Estimating Space and Space-Time Covariance Functions for Large Data Sets: A Weighted Composite Likelihood Approach

Moreno BEVILACQUA, Carlo GAETAN, Jorge MATEU, and Emilio PORCU

In this article, we propose two methods for estimating space and space-time covariance functions from a Gaussian random field, based on the composite likelihood idea. The first method relies on the maximization of a weighted version of the composite likelihood function, while the second one is based on the solution of a weighted composite score equation. This last scheme is quite general and could be applied to any kind of composite likelihood. An information criterion for model selection based on the first estimation method is also introduced. The methods are useful for practitioners looking for a good balance between computational complexity and statistical efficiency. The effectiveness of the methods is illustrated through examples, simulation experiments, and by analyzing a dataset on ozone measurements.

KEY WORDS: Composite likelihood; Estimating equations; Godambe information; Identification; Space-time geostatistics.

1. INTRODUCTION

The estimation of covariance functions sets an important problem for space and space-time random fields, especially in environmental, meteorological, or geophysical contexts. For instance, the kriging predictor (Cressie 1993) depends on knowledge of the covariance function of the random field. Since a covariance function must be positive definite, practical estimation generally requires the selection of some parametric classes of covariances and the corresponding estimation of these parameters.

The maximum likelihood (ML) and restricted maximum likelihood (REML) methods are generally considered the best options for estimating the covariance model parameters. Nevertheless, evaluation of the objective function under the Gaussian assumption requires to solve a system of linear equations. For a dataset of \( n \) observations, the computational burden is \( O(n^3) \), making this method computationally impractical for large datasets. This fact motivates the search for estimation methods with a good balance between computational complexity and statistical efficiency.

The alternatives can be classified in two sets: methods that try to build an approximation of the covariance matrix of the observations, which is easy to invert and methods that consider a different objective function.

Significant computational gain is achieved when the sampling locations form a regular lattice. In this case, the spectral representation of the Gaussian process allows us to build a computationally tractable approximation of the log-likelihood (Whittle 1954). 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.

For irregularly spaced data, Fuentes (2007) extended Whittle’s idea and suggested to integrate the spatial process over grid cells, obtaining an approximation to the likelihood on a lattice structure. Another idea (Cressie and Johannesson 2008; Stein 2008) is considering a low rank structure on the covariance matrix. This allows us to calculate the inverse and the determinant of a large covariance matrix by inverting and calculating the determinant of a matrix of lower dimension.

Since inversion is less demanding when using algorithms for sparse matrices, part of the literature has been focused on methods allowing for sparse matrices. In the tapering approach, suggested by Kaufman, Schervish, and Nychka (2008), these are obtained through the Schur product of the covariance matrix with a matrix coming from a compactly supported correlation function.

Other methods use a different objective function whose optimization leads to estimators that keep satisfactory statistical properties. The composite likelihood (CL) (Lindsay 1988) is a likelihood-based approximation, indicating a general class of pseudo-likelihoods based on the likelihood of marginal or conditional events. In general, a loss of statistical efficiency is expected from CL estimation with respect to likelihood methods, but its computational tractability makes the CL methods appropriate when working with large datasets. Known examples can be found in Vecchia (1988), Stein et al. (2004), Stein (2005), and Caragea and Smith (2006).

A particular form of CL based on differences has been introduced by Curriero and Lele (1999) in a spatial framework. This kind of CL is the natural approximation of the REML estimation since both are based on the constraints distribution and cope with the covariance parameters only.

In this article, we modify and extend this approach to a spatiotemporal context. More precisely, we propose two estimation methods. The former relies on the maximization of a weighted version of the CL function where we use cutoff weights based on the distance in space and in time. We prove that the associated estimator is consistent and asymptotically Gaussian under increasing domain asymptotics (Cressie 1993). The latter

Moreno Bevilacqua, Dipartimento di Statistica, Università Ca’ Foscari, Venezia, Italy. Carlo Gaetan, Dipartimento di Statistica, Università Ca’ Foscari, Venezia, Italy. Jorge Mateu, Department of Mathematics, Universitat Jaume I - Castellón, Spain. Emilio Porcu, Institut für Mathematische Stochastik, University of Göttingen, Göttingen, Germany (E-mail for correspondence: eporcu@uni-goettingen.de). Work partially funded by grant MTM2010-14961 from the Spanish Ministry of Science and Education, and by MIUR (Italy) under grant 2008MRFM2H.
is based on the solution of a weighted CL equation. The weights are obtained by minimizing an upper bound for the asymptotic variance of the estimator. The weighting scheme is completely general and can be applied to a nonspatial framework.

We then address an important issue, for example, how to select an appropriate space-time model by taking into account the trade-off between goodness-of-fit and model complexity. The methods that are available in the literature (see, for instance, Huang et al. 2007) use information criteria such as the Akaike information criterion and the Bayesian information criterion, both based on the evaluation of the likelihood. We propose an information criterion for covariance model selection based on weighted composite likelihood (WCL) function that exploits recent advances in model selection criteria (Varin and Vidoni 2005).

Moreover, a distinctive feature of the methods is that they are less reliant on strong distributional assumptions of the random fields. The remainder of the article is organized as follows. In Section 2, we introduce our proposals for space and space-time covariance estimation while in Section 3 we describe the asymptotic properties of the WCL estimation method. The information criteria based on WCL is proposed in Section 4. In Section 5, we show through simulation experiments that our methods represent a satisfactory approximation to REML estimation. As a real data example, we apply our methods on a dataset of ozone hourly measurements across the eastern United States (Fuentes 2003). Finally, in Section 6, we give some conclusions.

2. CL ESTIMATION

Let \( \{Z(s, t), s \in \mathbb{R}^d, \ t \in \mathbb{R}\} \) be a real-valued space-time random field with finite and constant variance. The high order of complexity of space-time interactions calls for simplifying assumptions, such as those of intrinsic or weak stationarity, that have implications on the existence of the moments of the random field. A random field is weakly stationary if \( E[Z(s, t)] = \mu \), for all \( s, t \), and the covariance between \( Z(s, t) \) and \( Z(s', t') \) depends only on \( h = s - s' \in \mathbb{R}^d \) and \( u = t - t' \in \mathbb{R} \), the spatial and temporal separation vectors, namely, \( \text{cov}[Z(s, t), Z(s', t')] = C(h, u) \). It is intrinsic stationary if \( E[Z(s, t)] = \mu \), and the variance of the difference between \( Z(s, t) \) and \( Z(s', t') \) depends only on \( h \) and \( u \), namely, \( \text{var}([Z(s, t) - Z(s', t')]) = 2\gamma(h, u) \). The function \( 2\gamma \) is called variogram and under weak stationarity we have \( 2\gamma(h, u) = 2(C(0, 0) - C(h, u)) \). In the sequel, we consider weakly stationary random fields with a covariance function \( C(h, u; \theta) \) known up to the parameter vector \( \theta \in \mathbb{R}^p \) that must be estimated on the basis of a finite number of \( n \) observations \( Z = (Z(s_1, t_1), \ldots, Z(s_n, t_n))' \).

Curriero and Lele (1999) proposed a special case of CL estimator (Lindsay 1988) based on the differences between the observations. The reason for considering such differences is that, as the REML estimation method, it allows us to remove the mean which is a nuisance parameter. We follow that article and first consider all possible space-time differences \( U_{ij} = Z(s_i, t_j) - Z(s_j, t_i), i \neq j \) with log-likelihood \( l_{ij}(\theta) \). Assuming independence between the differences, we get the composite (log) likelihood function

\[
cl(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{n} l_{ij}(\theta). \tag{1}
\]

The maximum CL estimator \( \hat{\theta}_{CL} \) locates the maximum of (1). If we assume that the differences are Gaussian distributed with zero mean, that is, \( U_{ij} \sim N(0, 2\gamma_{ij}(\theta)) \), we get

\[
cl(\theta) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left( \log \gamma_{ij}(\theta) + \frac{U_{ij}^2}{2\gamma_{ij}(\theta)} \right), \tag{2}
\]

where \( \gamma_{ij}(\theta) = \gamma(s_i - s_j, t_j - t_i; \theta) \). A clear advantage is that evaluation of (2) does not require any inversion matrix and the order of computation for CL estimator is \( O(n^2) \).

