Journal of the
Royal Statistical Society

SERIES B
Statistical
Methodology

J. R. Statist. Soc. B (2012)
74, *Part* 5, *pp.*

Joint composite estimating functions in spatiotemporal models

Yun Bai, Peter X.-K. Song and T. E. Raghunathan

University of Michigan, Ann Arbor, USA

[Received September 2010. Final revision February 2012]

Summary. Modelling of spatiotemporal processes has received considerable attention in recent statistical research. However, owing to the high dimensionality of the data, the joint modelling of spatial and temporal processes presents a great computational challenge, in both likelihood-based and Bayesian approaches. We propose a joint composite estimating function approach to estimating spatiotemporal covariance structures. This substantially reduces the computational complexity and is more efficient than existing composite likelihood methods. The novelty of the proposed joint composite estimating function is rooted in the construction of three sets of estimating functions from spatial, temporal and spatiotemporal cross-pairs, which results in overidentified estimating functions. Thus, we form a joint inference function in a spirit that is similar to Hansen's generalized method of moments. We show that under practical scenarios the estimator proposed is consistent and asymptotically normal. Simulation studies prove that our method performs well in finite samples. Finally, we illustrate the joint composite estimating function method by estimating the spatiotemporal dependence structure of airborne particulates (PM10) in the north-eastern USA over a 32-month period.

Keywords: Asymptotics; Correlated data; Dimension reduction; Generalized method of moments; Quadratic inference function

1. Introduction

Spatiotemporal data arise from many scientific disciplines such as environmental sciences, climatology, geology and epidemiology among others. Through data analysis, scientists are interested in understanding important factors that are associated with the underlying process and in predicting the process at unobserved locations and time points. Both of these tasks require modelling the intrinsic dependence structure of the data, which is usually depicted by the spatiotemporal covariance structure. During recent decades, much effort has been made in developing valid yet flexible spatiotemporal covariance models. For example, Cressie and Huang (1999) introduced a class of non-separable, stationary covariance functions that address space–time interactions. Gneiting (2002) later expanded their work to larger classes of space–time covariance structures that do not depend on closed form Fourier inversions. Stein (2005) derived space–time covariance functions that are spatially isotropic and not fully symmetric. Porcu *et al.* (2007) proposed another class of non-separable space–time covariance structures that are spatially anisotropic; these allow us to formulate temporally asymmetric covariance functions. Unfortunately, most of these useful covariance models are seldom applied in practical studies collecting large-scale data sets. This is largely because of the tremendous computational burden in handling high dimensional covariance matrices for either likelihood-based or Bayesian approaches.

Address for correspondence: Peter X.-K. Song, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.
E-mail: pxsong@umich.edu

1369–7412/12/74000

This difficulty has long been recognized in spatial statistics, where two types of approaches have been developed to speed computation. The first approach is based on simplifying covariance structures. For stationary spatial processes on regular grids, Zimmerman (1989) showed that covariance structures have patterns that can be used to reduce the computational burden. Cressie and Johannesson (2008) proposed fixed rank kriging for very large spatial data sets, where the covariance matrices were specially designed so that the matrix manipulations were of a fixed magnitude. A similar idea was exploited in Banerjee *et al.* (2008). However, these approaches either require the spatial process to be stationary or impose oversimplified structures for the covariance matrices. Thus they may not work well with real data analysis.

Another approach is based on likelihood approximations, where simplified versions of the full likelihood are considered. For example, composite likelihood (CL) methods (Lindsay, 1988) have been proposed to model spatial data. As a general class of pseudolikelihoods, CL is based on valid marginal or conditional likelihood functions. Curriero and Lele (1999), Heagerty and Lele (1998) and Li and Lin (2006) all used pairwise marginal densities to build CL estimation functions, whereas Vecchia (1988) and Stein *et al.* (2004) suggested approximating the likelihood by a product of conditional densities with truncated conditioning sets. Apart from CL approaches, Furrer *et al.* (2006) and Kaufman *et al.* (2008) used covariance tapering to shrink small values of covariance entries to 0 so that sparse matrix algorithms could be used to speed up computation. Fuentes (2007) proposed an approximation by modelling the covariance structures in the spectral domain. This appears to be quite involved and hence less useful.

Additional challenges arise in spatiotemporal settings. With the addition of time, the data scale becomes much larger. Also, the distinct, yet intricately involved, nature of space and time further complicates the data analysis. To simplify covariance structures, people usually separately model spatial and temporal dependences (Sahu *et al.*, 2007; Smith and Kolenikov, 2003) or apply a separable spatiotemporal covariance function for ease of computation (Haas, 1995; Genton, 2007). Although these approaches have many desirable properties, they all ignore a crucial model component: the spatiotemporal interaction effect. Paciorek *et al.* (2009) attempted to capture the spatiotemporal interaction of PM10- and PM2.5-particles by using monthly varying spatial surfaces. However, to reduce computational difficulty, they assumed independence across spatial residual surfaces at each time point. This hampered their ability to quantify spatiotemporal interaction.

The objective of this paper is to develop an efficient CL approach for the joint analysis of spatiotemporal processes. We propose to use pairwise marginal densities as the building blocks of our estimating function, for the following reasons.

(a) Pairwise CL is both analytically and numerically simple to work with.

(b) It requires only the correct specification of bivariate densities; hence the resulting estimation and inference are robust to misspecification of higher dimensional moment structures (Varin *et al.*, 2011). In contrast, conditional CL approaches (e.g. Vecchia (1988) and Stein *et al.* (2004)) are usually vulnerable to model misspecification, because they require the formulation of higher dimensional distributions. In addition, it is easier to check assumptions on bivariate distributions than on higher dimensional distributions.

(c) The pairwise CL approach does not require a distance metric accommodating both space and time, in contrast with the unified distance norm that is needed by the tapering approach (Kaufman *et al.*, 2008). As we shall see in the simulation experiments (see Section 4.2), tapering expedites computing time only when the number of non-zero covariance elements is small.

Pairwise CL does seem appealing in modelling large-scale spatiotemporal data, owing to its simplicity, flexibility and feasibility for numerical computation.

Another contribution of our proposed method is that it overcomes a major shortcoming of the conventional CL estimation method, which treats pairwise observations as independent. We propose to account for correlations between these composite pairs in a computationally feasible manner, leading to significant gains in efficiency. Nott and Rydén (1999) and Kuk and Nott (2000) are among those researchers who have advocated the incorporation of correlations between composite pairs into the estimation. However, their methods appear difficult to implement in spatiotemporal settings, owing to the enormous number of possible pairs.

Our idea is to take advantage of the distinct characteristics of space and time by dividing all pairs into spatial, temporal and spatiotemporal cross-groups, and then to form group-based estimating functions. After that we construct a joint inference function with different weights for the groups to improve efficiency. This approach is similar to the generalized method of moments (GMM) (Hansen, 1982) and the quadratic inference function (Qu *et al.*, 2000). The weighting scheme is designed to give larger weights to more informative pairs and to downweight noisy pairs, leading to gains in efficiency.

The rest of the paper is structured as follows. In Section 2, we present the joint composite estimating function approach for spatiotemporal processes. In Section 3, we discuss large sample properties of the estimator proposed. In Section 4, we detail simulation studies comparing our method with some of the popular likelihood approximation methods. In Section 5, we illustrate an application of our method to study the spatiotemporal dependence structure of airborne PM10-particles in the north-eastern USA. A discussion follows in Section 6. Some technical details are listed in Appendix A.

2. Methodology

2.1. Model

Consider a realization of a spatiotemporal process $\{Y(s,t): s\in\mathcal{S}, t\in\mathcal{T}, \mathcal{S}\subset\mathbb{R}^2, \mathcal{T}\subset\mathbb{R}^+\}$, where $\mathcal{S}$ denotes the set of spatial locations and $\mathcal{T}$ stands for the collection of time points. Assume that $Y(s,t)$ can be decomposed into a deterministic mean function $\mu(s,t)$ and a random component $X(s,t)$ as follows:

$$Y(s,t)=\mu(s,t)+X(s,t), \qquad s\in\mathcal{S}, \quad t\in\mathcal{T}.$$

Suppose that $X(s,t)$ can be further modelled as

$$X(s,t)=\alpha(s,t)+\varepsilon(s,t), \qquad s\in\mathcal{S}, \quad t\in\mathcal{T},$$

where the process $\alpha(s,t)$ characterizes the spatiotemporal variations, and $\varepsilon(s,t)$ is a normally distributed measurement error with mean 0 and variance σ_ε^2, independent of each other and independent of $\alpha(s,t)$. In geostatistics, the variance σ_ε^2 is called the 'nugget effect'. Assume that $\{\alpha(s,t)$: $s\in\mathcal{S}$, $t\in\mathcal{T}\}$ follows a multivariate Gaussian process with mean 0 and covariance function C which, for any two observations at spatiotemporal co-ordinates (s_1,t_1) and (s_2,t_2), is given by

$$\operatorname{cov}\{\alpha(s_1,t_1),\alpha(s_2,t_2)\}=C(s_1,s_2,t_1,t_2;\boldsymbol{\theta}').$$

Let $\boldsymbol{\theta}=(\boldsymbol{\theta}',\sigma_\varepsilon^2)$ be an r-element vector of parameters of interest. We shall focus on estimating the covariance structure of $X(s,t)$ in the rest of this paper, provided that the observed process has first been properly detrended; otherwise, it is relatively straightforward to extend the method proposed by including a mean model for $\mu(s,t)$ (Cressie, 1993).

2.2. Composite estimating functions

To apply the CL method, we consider pairwise differences following Curriero and Lele (1999). Let

$$d(k) \equiv d(s_1,t_1,s_2,t_2) = X(s_1,t_1) - X(s_2,t_2), \qquad k \in D_n(p,q), \tag{1}$$

where

$$D_n(p,q) = \left\{ (s_1,t_1,s_2,t_2) : \begin{array}{l} s_2 \geqslant s_1, \|s_1 - s_2\| \leqslant p, \\ t_2 \geqslant t_1, |t_1 - t_2| \leqslant q, \\ t_1 \neq t_2 \text{ if } s_1 = s_2, \\ s_1 \neq s_2 \text{ if } t_1 = t_2 \end{array} \right.$$
$$\subset \mathcal{S} \times \mathcal{T} \times \mathcal{S} \times \mathcal{T} \subseteq \mathbb{R}^2 \times \mathbb{R}^+ \times \mathbb{R}^2 \times \mathbb{R}^+.$$

Here n is the length of the realized process $X(s,t)$, and $\|\cdot\|$ is the Euclidean distance between two points in a d-dimensional space with $d \geqslant 2$. The ordering of spatial locations is defined as follows: for two locations $s_1 = (a_1,b_1)$ and $s_2 = (a_2,b_2)$, we say that $s_1 > s_2$ if $a_1 > a_2$ or if $a_1 = a_2$ and $b_2 > b_1$, where (a,b) are the co-ordinates for a location. The set $D_n(p,q)$ contains all pairs of observations within p units of space and q units of time lags in the co-ordinate space $\mathcal{S} \times \mathcal{T}$. When both p and q are infinite, the set includes all possible pairs of observations. For simplicity of exposition, we drop the two indices and write $D_n(p,q)$ as D_n.

The values of p and q may be determined according to different criteria. They can be chosen by practical considerations, such as sample size or boundary limits. They can also be determined by some preliminary evaluations (e.g. empirical variograms) of the spatial and temporal dependence decay rate and set to ranges that sustain a fairly high level of correlation. Or we may choose such p and q as to maximize the Godambe information (Godambe and Heyde, 1987) of the corresponding composite estimating functions, so that the resulting estimator will have minimum variance estimates (Bevilacqua *et al.*, 2011). Clearly, the latter approach requires the evaluation of the sandwich information matrix for different combinations of cut-off lags, which is computationally demanding. Many simulation results reported in the literature (e.g. Varin *et al.* (2005), Bevilacqua *et al.* (2011) and Davis and Yau (2011)) have suggested choosing p and q to include only pairs that are within some short distance, for better estimation efficiency. If we do that, we can exclude a substantial number of pairs from D_n that are far apart in either space or time to reduce the computational burden.

