A General Class of Nonseparable Space-time Covariance Models

Thaís C.O. Fonseca and Mark F.J. Steel
Department of Statistics, University of Warwick, U.K.

Summary.
The aim of this work is to construct nonseparable, stationary covariance functions for processes that vary continuously in space and time. Stochastic modelling of phenomena over space and time is important in many areas of application. But choice of an appropriate model can be difficult as we need to ensure that we use valid covariance structures. A common choice for the process is a product of purely spatial and temporal random processes. In this case, the resulting process possesses a separable covariance function. Although these models are guaranteed to be valid, they are severely limited, since they do not allow space-time interactions. We propose a general and flexible class of valid nonseparable covariance functions based on the mixture approach presented in Ma (2002). The proposed model allows for different degrees of smoothness across space and time and long-range dependence in time. Moreover, the proposed class has as particular cases several popular covariance models proposed in the literature such as the Matérn and the Cauchy Class. We use a Markov chain Monte Carlo sampler for Bayesian inference and apply our modelling approach to the Irish wind data.

Key words: Bayesian Inference; Irish wind data; Mixtures; Spatiotemporal modelling.

1. Introduction

Stochastic modelling of phenomena over space and time is in great demand in many areas of applications such as environmental sciences, agriculture and meteorology. The aim of this work is to introduce a general and flexible spatiotemporal nonseparable covariance model based on a mixture of separable covariance functions as presented in Ma (2002). We show that adequate choices of the mixing structure generate closed forms for the covariance function and ensure some desirable properties of the resulting covariance structure. Moreover, many
commonly used models proposed in the recent literature are subclasses of the general model presented here. Suppose that \((s, t) \in D \times T, D \subseteq \mathbb{R}^d, T \subseteq \mathbb{R}\) are space-time coordinates that vary continuously in \(D \times T\) and we seek to define a spatiotemporal stochastic process \(\{Z(s, t) : s \in D; t \in T\}\). In order to specify this process we need to determine the space-time covariance structure \(C(s_1, s_2; t_1, t_2), \) for \(s_1, s_2 \in D\) and \(t_1, t_2 \in T\). In practice, it is often necessary to consider simplifying assumptions such as stationarity, isotropy, Gaussianity and separability. In what follows, we assume \(\text{Var}(Z(s, t)) < \infty\), for all \((s, t) \in D \times T\) and stationary covariance functions, that is, \(\text{Cov}(Z(s_0, t_0); Z(s_0 + s, t_0 + t)) = C(s, t), s \in D, t \in T\) depends on the space-time lag \((s, t)\) only, for any \(s_0 \in D, t_0 \in T\). Choice of an appropriate model can be difficult as we must use valid covariance structures, that is, for any \((s_1, t_1), \ldots , (s_m, t_m)\), any real \(a_1, \ldots , a_m\), and any positive integer \(m\), \(C\) must satisfy \(\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j C(s_i - s_j, t_i - t_j) \geq 0\), as this is the covariance function \(\text{Var}\left(\sum_{i=1}^{m} a_i Z(s_i, t_i)\right)\) of real linear combinations of \(Z(s, t)\). One way to verify the validity of a covariance function is by using Bochner’s Theorem which states that an integrable continuous function \(C(s, t)\) for \((s, t) \in \mathbb{R}^{d+1}\) is positive definite if and only if it is of the form

\[
C(s, t) = \int \int \exp\{i s^T h_1 + i t h_2\} f(h_1, h_2) dh_1 dh_2,
\]

where \(f(h_1, h_2)\) is the positive spectral density. The latter is then linked to the covariance through the Fourier inversion formula and given by \(f(h_1, h_2) = (2\pi)^{-(d+1)} \int \int \exp\{-i s^T h_1 - i t h_2\} C(s, t) ds dt\). In general, it is quite difficult to check whether a function is positive definite and this is one of the main difficulties in the construction of new covariance functions. There is a vast literature that uses the spectral approach in the development of new class of covariance functions. Many of them do not lead to a closed-form expression for the resulting covariance function. Fuentes et al. (2008) propose an interesting example of nonseparable covariance functions generated using spectral densities. A nonparametric test for separability is derived in Crujeiras et al. (2009) on the basis of an estimator of the spectral density. In this paper, we do not use spectral methods in the constructions of valid nonseparable covariances functions, but adopt an approach based on mixing over separable covariance structures, described in detail in Section 2.

A common choice for the process \(\{Z(s, t), (s, t) \in D \times T\}\) is given by

\[
Z(s, t) = Z_1(s)Z_2(t), (s; t) \in D \times T,
\]

where \(\{Z_1(s) : s \in D\}\) is a purely spatial random process with covariance function \(C_1(s)\) and \(\{Z_2(t) : t \in T\}\) is a purely temporal random process with covariance function \(C_2(t)\). The processes \(Z_1(s)\) and \(Z_2(t)\) are uncorrelated. The resulting process \(Z(s, t)\) possesses a
separable covariance function given by

\[ C(s, t) = C_1(s)C_2(t), \quad (s, t) \in D \times T, \tag{3} \]

where \( C_1 \) and \( C_2 \) are valid covariance functions in \( \mathbb{R}^d \) and \( \mathbb{R} \), respectively. The validity of the resulting covariance function in (3) comes from the property that sums, products, convex combinations and limits of positive definite functions are positive definite. Note that the validity of (3) depends on the choice of valid covariance functions in \( \mathbb{R}^d \) and \( \mathbb{R} \). For instance, the Matérn class is valid in any number of dimensions (Stein, 1999, page 49), as is its special case, the exponential. On the other hand, the spherical model is valid in up to 3 dimensions but it fails to correspond to a variance matrix that is positive definite for \( d \geq 4 \) (Stein, 1999, page 52).