By construction, the associated composite score

\[
cl^{(1)}(\theta) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{ij}(\theta) \frac{U_{ij}^2}{2\gamma_{ij}(\theta)} \left( 1 - \frac{U_{ij}^2}{2\gamma_{ij}(\theta)} \right) \tag{3}
\]

is unbiased. Here, \( f^{(1)} \) is the vector of partial derivatives of a function \( f \) with respect to \( \theta \). Actually, the previous result does not depend on distributional assumptions about \( U_{ij} \).

Since the CL method results in an unbiased estimating function, in principle, it is possible to improve the efficiency of the estimator, exploiting the theory of optimal estimating equations (see, e.g., Andersen 2004; Kuk 2007).

If we stack all the individual scores \( f^{(1)}_{ij}, j > i \) into a vector \( S(\theta) \), then the optimal estimating function is (Heyde 1997)

\[
e^{*}(\theta) = -E[S(\theta)]'(\text{cov}[S(\theta)])^{-1}S(\theta), \tag{4}
\]

provided \( \text{cov}[S(\theta)] \) is nonsingular.

Simple algebra shows that an entry of the matrix \( \text{cov}[S(\theta)] \) depends on \( \text{cov}[U_{ij}^2, U_{ik}^2] \) whose expression is, for the case of a stationary zero-mean Gaussian random field,

\[
\text{cov}[U_{ij}^2, U_{ik}^2] = \frac{(\gamma_{ij}(\theta) - \gamma_{ik}(\theta) + \gamma_{jk}(\theta) - \gamma_{ij}(\theta))^2}{4\gamma_{ij}(\theta)\gamma_{jk}(\theta)}. \tag{5}
\]

However, evaluation of \( \text{cov}[S(\theta)] \) is a key issue because of its dimension, making it hard to improve the efficiency of the estimator in this way.

A simpler solution can be obtained by downweighting those couples of observations that are far apart in space by a weight \( w_{ij} \geq 0 \) and obtaining the weighted composite (log) likelihood (WCL)

\[
wcl(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{n} l_{ij}(\theta)w_{ij}. \tag{5}
\]

There are many possible choices for the weights. The simplest case is when \( w_{ij} = 1 \), if \( |s_i - s_j| \leq d_S \) and \( |t_i - t_j| \leq d_T \), 0 otherwise, for a fixed spatiotemporal lag \( d = (d_S, d_T)' \). This choice has evident computational advantages, but can also improve the statistical efficiency. Let us illustrate this fact in the spatial case. We first introduce the Godambe information \( G_n(\theta) \) associated to the WCL in Equation (5):

\[
G_n(\theta) = H_n(\theta)J_n(\theta)^{-1}H_n(\theta)', \tag{6}
\]
where
\[ H_n(\theta) = -IE[wcl^{(2)}(\theta)] = \frac{1}{2} \sum_{i=1}^{n} \sum_{j>i} (\gamma_{ij}(\theta)')(\gamma_{ij}(\theta))^{'2} w_{ij}, \tag{7} \]
and
\[ J_n(\theta) = IE[wcl^{(1)}(\theta)wcl^{(1)}(\theta)'] = \frac{1}{2} \sum_{i=1}^{n} \sum_{j>i} \sum_{k>l} \gamma_{ij}(\theta)\gamma_{kl}(\theta) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j>i} \sum_{k>l} \gamma_{ij}(\theta)\gamma_{kl}(\theta) \mathrm{cor}[U_{ij}, U_{lk}] w_{ij} w_{lk}. \tag{8} \]

Here, \( f^{(2)} \) is the Hessian matrix of the function \( f \) with respect to \( \theta \). The inverse of \( G_n(\theta) \) is an approximation of the asymptotic variance of the WCL estimator (see Section 3). Note that the Godambe information depends on \( \mathbf{d} \) by means of \( w_{ij} \) and in the sequel we shall stress this through the abuse of notation \( G_n(\theta; \mathbf{d}) \).

For illustration purposes, we consider two examples of covariance models for a spatial random field observed on 64 points located over a \( 8 \times 8 \) regular grid \( \{0, 1, \ldots, 7\}^2 \):

1. the exponential model
   \[ C(h) = \sigma^2 \exp \left( -\frac{\|h\|}{\phi} \right), \quad \phi > 0, \quad h \in \mathbb{R}^d; \tag{9} \]
2. the wave model
   \[ C(h) = \sigma^2 \frac{\phi}{6.5 \pi \|h\|} \sin \left( \frac{6.5 \pi \|h\|}{\phi} \right), \quad \phi > 0, \quad h \in \mathbb{R}^d. \tag{10} \]

The models are parameterized in terms of the practical range, that is, \( C(h)/C(0) \approx 0.05 \) when \( \|h\| = \phi \). We use and compare these two models since the former is an example of correlation function attaining only positive values and having a monotonic decay while the latter can attain positive or negative values with a decay being not monotonic.

We consider \( \sigma^2 \) is known and we estimate \( \phi \). In Figure 1, we show the asymptotic relative efficiency (ARE) of the estimator derived from (5) with respect to the REML estimator for different values of the distance \( d_S \).

Note that, as expected, the efficiency decreases when the range \( \phi \) increases, but the efficiency decreases with the distance \( d_S \). In practice, if we use all differences we may lose efficiency.
since too many redundant pairs of observations can skew the information confined in pairs of nearby observations. However, the same comments do not apply to the wave model. In this case, we need to consider the whole information in the pairs. Note also that for $\phi = 8$, the ARE is very low.

The example suggests that choosing $d$ in a suitable way under specific models can improve significantly the statistical efficiency of the estimate. Moreover, our findings agree with the results by Davis and Yau (2010) for time series models. These authors found that choosing weights so that the CL is constructed only from adjacent pairs improves the efficiency preferable to the “full” CL.

Therefore, using a scalar equivalence result (Heyde 1997), a sensible strategy for improving the efficiency of the CL estimator can be seeking the “optimal” $d^*$ such that

$$d^* = \arg\min_{d \in D} \text{tr} \left( G_n^{-1}(\hat{\theta}; d) \right),$$

where $D$ is a finite set of space-time lags. If the data are observed on a regular grid on space and time, the definition of $D$ is simple, but, in general, an inspection of the variogram cloud could help with the choice of the lags in $D$.

The lag $d$ controls both the statistical efficiency of the estimation and the computational efficiency. In general, there is no trade-off between statistical and computational efficiency as highlighted in the examples. In this sense, the method is different from other estimation methods of covariance functions (Stein, Chi, and Welty 2004; Kaufman et al. 2008).

Summarizing, evaluation of the WCL estimate involves three steps:

1. Get a preliminary and consistent estimate $\hat{\theta}$ of $\theta$. In our implementation, we have used the weighted least squares estimates (Cressie 1993, p. 91) based on empirical estimation of the variogram function. This method is one of the most widely used estimation methods for space-time data (see, e.g., Cressie and Huang 1999; Gneiting, Genton, and Guttorp 2007);

2. Minimize tr($G_n^{-1}(\hat{\theta}; d)$) with respect to $d$ in order to get $d^*$. An estimate of $G_n$ is computed using subsampling techniques as explained in Section 3.4;

3. Find the minimum of $w(\theta)$.

The 0/1 weights can be a simple and effective solution but we can use smoother weights. Formula (4) has also shown the need of adaptive weights, that is, weights depending on the covariance model. However, the evaluation of $\text{cov}[S(\theta)]$ is the point at issue.

A possible strategy for circumventing this difficulty is the following. Assume that $\theta$ is scalar and consider the estimating function

$$e(\theta, w) = w S(\theta),$$

where $w$ is the vector of the weights $w_{ij}$. The asymptotic variance of the estimator associated to this estimating function is

$$\text{avar}(\theta) = \frac{\text{w} \text{cov}[S(\theta)] \text{w}}{\left(\text{w}h(\theta)\right)^2},$$

with $h(\theta) = \mathbb{E}[S^{(1)}(\theta)]$. This variance is bounded by

$$\lambda_{\min}(\theta) B(w) \leq \text{avar}(\theta) \leq \lambda_{\max}(\theta) B(w)$$

where $\lambda_{\min}(\theta)$ and $\lambda_{\max}(\theta)$ are the minimum and maximum eigenvalues of $\text{cov}[S(\theta)]$, respectively, and $B(w) = \frac{\text{w} \text{w}}{\left(\text{w}h(\theta)\right)^2}$.