It is easy to see that the difference process $d(k)$ in equation (1) follows a univariate normal distribution with mean 0 and variance given by

$$\begin{aligned} \operatorname{var}\{d(k)\} &= C(s_1,s_1,t_1,t_1;\boldsymbol{\theta}) + C(s_2,s_2,t_2,t_2;\boldsymbol{\theta}) + 2\sigma_\varepsilon^2 - 2C(s_1,s_2,t_1,t_2;\boldsymbol{\theta}) \\ &\equiv 2\gamma_k(\boldsymbol{\theta}). \end{aligned}$$

Denote the composite score function for the observed $d(k)$ as $f_k\{d(k);\boldsymbol{\theta}\}$. Then

$$f_k\{d(k);\boldsymbol{\theta}\} = \frac{\dot{\gamma}_k(\boldsymbol{\theta})}{2\gamma_k(\boldsymbol{\theta})} \left\{ 1 - \frac{d^2(k)}{2\gamma_k(\boldsymbol{\theta})} \right\},$$

where, for any function f, $\dot{f}$ denotes the vector of gradients of f with respect to the parameter vector $\boldsymbol{\theta}$. It is clear that $f_k\{d(k);\boldsymbol{\theta}\}$ is an unbiased estimating function for $\boldsymbol{\theta}$ since it is derived from a valid density function.

According to the CL literature (Reid and Cox, 2004; Varin *et al.*, 2011), a common version of composite estimating functions is

$$\Psi_{\mathrm{CL}}(\boldsymbol{\theta}) = \sum_{k \in D_n} f_k\{d(k); \boldsymbol{\theta}\},$$

where $d(k)$ are implicitly treated as being independent.

Alternatively, one may stack the individual composite score function terms into a column vector $\nu(\boldsymbol{\theta}) = \{f_k\{d(k); \boldsymbol{\theta}\}\}_{k \in D_n}$, from which the estimating function is given by

$$E\{\dot{\nu}(\boldsymbol{\theta})\}^{\mathrm{T}} \operatorname{cov}\{\nu(\boldsymbol{\theta})\}^{-1} \nu(\boldsymbol{\theta}) = 0.$$

As pointed out by Kuk (2007), this version of composite estimating equations effectively accounts for the correlations between the differences. However, the calculation of $\operatorname{cov}\{\nu(\boldsymbol{\theta})\}$ and its inverse is computationally prohibitive when the number of pairs (or differences) is large.

To improve on the existing CL methods and to incorporate correlations between the pairs in the estimation, we propose a new approach, i.e. we construct three sets of estimating functions by using the spatiotemporal characteristics of the data. Specifically, we first partition D_n into three subsets, namely $D_{S,n}$, with pairs differing only in locations, $D_{T,n}$, with pairs differing only in time and $D_{C,n}$, with pairs differing in both locations and time. Hence $D_n = D_{S,n} \cup D_{T,n} \cup D_{C,n}$. Fig. 1 displays such a partition with the three types of pairs,

(a) for a spatial pair,
(b) for a temporal pair and
(c) for a spatiotemporal cross-pair, in a typical spatiotemporal setting with four locations observed at two time points.

Summing over all pairwise differences of spatial pairs across all time points, we obtain the following spatial composite estimating function (CEF):

$$\Psi_{S,n}(\boldsymbol{\theta}) = \frac{1}{|D_{S,n}|} \sum_{i \in D_{S,n}} f_i\{d(i); \boldsymbol{\theta}\},$$

where, for any set $\mathcal{A}$, $|\mathcal{A}|$ denotes the number of elements in $\mathcal{A}$. In a similar fashion, we construct the temporal CEF:

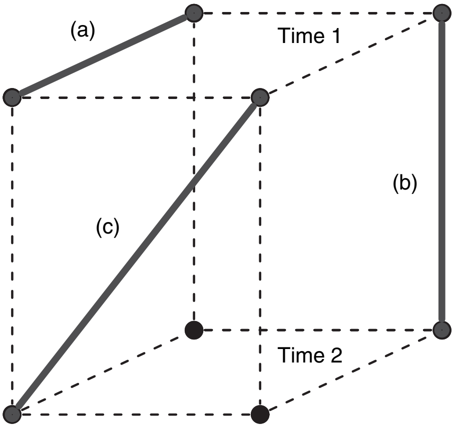

Fig. 1. Configurations of spatiotemporal pairs: the upper plane represents four locations observed at time 1, and the lower plane represents the same four locations observed at time 2; (a) is the spatial pair, (b) the temporal pair and (c) the spatiotemporal cross-pair

$$\Psi_{T,n}(\boldsymbol{\theta}) = \frac{1}{|D_{T,n}|} \sum_{j \in D_{T,n}} f_j\{d(j); \boldsymbol{\theta}\}.$$

Likewise, the third CEF is formed by using spatiotemporal cross-pairs:

$$\Psi_{C,n}(\boldsymbol{\theta}) = \frac{1}{|D_{C,n}|} \sum_{l \in D_{C,n}} f_l\{d(l); \boldsymbol{\theta}\}.$$

Note that the resulting estimating functions constructed by using the group-specific pairs characterize different profiles of the spatiotemporal process. The spatial piece $\Psi_{S,n}(\boldsymbol{\theta})$ provides paramount information about the spatial dependence; the temporal piece $\Psi_{T,n}(\boldsymbol{\theta})$ contains key information about the temporal dependence; and the spatiotemporal cross-piece $\Psi_{C,n}(\boldsymbol{\theta})$ is more relevant for information about the spatiotemporal interaction. The total number of equations, when the three pieces are combined as $(\Psi^{\mathrm{T}}_{S,n}(\boldsymbol{\theta}), \Psi^{\mathrm{T}}_{T,n}(\boldsymbol{\theta}), \Psi^{\mathrm{T}}_{C,n}(\boldsymbol{\theta}))^{\mathrm{T}}$, is larger than the number of parameters. As a result, owing to overidentification, parameters cannot be estimated by directly solving these equations. Thus, we form a weighted quadratic objective function in a spirit similar to the GMM (Hansen, 1982), so that estimates can be obtained by minimizing this objective function.

More precisely, let W be a positive definite matrix, and let

$$\Gamma_n(\boldsymbol{\theta}) = (\Psi^{\mathrm{T}}_{S,n}(\boldsymbol{\theta}), \Psi^{\mathrm{T}}_{T,n}(\boldsymbol{\theta}), \Psi^{\mathrm{T}}_{C,n}(\boldsymbol{\theta}))^{\mathrm{T}}.$$

A quadratic inference function takes the form

$$Q_n(\boldsymbol{\theta}) = \Gamma^{\mathrm{T}}_n(\boldsymbol{\theta}) W^{-1} \Gamma_n(\boldsymbol{\theta}),$$

and the estimator is given by

$$\hat{\boldsymbol{\theta}}_n = \arg \min_{\boldsymbol{\theta} \in \Theta} Q_n(\boldsymbol{\theta}). \tag{2}$$

We call this $\hat{\boldsymbol{\theta}}_n$ the joint composite estimating function (JCEF) estimator.

Classical GMM theory indicates that the optimal weight matrix is the asymptotic covariance matrix of CEFs, namely $\mathrm{cov}\{n \Gamma_n(\boldsymbol{\theta})\}$. However, this result cannot be directly applied here, because our objective function $Q_n(\boldsymbol{\theta})$ is special in two aspects. First, the three estimating functions $\Psi_{S,n}(\boldsymbol{\theta})$, $\Psi_{T,n}(\boldsymbol{\theta})$ and $\Psi_{C,n}(\boldsymbol{\theta})$ are constructed from different sets of observations, whereas, in the standard GMM, different moment conditions are based on the same set of observations. Second, the numbers of terms in the three CEFs are different, because the numbers of spatial, temporal and cross-pairs are different. When one CEF consists of significantly more pairs, it will attain a higher weight in the objective function due to its larger stratum size. So it is necessary to adjust for such stratum effects by using a normalized weight matrix, in a spirit similar to stratified sampling.

To proceed, let I_r be the $r \times r$ identity matrix. Write

$$\sqrt{}\mathcal{N} = \mathrm{diag}(\sqrt{}|D_{S,n}|, \sqrt{}|D_{T,n}|, \sqrt{}|D_{C,n}|) \otimes I_r,$$

where '⊗' denotes the Kronecker product of two matrices. This defines a block diagonal matrix with the first r diagonals being $\sqrt{}|D_{S,n}|$, the next r diagonals $\sqrt{}|D_{T,n}|$ and the last r diagonals $\sqrt{}|D_{C,n}|$. The normalized weight matrix is given by

$$W = \sqrt{}\mathcal{N} \,\mathrm{cov}\{\Gamma_n(\boldsymbol{\theta})\} \sqrt{}\mathcal{N}.$$

When $|D_{S,n}|$, $|D_{T,n}|$ and $|D_{C,n}|$ are approximately the same, W and $\mathrm{cov}\{n \Gamma_n(\boldsymbol{\theta})\}$ will play the same role in weighting. However, when one of $|D_{S,n}|$, $|D_{T,n}|$ or $|D_{C,n}|$ is considerably larger than the rest, W will help to adjust for the unbalanced stratum sizes, so the smaller stratum will make a comparable contribution to the estimation. Zhao and Joe (2005) used a similar approach to account for different cluster sizes in their CL formulation for familial data. See also Joe and Lee (2009) for a more detailed discussion.

2.3. Estimation of the weight matrix

Although $\text{cov}\{\Gamma_n(\boldsymbol{\theta})\}$ can be derived analytically by using multivariate Gaussian quadrant probabilities, given the large number of possible pairs, computing it on the basis of analytic formulae is not practically feasible. Alternatively, in spatial data analysis, estimation of this covariance matrix is typically achieved by subsampling techniques as done in Heagerty and Lele (1998), Heagerty and Lumley (2000), Lee and Lahiri (2002) and Li and Lin (2006). Specifically, let the sampling region $A_n = \mathcal{S} \times \mathcal{T}$, where $|A_n| = n$. Under the assumption that asymptotically $|A_n|\, E\{\Gamma_n(\boldsymbol{\theta})\,\Gamma_n^{\mathrm{T}}(\boldsymbol{\theta})\} \rightarrow \Sigma$, we can estimate Σ by using the sample covariance matrix of statistics computed on subshapes of the sampling region A_n, i.e.

$$\hat{\Sigma}_n = k_n^{-1} \sum_{i=1}^{k_n} |A^i_{l(n)}| \{\Gamma^i_n(\boldsymbol{\theta}) - \bar{\Gamma}_n(\boldsymbol{\theta})\}^2, \tag{3}$$

with $\bar{\Gamma}_n(\boldsymbol{\theta}) = \Sigma_{i=1}^{k_n} \Gamma^i_n(\boldsymbol{\theta})/k_n$, where $\Gamma^i_n(\boldsymbol{\theta})$ is vector $\Gamma_n(\boldsymbol{\theta})$ evaluated in $A^i_{l(n)}$, $i = 1, \ldots, k_n$, a collection of (non-)overlapping subshapes of A_n and k_n is the number of subshapes.

This subsampling method was first introduced by Carlstein (1987) for strictly stationary time series. Sherman (1996) later showed that it could be used to estimate the moments of a general statistic for random fields on a lattice. Moreover, Kunsch (1989) demonstrated that the use of overlapping replicates led to a more stable variance estimate than non-overlapping replicates. The optimal subsample size was given by Politis and Romano (1994) for a stationary random field on a d-dimensional lattice as $Mn^{d/(d+2)}$, where M is a certain tuning constant. Heagerty and Lumley (2000) studied the effect of various choices of M for regression models. Sherman (1996) pointed out that it was useful to gather some empirical evidence about the range of correlation in determining M. If the correlation decays fast, small subsamples can be used; otherwise large subsamples should be considered.