Separability is a convenient property since the covariance matrix can be expressed as a Kronecker product of smaller matrices that come from the purely temporal and purely spatial processes. Thus, determinants and inverses are easily obtained providing a potentially large computational benefit. Although the assumption of separable processes in time and space is very convenient, it is usually unrealistic. Cressie and Huang (1999) discuss some shortcomings of separable models. While these models are guaranteed to be valid, they are severely limited, since they do not allow space-time interactions. Stein (2005) points out that separable models have lack of smoothness away from the origin, that is, small changes in the location of observations can lead to large changes in the correlation between certain linear combinations of observations. This will not happen for analytic functions such as \( c \exp(at^2) \) that are infinitely differentiable, but these functions are not adequate for physical processes and the Matérn Class is suggested as an alternative for differentiable processes. Some recent advances were made in developing valid nonseparable models. Carroll et al. (1997), Cressie and Huang (1999), Gneiting (2002), Stein (2005) and Rodrigues and Diggle (2009) all suggest ways of constructing nonseparable covariance models. Carroll et al. (1997) propose a nonseparable spatiotemporal model to reconstruct ozone surfaces and estimate the population exposure in Harris county. They do not show analytically that the covariance function is positive definite and concerns about the validity of the model were raised in comments by Cressie (1997) and Guttorp et al. (1997). Cressie and Huang (1999) introduce new classes of nonseparable, stationary covariance functions that allow for space-time interaction but the approach is restricted to a small class of functions for which a closed-form solution to a \( d \)-variate Fourier integral is known. Gneiting (2002) proposes a new class of valid covariance models. The same approach as Cressie and Huang (1999) is adopted but the Fourier integral limitation is avoided. A criterion for positive definiteness is proposed to validate covariance functions and it is used to show that some of the space-time covariance functions presented by Carroll et al. (1997) and Cressie and Huang (1999) are not valid. Stein (2005) considers stationary covariance functions and
discusses what space-time covariances imply about the corresponding processes. In particular, he points out that the examples provided by Cressie and Huang (1999) are analytic, that is, do not have lack of smoothness away from the origin, but the general approach can yield covariance functions without this property. Also, the nonseparable functions proposed by Gneiting (2002) are possibly not smoother along their axes than at the origin. Stein (2005) provides an example of space-time covariance functions that can achieve any degree of differentiability in space and in time, but the general approach does not provide explicit expressions for the covariance functions. Rodrigues and Diggle (2009) propose nonseparable spatiotemporal covariance functions constructed using convolution which allows for computationally efficient methods of parameter estimation. They define positive and negative nonseparability and show that their proposal is able to capture both types by specification of one parameter in the model. They also show that the valid models in Cressie and Huang (1999) and the models in Gneiting (2002) just allow for positive nonseparability.

In the context of environmental applications, models that are able to capture the simultaneous behaviour of the spatial and temporal components are crucial as predictions will strongly depend on the choice of the covariance associated with the spatiotemporal random function. For instance, separability is usually an unrealistic assumption for ozone. Fuentes et al. (2008) explore spatiotemporal aspects of ozone modeling and propose a nonseparable model which has a unique parameter responsible for indicating the strength of dependence between spatial and temporal components. However, the covariance function is not obtained in closed form. For an overview of stochastic modeling of environmental processes observed in space and time see Porcu and Mateu (2007) and references therein. They present examples of spatiotemporal models generated through different methods such as linear combinations of covariances, bivariate Laplace transformations, copulas and completely monotone functions.

In this work we propose a general and flexible class of valid nonseparable covariance functions derived through mixing over separable covariance functions. Section 2 provides the general mixing approach that guarantees positive definiteness for the class. Section 3 provides the proposed general class of covariance models that allows for different degrees of smoothness across space and time. The purely temporal process can achieve different degrees of smoothness while the purely spatial process can possess a covariance function with the same differentiability properties as the Matérn Class. Moreover, for any given location \( s_0 \in D \), the purely temporal process \( Z(s_0, \cdot) \) can have long-range dependence in time. Inference on these models will be conducted from a Bayesian perspective through Markov chain Monte Carlo (MCMC) methods, as described in Section 4. Section 5 illustrates through simulated examples that we can conduct sensible inference with these models. Code for the implementation of this inference is freely available on http://www.warwick.ac.uk/go/msteel/steel_homepage/software/.
A General Class of Nonseparable Space-time Covariance Models

application to a well-known data set, the Irish wind data, is provided in Section 6. The final section concludes. Proofs will be deferred to Appendix B without mention in the text.

2. Mixture representation

Let \((U, V)\) be a bivariate nonnegative random vector with distribution \(\mu(u, v)\) and independent of \(\{Z_1(s), s \in D\}\) and \(\{Z_2(t), t \in T\}\) as presented in (2). Define the process

\[
Z(s, t) = Z_1(s; U)Z_2(t; V), \ (s, t) \in D \times T,
\]

where \(Z_1(s; u)\) is a purely spatial random process for every \(u \in \mathbb{R}_+\) with covariance function \(C_1(s; u)\) that is a stationary covariance for \(s \in D\) and every \(u \in \mathbb{R}_+\) and a measurable function of \(u \in \mathbb{R}_+\) for every \(s \in D\). And \(Z_2(t; v)\) is a purely temporal random process for every \(v \in \mathbb{R}_+\) with covariance function \(C_2(t; v)\) that is a stationary covariance for \(t \in T\) and every \(v \in \mathbb{R}_+\) and a measurable function of \(v \in \mathbb{R}_+\) for every \(t \in T\). In particular, \(Z_1(s; U)\) and \(Z_2(t; V)\) are defined as the purely spatial and purely temporal processes \(Z_1(s)\) and \(Z_2(t)\), respectively, applied to functions of both arguments. For example, Ma (2002) considers \(Z_1(s; U) = Z_1(sU)\) and \(Z_2(t; V) = Z_2(tV)\) which results in a scale mixture covariance. The corresponding covariance function of \(Z(s, t)\) as in (4) is a convex combination of valid separable covariance functions. It is valid (see Ma, 2002, 2003b), generally non-separable, and is given by

\[
C(s, t) = \int C_1(s; u)C_2(t; v)d\mu(u, v).
\]