We can improve the upper bound minimizing the function $B(w)$. It is easy to show that the minimum of $B(w)$ is achieved when $w = c h(\theta)$, for $c \neq 0$, and we put $c = 1$ without loss of generality. Obviously, the weighted version of the CL, with $w = h(\theta)$, can assure an effective improvement in the efficiency with respect to the unweighted one (1) if the intervals $[\lambda_{\min}(\theta) B(h(\theta)), \lambda_{\max}(\theta) B(h(\theta))]$ and $[\lambda_{\min}(\theta) B(1), \lambda_{\max}(\theta) B(1)]$ are not overlapping, that is, if

$$\frac{\lambda_{\max}(\theta)}{\lambda_{\min}(\theta)} B(1) < B(h(\theta)).$$

Actually, a less restrictive condition is that

$$h(\theta) \text{cov}[S(\theta)] h(\theta) < \frac{1}{(\text{IE} h(\theta))^2},$$

This discussion suggests that we can use a weighted composite score (WCS)

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

where $w_{ij}(\theta) = -\mathbb{E}[l_{ij}^{(2)}(\theta)]$, to build an estimator that can improve the statistical efficiency.

Going back to our previous example, Figure 1 shows the gain in efficiency using a WCS. Note that we have an effective improvement only for the exponential model, namely, when we consider a wave model, the simple 0/1 weight scheme outperforms the adaptive ones. The extension of the previous argument to the multiparameter case is overlay complicated. A simple solution comes from using different weights for each composite score component and getting a simplified version of that defined in Equation (14), allowing to obtain the WCS

$$e_w(\theta) = \sum_{i=1}^{n} \sum_{j > i} W_{ij}(\theta) l_{ij}^{(1)}(\theta),$$

where $W_{ij}(\theta) = \text{diag}([w_{ij}(\theta)])$.

The weight function depends on the parameters to estimate. For instance, if we consider only the variance parameter, then from (7) we obtain $w_{ij}(\sigma^2) = \mathbb{E}[l_{ij}^{(2)}(\sigma^2)] = \frac{1}{\sigma^2}$.

In this case, WCS method does not modify the solution with respect to the unweighted version since the weight function is constant.

If we consider a range parameter of an isotropic covariance function that is completely monotone on the positive real line, then we can prove that the weight function is monotonically decreasing with the distance. In this case, pairs of observations that are far apart in space are smoothly downweighted.

For the multiparameter case, our proposal is shown to be effective through some examples. Figure 2 compares the ARE when we estimate the range and variance parameter in models (9) and (10). As an overall summary, we use the square-root of the ratio of the determinants of the asymptotic variance matrices. Again, the gain in efficiency is clear for the exponential model. Note that, since the score component relative to the variance is unweighted, the gain is mainly due to the improvement in estimating the range parameter.
Figure 2. ARE’s of WCL (•) and WCS (◦) estimators with respect to the REML estimator in the multi-parameter case. We set $\sigma^2 = 1$ and different values of the practical range, $\phi = 2, 4, 8$. Respectively (a)–(c) for the exponential covariance model (9); (d)–(f) for the wave covariance model (10).

Summing up, WCL (5) or WCS (15) are two possible ways for estimating covariance models. If the efficiency is not a big issue, in the first case, we can build an inference method based on a contrast function that allows simpler asymptotic results and derivation of a model identification criterion (see Sections 3 and 4). Moreover, WCLs lead to optimization algorithms that are faster, by setting the maximum number of pairs allowed for each site, and that do not show multiple solution problems that can arise in solving estimating equations (Heyde and Morton 1998).

3. ASYMPTOTIC PROPERTIES OF THE WEIGHTED CL ESTIMATOR

3.1 Setup

Two types of asymptotic frameworks are considered in spatial statistics. In the increasing domain framework, the domain of observation expands as the distance between any pair of locations is bounded away from zero. Instead, fixed domain asymptotics deals with a fixed region within which the number of sampling locations increases.

Note that the asymptotic behavior of spatial covariance parameter estimators can be quite different under the two frameworks (Stein 1999; Zhang 2004). For a comparison of these two frameworks, see Zhang and Zimmerman (2005). We focus on increasing domain asymptotics and we use the same setup as in Li et al. (2008).

In order to explain with clarity the results following throughout this section, we need to adopt some abuse of notation that is to be considered as independent with respect to the notation adopted throughout the remainder of the article. We suppose that the sampling region $R_n$ is a region in $\mathbb{R}^{d+1}$ and we factorize $R_n$ into $R_n = I_n \times F$, where $I_n \subset \mathbb{R}^{d_1}$, $F \subset \mathbb{R}^{d_2}$ and $d_1 + d_2 = d + 1$. This notation reflects two space-time settings:

1. if $d_1 = 1$ and $d_2 = d$, we are considering an increasing in time domain, $I_n$, with the spatial domain, $F$, fixed.
2. if $d_1 = d$ and $d_2 = 1$, we are considering an increasing spatial domain with the temporal one fixed.

The set $I_n = A_n \cap \mathbb{Z}^{d_1}$ is a part of the integer grid $\mathbb{Z}^{d_1}$ that lies inside a sampling region $A_n$ obtained by “inflating” a prototype.
set $A_0$ by the scaling factor $\lambda_n$, that is, $A_n = \lambda_n A_0$. Here, $A_0 \subset (-1/2, 1/2]^d$ is an open connected neighborhood of the origin and $\{\lambda_n\}_{n=0}^\infty$ is a sequence of positive real numbers increasing to $\infty$. Setting (1) (a small number of locations and a large number of measures in time) reflects a common setting in environmental studies, and this is the case of our real data example. Setting (2) is useful in a pure spatial setting (i.e., $|F| = 1$) when the sites are located on a regular lattice.

Next, we suppose that the set of the data sites is given by $\mathcal{R}_n = \{s_1, \ldots, s_n\} \subset \mathbb{R}^d$. When considering the $0/1$ weights in (5), it is not necessary to include all distinct pairs of observations. In fact, we introduce a finite set $\mathcal{M} = \{h_1, \ldots, h_q\}, h_i \in \mathbb{R}^{d+1}$ that does not contain the origin and determine which pairs of observations contribute to the product for each considered site $s_i$. The set $\mathcal{M}$ defines the neighborhood of $s_i$ and we denote $\mathcal{R} = \{s \in \mathcal{R}_n, s + h_i \subseteq \mathcal{R}_n \forall h_i \in \mathcal{M}\}$, the $\mathcal{M}$-interior of $\mathcal{R}_n$ and $\partial \mathcal{R}_n = \mathcal{R}_n \setminus \mathcal{R}$, the $\mathcal{M}$-boundary of $\mathcal{R}_n$. Note that the size $|\mathcal{R}_n|$ of the sampling region $\mathcal{R}_n$ is $O(\lambda_n^d)$ and $|\mathcal{R}_n|/|\mathcal{R}_n| \to 1$, as $n \to \infty$.

Instead of considering the CL
\[
-\frac{1}{2} \sum_{s \in \mathcal{R}_n} \sum_{h \in \mathcal{M}} \left\{ \log \gamma(h; \theta) + \frac{(Z(s) - Z(s + h))^2}{2\gamma(h; \theta)} \right\},
\]
we consider a modified version of (16), namely,
\[
-\frac{1}{2} \sum_{s \in \mathcal{R}_n} \sum_{h \in \mathcal{M}} \left\{ \log \gamma(h; \theta) + \frac{(Z(s) - Z(s + h))^2}{2\gamma(h; \theta)} \right\}
\]
which differs from (16) by an edge-effect and allows for interpreting (17) as a sum of independently distributed random variables, simplifying the proofs reported in the Appendices.