We shall apply this subsampling technique to estimate our weight matrix and later investigate its performance in the standard error calculation. Other and more sophisticated resampling schemes for spatial data analysis can be found in Lele (1991), Lahiri *et al.* (1999) and Zhu and Morgan (2004) among others. To calculate $\text{cov}\,\{\Gamma_n(\boldsymbol{\theta})\}$ for each subsample, parameter values must be given. We propose to generate some simple consistent estimates either by setting the weight matrix to the identity matrix in the JCEF method or by using estimates from the empirical variogram.

Many established numerical optimization methods can be used to obtain parameter estimates that minimize $Q_n(\boldsymbol{\theta})$. However, given the complex nature of the parametric covariance structure $C(\cdot;\boldsymbol{\theta})$, algorithms that do not require calculations of the Hessian matrix are desirable. Quasi-Newton, Nelder–Mead and conjugate gradient methods are possible choices. These optimization routines are offered by many mathematical and statistical software packages, including MATLAB and R. To ensure that the true minimum of the target function is found, a set of good starting values is very important. In our case, this can be found by fitting the corresponding parametric variogram to the empirical variogram (Cressie, 1993). Details are illustrated in Section 5.

3. Large sample properties

The asymptotic properties of the JCEF estimator that is defined in equation (2) are mainly governed by the asymptotic behaviour of $\Gamma_n(\boldsymbol{\theta})$. Once we establish a uniform law of large numbers and a central limit theorem for $\Gamma_n(\boldsymbol{\theta})$, the consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_n$ will follow from standard GMM arguments. We derive these large sample results for fixed spatial and temporal lags p and q under increasing domain asymptotics, i.e. the increase in sample size is achieved by the expansion of the sampling domain in space or time or simultaneously. As a result of fixing p and q, the numbers of pairs in the spatial, temporal and spatiotemporal cross-groups are proportional to the total number of data points n for the observed random process, i.e. $|D_{S,n}|$, $|D_{T,n}|$ and $|D_{C,n}|$ are of the same order $O(n)$. For simplicity, we assume that the weight matrix W is known. Otherwise, a root-n-consistent $\hat{W}$ would be sufficient for us to modify our justifications.

3.1. Assumptions

Jenish and Prucha (2009) developed a set of limit theorems for random processes under rather general conditions of non-stationarity, unevenly spaced locations and general forms of sample regions. We tailor the relevant regularity conditions to establish large sample properties for our JCEF estimator as follows.

Assumption 1. The (possibly unevenly spaced) lattice $D \subset \mathbb{R}^2 \times \mathbb{R}^+ \times \mathbb{R}^2 \times \mathbb{R}^+$ is infinitely countable. All elements in D are at distances of at least $d_0 > 0$ from each other, i.e. $\rho(i, j) \geqslant d_0$, for all $i, j \in D$, where $\rho(i, j)$ is a distance metric for any two points $i, j \in D$. See a detailed definition of the distance metric in Appendix A.

Assumption 2. $\{D_{\mathcal{A},n} : n \in \mathbb{N}\}$ is a sequence of arbitrary finite subsets of D, satisfying $|D_{\mathcal{A},n}| \rightarrow \infty$ as $n \rightarrow \infty$, for $\mathcal{A} \in \{S, T, C\}$.

Assumption 3. $(\boldsymbol{\Theta}, \upsilon)$ is a totally bounded parameter space with metric υ.

Assumption 4 (uniform $L_{2+\delta}$ integrability). Let $q_k = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \| f_k\{d(k); \boldsymbol{\theta}\}\|$. Then, for some $\delta > 0$, $\lim_{e\rightarrow\infty} E\{q_k^{2+\delta}\mathbf{1}(\|q_k\| > e)\} = 0$, for all $k \in D_n$.

Assumption 5. $E \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\dot{f}_k\{d(k); \boldsymbol{\theta}\}\| < \infty$, for all $k \in D_n$.

Assumption 1 ensures that the increase in sample size is achieved by an expanding domain; thus, it rules out the infill asymptotics. Assumption 2 guarantees that sequences of subsets $D_{S,n}$, $D_{T,n}$ and $D_{C,n}$, on which the process is generated, increases in cardinality. Assumption 3 regulates the parameter space. Assumptions 4 and 5 are regularity conditions for score functions. The uniform integrability condition in assumption 4 is a standard moment assumption postulated in central limit theorems for one-dimensional processes. A sufficient condition for the uniform $L_{2+\delta}$ integrability of f_k is its uniform L_γ-boundedness for some $\gamma > 2 + \delta$. A weaker assumption of L_1 integrability is sufficient for a law of large numbers for f_k. Assumption 5 is a Lipschitz-type condition, implying that the score functions are L_0 stochastically equicontinuous, so that a uniform law of large numbers can be obtained.

The difference process $d(k)$ is usually not stationary. To regulate its dependence structure, we impose some α-mixing conditions on $d(k)$. Let U and V be two subsets of D_n, and let $\sigma(U) = \sigma\{d(k); k \in U\}$ be the σ-algebra generated by random variables $d(k)$, $k \in U$. Define

$$\alpha(U, V) = \sup\{|P(A \cap B) - P(A)\,P(B)|; A \in \sigma(U), B \in \sigma(V)\}.$$

Then the α-mixing coefficient for the random field $\{d(k), k \in D_n\}$ is defined as

$$\alpha(k, l, m) = \sup\{\alpha(U, V), |U| < k, |V| < l, \rho(U, V) \geqslant m\},$$

with $k, l, m \in \mathbb{N}$ and $\rho(U, V)$ the distance between sets U and V; see Appendix A for the definition of ρ. In addition, we need the following conditions which are similar to those stated in assumption 3 (Jenish and Prucha, 2009).

Assumption 6. The process $\{d(k), k \in D_n\}$ satisfies the following mixing conditions in an a-dimensional space:

(a) $\Sigma_{m=1}^{\infty} m^{a-1} \alpha(1, 1, m)^{\delta/(2+\delta)} < \infty$, for some $\delta > 0$;
(b) $\Sigma_{m=1}^{\infty} m^{a-1} \alpha(k, l, m) < \infty$ for $k + l \leqslant 4$;
(c) $\alpha(1, \infty, m) = O(m^{-a-\varepsilon})$ for some $\varepsilon > 0$.

Assumption 6 requires a polynomial decay of the α-mixing coefficient, which can be shown to hold for Gaussian processes, a special case of a Gibbs field (Winkler, 1995; Doukhan, 1994).

3.2. Consistency

Consider a generic case of

$$\Psi_{\mathcal{A},n}(\boldsymbol{\theta}) = \frac{1}{|D_{\mathcal{A},n}|} \sum_{k \in D_{\mathcal{A},n}} f_k\{d(k); \boldsymbol{\theta}\},$$

where $\mathcal{A} \in \{S, T, C\}$.

On the basis of theorems 2 and 3 in Jenish and Prucha (2009), assumptions 1, 2, 4 and 6 ensure a pointwise law of large numbers for f_k based on subseries $\{d(k), k \in D_{\mathcal{A},n}\}$; with additional assumption 5 on stochastic equicontinuity of f_k, a uniform version of the law of large numbers is warranted. Thus, we have the following lemma.

Lemma 1. Given assumptions 1–6,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \| \Psi_{\mathcal{A},n}(\boldsymbol{\theta}) - E\{\Psi_{\mathcal{A},n}(\boldsymbol{\theta})\} \| \overset{\mathrm{p}}{\rightarrow} 0, \qquad \text{as } n \rightarrow \infty.$$

Lemma 1 holds for $\Psi_{S,n}(\boldsymbol{\theta})$, $\Psi_{T,n}(\boldsymbol{\theta})$ and $\Psi_{C,n}(\boldsymbol{\theta})$, so we can show easily that, for any given positive definite weight matrix W,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} | Q_n(\boldsymbol{\theta}) - E\{Q_n(\boldsymbol{\theta})\} | \overset{\mathrm{p}}{\rightarrow} 0, \qquad \text{as } n \rightarrow \infty.$$

Consequently, we establish the consistency of the JCEF estimator in theorem 1.

Theorem 1. Under the same conditions stated in lemma 1, if the true parameter value $\boldsymbol{\theta}_0$ is the unique minimizer of $E\{Q_n(\boldsymbol{\theta})\}$, and $\hat{\boldsymbol{\theta}}_n$ minimizes $Q_n(\boldsymbol{\theta})$, then $\hat{\boldsymbol{\theta}}_n \rightarrow^{\mathrm{p}} \boldsymbol{\theta}_0$, as $n \rightarrow \infty$.

3.3. Asymptotic normality

To derive the asymptotic distribution of the JCEF estimator, the following additional regularity conditions are needed.

Assumption 7. Let $\Sigma_n(\boldsymbol{\theta}) = \mathrm{var}\{\Gamma_n(\boldsymbol{\theta})\}$, $\lim_{n \rightarrow \infty} n \Sigma_n(\boldsymbol{\theta}) = \Sigma(\boldsymbol{\theta})$, where $\Sigma(\boldsymbol{\theta})$ is a positive definite matrix.

Assumption 8. $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \| \dot{\Gamma}_n(\boldsymbol{\theta}) - E\{\dot{\Gamma}_n(\boldsymbol{\theta})\} \| \rightarrow^{\mathrm{p}} 0$. Write $\lim_{n \rightarrow \infty} E\{\dot{\Gamma}_n(\boldsymbol{\theta})\} = I(\boldsymbol{\theta})$, where $I(\boldsymbol{\theta})$ is a positive definite matrix.

Assumption 7 assumes that the variance of $\Gamma_n(\boldsymbol{\theta})$ is of order $O(n^{-1})$, which is also a standard assumption for the subsampling estimation of the covariance. Assumption 8 is a uniform law of large numbers for $\dot{\Gamma}_n(\boldsymbol{\theta})$, which regulates the asymptotic variance of the estimator and can be obtained with the same regularity conditions on $\dot{\Gamma}_n(\boldsymbol{\theta})$ as those in lemma 1.

Lemma 2. Given assumptions 1–4, 6 and 7, we have

$$\Gamma_n(\boldsymbol{\theta})\surd n \xrightarrow{\mathrm{d}} N\{0, \Sigma(\boldsymbol{\theta})\}, \qquad \text{as } n \to \infty.$$

A sketch of the proof for lemma 2 is given in Appendix A. Then, on the basis of standard GMM arguments (Hansen, 1982), we establish the following theorem.

Theorem 2. Given assumptions 1–4 and 6–8, we have

$$(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\surd n \xrightarrow{\mathrm{d}} N\{0, \Omega(\boldsymbol{\theta}_0)\,\Sigma(\boldsymbol{\theta}_0)\Omega^{\mathrm{T}}(\boldsymbol{\theta}_0)\}, \qquad \text{as } n \to \infty,$$

where $\Omega(\boldsymbol{\theta}_0) = -\{I^{\mathrm{T}}(\boldsymbol{\theta}_0)W^{-1}\,I(\boldsymbol{\theta}_0)\}^{-1}I^{\mathrm{T}}(\boldsymbol{\theta}_0)W^{-1}$.

The above results are applicable to more general settings than those considered in Heagerty and Lele (1998), who derived their asymptotic results for spatial data on the basis of the theory in Guyon (1995), which requires that the sample regions form a strictly increasing sequence on evenly spaced lattices. In contrast, we do not impose any restrictions on the geometry or growth behaviour of the sample regions and allow for unevenly spaced locations, which is a situation that is frequently encountered in real data analysis. Moreover, our results accommodate sampling domain expansions both in space and time, whereas results in Li *et al.* (2007) deal only with the expansion in time. In fact, our results even apply to processes with unbounded moments, which arise in many real world applications. For more discussion, refer to Jenish and Prucha (2009). It is also worth pointing out that asymptotic results for infinite spatial or/and temporal lags are slightly different, because the convergence rates of $\Psi_{S,n}(\boldsymbol{\theta})$, $\Psi_{T,n}(\boldsymbol{\theta})$ and $\Psi_{C,n}(\boldsymbol{\theta})$ may be of different orders, owing to the differences on expansion rates in space and time. In addition, Davis and Yau (2011) have pointed out that using all possible pairs may even destroy the consistency of the maximum CL estimators, which corresponds to infinite spatial and/or temporal lags in our situation.