This reduces to a separable covariance function if \(U\) and \(V\) are independent. This covariance function is a mixture of separable covariance functions and conditional on \(U = u_0\) and \(V = v_0\) the process \(Z(s, t)\) possesses a separable covariance \(C_1(s; u_0)C_2(t; v_0)\). Note that if \(U = V\) then we obtain the model proposed in Ma (2003a) and also in De Iaco et al. (2002). Using the Fourier representation corresponding to (1) we obtain

\[
f(h_1, h_2) = (2\pi)^{-(d+1)} \int \int f_1(h_1; u)f_2(h_2; v)g(u, v)du dv,
\]

that is, the spectral density is also written as the mixture of a spatial and a temporal spectral density. Clearly, if \((U, V)\) are independent the spectral density is just the product of the purely spatial and the purely temporal spectral densities.

One feature of the mixing approach is that it generates a large variety of valid nonseparable, spatiotemporal covariance models, by using appropriate choices of the mixing function and the purely spatial and temporal covariances. Moreover, it inherits the advantages of the well-known theory developed for purely spatial and purely temporal processes in the joint modelling of space-time interactions. Spatial statistics methods have been available for some years and substantial effort has been made to understand the properties of the processes and to abandon
convenient but inappropriate assumptions in the modelling of spatially referenced data. For instance, Stein (1999) presents a very complete theoretical review of spatial methods. In a more applied setting Banerjee et al. (2003) present applications of spatial methods to real data sets. 

In attempts to abandon the assumption of stationarity in spatial processes, Higdon et al. (1999), Fuentes and Smith (2001) and Sampson and Guttorp (1992) propose approaches that accommodate non-stationary features of spatial datasets. To allow for deviations from Gaussianity, De Oliveira et al. (1997) and Palacios and Steel (2006) propose models that accommodate non-Gaussian tail behaviour. Diggle et al. (1998) propose the use of Generalized Linear Models for spatial processes. Also, in the context of time series, the use of non-stationary and non-Gaussian models is common practice. We believe that the efforts to build complex spatial and also temporal models should be exploited in the development of spatiotemporal models. The mixture approach used here allows this in a very natural way, as decisions regarding the modelling of spatial and temporal components may be taken separately. In principle, these components can be used directly as building blocks for the construction of spatio-temporal models.

Various authors have derived covariance models using the mixing representation, such as De Cesare et al. (2001), Ma (2002), Ma (2003b) and Porcu et al. (2007), leading to a number of (partially overlapping) models. The simplest special case of (5) is given by a discrete mixture

\[ C(s, t) = p_{11}C_1(s)C_2(t) + p_{10}C_1(s)C_2(0) + p_{01}C_1(0)C_2(t) + p_{00}C_1(0)C_2(0), \]

where \( C_1(s; u) = C_1(sw), \) \( C_2(t; v) = C_2(tv) \) and \( P(U = i, V = j) = p_{ij}, \) \( i, j \in \{0, 1\} \). A closely related model is the product-sum model of De Cesare et al. (2001) where some constraints are imposed in order to guarantee positive definiteness. Ma (2002, 2003b) develops nonseparable covariances using two approaches. One is a positive power mixture where \( C_1(s; u) = C_1(s)^u, \) \( C_2(t; v) = C_2(t)^v \) and \( (U, V) \) is a nonnegative bivariate discrete random vector with probability function \( \{p_{ij}, (i, j) \in \mathbb{Z}_+^2\} \). The other is a scale mixture that uses a nonstationary version of (5) where \( C_1(s_1, s_2; u) = C_1(s_1u, s_2u) \) depends on \( s_1 \) and \( s_2 \) and \( C_2(t_1, t_2; v) = C_2(t_1v, t_2v) \) depends on \( t_1 \) and \( t_2 \).

In order to guarantee positive definiteness of the covariance function we now assume \( C_1(s; u) = \exp(-\gamma_1 u) \) and \( C_2(t; v) = \exp(-\gamma_2 v) \). This is equivalent to \( C_1(s; u) \) and \( C_2(t; v) \) being valid covariance functions for any \( u, v > 0 \) (see e.g. Chilès and Delfiner, 1999, p. 66-67) and ensures validity of the resulting covariance function in (5). Moreover, this choice will lead to a closed form integral and a mathematically convenient class of functions.

**Proposition 2.1.** Consider a bivariate nonnegative random vector \((U, V)\) with joint moment generating function \( M(., .) \). If \( \gamma_1 = \gamma_1(s) \) is a purely spatial variogram on \( D \), \( \gamma_2 = \gamma_2(t) \) is a purely temporal variogram on \( T \) and \( C_1(s; u) = \exp(-\gamma_1 u) \) and \( C_2(t; v) = \exp(-\gamma_2 v) \), then (5)
becomes

\[ C(s, t) = M(-\gamma_1, -\gamma_2), \quad (s, t) \in D \times T, \quad (7) \]

which is a valid spatiotemporal covariance function on \( D \times T \).

This is the result presented in Theorem 3 of Ma (2003b). Note that Porcu et al. (2007) also propose nonseparable spatiotemporal covariance functions based on the mixture approach as in Ma (2002) and Ma (2003b). Their class is also generated as a particular case of Theorem 3 in Ma (2003b). However, they tend to deal with each dimension in space separately, which makes it easier to accommodate anisotropy but harder to replicate popular covariance structures (such as a Matérn) in space.