### 3.2 Consistency

In order to show consistency and asymptotically normality of the composite log-likelihood estimator, we use standard results on minimum contrast estimators (Guyon 1995, p. 119). We write the composite log-likelihood (17) as a contrast process
\[
\hat{Q}_n(\theta) = \frac{1}{|\mathcal{R}_n|} \sum_{s \in \mathcal{R}_n} \sum_{h \in \mathcal{M}} \left\{ \log \gamma(h; \theta) + \frac{(Z(s) - Z(s + h))^2}{2\gamma(h; \theta)} \right\},
\]
and the CL estimator is the value $\hat{\theta}_n$, such that
\[
\hat{\theta}_n = \arg\min_{\theta \in \Theta} \hat{Q}_n(\theta).
\]

**Proposition 1.** Assume that:

- (C1) $Z$ is a stationary and ergodic random field, with $E[Z(s)]^2 < \infty$ for all $s \in \mathbb{R}^{d+1}$;
- (C2) $\Theta$ is a compact subset of $\mathbb{R}^d$ and the true value $\theta_0$ belongs to $\Theta$; the interior of $\Theta$;
- (C3) for each fixed $h \in \mathcal{M}$, the mapping $\theta \mapsto \gamma(h; \theta)$ has continuous partial derivatives with respect to $\theta$, and $\inf_{\theta \in \Theta} \gamma(h; \theta) > 0$;
- (C4) the function
\[
\theta \mapsto K(\theta, \theta_0) = \sum_{h \in \mathcal{M}} \left[ \log \gamma(h; \theta) + \frac{\gamma(h; \theta_0) - \gamma(h; \theta)}{\gamma(h; \theta)} \right]
\]
has a unique global minimum over $\Theta$ at $\theta_0$

Then, $\hat{\theta}_n \to \theta_0, n \to \infty$, in $P_{\theta_0}$-probability.

**Remark.** (C4) is an identifiability condition. Under Gaussianity, for each $h \in \mathcal{M}$, the function $\theta \mapsto \log \gamma(h; \theta)$ has a global minimum at $\theta_0$ (because essentially it is a Kullback–Leibler divergence), but in the multidimensional case ($p > 1$) $\theta_0$ fails, in general, to be the unique minimizer. A necessary condition for the unicity is that $p \leq q$, where $q = |\mathcal{M}|$.

### 3.3 Asymptotic Normality

For the asymptotic normality, we need to quantify the strength of dependence in the random field. The strong mixing coefficient for a random field is defined as
\[
a(a, b) = \sup_{E_1, E_2} \{P(A_1 \cap A_2) - P(A_1)P(A_2)\}:
\]
\[
A_j \in \mathcal{F}(E_j), |A_j| \leq b,
\]
\[
i = 1, 2, d(E_1, E_2) \geq a,
\]
for $a > 0, b > 0$, and where the supremum is taken over rectangles $E_1, E_2$ in $\mathbb{R}^{d+1}, \mathcal{F}(E)$ denotes the $\sigma$-algebra generated by the random variables $\{Z(s), s \in E\}, d(E_1, E_2)$ is the distance between the sets $E_1, E_2$, defined as $d(E_1, E_2) = \inf_{x \in E_1, y \in E_2} \sup_j|x_j - y_j|$, $j = 1, \ldots, d + 1$. Under the fixed notation and stated conditions, we are now ready to state our result about the asymptotic distribution of the CL estimator.

Let $f^{(2)}(\theta)$ be the $p \times p$ matrix of second mixed derivatives
\[
f^{(2)}_{ij}(\theta) = \frac{\partial^2 f(\theta)}{\partial \theta_i \partial \theta_j}.
\]

**Proposition 2.** Assume that (C1) to (C4) are satisfied. In addition, assume:

- (N1) $E[Z(s)]^{4+\epsilon} < \infty$ and for all $a \geq b \geq 1, a(b, b) \leq Ca^{-\kappa} b^{\kappa}$ for some $0 < \epsilon \leq 4, C > 0, \kappa_1 > (4 + \epsilon)d_1/2$ and $0 \leq \kappa < 2 \kappa_1/d_1$;
- (N2) $p \leq q = |\mathcal{M}|$ and for each fixed $h \in \mathcal{M}$, the function $\gamma(h; \theta)$ has continuous partial derivatives of order 2 with respect to $\theta$;
- (N3) the $p \times q$ matrix $\Gamma(\theta)$ with the $(i, j)$th element given by $\gamma_{ij}(h_1, \ldots, h_q, \theta, \theta_0), i = 1, \ldots, p, j = 1, \ldots, q$, is such that $\Gamma(\theta_0)$ has rank $p$.

Then,
\[
|\hat{\mathcal{R}}_n|^{1/2} \mathcal{J}_n^{-1/2} \mathcal{H}_n(\theta_0) \hat{\theta}_n - \theta_0 \overset{d}{\to} \mathcal{N}(0, I_p)
\]
as $n \to \infty$, where $\mathcal{J}_n(\theta) = |\hat{\mathcal{R}}_n|E_{\theta_0}[Q^{(1)}(\theta) \hat{Q}^{(1)}(\theta)]$ and $\mathcal{H}_n(\theta) = E_{\theta_0}[Q^{(2)}(\theta)]$, with $Q^{(1)}(\cdot)$ the $p$-dimensional column vector of partial derivatives of $Q(\cdot)$ with respect to $\theta$, and $Q^{(2)}(\cdot)$, the Hessian matrix.

**Remarks:**

1. Condition (N1) is satisfied if $Z$ is a stationary Gaussian random field with correlation function that decays at a polynomial rate faster than $d_1$ and its spectral density is bounded below due to Corollary 2 of Doukhan (1994, p. 59).

2. The finite sample approximation to the asymptotic variance
\[
\text{avar}(\hat{\theta}_n) \approx |\hat{\mathcal{R}}_n|^{-1} \mathcal{H}_n^{-1}(\theta_0) \mathcal{J}_n(\theta_0) \mathcal{H}_n^{-1}(\theta_0)'
\]
...the inverse of the Godambe information associated to the estimating function $\hat{Q}^{(1)}(\theta)$ and evaluated at $\theta_0$. We can also prove that the edge-effects are negligible, so we have

$$\text{avar}(\hat{\theta}_n) \approx H_n^{-1}(\theta_0)J_n(\theta_0)H_n^{-1}(\theta_0)^t,$$

where $H_n(\theta)$ and $J_n(\theta)$ are defined by (7) and (8), respectively.

### 3.4 Standard Error Evaluation

The evaluation of the standard error requires consistent estimation of the matrices $H_n(\theta_0)$ and $J_n(\theta_0)$. However, the evaluations of the plug-in estimates $H_n(\hat{\theta}_n)$ and $J_n(\hat{\theta}_n)$ are of order $O(n^2)$ and $O(n^4)$, respectively. The latter becomes computationally unfeasible for large datasets. Note that asymptotic normality entails that $W_n^{-1}J_n(\theta_0)\to J(\theta_0)$, as $n \to \infty$, where $J(\theta_0)$ is a positive definite matrix (see the Appendix for its definition) and $W_n = \sum_{i=1}^n I_{i,j}$.

Our strategy here is to use subsampling techniques to obtain an estimate $\hat{J}_n(\theta)$ of $J(\theta_0)$ and then estimate $J_n(\theta_0)$ by $W_n \hat{J}_n(\theta_n)$.

For $S_1, \ldots, S_m, m_n$ overlapping subsets of $s_1, \ldots, s_n$, Heagerty and Lumley (2000) proposed as an estimator of $J(\theta)$ the following expression:

$$\hat{J}_n(\theta) = \frac{1}{m_n} \sum_{k=1}^{m_n} \frac{1}{W_n^{(k)}} \sum_{i,j \in S_k}^\prime y_{ij}(\theta)y_{ij}(\theta)^t \gamma_{ij}(\theta)\gamma_{ij}(\theta)^t$$

$$\times \left( 1 - \frac{U_{ij}^2}{2\gamma_{ij}(\theta)} \right) \left( 1 - \frac{U_{ij}^2}{2\gamma_{ij}(\theta)} \right) u_{ij}w_{ij} v_{ij},$$

where $W_n^{(k)} = \sum_{i,j \in S_k} w_{ij}$.

Finally, the asymptotic variance of $\hat{J}_n(\theta)$ can then be estimated using the subsampling approximation

$$\text{avar}(\hat{J}_n(\theta)) = \frac{H_n^{-1}(\theta_0)W_n \hat{J}_n(\theta_n)H_n^{-1}(\theta_0)}{m_n}.$$  

In general, the choice of the optimal (in the sense of minimizing the mean square error (MSE) of the estimator) window size is not an easy task since it depends on several aspects such as the dimension of the lattice and the underlying dependence structure.