4. Simulation experiments

To assess the performance of the JCEF method proposed we conduct simulation experiments to compare it with some of the other available methods in the literature, including

(a) weighted composite likelihood (WCL), the current available CL approach (Bevilacqua *et al.*, 2011),
(b) the tapering method (taper) based on covariance regularization (Kaufman *et al.*, 2008),
(c) conditional pseudolikelihood methods (Stein) proposed in Stein *et al.* (2004) and Vecchia (1988), which are variants of the CL formulation based on conditional density functions,
(d) a weighted least square (WLS) approach, which is the method that is used most often by practitioners in spatial statistics (Cressie 1993), and
(e) maximum likelihood estimation (MLE), which is the gold standard.

We compare their performances in terms of the mean-squared error (MSE) of parameter estimates. We also scale parameter-specific MSEs by their corresponding parameter values and sum them to obtain an overall efficiency measure, called the *total scaled MSE*. This scaling balances different scales of parameter values, so that a fair comparison can be made. The relative efficiency (RE) is then computed as the ratio of the total scaled MSEs between two methods under comparison. All simulations are coded in R 2.11.1 (R Development Core Team, 2010) and executed on a LINUX cluster with Intel Xeon X5680 processors (3.33 GHz central processor unit and 1.5 Gbytes memory for each of 16 nodes).

The spatiotemporal covariance function that was used in the data generation is a non-separable spatiotemporal covariance structure proposed in Cressie and Huang (1999):

$$C(h,u;\boldsymbol{\theta})=\begin{cases}\dfrac{\sigma^2(2\beta)}{(a^2u^2+1)^{\nu}(a^2u^2+\beta)\Gamma(\nu)}\left\{\dfrac{b}{2}\left(\dfrac{a^2u^2+1}{a^2u^2+\beta}\right)^{1/2}h\right\}^{\nu}K_{\nu}\left\{b\left(\dfrac{a^2u^2+1}{a^2u^2+\beta}\right)^{1/2}h\right\}, & \text{if } h>0,\\[2ex] \dfrac{\sigma^2(2\beta)}{(a^2u^2+1)^{\nu}(a^2u^2+\beta)}, & \text{if } h=0,\end{cases} \tag{4}$$

where $u=|t_1-t_2|$ is the time lag and $h=\|s_1-s_2\|$ is the Euclidean distance between two locations. K_{ν} is the modified Bessel function of the second kind of order ν (Abramowitz and Stegun (1972), page 374), where $\nu>0$ is a smoothness parameter characterizing the behaviour of the correlation function near the origin. If $u=0$, $C(h,0;\boldsymbol{\theta})$ degenerates into a purely spatial covariance, which is the popular Matérn class that is used in spatial statistics. When $\nu=0.5$, this spatial correlation is an exponential function of h; when $\nu\rightarrow\infty$, the Gaussian correlation function. In practice, ν is difficult to estimate, because it requires dense space data and may run into identifiability problems (Stein, 1999). We shall discuss a profile quadratic inference function approach for estimating ν in Section 6.

For the rest of the parameters, $a\geqslant 0$ is the scaling parameter of time, $b\geqslant 0$ is the scaling parameter of space, $\beta>0$ is a space–time interaction parameter and $\sigma^2=\frac{1}{2}C(0,0)>0$. Note that a separable covariance function is obtained when $\beta=1$. We also study the presence of a nugget effect in our simulation comparison, and we denote its variance by σ_{ε}^2. As a result, the parameter vector of interest is $\boldsymbol{\theta}\equiv(a,b,\beta,\sigma^2,\sigma_{\varepsilon}^2)$.

4.1. Comparison with weighted composite likelihood

We first compare the JCEF with WCL. We form our CEFs on the basis of neighbouring pairs for both WCL and the JCEF, following the suggestion given in Bevilacqua *et al.* (2011). We note that tuning the distance lag according to a certain optimality criterion (e.g. minimizing the trace of the inverse of the Godambe information) for each specific case can result in better efficiency. However, using a common distance lag in the simulation study serves the purpose of comparison and keeps the computational burden manageable.

We generate $X(s,t)$ on a regular grid of $7\times 7\times 30$ space–time points, with spatial co-ordinates being set at $(1,1.5,\ldots,4)\times(1,1.5,\ldots,4)$ and $\mathcal{T}=(1,2,\ldots,30)$. Table 1 includes three simulation set-ups. We vary β-values as 0.5, 1 and 5, corresponding to negative, none and positive spatiotemporal interaction effects respectively. Each columnwise plot in Fig. 2 shows the marginal spatial and temporal correlation patterns. It is clear that the decay rate of spatial or temporal correlation given different temporal or spatial lags changes with different β-values. Parameter ν is fixed at 0.5 in the simulation.

Estimation of the weight matrix is achieved by subgroup sampling on overlapping sub-blocks of size $4\times 4\times 15$, following the rule that was suggested in Politis and Romano (1994). We use estimates from WCL to evaluate individual score functions in each sub-block. A total of 200 simulated data sets are generated for each set-up.

Table 1. (With nugget) MSEs of parameter estimates†

Scenario	*Method*	*MSEs*					*Total scaled MSE*	*RE*
		a	b	β	σ^2	σ_ε^2		
Set-up 1		1	3	0.5	1	0.5		
	WCL	0.0122	0.2915	0.0060	0.0060	0.0039	0.0901	
	JCEF	0.0122	0.2651	0.0039	0.0051	0.0025	0.0724	1.25
Set-up 2		1	3	1	1	0.5		
	WCL	0.0086	0.1492	0.0186	0.0047	0.0027	0.0594	
	JCEF	0.0070	0.1376	0.0107	0.0051	0.0016	0.0447	1.33
Set-up 3		1	3	5	1	0.5		
	WCL	0.0078	0.1593	0.3855	0.0133	0.0014	0.0599	
	JCEF	0.0074	0.1065	0.2576	0.0125	0.0013	0.0471	1.27
Average MSE reduction (%)		12.24	9.09	11.04	3.76	10.59		

†The results are from 200 rounds of simulations based on the covariance structure in equation (4) and a nugget effect σ_ε^2. Total scaled MSE is the sum of MSEs for four parameters scaled by parameter means. RE is the RE defined as the total scaled MSE of WCL over that of the JCEF.

We first compare JCEF and WCL in the presence of a nugget effect σ_ε^2. The results, which are summarized in Table 1, show that the JCEF method clearly outperforms WCL in all three simulation set-ups in terms of the total scaled MSE. The resulting REs show that, for all three scenarios, the JCEF clearly reaches 25% or higher efficiency improvement compared with WCL. Unscaled parameter-specific MSEs indicate that, on average, an approximately 10% reduction in MSE is achieved for parameters a, b, β and σ_ε^2.

We then compare the two methods without the nugget effect in the covariance structures. Similar summary statistics are listed in Table 2. It appears that, in this case, the JCEF gains even more efficiency for a, b and β. On average, the reduction in MSE is 40.5% for β, followed by 26.1% for a (the temporal scaling parameter), and then 13.5% for b (the spatial scaling parameter). The estimates for the variance parameter σ^2 are comparable between the two methods. The significant improvement in efficiency for β, a and b is very desirable, since these are important parameters pertaining to the dependence structure. In addition, for the interaction parameter β, valid parameter and standard error estimates will help researchers to make inferences about whether a simpler and separable spatiotemporal covariance is supported by the data.

4.2. Comparison with tapering

Tapering (Furrer *et al.*, 2006; Kaufman *et al.*, 2008) is becoming increasingly popular in spatial statistics because of its simplicity both in concept and in implementation. The idea is to set certain elements of the covariance matrix to 0, such that the resulting matrix is positive definite and retains the original properties for proximate locations. Specifically, let $C(h;\boldsymbol{\theta})$ be the covariance function for two observations with distance h in space, and $K_{\text{taper}}(h;\eta)$ be the tapering function that is identically 0 whenever $h \geqslant \eta$, where η is a prespecified cut-off. Then the tapered covariance function is given by

$$C_{\text{taper}}(h;\boldsymbol{\theta}) = C(h;\boldsymbol{\theta})\,K_{\text{taper}}(h;\eta).$$

In our spatiotemporal setting, applying the tapering technique requires the specification of a joint distance metric that will accommodate both space and time co-ordinates. This is generally difficult, as space and time are distinct with respect to distance. Nevertheless, for simulation

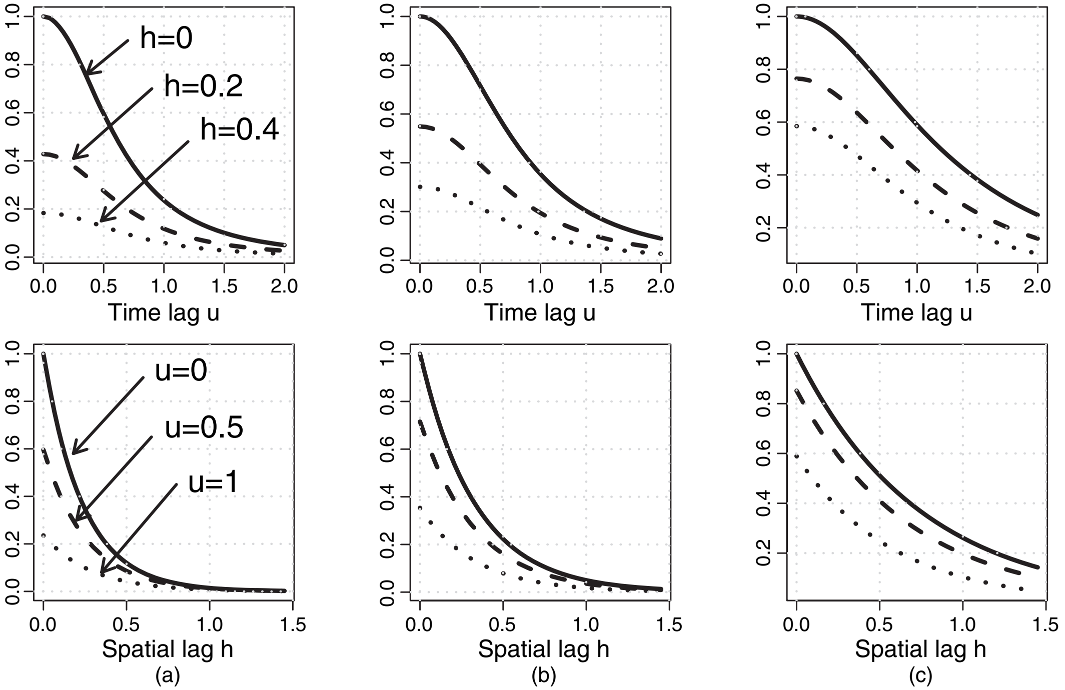

Fig. 2. Plot of $C(h, u; \theta)$ in equation (4) (the parameter ν is fixed at 0.5): (a) set-up 1, $\beta = 0.5$; (b) set-up 2, $\beta = 1$; (c) set-up 3, $\beta = 5$

Table 2. (Without nugget) MSEs of parameter estimates†

Scenario	*Method*	*MSEs*				*Total scaled MSE*	*RE*
		a	*b*	β	σ^2		
Set-up 1		1	3	0.5	1		
	WCL	0.0047	0.1046	0.0030	0.0018	0.11	
	JCEF	0.0037	0.0703	0.0023	0.0019	0.08	1.46
Set-up 2		1	3	1	1		
	WCL	0.0029	0.0599	0.0073	0.0028	0.07	
	JCEF	0.0022	0.0527	0.0037	0.0027	0.06	1.19
Set-up 3		1	3	5	1		
	WCL	0.0031	0.1157	0.2439	0.0073	0.37	
	JCEF	0.0025	0.1080	0.1288	0.0072	0.25	1.50
Average MSE reduction (%)		22.38	17.17	40.18	−1.08		

†The results are from 200 rounds of simulations based on the covariance structure in equation (4). Total scaled MSE is the sum of MSEs for four parameters scaled by parameter means. RE is the RE defined as the total scaled MSE of WCL over that of the JCEF.

purposes, we use the Euclidean norm on standardized spatial and temporal co-ordinates. Note that MLE is a special case of the tapering method when the taper range η is set at ∞.