In the sequel, we follow the specification (7) and propose a general way to define the nonnegative bivariate random vector \((U, V)\) that leads to flexible nonseparable covariance functions with very useful properties. The joint distribution of \((U, V)\) is chosen in order to satisfy certain requirements based on a review of the literature. Most of the proposed classes result in very similar structures that lead to the same kind of margins in space and time: either Cauchy or Matérn. We are aware that the analysis of margins is not everything that matters in the modeling of spatiotemporal processes but it is a starting point to understand new classes proposed in the literature. Thus, we would like the proposed class to generate a range of different margins, including the Cauchy and the Matérn classes. Also, we would like the new class to have a closed form expression to facilitate computation. This new class is developed and its properties are studied in the next Section.

3. A nonseparable covariance structure

**Proposition 3.1.** Consider \(X_0, X_1, \text{ and } X_2\), which are independent nonnegative random variables with finite moment generating functions \(M_0, M_1, \text{ and } M_2\), respectively and define \(U = X_0 + X_1 \text{ and } V = X_0 + X_2\). Let \(C_1(s; u) = \sigma \exp\{-\gamma_1 u\} \text{ and } C_2(t; v) = \sigma \exp\{-\gamma_2 v\}\), with \(\gamma_1 \text{ and } \gamma_2\) as in Proposition 2.1. Then the resulting space-time covariance function from (5) is

\[ C(s, t) = \sigma^2 M_0(-\gamma_1 - \gamma_2) M_1(-\gamma_1) M_2(-\gamma_2), \quad (s, t) \in D \times T, \quad (8) \]

where \(\sigma^2\) is the space-time variance.

Notice that if \(U \text{ and } V\) are uncorrelated then the separable case is obtained and \(C(s, t) = \sigma^2 M_1(-\gamma_1) M_2(-\gamma_2)\). The class generated as in Proposition 3.1 is a very flexible class since it allows for different spatial and temporal marginals and space-time interactions, depending on the distributions of \(X_0, X_1, \text{ and } X_2\). As a consequence of the construction, any nonzero correlation between \(U \text{ and } V\) will always be positive.
We now choose particular univariate distributions for $X_0$, $X_1$ and $X_2$. Notice that this choice will define the kind of dependence obtained in space, time and space-time. It is usual practice in applications to use the Cauchy or the Matérn classes of covariance functions. Stein (1999) strongly recommend the use of the Matérn function for spatial data sets due to its flexibility in terms of smoothness properties. Gneiting and Schlather (2004) recommend the Cauchy class to model time series as it generates different dependence behaviour such as short and long range dependence. Thus, we define $(X_0, X_1, X_2)$ so that we can obtain Cauchy and Matérn classes in the temporal and spatial components, respectively. For that purpose we use the generalized inverse Gaussian (GIG) distributions, which are described in Appendix A. A gamma distribution with shape parameter $\lambda$ and scale $a$ (with mean $\lambda/a$) will be denoted as $\text{Ga}(\lambda, a)$.

**Theorem 3.1.** Consider $X_i \sim \text{Ga}(\lambda_i, a_i)$ for $i = 0, 2$ and $X_1 \sim \text{GIG}(\lambda_1, \delta, a_1)$, then the corresponding space-time covariance function generated through Proposition 3.1 is

$$C(s, t) = \sigma^2 \left\{ 1 + \frac{\gamma_1 + \gamma_2}{a_0} \right\}^{-\lambda_0} \left\{ 1 + \frac{\gamma_1}{a_1} \right\}^{-\lambda_1 - \frac{\lambda}{2} \frac{K_{\lambda_1}(2\sqrt{(a_1 + \gamma_1)\delta})}{K_{\lambda_1}(2\sqrt{a_1\delta})}} \left\{ 1 + \frac{\gamma_2}{a_2} \right\}^{-\gamma_2},$$

where $K_{\lambda}(.)$ is the modified Bessel function of the second kind of order $\lambda$. Permitted parameter values are $\sigma^2 > 0$, $\lambda_0 \geq 0$, $a_0 > 0$, $\lambda_2 > 0$, $a_2 > 0$, and we allow for $a_1 > 0$, $\delta \geq 0$ if $\lambda_1 > 0$, whereas $a_1 > 0$, $\delta > 0$ if $\lambda_1 = 0$ and $a_1 \geq 0$, $\delta > 0$ if $\lambda_1 < 0$.

Notice that in Theorem 3.1 the structure derived in space is more complex than the one derived in time. If the main interest were about time, we could have put a GIG distribution on $X_2$ instead to generate more complex structures in time (in principle, we could even use GIG distributions for both $X_1$ and $X_2$). When $a_1 = 0$ we use the asymptotic formula $K_{\lambda_1}(x) = 2^{\lambda_1 - 1} \Gamma(\lambda_1)x^{-\lambda_1}$ resulting in an inverse gamma moment generating function as will be illustrated in Model 2 (Subsection 3.1). In the representation (9), the parameter $\sigma^2$ is the space-time variance, that is, $\sigma^2 = C(0, 0)$. The parameters $a_1$ and $\delta$ parameterize the rate of decay for the spatial correlation and $a_2$ has the same role in the temporal dimension. To avoid lack of identifiability in the model we fix $a_0 = 1$. Note that if $a_0 \neq 1$ the resulting class would not change but we would have a superfluous scale parameter. Contour plots of some spatiotemporal covariance functions for this class are illustrated in Figure 1, using isotropic choices for $\gamma_1$ and $\gamma_2$.