Assuming that the data are observed in the region $R_n$ defined in Section 3.1, under increasing in time domain the overlapping subsets are $S_i = I_n^{(i)} \times F$, where $I_n^{(i)}$ are overlapping temporal subwindows. The size of the $i$th subwindow $|I_n^{(i)}|$ can be determined following Carlstein (1986) as

$$|I_n^{(i)}| = \left( \frac{2\beta}{1 - \beta^2} \right)^{2/3} \left( \frac{3n}{2} \right)^{1/3}.$$

In that article, the author assumes that the statistic of interest is the simple mean and that the temporal correlation follows an $AR(1)$ process with parameter $\beta$. An estimate of $\beta$ is given by $\hat{\beta} = \hat{C}(0, 1)/\hat{C}(0, 0)$. Li, Genton, and Sherman (2008) used this procedure to estimate the covariance matrix for testing some properties of space-time covariance functions.

Under an increasing spatial domain for a set of overlapping spatial subwindows on a regular lattice (see, e.g., Sherman 1996; Lee and Lahiri 2002; Li et al. 2008), the optimal size of the subwindow is $|S_n| = Cn^{d/(d+2)}$. The constant $C$ is difficult to specify and in this context a careful sensitivity analysis is required for establishing its value.

### 4. A MODEL SELECTION CRITERION BASED ON THE WEIGHTED CL ESTIMATOR

Model selection is a major problem in spatial and space-time statistics and some few criteria have been proposed in the literature (see, e.g., Huang and Chen 2007). Hoeting et al. (2006) proposed a model selection criteria based on AIC and BIC in the spatial and space-time setting. Their proposals rely on the evaluation of the likelihood for a Gaussian random field. Unfortunately, in the case of massive datasets, the selection criteria based on the full likelihood fail due to their computational burden. This drawback can be overcome by using the CL framework, and recasting the approach proposed by Varin and Vidoni (2005).

We consider a parametric family of covariance functions $\mathcal{F} = \{ \mathcal{C}(\cdot; \theta), \theta \in \Theta \}$, with $\Theta$ a compact subset of $\mathbb{R}^p$. This family may or may not contain the true covariance function $\mathcal{C}(\cdot; \theta_0)$. Thus, we face two possible sources of misspecification: (a) we may choose a parametric class of covariance functions that does not contain the true covariance function and we are using a likelihood approximation and (b) the misspecification is endemic to the proposed methodology that constitutes an approximation to the full likelihood.

Our goal is to compare different models in terms of their predictive ability at forecasting a new set of data $\tilde{Z}$. Varin and Vidoni (2005) suggested to use the composite Kullback–Leibler divergence defined as the linear combination of the Kullback–Leibler divergences associated to each component of the CL. In our setup, the composite Kullback–Leibler divergence of a model, identified with $\theta$, with respect to the true distribution is given by

$$\text{KL}_\mathcal{C}(\theta_0, \theta) = \sum_{i=1}^n \sum_{j=1}^n \text{E}_{\theta_0}[|l_{ij}(\theta_0) - l_{ij}(\theta)|]w_{ij}. \quad (21)$$

The model selection can be performed by minimizing the expected composite Kullback–Leibler divergence of the estimated model using WCL relative to the true distribution, namely, $\text{E}_{\theta_0}[\text{KL}_\mathcal{C}(\theta_0, \hat{\theta}_n)]$, or, equivalently, maximizing

$$\text{E}_{\theta_0}[\text{KL}_\mathcal{C}(\theta_0, \hat{\theta}_n)] = \sum_{i=1}^n \sum_{j=1}^n \text{E}_{\theta_0}[|l_{ij}(\hat{\theta}_n)|]w_{ij}, \quad (22)$$

where the second expectation is taken assuming $\hat{\theta}_n$ as fixed.

A natural estimator of $\hat{\theta}_n$ comes from the maximized composite log-likelihood $\text{wcl}(\hat{\theta}_n)$, which is a biased estimator, and the bias is of the order of the trace of $J_n(\theta_0)H_n(\theta_0)^{-1}$, where $\theta_0 \in \Theta$ is the pseudo-true value, that is, the value minimizing the Kullback–Leibler divergence of the models of $\mathcal{F}$ with respect to the true distribution. Following Varin and Vidoni (2005), the CL information criterion selects the model maximizing the composite likelihood criterion (CLIC)

$$\text{CLIC}(\hat{\theta}_n) = \text{wcl}(\hat{\theta}_n) + \text{tr}(\hat{J}_nH_n^{-1}), \quad (23)$$
Table 1. ARE for CL, WCL, and WCS in estimating the range and variance of 1000 Gaussian random fields with exponential covariance function increasing the locations sites and the range

| n  | CL     | WCL    | WCS    | CL     | WCL    | WCS    | CL     | WCL    | WCS    |
|----|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 100| 0.57   | 0.68   | 0.93   | 0.70   | 0.69   | 0.73   | 0.66   | 0.70   | 0.73   |
| 256| 0.98   | 0.91   | 0.99   | 0.92   | 0.98   | 0.91   | 0.96   | 0.98   | 0.99   |
| 512| 0.41   | 0.52   | 0.63   | 0.44   | 0.54   | 0.77   | 0.47   | 0.62   | 0.83   |
| 1   | 0.96   | 0.92   | 0.96   | 0.97   | 0.99   | 0.96   | 0.96   | 0.98   | 0.99   |
| 2   | 0.38   | 0.46   | 0.59   | 0.41   | 0.50   | 0.72   | 0.44   | 0.57   | 0.78   |
| 4   | 0.96   | 0.91   | 0.94   | 0.94   | 0.95   | 0.96   | 0.96   | 0.92   | 0.97   |

where the routine for estimation of $J_n$ and $H_n$ has been previously illustrated. The effectiveness of this criterion will be illustrated in the next section.

5. NUMERICAL RESULTS

5.1 Simulated Data

In this section, we compare the statistical and computational efficiency of CL, WCL, and WCS estimation methods with respect to the classical REML estimation for two models in space and in space time with different number of observations. In both cases, the number of observations was chosen in order to make the evaluation of the restricted likelihood rather simple. Then, the estimates were compared in terms of the MSE evaluated by a Monte Carlo method. For this, we simulated 1000 replications from a Gaussian random field with a given covariance function.

All calculations were carried out on a 3.2 GHz dual processor with 16 GB of memory and the estimation procedures were implemented coupling R functions and C routines. Some of these routines are collected in the R package BLINDED.

We first considered a spatial example with exponential covariance function (9) and parameters $\sigma^2 = 1$ and $\phi = 2, 4, 6$. We assumed to observe the random field over different spatial grids, $R_n = [-n + 1, \ldots, n]^2, n = 5, 8, 12$. This setting reflects scenario (2) in Section 3.

Table 1 shows the relative statistical efficiency of the CL, WCL, and WCS estimates with respect to the REML ones. On general grounds, the most critical part is in estimating the parameter driving the spatial dependence ($\phi$). The WCL estimates show better performances with respect to the unweighted version. Moreover, WCS always outperforms the WCL, giving efficiencies close to the REML method. All methods tend to get worse when the spatial dependence increases, although the loss in efficiency of WCS is slower, always keeping a good compromise.

We then used the spatial grid of the previous example but with $N = 20, 25, \ldots, 75$, that is, 400, 625, \ldots, 5,625 observations, to show an example of the practical computational performances of the WCL methods. Figure 3 reports the computational time (in terms of seconds of real elapsed time) in evaluating the CL function and the WCL function taking the difference whose distance is lower than half the maximum distance on the grid, and compare it with the computational time of the REML method. The computational gains are substantial and increase when increasing the number of observations. For instance, for 900 observations, the ratio between CL and REML times is 0.1586, while the ratio between WCL and REML is 0.0689. When considering 5,625 observations, the ratios are 0.0301 and 0.0212, respectively. We do not report the WCS time since it is the same of the CL method.