We compare the tapering method and JCEF with varying distance lags and taper ranges. For each combination of the spatial and temporal lags (p, q) used for the JCEF, we select an appropriate taper range, so that the same pairs of observations are included in the latter method.

Given that each pair of observations corresponds to two entries in the full covariance matrix, we quantify the shared amount of information by both methods in terms of the percentage of covariance elements utilized in estimation for each set of spatial and temporal lags (p, q) and the respective taper range η. These percentages are marked below the horizontal axis label in Fig. 3, where boxplots of estimates of $\log(\beta)$ and the averaged computing times (the dots for MLE and the full line for the method used) are presented. Results are based on simulation set-up 3 presented in Table 2.

In terms of parameter estimates, boxplots in Fig. 3(b) show that, for the JCEF, increasing spatial and temporal lags do not improve the estimates, which is consistent with findings in the current literature (e.g. Varin *et al.* (2005) and Davis and Yau (2011)). This is because pairs that are

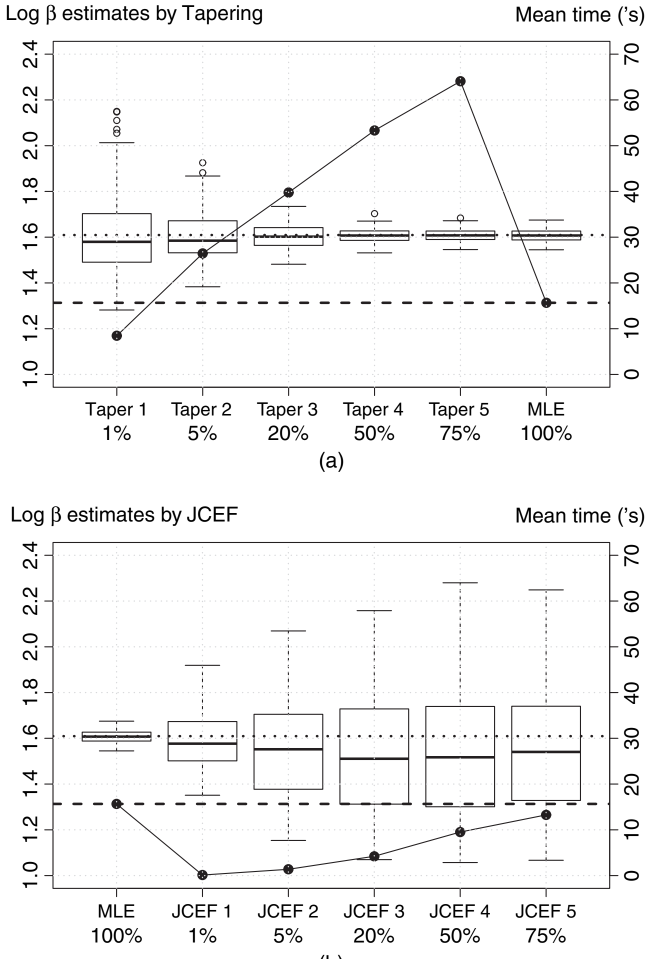

Fig. 3. Boxplots of $\log(\beta)$ estimates from (a) tapering and (b) the JCEF for set-up 3 considered in Table 2 with a spatial grid of 7 × 7 and 30 time points (five sets of spatial and temporal lag combinations (p,q) with increasing values are considered for the JCEF, corresponding to JCEF 1–JCEF 5; the percentage of information utilized by each (p,q) is marked below the horizontal axis label ranging from 1% to 75%; five taper ranges η are chosen with respect to each (p,q), and are labelled as Taper 1–Taper 5; the same percentages are marked for tapering accordingly; MLE is the special case when $p = q = \infty$ for the JCEF and $\eta = \infty$ for tapering): -------, mean computing time; – – –, mean time used by MLE; ·······, true $\log(\beta)$ value

farther apart are less likely to be correlated and will contain little information about dependence. Including them in the estimation will add more noise in the estimation of covariance structures.

In contrast, boxplots in Fig. 3(a) show that increasing the taper range from nearest neighbours (1%) to the maximum distance (100%) leads to improved estimation. This is because tapering works on the covariance matrix. Including pairs farther apart increases the non-zero covariance elements that are used in estimation. This in turn brings in high order correlations between the covariance elements included and leads to a gain in efficiency. This explanation does not apply to pairwise CL methods, since no high order correlations are contained in pairs. However, this weakness is overcome, to some extent, by the JCEF, because the weight matrix effectively accounts for some of the correlations beyond pairwise dependences. Obviously, WCL does not incorporate such high order correlation information, so it is less efficient than the JCEF as shown in Section 4.1.

In terms of computing time needed for the optimization to converge, the full curve in Fig. 3(a) shows that, as the taper range increases, tapering requires a much longer time than MLE (the dots). Tapering is faster when only 1% of the covariance elements (nearing neighbours) are used in estimation. Note that we use the R code at `http://www.image.ucar.edu/Data/precip_tapering/` for executing the tapering method (with minor changes), which is the same code as used by Kaufman *et al.* (2008). So the comparison of computing time is based on the same sparse matrix algorithm. What makes tapering run slowly is the time spent in indexing and retrieving non-zero entries; this can be a substantial workload for a larger taper range. Fig. 3 clearly indicates that tapering is only competitive when the taper range is small. However, in this case, the JCEF is superior to tapering in both estimation and computational efficiency.

4.3. *Comparison with weighted least squares and maximum likelihood estimation*

We now consider WLS and MLE in the comparison. WLS is probably the most commonly used method in spatial data analysis. It estimates dependence parameters by fitting a parametric covariance function to the computed empirical spatiotemporal variogram. As already shown in Lele and Taper (2002), WLS is less efficient than WCL which, as we have shown, is less efficient than the JCEF.

We use set-up 3, considered in Table 2, for comparing the five methods. Table 3 lists results of β-estimates from three increasing grids. In particular, we choose the taper range so that it is computationally competitive with MLE. Then the spatial and temporal lags in the JCEF are set at values that are comparable with the tapering method.

As the gold standard, MLE is the most accurate and has the smallest MSEs, but the price is high in computing time. By contrast, WLS is the fastest, at the cost of being least accurate. It is clear that the JCEF achieves a good balance between time and MSE, and is the best among all methods in this simulation set-up.

4.4. *Comparison with conditional pseudolikelihood*

As an alternative to the marginal bivariate distributions that are used in the JCEF, estimations based on conditional density functions are also extensively considered in the literature. See Vecchia (1988) and Stein *et al.* (2004) among others.

Given innumerable ways of constructing the conditioning sets, in the simulation study, we follow Stein (2005) and select half of the conditioning set from the nearest neighbours, and the other half from observations farther apart. We vary the number of conditioning observations as 1, 2, 4, 6 and 8, and term them Stein 1, Stein 2, Stein 4, Stein 6 and Stein 8 respectively. Results shown in Fig. 4 are obtained on the basis of set-up 3 considered in Table 2: the same setting as used in Table 3 and Fig. 3. Fig. 4 displays boxplots of log(β) estimates and mean computing

Table 3. Comparison of MSEs and computing time for the MLE, JCEF, WCL, tapering and WLS methods for set-up 3 considered in Table 2†

Method	*Results for the following grids:*					
	5 × 5 × 15		*6 × 6 × 20*		*7 × 7 × 30*	
	MSE	*Time (s)*	*MSE*	*Time (s)*	*MSE*	*Time (s)*
MLE	0.04	2.61	0.02	4.11	0.01	15.66
JCEF	0.55	0.03	0.33	0.06	0.20	0.12
WCL	2.02	0.03	0.87	0.05	0.43	0.10
Taper	1.30	3.40	0.65	3.94	0.31	8.45
WLS	6.02	0.00	3.49	0.00	2.03	0.01

†Data are generated from three increasing grids of 5 × 5 × 15, 6 × 6 × 20 and 7 × 7 × 30.

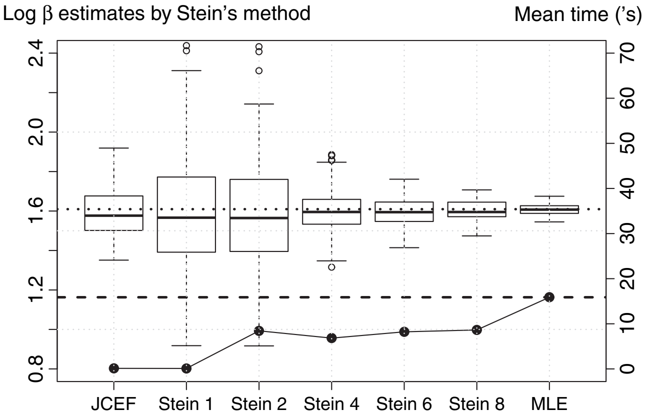

Fig. 4. Boxplots for estimates of log(β) by Stein's method with varying lengths of conditioning sets (Stein 1 refers to Stein's method with one conditioning observation etc.; estimates by using the JCEF based on neighbouring pairs and by MLE are also plotted for comparison): ———, mean computing time; – – –, mean time used by MLE; · · · · · · ·, true log(β) value

time for five versions of Stein's method and for our JCEF based on neighbouring pairs. Results from MLE are included as the gold standard. From Fig. 4, we learn the following results.

(a) As the size of the conditioning sets increases, Stein's method yields improved efficiency as a result of including high order conditional dependence.

(b) When the size of the conditioning set is 1, the Stein 1 method uses bivariate density functions and, hence, similar pairs are used in both the Stein 1 and the JCEF methods. Clearly, the JCEF performs much better in estimation efficiency. Interestingly, the JCEF is shown to be comparable with Stein's method, up to the size of four conditioning observations. This suggests that the weight matrix that is used in the JCEF incorporates additional information beyond pairwise correlation and is comparable with the four-dimensional conditional density functions.

(c) Although Stein's method is always faster than MLE, it is clearly slower than the JCEF method. Thus, as far as computing time is concerned, the JCEF will be advantageous for large-size data problems and for analyses on ordinary personal computers.

In summary, we conclude that, compared with Stein's method, the JCEF is a desirable compromise between estimation and computational efficiency. In addition, unlike Stein's method, the JCEF does not require an explicit specification and evaluation of multivariate density functions. It can be a considerable challenge to generalize the conditional pseudolikelihood approach for non-normal data, such as binary and Poisson data.

4.5. Standard error estimation

We now evaluate the standard error estimation for the JCEF. The key to obtaining valid standard error estimates is to create proper replicates of the data. As we did in the step of weight matrix estimation, we invoke the subsampling method to calculate standard errors for data generated on regular grids. A similar formula to that in equation (3) is used with $\boldsymbol{\theta}_n^i$ replacing $\Gamma_n^i(\boldsymbol{\theta})$. The subsample size is determined by $Mn^{d/(d+2)}$, where $d=3$ in the spatiotemporal setting. Following Heagerty and Lumley (2000), we vary the tuning constant M from 2 to 4 to assess the effects of different subsample sizes on the standard error estimation, resulting in three subsampling schemes: 3 × 3 × 15, 4 × 4 × 15 and 4 × 4 × 20. The same weight matrix as used for the JCEF on the entire grid is used in each subsample evaluation.

