Environmental Statistics*, 4, 9–28.

Davis, R., and Yau, C.-Y. (2011), "Comments on pairwise likelihood in time series models," *Statistica Sinica*, 21, 255–277.

Doukhan, P. (1994), *Mixing. Properties and Examples*, New York: Springer-Verlag.

Du, J., Zhang, H., and Mandrekar, V. S. (2009), "Fixed-domain asymptotic properties of tapered maximum likelihood estimators," *The Annals of Statistics*, 37, 3330–3361.

Eidsvik, J., Shaby, B., Reich, B., Wheeler, M., and Niemi, J. (2013), "Estimation and prediction in spatial models with block composite likelihoods," *Journal of Computational and Graphical Statistics*, to appear.

Fuentes, M. (2007), "Approximate likelihood for large irregularly spaced spatial data," *Journal of the American Statistical Association*, 102, 321–331.

Furrer, R., and Sain, S. R. (2010), "spam: a sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random rields," *Journal of Statistical Software*, 36, 1–25.

Gneiting, T. (2002), "Compactly supported correlation functions," *Journal of Multivariate Analysis*, 83, 493–508.

Gneiting, T., and Raftery, A. E. (2007), "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, 102, 359–378.

Guyon, X. (1995), *Random Fields on a Network*, New York: Springer.

Heagerty, P., and Lumley, T. (2000), "Window subsampling of estimating functions with application to regression models," *Journal of the American Statistical Association*, 95, 197–211.

Huang, C., Zhang, H., and Robeson, S. (2011), "On the validity of commonly used covariance and variogram functions on the sphere," *Mathematical Geosciences*, 43, 721–733.

Jenish, N., and Prucha, I. R. (2009), "Central limit theorems and uniform laws of large numbers for arrays of random fields," *Journal of Econometrics*, 150, 86–98.

Joe, H., and Lee, Y. (2009), "On weighting of bivariate margins in pairwise likelihood," *Journal of Multivariate Analysis*, 100, 670–685.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008), "Covariance tapering for likelihood-based estimation in large spatial data sets," *Journal of the American Statistical Association*, 103, 1545–1555.

Lee, Y., and Lahiri, S. (2002), "Least squares variogram fitting by spatial subsampling," *Journal of the Royal Statistical Society B*, 64, 837–854.

Lindgren, F., Rue, H., and Lindström, J. (2011), "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach," *Journal of the Royal Statistical Society: Series B*, 73, 423–498.

Lindsay, B. (1988), "Composite likelihood methods," *Contemporary Mathematics*, 80, 221–239.

Lindsay, B. G., Yi, G. Y., and Sun, J. (2011), "Issues and strategies in the selection of composite likelihoods," *Statistica Sinica*, 21, 71–105.

Mardia, K. V., and Marshall, J. (1984), "Maximum likelihood estimation of models for residual covariance in spatial regression," *Biometrika*, 71, 135–146.

Newey, W. K. (1991), "Uniform convergence in probability and stochastic equicontinuity," *Econometrica*, 59, pp. 1161–1167.

Padoan, S., and Bevilacqua, M. (2013), *CompRandFld: Composite-likelihood based Analysis of Random Fields.* R package version 1.0.3.

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

Shaby, B., and Ruppert, D. (2012), "Tapered covariance: Bayesian estimation and asymptotics," *Journal of Computational and Graphical Statistics*, 21, 433–452.

Stein, M. (2008), "A modeling approach for large spatial datasets," *Journal of the Korean Statistical Society*, 37, 3–10.

Stein, M. (2013), "Statistical properties of covariance tapers," *Journal of Computational and Graphical Statistics*, to appear.

Stein, M., Chi, Z., and Welty, L. (2004), "Approximating likelihoods for large spatial data sets," *Journal of the Royal Statistical Society B*, 66, 275–296.

Varin, C., Reid, N., and Firth, D. (2011), "An overview of composite likelihood methods," *Statistica Sinica*, 21, 5–42.

Vecchia, A. (1988), “Estimation and model identification for continuous spatial processes,” *Journal of the Royal Statistical Society B*, 50, 297–312.

Wendland, H. (1995), “Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree,” *Adv. Comput. Math.*, 4, 389–396.

Whittle, P. (1954), “On stationary processes in the plane,” *Biometrika*, 49, 305–314.

Zhang, H., and Wang, Y. (2010), “Kriging and cross-validation for massive spatial data,” *Environmetrics*, 21, 290–304.