![Figure 3. Time in seconds when evaluating REML, CL, and WCL functions increasing the spatial observations.](image-url)
Table 2. ARE in terms of the square-root of the ratio of the determinants of the asymptotic variance matrices for CL, WCL, and WCS method. The estimates are spatial and time range parameters, variance, and separability parameter of 1,000 Gaussian random fields with Gneiting model (24), increasing the time and the space-time range.

| $T_n$ | $c = 2$ | $c = 3$ | $c = 4$ | $c = 2$ | $c = 3$ | $c = 4$ |
|------|--------|--------|--------|--------|--------|--------|
| 40   | 0.55   | 0.47   | 0.42   | 0.57   | 0.50   | 0.44   |
| 80   | 0.66   | 0.59   | 0.50   | 0.67   | 0.62   | 0.54   |
| 100  | 0.83   | 0.78   | 0.73   | 0.84   | 0.80   | 0.75   |
| 100  | 0.41   | 0.37   | 0.32   | 0.42   | 0.38   | 0.34   |
| 100  | 0.63   | 0.55   | 0.50   | 0.64   | 0.57   | 0.51   |
| 100  | 0.74   | 0.70   | 0.65   | 0.75   | 0.71   | 0.67   |
| 100  | 0.30   | 0.25   | 0.20   | 0.33   | 0.28   | 0.24   |
| 100  | 0.43   | 0.35   | 0.24   | 0.52   | 0.45   | 0.38   |
| 100  | 0.67   | 0.62   | 0.59   | 0.71   | 0.65   | 0.60   |

In a further step, we performed a simulation study for the space-time covariance function proposed by Gneiting (2002):

$$C(h, u; \theta) = \frac{\sigma^2}{\left(\frac{20a|b|}{a} + 1\right)} \exp\left(-\frac{3\|h\|}{c\left(\frac{20a|b|}{a} + 1\right)^{1/2}}\right).$$

with $\theta = (\sigma^2, a, c, \alpha, \beta)$. Here, we set $\sigma^2 = 1$, $c = 2, 3, 4$, $a = 5, 10, 20$, and $\alpha = \beta = 0.5$, mirroring a situation with increasing space and time correlation. In our experiment, $a$ has been kept fixed.

We considered space-time grids increasing in time $R_n = I_n \times F$ with $F = [-1, \ldots, 2]^2$ and $I_n = \{1, \ldots, T_n\}$, $T_n = 40, 80$. This setting reflects scenario (1) in Section 3 and tries to mimic the real example presented in Section 5.2.

Table 2 reports the ARE for the three methods when applied to the previous space-time context. In this case, we report for shortness only the trace but the crucial parameters are the space and time range. As is the former example when increasing the spatial and/or temporal dependence, the ARE get worse and WCS always outperforms the WCL and the unweighted version.

Finally, we considered an example of model identification using the information criterion defined in (23). We considered 100 independent simulations from a stationary Gaussian random field with zero mean observed over two space-time grids: $R_n = I_n \times F$, where $F = [-2, \ldots, 3]^2$ and $I_n = \{1, \ldots, T_n\}$ with $T_n = 100, 200$, and three different covariance models:

1. the double exponential model covariance model:

$$C(h, u; \theta) = \sigma^2 \exp\left(-\frac{3\|h\|}{a} - \frac{3|u|}{b}\right),$$

with $\theta = (\sigma^2, a, b)' = (1, 3, 3)'$;

2. the model in Equation (24) with $\theta = (1, 10, 3, 0.5, 0)'$;

3. the model in equation (24) with $\theta = (1, 10, 3, 0.5, 0.7)'$.

Again, parameter $a$ has been kept fixed in models (2) and (3). Models (1) and (2) are separable models, that is, $C(h, u) = \frac{C(h, 0)C(0, u)}{C(0, 0)}$, and models (2) and (3) are nested.

Table 3. Number of cases of identification of a model with respect to the true one, using CLIC

| Model | $T = 100$ | $T = 200$ |
|-------|-----------|-----------|
|       | (1) | (2) | (3) | (1) | (2) | (3) |
| True  | 84  | 10  | 6   | 96  | 2   | 2   |
| (2)   | 8   | 70  | 22  | 5   | 82  | 13  |
| (3)   | 9   | 19  | 72  | 3   | 12  | 85  |

Note that when comparing models with the model selection criteria, the same common distance must be used in the WCL estimation.

Estimation of $J$ and $H$ are performed as described in Section 3.4 and in estimating $J$ we used overlapping subwindows in time. The results are summarized in Table 3, reporting the number of cases in which a particular model is identified with respect to the model chosen for the simulations, that is, “true” model.

These results point out that the CLIC is promising identification criterion because at least 70% of the models have been correctly identified even for a relatively small number of observations $T_n = 100$.

5.2 A Real Data Example

We consider maxima of hourly means over eight consecutive hours of ozone levels measured at 513 stations in the United States (see Figure 3) from May 1, 1995, to October 31, 1999, giving 184 observations per site and per year. The whole dataset can be obtained from www.image.ucar.edu/Data. The dataset was first analyzed by Fuentes (2003) who suggested a nonstationary spatial model for the 3-year average of the annual fourth-highest daily maximum 8-hour average ozone concentration. Gillett and Nychka (2005) performed a spatiotemporal analysis on a subset of 72 monitoring stations in a study region centered at North Carolina. Very recently, Matsuo et al. (2011) fitted a nonstationary spatial model using data for 1997, and assumed temporal independence.

We restrict our attention on the last 2 years (1998–1999) and we keep apart the last 60 days for assessing the prediction performance. Here, we do not want to carry out a new data analysis but, instead, we want to offer an example with a dataset that is quite challenging for WCL and WCS methods because we deal with 158,004 observations. In our dataset, 8,845 values are missing, but this fact does not affect our estimation procedure.

Because of the seasonal dependence of the ozone, we standardized the data as in Gillett and Nychka (2005) and supposed that the daily maximum 8-h ozone measurement, at location $s$ and day $t$, $Z(s, t)$, can be decomposed into a large scale spatiotemporal $\mu(s, t)$

$$\mu(s, t) = a(s) + \sum_{j=1}^{3} b_j(s) \cos(2\pi j t/184) + c_j(s) \sin(2\pi j t/184)$$

and a small-scale stochastic component, $\varepsilon(s, t)$, with variance that varies over the space, namely,

$$Z(s, t) = \mu(s, t) + \sigma(s) \varepsilon(s, t).$$
The parameters in the seasonal component were estimated by ordinary least squares and $\sigma(s)$ was estimated from the residuals. Moreover, the large-scale components were smoothed over the space using a singular value decomposition approach (see Gilleland and Nychka 2005, for more details). We modeled the small-scale component $\varepsilon(s, t)$ as a Gaussian random field with zero mean, unit variance, and correlation (i.e., $\sigma^2 = 1$):

**model 1.** the double exponential model (25);
**model 2.** the separable model (24) with $\alpha = 0.5, \beta = 0$;
**model 3.** the separable model (24) with $\beta = 0$;
**model 4.** the nonseparable model (24).

We started from a preliminary weighted least squares (WLS) estimate (Cressie 1993), fixing a small number of relevant lags in space and in time in order to identify the distance $d = (d_s, d_f)$ for the WCL estimate. However, because we want to select the best model on the basis of the CLIC, we used the most complex model D to identify a common distance $d$. According to the criterion (11), we found $d = (400, 3)$. Moreover, practical implementation of WCS for a dataset of such dimension exploits the monotonic behavior of the weights of the different weighted estimating equations for the correlation parameters. The preliminary estimates of the weights, calculated again by WLS method, are nearly zero for a couple of observations such that $|h| > 800, |u| > 6$. So, these pairs were dropped in the estimation procedure, allowing a remarkable computational gain.

The estimation results for models 1–4 are summarized in Table 4 for WLS, WCL, and WCS, where standard errors were estimated using subsampling techniques with a temporal

**Table 4.** WCL (upper), WCS (middle), and WLS (lower) estimates for the A–D correlation models. Standard errors are specified in parentheses.

| Parameter | Model | A    | B    | C    | D    |
|-----------|-------|------|------|------|------|
| $\alpha$  |       | 4.18(0.48) | 15.00(2.68) | 22.74(2.40) | 17.59(1.36) |
|           |       | 4.22(0.41) | 16.42(2.54) | 21.64(1.85) | 18.81(1.29) |
|           |       | 3.88 | 8.38 | 21.25 | 21.15 |
| $c$       |       | 1287.01(73.62) | 1281.25(72.83) | 1277.42(74.32) | 1281.25(74.46) |
|           |       | 1188.66(66.07) | 1182.61(66.42) | 1191.96(66.78) | 1182.86(64.61) |
|           |       | 1346.56 | 1304.06 | 1336.52 | 1336.06 |
| $\alpha$  |       | - | - | 1.00(0.20) | 1.00(0.21) |
|           |       | - | - | 1.00(0.29) | 1.00(0.33) |
|           |       | - | - | 1.00 | 1.00 |
| $\beta$   |       | - | - | - | 1.00(0.34) |
|           |       | - | - | - | 1.00(0.39) |
|           | CLIC  | -69797056 | -69908127 | -69777098 | -69766566 |

Figure 4. Locations of 513 ozone monitoring stations distributed across the eastern United States.
window as explained in Section 3.4. Estimations for WCL and WCS methods are similar and this fact is confirmed by Figure 5 where we compare the empirical variograms versus the estimated ones for model D.