Another popular approach to creating data replicates is the parametric bootstrap, i.e. after obtaining JCEF estimates, we generate data on the basis of the estimated model. The square root of the sample variance of the JCEF estimates across replicates is obtained as the estimate. This method involves more computation but is less prone to bias than subsampling, which is likely in finite samples to introduce extra bias with artificially created subsamples. Bevilacqua *et al.* (2010) adopted the parametric bootstrap approach for constructing tests of separability of space–time covariance functions. We also consider a comparison of subsampling with the parametric bootstrap with bootstrap sample size 200. Given the importance of the spatiotemporal interaction parameter β, we devote our attention to parameter β in the evaluation.

Table 4 lists results from 300 rounds of simulation for set-ups 1–3 that were considered in Table 2 with β equal to 0.5, 1 and 5 respectively. We can see that different subsample sizes do have an effect on standard error estimation. Smaller subsamples yield standard error estimates that are closer to the empirical standard deviations, whereas larger subsamples tend to underestimate the variations. The reason may be that we use all overlapping sub-blocks and that larger sub-blocks share more common observations, leading to less variation between blocks. However, truncation bias can occur if the subsamples are too small, because they may fail to account for correlations at longer distances. Parametric bootstrapped standard error estimates perform very well in all three settings, giving estimates that are very close to the empirical standard deviations. This is because, with consistent parameter estimates, each bootstrap procedure will yield a standard error estimate of the same distribution as the empirical distribution. In summary, if the parametric bootstrap is feasible computationally, it is recommended, especially for data that are collected on irregular grids; for data on regular grids, subsampling is recommended. Nevertheless, further investigation is needed to choose the tuning constant M.

To assess the validity of statistical inference, we computed the 95% coverage probabilities across replicates for the three set-ups. In Table 4, the parametric bootstrap and subsampling with a 3 × 3 × 15 partition scheme yield coverage probabilities that are close to the nominal 95%. The other two subsampling schemes have smaller coverage probabilities due to underestimated standard errors. As a by-product of this simulation, WCL estimates as inputs for the weight estimation are also recorded. The calculated MSE in Table 4 again shows that the JCEF method considerably lowers the MSE, leading to a gain in efficiency. The reduction in MSE is mainly due to the reduction in standard deviations. In other words, both methods produce consistent estimates, but those from the JCEF have smaller variances, which again corroborates the theory.

Table 4. Standard errors of parameter estimates for β†

			Results for JCEF				*Results for WCL*		
	Method	*SE*	*CP (%)*	*SE*$_e$	*Mean*	*MSE*	*SE*$_e$	*Mean*	*MSE*
Set-up 1	Subsampling $4\times4\times20$	0.0569	80.12	0.0797	0.4916	0.0064	0.1387	0.5031	0.0191
$\beta=0.5$	$4\times4\times15$	0.0695	89.76						
	$3\times3\times15$	0.0732	92.17						
	Parametric bootstrap	0.0748	94.67						
Set-up 2	Subsampling $4\times4\times20$	0.0728	87.67	0.0937	0.9998	0.0088	0.2126	1.0282	0.0458
$\beta=1$	$4\times4\times15$	0.0814	93.33						
	$3\times3\times15$	0.0929	95.67						
	Parametric bootstrap	0.0997	96.67						
Set-up 3	Subsampling $4\times4\times20$	0.4034	79.67	0.6125	5.0254	0.3746	1.1050	5.2141	1.2627
$\beta=5$	$4\times4\times15$	0.4696	86.00						
	$3\times3\times15$	0.6053	94.00						
	Parametric bootstrap	0.6221	94.67						

†The results are from 300 simulations based on the covariance structure in equation (4). SE is the mean standard error. CP is the mean 95% coverage probability. Subsampling and the parametric bootstrap are used to calculate SE for β. SE$_e$ is the empirical standard deviation of $\hat{\beta}$.

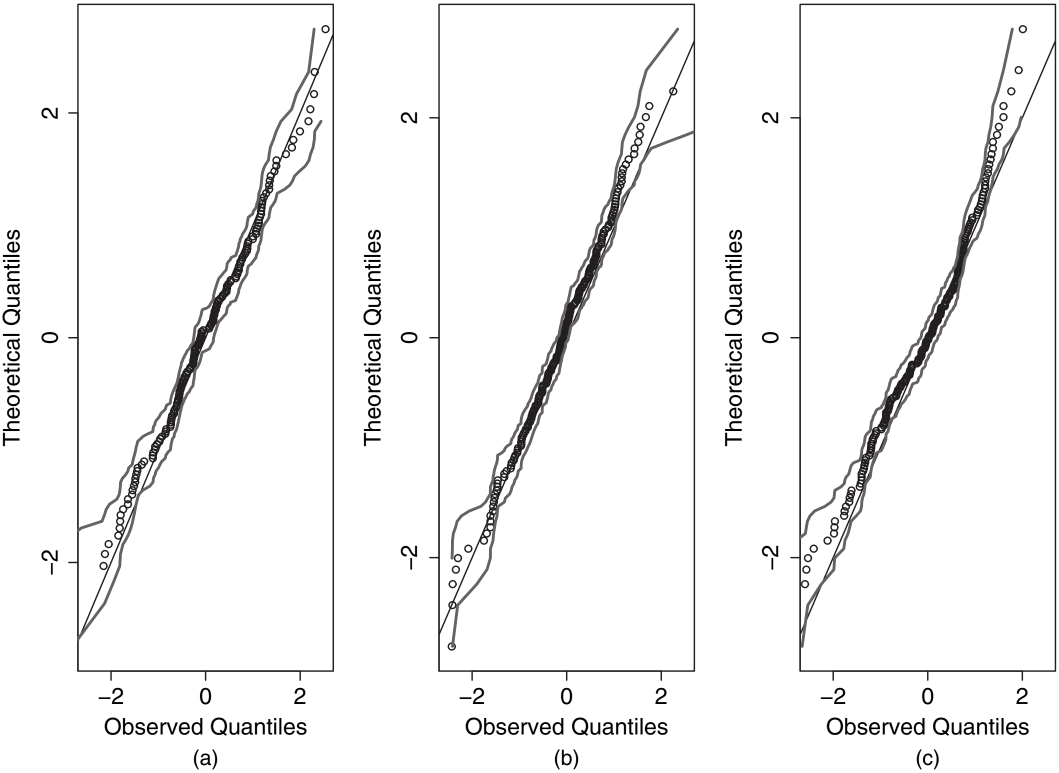

Fig. 5. Normal *QQ*-plots of the standardized estimates of $\hat{\beta}$ by using the JCEF, fixing other parameters (observed quantiles are ordered $(\hat{\beta}-\bar{\hat{\beta}})/\mathrm{SE}(\hat{\beta})$, where $\bar{\hat{\beta}}$ is the mean of $\hat{\beta}$ across simulation replicates and SE($\hat{\beta}$) is based on standard error estimates from the parametric bootstrap): (a) set-up 1; (b) set-up 2; (c) set-up 3

To assess the finite distribution of the β-estimates further, we plotted the QQ-plots with 95% confidence bands by using the R package `fBasics` (function `qqnormPlot`). To reflect common practical situations, we standardized the estimate by the sample mean and the corresponding bootstrap standard error estimates and plotted them against a standard normal random variable. From the QQ-plots in Fig. 5, we see a reasonable coverage of the 95% confidence bands over the 45° diagonal line, which means that the estimates can be regarded as (approximately) normally distributed. Scenarios from set-up 1 to set-up 3, representing fast, median and slow rates of spatial–temporal correlation decay respectively (see Fig. 2), show a slight deviating tendency. The 95% confidence band coverage deteriorates when the dependence decay rate becomes slower. This means that a larger sample size may be required to achieve the asymptotic normality for long memory processes.

5. Analysis of particulate matter data

To illustrate the JCEF method, we analyse 20-year airborne particulate matter data (PM10-data) across the north-eastern USA from August 1982 to August 2002. PM10-particles amount to fine soot that enters the atmosphere from fuel combustion sources, industrial processes and transportation sources. The goal is to study their spatiotemporal dependence, so that predictions can be made at specific locations and time points. Monthly mean PM10-measures are obtained by averaging all available readings for a given month and are log-transformed. Because not all monitors are observed all the time, the final data come from readings at 108 air pollution stations between Maine and Virginia during the months from January 2000 to August 2002. The layout of the monitor locations is displayed in Fig. 6. The distance between any two monitor locations ranges from 0.45 to 956 miles.

We first remove location and month effects by an analysis-of-variance model, treating each month and location as class variables (Diez Roux *et al.*, 2008). Separate spatial and temporal models can be developed for the estimated mean location and month effects. Our focus is on

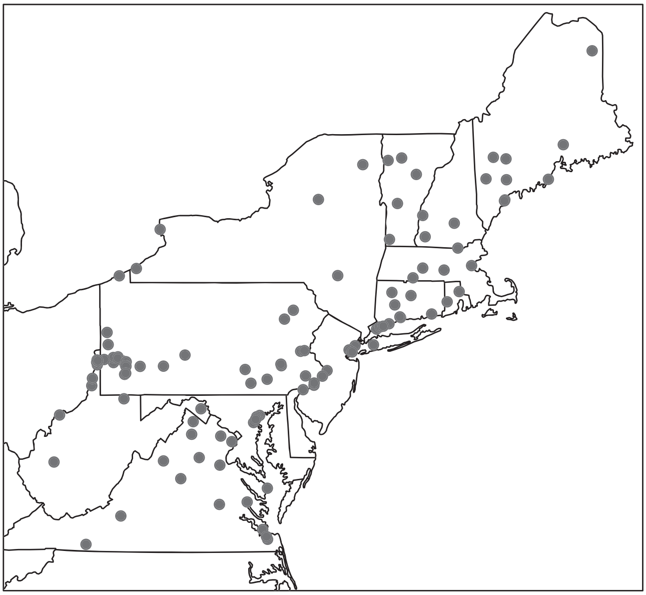

Fig. 6. Layout of PM10 monitoring stations in the north-eastern USA from January 2000 to August 2002

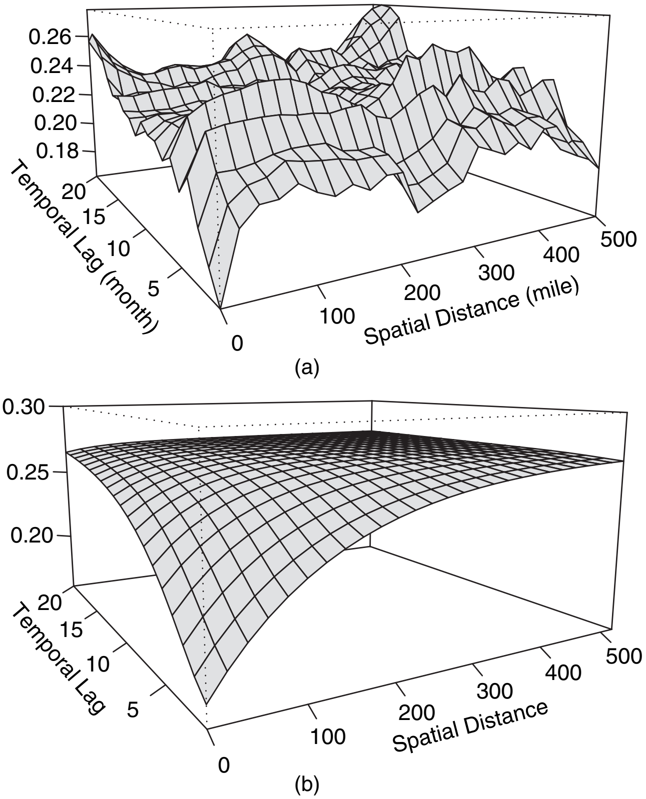

Fig. 7. (a) Empirical and (b) fitted spatiotemporal variograms for PM10-residuals: observation pairs are grouped by distance lags of 20–500 miles with a unit increase of 20 miles and temporal lags of 1–20 months, with a unit increase of 1 month

studying the spatiotemporal dependence structure of the resulting residuals. To visualize the spatiotemporal pattern, we plot the estimated spatiotemporal empirical variogram in Fig. 7(a). Observation pairs are grouped by distance lags of 20–500 miles with the unit of increase being 20 miles. Temporal pairs are grouped by time lags of 1–20 months, with the unit of increase being 1 month.