It is important to measure separability in space time in the proposed model. We suggest to use the correlation between the variables $U$ and $V$ as an indication of interaction between space and time components. This correlation is given by

$$c = \frac{\lambda_0}{\sqrt{(\lambda_0 + \lambda_1)(\lambda_0 + \lambda_2/a_2^2)}},$$

where $V_1 = \text{Var}(X_1)$ is defined in (23) for $a = a_1 > 0$, $\delta > 0$ and $\lambda = \lambda_1$. Thus, $0 \leq c \leq 1$ could be used as a measure of space-time interaction, with $c = 0$ indicating separability and $c = 1$
meaning high dependence between space and time. In Figure 1 we have indicated the implied values of \( c \). The parameter \( \lambda_0 \) plays the role of separability parameter and the separable case is obtained for \( \lambda_0 = 0 \). In this case we say that \( X_0 = 0 \), implying \( U = X_1 \) and \( V = X_2 \). On the other hand, if \( \lambda_0 \to \infty \), \( U \to V \) resulting in an extreme non-separable case. Plots of a separable \((c = 0)\) and a nonseparable \((c = 0.98)\) spatiotemporal covariance function for this class are given in Figure 2. Note from the plots of \( \rho(s, t)/\rho(0, t) = C(s, t)/C(0, t) \) that the decay of the correlations in space is less rapid for larger differences in time, \( t \). We can show that whenever \( \lambda_0 > 0 \) the ratio \( C(s, t)/C(0, t) \) is always a strictly increasing function of \( t \) for any \( s \), so this behaviour is a feature of the construction. We have investigated the issue of measuring the degree of nonseparability of spatiotemporal processes in Fonseca and Steel (2009a). The latter paper developed a general nonseparability measure which can be applied to any class of nonseparable covariance functions. According to this general measure, the model proposed here is able to reach the full range of nonseparability whereas that is not the case for other models proposed in the literature (Cressie and Huang, 1999; Gneiting, 2002). In addition, the correlation in (10) proves to be quite a good proxy for the general nonseparability measure. Just like most of the models in the literature (Cressie and Huang, 1999; Gneiting, 2002), it can be shown (Fonseca and Steel, 2009a) that our model excludes negative nonseparability in the sense of Rodrigues and Diggle (2009).

In the case where \( \lambda_0 = 0 \) and \( \gamma_1 = ||s||^2 \) the spatial margin is a generalization of the Matérn Class, proposed by Shkarofsky (1968) that allows two complementary positive scale parameters \( a_1 \) and \( \delta \). If also \( a_1 = 0 \) we obtain the Matérn Class as a particular case. For the temporal margin, if \( \gamma_2 = |t|^\beta \) and the process is separable the resulting covariance function is in the Cauchy Class (Gneiting and Schlather, 2004). This class is the temporal margin obtained for most of the non-separable models proposed in the literature providing flexible power-law correlations that generalize stochastic models used in several fields. When \( \lambda_0 \neq 0 \) we have a model that is similar to the generalized Matérn \((a_1 \neq 0)\) and to the Matérn \((a_1 = 0)\) in the space dimension and similar to the Cauchy class in the time dimension but it also allows for space-time interactions.

Another important feature of the class (9) is that for any given \( s_0 \in D \), the purely temporal process \( Z(s_0, \cdot) \) can have long-range dependence, a global characteristic associated with power law correlations. Consider \( \gamma_2 = |t|^\beta \), if \( 0 < \lambda_0 + \lambda_2 < 1/\beta \) then \( C(0, t) \sim |t|^{-\beta(\lambda_0+\lambda_2)} \) as \( t \to \infty \) and the purely temporal process is said to have long memory dependence. This characteristic implies that correlations between distant times decay much slower than for standard ARMA or Markov-type models. The closer \( \lambda_0 + \lambda_2 \) is to zero, the stronger the dependence of the process. To examine the smoothness properties of the process \( \{Z(s, t) : s \in D, \ t \in T\} \), we study the behaviour of \( C(s, t) \) across space for a fixed time \( t_0 \in T \) and across time for a fixed
Fig. 1. Contour plot of $C(s, t)$ in (9) for $\gamma_1(s) = ||s||^\alpha$, $\gamma_2(t) = |t|^\beta$, $\sigma^2 = 1$, $\delta = 1/50$, $\alpha = 1.5$, $\beta = 1.5$, $\lambda_0 = 1$, $\lambda_1 = 1/4$ and $\lambda_2 = 1/8$. The horizontal axis represents the spatial lag $h_s = ||s||$ and the vertical axis, the temporal lag $h_t = |t|$. The separability measure $c$ is as defined in (10).

location $s_0 \in D$. Generally, $f^{(q)}$ will denote the $q^{th}$ derivative of a function $f$. Consider the Taylor expansion of an isotropic $\gamma_1(s)$ about 0 given by $\sum_{k=0}^{\infty} c_k ||s||^k$. Define $l$, the smallest power of $||s||$ in the Taylor expansion such that $c_l \neq 0$. Some insight about the behaviour of positive definite functions in a neighborhood of zero can be found in Stein (1999, p. 28-33).

**Proposition 3.2.** Under the conditions of Theorem 3.1,
(a) the purely temporal process $\{Z(s_0, t) : t \text{ in an interval } \subseteq T\}$ at a fixed location $s_0 \in D$ is $m$ times mean square differentiable if and only if $\gamma_2^{(2m)}(0)$ exists and is finite.
(b) When $a_1 \neq 0$, the purely spatial process $\{Z(s, t_0) : s \text{ in a connected set } \subseteq D\}$ at a fixed time $t_0 \in T$ is $m$ times mean square differentiable if and only if $\gamma_1^{(2m)}(0)$ exists and is finite. When $a_1 = 0$, the purely spatial process is $m$ times mean square differentiable if and only if $2m < -l\lambda_1$ and $\gamma_1^{(2m)}(0)$ exists and is finite.