In all cases, the estimated parametric variograms show a good agreement with the empirical ones. The CLIC criterion (see the last row of Table 4) suggests evidence in favor of the nonseparable hypothesis for the covariance models. However, the estimated values for models 3 and 4 are on the boundary of the parameter space, weakening the inferential conclusions using the standard errors and the CLIC.

In spite of this remark, the predictive performances of the four models were in favor of model 4. We considered the ozone measures during the last 60 days and calculated for each station $60 \times 513 = 30,780$ 1 day ahead predictions, $\hat{Z}(s, t)$, using the simple kriging predictor (Cressie 1993).

Kriging prediction is challenging for large datasets (Cressie and Johannesson 2008) and here we used as a predictor the last four daily observations recorded at the 513 stations.
Table 5. RMSE of the ozone data when using WLS, WCL, and WCS estimates of (a)–(d) covariance models.

| Model | A   | B   | C   | D   |
|-------|-----|-----|-----|-----|
| WCL   | 0.826 | 0.849 | 0.816 | 0.749 |
| WCS   | 0.826 | 0.847 | 0.815 | 0.742 |
| WLS   | 0.827 | 0.880 | 0.815 | 0.815 |

In order to assess and rank the point forecasts, we used the root mean square error (RMSE) RMSE = \sqrt{\sum_{i=1}^{513} \sum_{j=306}^{536} (Z(s_i, t_j) - \tilde{Z}(s_i, t_j))^2}/30780.

Looking at the RMSE results in Table 5, the nonseparable model D showed the best results for both WCL and WCS methods, with a slight advantage for the last one.

6. CONCLUSIONS

We proposed two methods for estimating covariance functions for datasets, which are large in space or in space time.

The first method is based on a particular instance of the CL of Curriero and Lele (1999). We have gained more statistical efficiency in considering a weighted version of the composite likelihood (WCL) where the weights are based on distances among the observed locations. Moreover, we have dealt with the model selection task, suggesting a suitable criterion that fits in the framework of CL. The effectiveness of this criterion has been shown by simulations.

In the second method, we have abandoned the construction of likelihood-type functions to resort on weighting score-type functions (WCS). Here, the weights are chosen minimizing an upper bound of the asymptotic variance of the estimates. In this regard, the weighting scheme proposed can be considered general and can be even applied in a nonspatial framework.

We have shown through examples and simulations that both methods give satisfactory results with respect to the REML estimates when we look for a reasonable trade-off between statistical and computational efficiency in dealing with large datasets. Our findings show that WCS performs better than WCL CL. However, WCS still requires more theoretical evidence especially when we consider multiparameter cases and it is heavier from computational point of view. Moreover, model selection based on an information criterion is an open issue even if solutions can be founded in a different context (Pan 2001).

Finally, we remark that even if our methods provide a feasible alternative for the estimation of the covariance function in large datasets, they do not tackle the prediction task and this will be a topic for future research.

APPENDIX A: PROOF OF PROPOSITION 1

We can write

\[ \tilde{Q}_n(\mathbf{h}) = \sum_{\mathbf{h} \in M} \left[ \log \gamma(\mathbf{h}; \theta) + \frac{\partial^2 \gamma_n}{\partial \theta_i \partial \theta_j}(\mathbf{h}) \right], \]

where \( \gamma_n(\mathbf{h}) = \sum_{s \in B_n} (Z(s) - \bar{Z}(s + \mathbf{h}))^2 \). Following the arguments in Guyon (1995), Theorem 3.4.1, we need to prove that

1. \( K(\theta_1, \theta_2) \) is the contrast function of the process \( \tilde{Q}_n(\mathbf{h}) \);
2. the modulus of continuity of \( \tilde{Q}_n(\mathbf{h}) \),

\[ W_n(\eta) = \sup_{\|\alpha - \beta\| \leq \eta, \|\alpha, \beta\| \in \Theta} |\tilde{Q}_n(\alpha) - \tilde{Q}_n(\beta)|, \]

is such that

\[ \lim_{n \to \infty} P_{0_0}(W_n(1/k) \geq \varepsilon_k) = 0, \]

for a positive sequence \( \varepsilon_k \) decreasing to zero.

1. By the weak law of large numbers, as \( n \) goes to \( \infty \), \( \hat{\gamma}_n(\mathbf{h}) \) converges in \( P_{0_0} \)-probability to \( \gamma(\mathbf{h}; \theta_0) \) and thus \( \tilde{Q}_n(\mathbf{h}) \) converges in \( P_{0_0} \)-probability to \( K(\theta_0, \theta_0) \), which has a unique minimum at \( \theta = \theta_0 \) by (C4).

2. We have

\[ |\tilde{Q}_n(\alpha) - \tilde{Q}_n(\beta)| \leq \sum_{\mathbf{h} \in M} \left[ \left| \frac{\partial \gamma_n(\mathbf{h})}{\partial \theta_i}(\mathbf{h}) \right| \left| \frac{1}{2\gamma_n(\mathbf{h}; \alpha)} - \frac{1}{2\gamma_n(\mathbf{h}; \beta)} \right| \right] \]

\[ \leq \left( c_1 + c_2 \sum_{\mathbf{h} \in M} 2\gamma_n(\mathbf{h}) \right) |\alpha - \beta| \]

for some positive constants \( c_1, c_2 \) by (C2) and (C3). For \( \|\alpha - \beta\| \leq 1/k \), there exists a sequence of numbers decreasing to zero, for \( k \to \infty \) such that \( c_1 \|\alpha - \beta\| \leq \varepsilon_k \). By the weak law of large numbers, as \( n \) goes to \( \infty \), \( c_1 \sum_{\mathbf{h} \in M} 2\gamma_n(\mathbf{h}) \) converges in \( P_{0_0} \)-probability to \( c_1 = c_2 \sum_{\mathbf{h} \in M} 2\gamma_n(\mathbf{h}; \theta_0) \). Since \( W_n(1/k) \leq \varepsilon_k + c_2 \sum_{\mathbf{h} \in M} 2\gamma_n(\mathbf{h}; 1/k) \), we have

\[ P_{0_0}(W_n(1/k) \geq \varepsilon_k + c_2/k) \leq P_{0_0} \left( c_2 \sum_{\mathbf{h} \in M} 2\gamma_n(\mathbf{h}) - c_3 > c_3 \right) \]

which converges to zero for \( n \to \infty \) since \( c_3 > 0 \).

\[ \square \]

APPENDIX B: PROOF OF PROPOSITION 2

We proceed by verifying the conditions (H1-H2-H3) of Theorem 3.4.5 in Guyon (1995).

(H1) If \( f \) has continuous partial derivatives of order 2 at \( \theta \in \mathbb{R}^p \), we define \( f_{ij}(\theta) = \frac{\partial^2 f}{\partial \theta_i \partial \theta_j} \), \( i, j = 1, \ldots, p \).

Since

\[ -\hat{Q}_{n,\delta}(\theta) = \sum_{\mathbf{h} \in M} \gamma_{ij}(\mathbf{h}; \theta) 2\gamma_n(\mathbf{h}) - 2\gamma_n(\mathbf{h}, \theta) \]

and

\[ \hat{Q}_{n,\delta}(\mathbf{h}) = \sum_{\mathbf{h} \in M} \gamma_{ij}(\mathbf{h}; \theta) \gamma_n(\mathbf{h}) - \gamma_{ij}(\mathbf{h}; \theta) \gamma_n(\mathbf{h}, \theta) \]

\[ \times \left[ \frac{\gamma_{ij}(\mathbf{h}) |\gamma_n(\mathbf{h})|^2 - 2\gamma_n(\mathbf{h}, \theta) \gamma_{ij}(\mathbf{h}, \theta) \gamma_n(\mathbf{h}, \theta)}{|\gamma_n(\mathbf{h})|^2} \right], \]

we have

\[ \sup_{\theta \in \Theta} |\hat{Q}_{n,\delta}(\theta)| \leq c_3 + c_4 \sum_{\mathbf{h} \in M} \gamma_n(\mathbf{h}), \]
with $E_{\theta_0}\sum_{h\in M}\hat{y}_n(h) < \infty$, for some positive constant $c_3$, $c_4$, by (N2).