We fit the non-separable covariance structure in equation (4) with a nugget effect of variance σ_ε^2 to the data. A set of initial parameter values is obtained by using WLS, by minimizing the weighted difference between the empirical variogram and the parametric variogram at prespecified lags.

As pointed out previously, subsampling may not be appropriate because the spatial monitor grid is irregular, so we use the parametric bootstrap to create sample replicates for the subsequent determination of the optimal distance lag, the weight matrix estimation and standard error estimation.

We follow the method that was proposed in Bevilacqua *et al.* (2011) to determine the optimal distance lags. It is computationally prohibitive in practice to compute the Godambe information for all possible combinations of spatial and temporal lags. We use the grid search method to find the optimal lags from a pool of spatial and temporal lags with time ranging from 1 to 6 months with a 1-month increment and spatial distances ranging from 20 to 260 miles with a 20-mile increment. The optimal combination is 6 months in time and 100 miles in space, which

Table 5. Parameter and standard error estimates of the spatiotemporal covariance structure in equation (4) fitted to the PM10 data set†

Parameter	*Results for WCL*			*Results for JCEF*		
	Estimate	*95% confidence interval*		*Estimate*	*95% confidence interval*	
a	1.0112	0.6213	1.5048	1.1636	0.7833	1.7285
b	0.0382	0.0148	0.0981	0.0403	0.0173	0.0939
β	4.1129	0.9423	20.7327	6.4341	1.7373	23.8292
σ^2	0.0219	0.0173	0.0265	0.0224	0.0180	0.0277
σ_ε^2	0.0194	0.0167	0.0231	0.0199	0.0173	0.0229

†The standard error estimates were obtained by subsampling.

means that we shall include pairs that are within $p=100$ miles in distance and $q=6$ months in time to specify our composite estimating functions. Then WCL is carried out for estimation and its estimates are used for weight matrix calculation. Finally, the JCEF method is applied to estimate the model parameters.

Parameter estimates, standard errors and 95% confidence intervals from the JCEF and WCL are listed in Table 5. Point estimates from the two methods are similar, but the JCEF yields smaller standard error estimates, especially for the interaction parameter β, which is consistent with the simulation results. For the JCEF method, given $\hat{\beta}=6.4341$, $\hat{a}=1.1636$ means that the marginal temporal correlation decays by around 40% with a 1-month increase in time, and $\hat{b}=0.0403$ indicates that the marginal spatial correlation decays by approximately 15% with a 10-mile increase in space. $\hat{\beta}=6.4341$ indicates that the temporal correlation decays approximately 1.5% faster with 10 miles farther away in space, whereas the spatial correlation decays about 2.5% faster with 1 month further apart in time. The confidence interval for $\hat{\beta}$ does not cover 1, indicating that there is a significant spatiotemporal interaction effect. As a result, a separable covariance structure is not applicable to this data set if the covariance function in equation (4) is used to model the dependence structure.

We also compare the sums of squared differences between the fitted parametric variograms obtained by the JCEF (Fig. 7(b)) and WCL with the empirical variogram. The sum of squared differences ratio of the JCEF over WCL is 0.67, indicating that the surface fitted by the JCEF is 33% closer to the empirical surface than that fitted by WCL. In summary, the proposed JCEF outperforms WCL in point estimates and standard error estimates, as well as the goodness of fit. Some additional analysis of the data may be carried out. For example, one could use the tests that were proposed in Li *et al.* (2007) to test the symmetry and isotropy of the data dependence, and to fit the corresponding parametric covariance function to the data, to improve the fit of the overall model.

6. Discussion

In this paper, we have proposed a statistically efficient and computationally feasible approach to estimating spatiotemporal covariance models for large data sets. The JCEF method proposed constructs separate CLs based on spatial, temporal and spatiotemporal cross-pairs, and then joins them into a quadratic inference function. Through such GMM formulation, our method accounts for correlations between the pairs via the weight matrix and allocates higher weights to groups of pairs with more information, and hence it substantially improves the estimation efficiency over existing WCL methods. The JCEF estimator has also proven to be consistent and asymptotically Gaussian under the increasing domain asymptotics. Comprehensive simulation studies have shown that the JCEF is advantageous over MLE, tapering, Stein's method, WCL and WLS in terms of balancing estimation and computational efficiency for large data sets.

Another advantage of the JCEF method is the possibility of deriving a goodness-of-fit statistic to test the zero-mean model assumption, $H_0 : E\{\Gamma_n(\boldsymbol{\theta})\} = 0$. This can be used for testing the separability structure of the covariance matrix. Since $\hat{\boldsymbol{\theta}}_n$ is obtained by an overidentified estimating function $\Gamma_n(\boldsymbol{\theta})$, $Q_n(\hat{\boldsymbol{\theta}}_n)$ falls in the 'overidentifying restriction' test by Hansen (1982), who proved that the asymptotic distribution of $Q_n(\hat{\boldsymbol{\theta}}_n)$ is χ^2 with degrees of freedom equal to the number of estimating functions minus the number of parameters, which in our case is $2r$. However, many researchers have pointed out that the first-order asymptotic theory often provides inadequate approximations to the distributions of the test statistics that are obtained from GMM estimators; see, for example, the special issue of the *Journal of Business and Economics Statistics* in July 1996. To improve inference, various alternative estimators have been suggested. These include empirical likelihood (Qin and Lawless, 1994; Owen, 1988, Imbens, 1997), modified bootstrap procedures (Hall and Horowitz, 1996) and the continuous updating estimator (Hansen *et al.*, 1996). Qu *et al.* (2000) used the last approach to construct the quadratic inference function and showed that the finite sample distribution of the objective function agrees well with the asymptotic counterpart. The performances of these goodness-of-fit methods under the JCEF framework for spatiotemporal data are worth further exploration.

As noted previously, the smoothness parameter ν is usually difficult to estimate. However, the quadratic objective function in the JCEF is analogous to profile likelihood and could be used as a tool to determine its value. Specifically, given a range of ν-values, we perform the JCEF estimation procedure for each ν and record the parameter values and the target function value. Then ν is estimated to be the one with the smallest target function value, and the corresponding parameter estimates are used as the final estimates. We plot $\log\{Q(\boldsymbol{\theta})\}$ and ν in Fig. 8 for simulation set-up 3 considered in Table 2. The true value for ν is 0.5. We

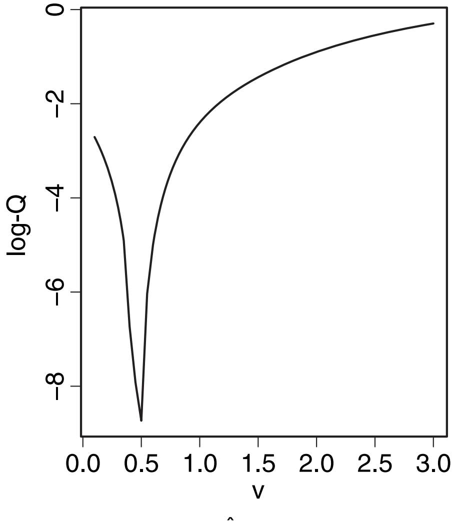

Fig. 8. Estimated log-quadratic objective function $Q(\hat{\theta})$ *versus* smoothness parameter ν, evaluated in set-up 3 in Table 2: the true value for ν is 0.5

can see that this profile approach provides an accurate estimate of ν. Hence it seems a very promising method for estimating the parameter ν. Further detailed work is needed to develop this.

We have considered covariance estimation from a detrended process. As is known, detrending may introduce artificial correlation into the residuals, which may distort the intrinsic correlation of the data. In fact, this is a common concern when a two-stage procedure is used to estimate covariance structures. A simple solution would be to estimate the mean and covariance parameters jointly. From a large sample point of view, as long as the mean parameter is consistently estimated, covariance estimates can be consistently estimated under some mild conditions. However, in actual applications, finite sample performances matter more. Our experiments in both theory and computation suggested that two factors are crucial to ensure similar performances between the two-stage approach and the joint estimation method:

(a) the strength of the intrinsic spatiotemporal dependence and
(b) the sample size.

For large data sets, the two-stage procedure is usually favoured.

It is worth noting that variance estimates of the JCEF estimator do not account for uncertainty in the weight matrix estimation. This uncertainty results from the plugged-in parameter estimates in the evaluation of the weight matrix. According to Windmeijer (2005), such variation is known to be of order $O(n^{-1})$, which is a lower order term than $O(n^{-1/2})$ and thus may be ignorable when n is large. In addition, this issue concerning the finite sample performance has been well studied in the GMM literature. Several methods have been proposed to correct for the downward bias that occurs in parameter standard error estimates when the sample size is inadequate. They include adding a variance correction term (Windmeijer, 2005) or using a parametric bootstrap procedure (Hall and Horowitz, 1996).

We have focused our attention in this paper on evaluating the gain in efficiency of the JCEF over the existing methods of covariance estimation. Kriging, which is one of the popular approaches that are used for prediction in geostatistics, relies heavily on covariance functions, as the kriging predictor is the best linear unbiased estimator on the basis of the covariance model that is specified for the process. It may also be interesting to study whether more efficient covariance estimators will yield more efficient predictors.

Acknowledgements

We are grateful to the Associate Editor and two referees for providing insightful comments and constructive suggestions that led to a substantial improvement over the earlier manuscript. We also thank Dr Ana Diez Roux for providing the PM10 data set that was used in Section 5 of the paper. The research of the second author is partly supported by National Science Foundation grant 0904177.

Appendix A: Definition of the distance metric ρ

The distance between two pairwise differences $d(k_1)$ and $d(k_2)$ defined in equation (1) depends on configurations of four points in the spatiotemporal domain $\mathbb{R}^2 \times \mathbb{R}^+$. Denote the co-ordinates of one point by (s, t). The distance between two points $p_1 = (s_1, t_1)$ and $p_2 = (s_2, t_2)$ in $\mathbb{R}^2 \times \mathbb{R}^+$ is defined as $\tau(p_1, p_2) = \max\{\|s_1 - s_2\|, |t_1 - t_2|\}$. Let $k_1 = (p_1, p_1')$ and $k_2 = (p_2, p_2')$. Then the distance between two points in $D \subset \mathbb{R}^2 \times \mathbb{R}^+ \times \mathbb{R}^2 \times \mathbb{R}^+$ is defined as $\rho(k_1, k_2) = \min\{\tau(p_1, p_2), \tau(p_1, p_2'), \tau(p_1', p_2), \tau(p_1', p_2')\}$, i.e. the minimum distance of two points in sets (p_1, p_1') and (p_2, p_2'). The distance between any subsets $U, V \subset D$ is defined as $\rho(U, V) = \min\{\rho(i, j) : i \in U, j \in V\}$.

Appendix B: Proof of lemma 2

Lemma 2 states a central limit theorem for $\Gamma_n(\boldsymbol{\theta})$, which is comprised of three estimating functions based on different groups of pairwise differences with varying numbers of terms. The three groups of pairwise differences are subseries of $d(k)$ and hence satisfy the same mixing conditions in assumption 6 imposed on $d(k)$. In addition, $|D_{S,n}|$, $|D_{T,n}|$ and $|D_{C,n}|$ are of the same order $O(n)$, making it possible to use a common scaling factor to unify the convergence rates.