Note that we are here assuming isotropic variograms $\gamma_1(s)$ and $\gamma_2(t)$ throughout, and thus it
is straightforward to define the notion of derivatives at 0 (as used in Stein, 1999, page 20-22). Otherwise more conditions would be required in order to define the degree of smoothness, as explained in Gelfand and Banerjee (2003). The proposed covariance model (9) allows for different degrees of smoothness across space and time obtained by choosing the parameter $\lambda_1$ and the functions $\gamma_1$ and $\gamma_2$. When $a_1 = 0$ (Matérn Class), the parameter $\lambda_1 < 0$ has a direct effect on the smoothness of the spatial process, and larger values of $-\lambda_1$ correspond to smoother processes. The Matérn Class is very flexible, in the sense that the model allows for the degree of smoothness to be estimated from the data rather than restricted a priori. But some characteristics, such as no cusp at the origin and negative second derivative, required in e.g. turbulence applications, are not fulfilled by this class. On the other hand, the covariance function for the purely spatial process obtained when $a_1 \neq 0$ has no cusp at the origin, that is, $\frac{d}{ds} C(s, 0)$ goes to zero as $s$ approaches zero. Furthermore, its second derivative always exists.
is finite and is negative if \( \gamma_1^{(2)}(0) \) exists and is finite. If we take \( \gamma_1 = ||s||^\alpha \), \( \gamma_2 = |t|^\beta \), \( \alpha \in (0, 2] \) and \( \beta \in (0, 2] \), we obtain the following.

**Corollary 3.1.** Suppose \( \gamma_1 = ||s||^\alpha \), \( \gamma_2 = |t|^\beta \), \( \alpha \in (0, 2] \) and \( \beta \in (0, 2] \) in addition to the conditions of Proposition 3.2. The temporal process is mean square continuous for \( \beta \in (0, 2) \) and it is infinitely mean square differentiable for \( \beta = 2 \). The spatial process is mean square continuous for \( \alpha \in (0, 2) \). When \( \alpha = 2 \), the process is \( m \) times mean square differentiable for \( a_1 = 0 \) and \( -\lambda_1 > m \) and it is infinitely mean square differentiable for \( a_1 \neq 0 \).

### 3.1. Parameterisation

We have specified a rich class of covariance structures in (9), and we now discuss useful parameterisations and interesting subclasses. As mentioned above, \( \sigma^2 \) is the space-time variance, and the parameters \( a_1 \) and \( \delta \) are spatial scales, while \( a_2 \) is a scale parameter in the temporal dimension. We now introduce extra scale parameters in the variograms \( \gamma_1 \) and \( \gamma_2 \) by taking \( \gamma_1(s) = ||s/a||^\alpha \) and \( \gamma_2(t) = |t/b|^\beta \), where \( \alpha \in (0, 2] \) and \( \beta \in (0, 2] \). Note that these extra scale parameters do not change any of the results on smoothness or temporal dependence explained above.

We have to introduce restrictions on the scale parameters, since the proposed class would now have \((a_0, a_1, a_2, \delta, a, b)\) as scales. As stated before, we fix \( a_0 = 1 \) because we already have scales in space and time and this extra one would be superfluous. Another reason for that choice is that \( a_0 \) would also influence the degree of separability in space and time. In particular, for \( a_0 \to 0 \) we would obtain \( c \to 1 \) (strong nonseparability) and when \( a_0 \to \infty \) then \( c \to 0 \) (separability). By fixing \( a_0 \) we keep just \( \lambda_0 \) as a separability parameter. The same motivation leads us to consider \((a, b)\) instead of \((a_1, a_2, \delta)\) as scales. Since \((a_1, a_2, \delta)\) also appears in expression (10) for the degree of dependence in space and time, it would again be confounded with \( \lambda_0 \). On the other hand, \((a, b)\) do not enter in this measure of separability. To avoid redundancy, Cressie and Huang (1999) use the same scales as suggested here. The main difference is that in their example 5, for instance, the extra scales are actually necessary for the existence of a flexible separability parameter in the model while here \((a_1, \delta, a_2)\) and \((a, b)\) are always redundant if used at the same time.

As discussed before, the functions \( \gamma_1(s) \) and \( \gamma_2(t) \) determine the smoothness of the random process \( Z(s, t) \). It is difficult to decide whether to fix \( \beta \) at a particular value or not since for \( \beta = 2 \) the process is infinitely smooth and for \( \beta < 2 \) it is not even once mean square differentiable. The same holds for \( \alpha \) when \( a_1 \neq 0 \). Depending on the application it might be appropriate to estimate these parameters. If we do set \( \alpha = 2 \) then this generates the subclass that gives the Matérn covariance function in space when \( a_1 = 0 \) and under separability. The Matérn class is
important for spatial applications since, besides the scale parameter, it also has a smoothness parameter controlling the differentiability of the random field (see Corollary 3.1).

In the following we present some interesting parametric families of spatiotemporal covariance functions adhering to these parameter restrictions. Throughout, \(a_2 = 1\) and we take \(a > 0\) as the scale in space and \(b > 0\) as the scale in time. The models below differ in how we constrain \(\delta\) and \(a_1\).

**Model 1**

An interesting model is obtained by setting \(\delta = a_1\) which means that \(\delta\) is now a concentration parameter (see Jørgensen, 1982). The resulting covariance function is

\[
C(s, t) = \sigma^2 \left\{ 1 + \frac{|s/a|^\alpha + |t/b|^\beta}{\delta} \right\}^{-\lambda_0} \left\{ 1 + \frac{|s/a|^\alpha}{\delta} \right\}^{\lambda_1/2} \frac{K_{\lambda_1} \left( 2\delta \sqrt{1 + \frac{|s/a|^\alpha}{\delta}} \right)}{K_{\lambda_1}(2\delta)} \left\{ 1 + |t/b|^\beta \right\}^{-\lambda_2}. \tag{11}
\]

The dependence in space and time is given by (10) where \(V_1(\lambda_1, \delta) = \text{Var}(X_1)\) and \(X_1 \sim \text{GIG}(\lambda_1, \delta, \delta)\) and \(a_2 = 1\). If \(\lambda_0 = 0\), we have independence between \(U\) and \(V\) and the purely spatial covariance is a generalized Matérn for \(\alpha = 2\) while the purely temporal covariance is in the Cauchy Class.