(H2) Let $g_n(\theta_0) = (2\hat{y}_n(h_1) - 2\gamma(h_1, \theta_0), \ldots, 2\hat{y}_n(h_p, \theta_0))$, and $B(\theta)$ the diagonal matrix $q \times q$ with diagonal element $|\gamma(h, \theta)|^2$. Then we can write

$$\hat{Q}_n^{(1)}(\theta_0) = -\frac{1}{2} \Gamma(\theta_0) B(\theta_0) g_n(\theta_0).$$

By (N1) and Lemma A.1 and Theorem 3.3 in Lahiri, Lee, and Cressie (2002), we have

$$\mathcal{J}_n(\theta_0) = \text{var}(\hat{\gamma}_n^{(1)}(\theta_0)) \rightarrow \mathcal{J}(\theta_0) = \frac{1}{4} \Gamma(\theta_0) B(\theta_0) \Sigma(\theta_0) B(\theta_0) \Gamma(\theta_0),$$

where the $(i, j)$th element of $\Sigma(\theta_0)$ is given by

$$\Sigma_{ij}(\theta_0) = \sum_{h \in \mathbb{Z}^{p+1}} \text{cov}_{\theta_0}(Z(\theta) - Z(h))^2, (Z(s) - Z(s + h))^2),$$

and

$$|\mathcal{R}_n|^{-1/2} \sum_{i, j = 1}^{n} \hat{Q}_n^{(1)}(\theta_0) \Gamma^{-1}(\theta_0) \rightarrow \mathcal{N}(0, I_p),$$

as $n \rightarrow \infty$.

(H3) We have

$$E_{\theta_0}[|\hat{Q}_n^{(2)}(\theta)|^2] = H_{\theta_{ij}}(\theta_0) = \sum_{h \in \mathbb{Z}^{p+1}} \gamma_n^{(1)}(h, \theta_0) \gamma_n^{(1)}(h, \theta_0) |\gamma(h, \theta_0)|^2$$

$$= H_{\theta_{ij}}(\theta_0).$$

Now $H_{\theta_{ij}}(\theta_0)$ is the $(i, j)$th element of $H(\theta) = \Gamma(\theta_0) B(\theta_0) \Gamma(\theta_0)$, that by (N3) turns out to be positive definite. Finally by the weak law of large numbers, $(\hat{Q}_n^{(2)}(\theta_0)) - H_{\theta(\theta_0)}$ converges in distribution to $\theta$ in $P_0$—probability.

[Received February 2008, Revised April 2011]

REFERENCES

Andersen, E. W. (2004), “Composite Likelihood and Two-Stage Estimation in Family Studies,” Biostatistics, 5, 15–30. [269]

Caragea, P., and Smith, R. (2006), “Approximate Likelihoods for Spatial Processes,” Technical report, Iowa State University, Department of Statistics. [268]

Carlstein, A. (1986), “The Use of Subseries Values for Estimating the Variance of a General Statistic from a Stationary Sequence,” The Annals of Statistics, 14, 1171–1179. [274]

Cressie, N. (1993), Statistics for Spatial Data (revised ed.), New York: Wiley. [268,271,277,278]

Cressie, N., and Huang, H. (1999), “Classes of Nonseparable, Spatiotemporal Stationary Covariance Functions,” Journal of the American Statistical Association, 94, 1330–1340. [271]

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. [268,278]

Curriero, F., and Lele, S. (1999), “A Composite Likelihood Approach to Semivariogram Estimation,” Journal of Agricultural, Biological and Environmental Statistics, 4, 9–28. [268,269,279]

Davis, R., and Yau, C. Y. (2010), Comment on “Pairwise Likelihood in Time Series Models,” Statistica Sinica, 21, 255–277. [271]

Doukhan, P. (1994), Mixing. Properties and Examples, New York: Springer-Verlag. [273]

Fuentes, M. (2003), “Statistical Assessment of Geographic Areas of Compliance with Air Quality Standards,” Journal of Geophysical Research, 108(D24), 9002. [269,276]

Gilleland, E., and Nychka, D. (2005), “Statistical Models for Monitoring and Regulating Ground-Level Ozone,” Environmetrics, 16, 535–546. [276]

Gneiting, T. (2002), “Stationary Covariance Functions for Space-Time Data,” Journal of the American Statistical Association, 97, 590–600. [275]

Gneiting, T., Genton, M. G., and Guttorp, P. (2007), “Geostatistical Space-Time Models, Stationarity, Separability and Full Symmetry,” in Statistical Methods for Spatio-Temporal Systems, eds. B. Finkenstadt, L. Held, and V. Isham, Boca Raton: FL: Chapman & Hall/CRC, pp. 151–175. [271]

Guyon, X. (1995), Random Fields on a Network, New York: Springer. [273,279]

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. [274]

Heyde, C. (1997), Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation, New York: Springer. [269,271]

Heyde, C. C., and Morton, R. (1998), “Multiple Roots in General Estimating Equations,” Biometrika, 85, 954–959. [271]

Hoeting, A., Davis, A., Merton, A., and Thompson, S. (2006), “Model Selection for Geostatistical Models,” Ecological Applications, 16, 87–98. [274]

Huang, H., and Chen, C. (2007), “Optimal Geostatistical Model Selection,” Journal of American Statistical Association, 102, 1009–1023. [274]

Huang, H., Martinez, F., Mateu, J., and Montes, F. (2007), “Model Comparison and Selection for Spatial Stationary Time-Measures,” Computational Statistics and Data Analysis, 51, 4577–4596. [269]

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. [268,271]

Kuk, A. (2007), “A Hybrid Pairwise Likelihood Method,” Biometrika, 94, 939–952. [269]

Lahiri, S., Lee, Y., and Cressie, N. (2002), “Efficiency of Least Squares Estimators of Spatial Variogram Parameters,” Journal of Statistical Planning and Inference, 103, 65–85. [280]

Lee, Y., and Lahiri, S. (2002), “Least Squares Variogram Fitting by Spatial Subsampling,” Journal of the Royal Statistical Society, Series B, 64, 837–854. [274]

Li, B., Genton, M., and Sherman, M. (2008), “On the Asymptotic Joint Distribution of Sample Space-time Covariance Estimators,” Bernoulli, 14, 228–248. [272,274]

Lindsay, B. (1988), “Composite Likelihood Methods,” Contemporary Mathematics, 80, 221–239. [268,269]

Pan, W. (2001), “Model Selection in Estimating Equations,” Biometrics, 57, 529–534. [279]

Rue, H., and Tjelmeland, H. (2002), “Fitting Gaussian Markov Random Fields to Gaussian Fields,” Scandinavian Journal of Statistics, 29, 31–49. [268]

Sherman, M. (1996), “Variance Estimation for Statistics Computed from Spatial Lattice Data,” Journal of the Royal Statistical Society, Series B, 58, 509–523. [274]

Stein, M. (1999), Interpolation of Spatial Data. Some Theory of Kriging, New York: Springer-Verlag. [272]

——— (2005), “Statistical Methods for Regular Monitoring Data,” Journal of the Royal Statistical Society, Series B, 67, 667–687. [268]

——— (2008), “A Modeling Approach for Large Spatial Datasets,” Journal of the Korean Statistical Society, 37, 3–10. [268]

Stein, M., Chl. Z., and Wely, L. (2004), “Approximating Likelihoods for Large Spatial Data Sets,” Journal of the Royal Statistical Society, Series B, 66, 275–296. [268,271]

Varin, C., and Vidoni, P. (2005), “A Note on Composite Likelihood Inference and Model Selection,” Biometrika, 52, 519–528. [269,274]

Vecchia, A. (1988), “Estimation and Model Identification for Continuous Spatial Processes,” Journal of the Royal Statistical Society, Series B, 50, 297–312. [268]

Whittle, P. (1954), “On Stationary Processes in the Plane,” Biometrika, 49, 315–341. [268]

Zhang, H. (2004), “Inconsistent Estimation and Asymptotically Equivalent Interpolations in Model-Based Geostatistics,” Journal of the American Statistical Association, 99, 250–261. [272]

Zhang, H., and Zimmerman, D. (2005), “Towards Reconciling Two Asymptotic Frameworks in Spatial Statistics,” Biometrika, 92, 921–936. [272]