We prove the asymptotic normality of $\Gamma_n(\boldsymbol{\theta})$ through the Cramer–Wold device. For ease of argument, we work on sums of component score functions instead of means. Define $\Gamma_n^*(\boldsymbol{\theta}) = (\Psi_{s,n}^{*\mathrm{T}}(\boldsymbol{\theta}), \Psi_{T,n}^{*\mathrm{T}}(\boldsymbol{\theta}), \Psi_{C,n}^{*\mathrm{T}}(\boldsymbol{\theta}))^{\mathrm{T}}$, where $\Psi_{\mathcal{A},n}^*(\boldsymbol{\theta}) = |D_{\mathcal{A},n}|\,\Psi_{\mathcal{A},n}(\boldsymbol{\theta})$ for $\mathcal{A} \in \{S, T, C\}$. The aim is to prove that, for arbitrary constants c_1, c_2 and c_3, the linear combination

$$c_1\,\Psi_{S,n}^*(\boldsymbol{\theta}) + c_2\,\Psi_{T,n}^*(\boldsymbol{\theta}) + c_3\,\Psi_{C,n}^*(\boldsymbol{\theta})$$

is asymptotically Gaussian. Define

$$G_n(\boldsymbol{\theta}) \equiv c_1\,\Psi_{S,n}^*(\boldsymbol{\theta}) + c_2\,\Psi_{T,n}^*(\boldsymbol{\theta}) + c_3\,\Psi_{C,n}^*(\boldsymbol{\theta}) = \mathbf{c}^{\mathrm{T}}\,\Gamma_n^*(\boldsymbol{\theta}),$$

where $\mathbf{c} = (\mathbf{c}_1; \mathbf{c}_2; \mathbf{c}_3)^{\mathrm{T}}$, which is a $3r \times r$ matrix with $\mathbf{c}_i = c_i I_r$, $i = 1, 2, 3$, and I_r is the $r \times r$ identity matrix. Let $\mathrm{var}\{\Gamma_n^*(\boldsymbol{\theta})\} = \Sigma_n^*(\boldsymbol{\theta})$, $\Sigma_{G,n}(\boldsymbol{\theta}) \equiv \mathrm{var}\{G_n(\boldsymbol{\theta})\} = \mathbf{c}^{\mathrm{T}}\Sigma_n^*(\boldsymbol{\theta})\mathbf{c}$.

Write

$$\begin{aligned} G_n(\boldsymbol{\theta}) &= \sum_{i \in D_{S,n}} c_1\, f_i\{d(i); \boldsymbol{\theta}\} + \sum_{j \in D_{T,n}} c_2\, f_j\{d(j); \boldsymbol{\theta}\} + \sum_{l \in D_{C,n}} c_3\, f_l\{d(l); \boldsymbol{\theta}\} \\ &\equiv \sum_{k \in D_n} h_k\{d(k); \boldsymbol{\theta}\}, \end{aligned} \tag{5}$$

where

$$h_k\{d(k); \boldsymbol{\theta}\} = \begin{cases} c_1\, f_k\{d(k); \boldsymbol{\theta}\}, & \text{if } k \in D_{S,n}, \\ c_2\, f_k\{d(k); \boldsymbol{\theta}\}, & \text{if } k \in D_{T,n}, \\ c_3\, f_k\{d(k); \boldsymbol{\theta}\}, & \text{if } k \in D_{C,n}. \end{cases}$$

Equation (5) simply multiplies each set of estimating functions by a constant and sums them together. Then, given assumptions 1–4 and 6 and 7, according to theorem 1 in Jenish and Prucha (2009),

$$\Sigma_{G,n}^{-1/2}(\boldsymbol{\theta})\, G_n(\boldsymbol{\theta}) \sim N(0, I_r), \qquad \text{as } n \to \infty.$$

Note that assumption 4 is imposed on f_k, which also applies to h_k, since h_k differs from f_k by a multiplicative constant. Assumption 7 implies the convergence of $n^{-1}\,\Sigma_n^*(\boldsymbol{\theta})$ to a positive definite constant matrix, provided that $|D_{S,n}|$, $|D_{T,n}|$ and $|D_{C,n}|$ are of order $O(n)$.

Since c_1, c_2 and c_3 are arbitrary constants, by the Cramer–Wold device, we obtain

$$\Sigma_n^*(\boldsymbol{\theta})^{-1/2}\,\Gamma_n^*(\boldsymbol{\theta}) \sim N(0, I_{3r}), \qquad \text{as } n \to \infty.$$

Let $B = \mathrm{diag}\{(1/|D_{S,n}|)I_r, (1/|D_{T,n}|)I_r, (1/|D_{C,n}|)I_r\}$. Then $\Gamma_n(\boldsymbol{\theta}) = B\,\Gamma_n^*(\boldsymbol{\theta})$, whose asymptotic normality follows immediately.

References

Abramowitz, M. and Stegun, I. (1972) *Handbook of Mathematical Functions*. New York: Dover Publications.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *J. R. Statist. Soc.* B, **70**, 825–848.

Bevilacqua, M., Gaetan, C., Mateu, J. and Porcu, E. (2011) Estimating space and space-time covariance functions: a weighted composite likelihood approach. *J. Am. Statist. Ass.*, to be published, doi 10.1080/01621459.2011.646928.

Bevilacqua, M., Mateu, J., Porcu, E., Zhang, H. and Zini, A. (2010) Weighted composite likelihood-based tests for space-time separability of covariance functions. *Statist. Comput.*, **20**, 283–293.

Carlstein, E. (1987) The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.*, **14**, 1171–1179.

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

Cressie, N. and Huang, H.-C. (1999) Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Am. Statist. Ass.*, **94**, 1330–1340.

Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc.* B, **70**, 209–226.

Curriero, F. C. and Lele, S. (1999) A composite likelihood approach to semivariogram estimation. *J. Agric. Biol. Environ. Statist.*, **4**, 9–28.

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

Diez Roux, A. V., Auchincloss, A. H., Franklin, T. G., Raghunathan, T., Barr, R. G., Kaufman, J., Astor, B. and Keeler, J. (2008) Long-term exposure to ambient particulate matter and prevalence of subclinical atherosclerosis in the multi-ethnic study of atherosclerosis. *Am. J. Epidem.*, **167**, 667–675.

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

Fuentes, M. (2007) Approximate likelihood for large irregularly spaced spatial data. *J. Am. Statist. Ass.*, **102**, 321–331.

Furrer, R., Genton, M. G. and Nychka, D. (2006) Covariance tapering for interpolation of large spatial datasets. *J. Computnl Graph. Statist.*, **15**, 502–523.

Genton, M. G. (2007) Separable approximations of space-time covariance matrices. *Environmetrics*, **18**, 681–695.

Gneiting, T. (2002) Nonseparable, stationary covariance functions for space-time data. *J. Am. Statist. Ass.*, **97**, 590–600.

Godambe, V. P. and Heyde, C. (1987) Quasi-likelihood and optimal estimation. *Int. Statist. Rev.*, **55**, 231–244.

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

Haas, T. C. (1995) Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *J. Am. Statist. Ass.*, **90**, 1189–1199.

Hall, P. and Horowitz, J. L. (1996) Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica*, **64**, 891–916.

Hansen, L. P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–1054.

Hansen, L. P., Heaton, P. and Yaron, A. (1996) Finite-sample properties of some alternative gmm estimators. *J. Bus. Econ. Statist.*, **14**, 262–280.

Heagerty, P. J. and Lele, S. R. (1998) A composite likelihood approach to binary spatial data. *J. Am. Statist. Ass.*, **93**, 1099–1111.

Heagerty, P. J. and Lumley, T. (2000) Window subsampling of estimating functions with application to regression models. *J. Am. Statist. Ass.*, **95**, 197–211.

Imbens, G. (1997) One-step estimators for over-identified generalized method of moments models. *Rev. Econ. Stud.*, **64**, 359–383.

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

Joe, H. and Lee, Y. (2009) On weighting of bivariate margins in pairwise likelihood. *J. Multiv. Anal.*, **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. *J. Am. Statist. Ass.*, **103**, 1545–1555.

Kuk, A. Y. (2007) A hybrid pairwise likelihood method. *Biometrika*, **94**, 939–952.

Kuk, A. and Nott, D. (2000) A pairwise likelihood approach to analyzing correlated binary data. *Statist. Probab. Lett.*, **47**, 329–335.

Kunsch, H. (1989) The jackknife and bootstrap for general stationary observations. *Ann. Statist.*, **17**, 1217–1241.

Lahiri, S., Kaiser, M., Cressie, N. and Hsu, N. (1999) Prediction of spatial cumulative distribution functions using subsampling. *J. Am. Statist. Ass.*, **94**, 86–97.

Lee, Y. D. and Lahiri, S. N. (2002) Least squares variogram fitting by spatial subsampling. *J. R. Statist. Soc.* B, **64**, 837–854.

Lele, S. (1991) Jackknifing linear estimating equations: asymptotic theory and applications in stochastic processes. *J. R. Statist. Soc.* B, **53**, 253–267.

Lele, S. and Taper, M. L. (2002) A composite likelihood approach to (co)variance components estimation. *J. Statist. Planng Inf.*, **103**, 117–135.

Li, B., Genton, M. G. and Sherman, M. (2007) A nonparametric assessment of properties of space-time covariance functions. *J. Am. Statist. Ass.*, **102**, 736–744.

Li, Y. and Lin, X. (2006) Semiparametric normal transformation models for spatially correlated survival data. *J. Am. Statist. Ass.*, **101**, 591–603.

Lindsay, B. G. (1988) Composite likelihood methods. *Contemp. Math.*, **80**, 221–239.

Nott, D. and Rydén, T. (1999) Pairwise likelihood methods for inference in image models. *Biometrika*, **86**, 661–676.

Owen, A. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 234–249.

Paciorek, C. J., Yanosky, J. D. and Puett, R. C. (2009) Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Ann. Appl. Statist.*, **3**, 370–397.

Politis, D. and Romano, J. (1994) Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.*, **22**, 2031–2050.

Porcu, E., Mateu, J. and Bevilacqua, M. (2007) Covariance functions that are stationary or nonstationary in space and stationary in time. *Statist. Neerland.*, **61**, 358–382.

Qin, J. and Lawless, J. (1994) Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300–325.

Qu, A., Lindsay, B. and Li, B. (2000) Improving generalized estimating equations using quadratic inference functions. *Biometrika*, **87**, 823–836.

R Development Core Team (2010) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reid, N. and Cox, D. (2004) A note on pseudolikelihood constructed from marginal densities. *Biometrika*, **91**, 729–737.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2007) High-resolution space-time ozone modeling for assessing trends. *J. Am. Statist. Ass.*, **102**, 1221–1234.

Sherman, M. (1996) Variance estimation for statistics computed from spatial lattice data. *J. R. Statist. Soc.* B, **58**, 509–523.

Smith, R. L. and Kolenikov, S. (2003) Spatiotemporal modeling of $pm_{2.5}$ data with missing values. *J. Geophys. Res.*, **108**, 4–27.

Stein, M. L. (1999) *Interpolation of Spatial Data: Some Theory of Kriging*. New York: Springer.

Stein, M. L. (2005) Space-time covariance functions. *J. Am. Statist. Ass.*, **100**, 310–321.

Stein, M. L., Chi, Z. and Welty, L. J. (2004) Approximating likelihoods for large spatial data sets. *J. R. Statist. Soc.* B, **66**, 275–296.

Varin, C., Høst, G. and Skare, Ø. (2005) Pairwise likelihood inference in spatial generalized linear mixed models. *Computnl Statist. Data Anal.*, **49**, 1173–1191.

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

Vecchia, A. V. (1988) Estimation and model identification for continuous spatial processes. *J. R. Statist. Soc.* B, **50**, 297–312.

Windmeijer, F. (2005) A finite sample correction for the variance of linear efficient two-step gmm estimators. *J. Econmetr*, **126**, 25–51.

Winkler, G. (1995) *Image Analysis, Random Fields and Dynamic Monte Carlo Methods: a Mathematical Introduction*. New York: Springer.

Zhao, Y. and Joe, H. (2005) Composite likelihood estimation in multivariate analysis. *Can. J. Statist.*, **33**, 335–356.

Zhu, J. and Morgan, G. (2004) Comparison of spatial variables over subregions using a block bootstrap. *J. Agric. Biol. Environ. Statist.*, **9**, 91–104.

Zimmerman, D. L. (1989) Computationally exploitable structure of covariance matrices and generalized covariance matrices in spatial models. *J. Statist. Computn Simuln*, **32**, 1–15.