**Model 1a**

As a special case of Model 1, consider \(\lambda_1 = -1/2\) so that \(X_1 \sim \text{InvGaussian}(\delta, \delta)\) and the space-time covariance function is

\[
C(s, t) = \sigma^2 (1 + |s/a|^\alpha + |t/b|^\beta)^{-\lambda_0} (1 + |t/b|^\beta)^{-\lambda_2} \exp \left\{ -2\delta \left[ 1 + \frac{|s/a|^\alpha}{\delta} \right] - 1 \right\}. \tag{12}
\]

If \(\lambda_0 = 0\) the purely spatial covariance is a shifted version of the exponential covariance function for \(\alpha = 2\) and the purely temporal covariance is in the Cauchy Class.

**Model 2**

Consider \(\lambda_1 < 0, a_1 = 0\) and \(\delta = 1\). Then, \(X_1\) has an inverse gamma distribution \(X_1 \sim \text{InvGa}(\nu = -\lambda_1, 1)\) and the space-time covariance function is

\[
C(s, t) = \sigma^2 \left\{ 1 + |s/a|^\alpha + |t/b|^\beta \right\}^{-\lambda_0} \frac{2\delta \sqrt{1 + |s/a|^\alpha}}{2^{\nu-1} \Gamma(\nu)} K_{\nu} \left( 2|s/a|^\alpha \right)^\nu \left\{ 1 + |t/b|^\beta \right\}^{-\lambda_2}. \tag{13}
\]

As the variance of \(X_1\) does not exist (unless \(\nu > 2\) is imposed through the prior), the dependence in space and time is now measured by

\[
\tilde{c} = \frac{\lambda_0}{\sqrt{(\lambda_0 + V_1(\nu))(\lambda_0 + \lambda_2)}}, \tag{14}
\]
where $\tilde{V}_1(\nu) = (Q(0.75; \nu) - Q(0.25; \nu))^2$ and $Q(x; \nu)$ is the quantile of $X_1$ corresponding to $x$. Under independence of $U$ and $V$ the purely spatial covariance is in the Matérn Class (with smoothness parameter $\nu$) if we take $\alpha = 2$ and the purely temporal covariance is in the Cauchy Class. Finally, if we use $\lambda_0 = 0$ in combination with $\nu = 1/2$ we generate a powered exponential covariance structure in space.

Model 3
Consider $\lambda_1 > 0$ and the restrictions $\delta = 0$ and $a_1 = 1$. Then, $X_1 \sim Ga(\lambda_1, 1)$ and the resulting space-time covariance function is

$$C(s, t) = \sigma^2 \left[ 1 + \frac{|s/a|^{\alpha} + |t/b|^\beta}{1 - \tau^2} \right] - \lambda_0 \left[ 1 + \frac{|s/a|^{\alpha}}{1 - \tau^2} \right] - \lambda_1 \left[ 1 + \frac{|t/b|^\beta}{1 - \tau^2} \right] - \lambda_2. \quad (15)$$

In this example, the random vector $(U, V)$ has the bivariate gamma distribution of Cheriyan-Ramabhardran (see Kotz et al. 2000, p. 432) and a closely related model is Example 6 of Ma (2002) where a nonstationary function is considered. If $\lambda_0 = 0$, $U$ and $V$ are independent gamma random variables and both purely spatial and purely temporal covariances are in the Cauchy Class. Nonseparability can be measured as in (10) with $V_1 = \lambda_1$ and $a_2 = 1$.

3.2. Including a Nugget Effect
In practice, it is often useful to consider discontinuities at the origin (nugget effect), to capture measurement error and small scale variation. This can be done in a natural way, using the mixture construction. Let us focus on a spatial nugget effect in what follows, but a temporal nugget effect can be dealt with in a similar fashion. Instead of (4) we consider $Z^*(s, t) = Z_1^*(s; U)Z_2(t; V)$, $(s, t) \in D \times T$, where $Z_1^*(s; U) = \sqrt{1 - \tau^2}Z_1(s; U) + \tau \epsilon(s)$ with $\{\epsilon(s) : s \in D\}$ a process with zero mean, variance one and $\text{Cov}(\epsilon(s_1), \epsilon(s_2)) = 0$ if $s_1 \neq s_2$, and $0 < \tau < 1$. We assume that $\epsilon(s)$ is uncorrelated with the purely spatial and purely temporal processes. Under the conditions of Proposition 3.1, the resulting covariance function is then

$$C^*(s, t) = \sigma^2 M_0(-\gamma_1 - \gamma_2)M_1^*(-\gamma_1)M_2(-\gamma_2), \quad (16)$$

where $M_1^*(-\gamma_1) = (1 - \tau^2)M_1(-\gamma_1) + \tau^2 I(s = 0)$ is a convex combination of a valid covariance function and a nugget effect, rendering the expression in (16) a valid covariance function.

4. Bayesian model and inference
Consider that observations $z_{ij}$ are obtained at locations $s_i$, $i = 1, \ldots, I$ and time points $t_j$, $j = 1, \ldots, J$. We confine ourselves to Gaussian joint distributions and the likelihood function is
given by
\[ l(\theta, \sigma^2, \mu; z) = (2\pi)^{-\frac{N}{2}} |\Sigma(\theta)|^{-1/2} \exp \left\{ -\frac{1}{2} (\text{Vec}(z) - \mu)^T \Sigma(\theta)^{-1} (\text{Vec}(z) - \mu) \right\}, \] 
with \( N = IJ \) and \( \Sigma(\theta) \) has elements
\[ \Sigma(\theta)_{kk'} = C(s_k - s_{k'}, t_k - t_{k'}; \theta), \ k, k' = 1, \ldots, N, \]
where \( C(s, t; \theta) \) is either of the covariance functions described in Subsection 3.1, possibly including a nugget effect as in (16). To complete the Bayesian model, we specific a prior on the parameter vector \( \theta \), which always contains \( \sigma^2, a, b, \beta, \lambda_0, \lambda_1, \lambda_2 \) as well as \( \delta \) for Model 1 (while for Model 1a it contains \( \delta \) but excludes \( \lambda_1 \)). We consider independent priors in line with the more or less clear-cut different roles of the parameters. For the scale parameters in space and time we adopt gamma distributions, \( a \sim \text{Ga}(1, c_1/\text{med}(d_s)) \) and \( b \sim \text{Ga}(2, 2) \). Here we have defined \( \text{med}(d_s) \) as the median spatial distance in the data, so that the prior on \( a \) takes into account the scaling of \( s \). Strictly speaking, this is in violation of Bayesian principles, where the prior can not depend on observed data, but this is merely a device to make sure the prior is properly calibrated in terms of getting the right order of magnitude for \( a \); small changes to this prior do not have any noticeable effect but any change in the units of measurement will automatically be taken into account (see Palacios and Steel, 2006 for a discussion of such priors in the context of a prior sensitivity analysis). The prior distribution is chosen in order to imply a prior for \( c \) that gives ample mass to values close to 0. That is, we do not want a prior distribution for \( c \) that unduly favours nonseparable models, as we would like to keep using separable (simpler) models unless there is data evidence to the contrary. This consideration lead us to use \( \lambda_0 \sim \text{Ga}(0.5, 1) \) in combination with \( \lambda_2 \sim \text{Ga}(2, 1), \lambda_1 \sim N(0, 4), \delta \sim \text{Ga}(2, 1) \) for Model 1 (Generalized Matérn covariance in space). For Model 2 (Matérn covariance in space), we adopt \( \lambda_0 \sim \text{Ga}(1, 1), \lambda_2 \sim \text{Ga}(3, 1), \nu = -\lambda_1 \sim \text{Ga}(2, 1) \). And for Model 3 (Cauchy covariance in space), we adopt \( \lambda_0 \sim \text{Ga}(1, 1), \lambda_2 \sim \text{Ga}(3, 1) \) and \( \lambda_1 \sim \text{Ga}(2, 1) \). In the case where \( \lambda_2 \) and/or \( \lambda_1 \) is fixed we suggest the use of an Exponential prior with mean 1/5 for \( \lambda_0 \). The prior distribution for \( \sigma^{-2} \) is \( \text{Ga}(10^{-6}, 10^{-6}) \), while we use a uniform distribution on (0, 2) for \( \alpha \) and \( \beta/2 \sim \text{Beta}(3, 2) \). For the models with nugget effect, the prior on \( \tau^2 \) is \( \text{Ga}(2, 6) \). These prior distributions are rather noninformative, yet impose a reasonable amount of structure on the fairly high-dimensional parameter space. For more detailed discussion on the choice of prior distributions in a spatial context, see Palacios and Steel (2006).
We need to be able to deal with a non-constant mean surface \( \mu(s, t) \) in (17). Complex space-time trends are common in spatiotemporal data sets. For instance, temporal periodicity due to seasonal fluctuations may be combined with spatial trends due to geologic characteristics. The simplest case is when \( \mu(s, t) \) is specified by a linear function of location, time and possible
explanatory variables as follows

\[ \mu(s, t) = \sum_{k=1}^{p} \psi_k f_k(s, t), \]  

(18)

where \( f_k(s, t), \ k = 1, \ldots, p \) are functions of location, time or location and time, and \( \psi = (\psi_1, \ldots, \psi_p)' \) are unknown coefficients. As pointed out by Møller (2003, p. 54-56), linear or quadratic trend surfaces are useful descriptions of the spatial mean but more complicated polynomials are seldom useful since they lead to unrealistic extrapolations beyond the region of observed locations. We suggest standardizing the functions \( f_k(s, t) \) and using the prior

\[ \psi | \sigma^2 \sim N_p(0, \sigma^2 V_0) \]

with \( V_0 = \sigma_0^2 I_p \), where \( \sigma_0^2 \) is large and \( I_p \) is the identity matrix.

We use stochastic simulation via MCMC to obtain an approximation of the posterior distribution of \((\theta, \psi)\). We obtain samples from the target distribution \( p(\theta, \psi|z) \) by successive generations from the full conditional distributions. More specifically, we use a hybrid Gibbs sampler scheme with Metropolis-Hastings steps. The evaluation of the likelihood is the main computational requirement, and this is done through an efficient and accurate approximation as described in Appendix C.

Model comparison is conducted on the basis of Bayes factors. These are ratios of marginal likelihoods computed from the MCMC output through three different methods. We use the estimator \( p_4 \) of Newton and Raftery (1994) (with their \( d \) as small as 0.01), the optimal bridge sampling approach of Meng and Wong (1996), and the shifted gamma estimator proposed by Raftery et al. (2007) (with values of their \( \lambda \) close to one). Throughout, these three estimators lead to very similar results.

5. Simulated examples

Initially, we illustrate the proposed covariance model with some simulated examples. We generated data sets from a Gaussian process \( \{Z(s, t) : s \in D, \ t \in T\} \) with zero mean and covariance structure specified through Theorem 3.1. The data sets consist of \( I = 20 \) spatial locations and \( J = 20 \) time points, totalling \( N = 400 \) observations. Throughout, we generate data using \( \gamma_1 = ||s/a||^{\alpha} \) and \( \gamma_2 = |t/b|^{\beta} \), where we fix \( \alpha = \beta = 2 \) (these parameters are not estimated in this Section). Samples from the posterior distribution were obtained from MCMC chains using a burn-in of 20,000 draws and recording every 120th from the remaining 180,000 draws resulting in a sample of 1500 values for each free parameter. Priors on the scale parameters \( a, b, \sigma^2 \) are as explained in Section 4.
Table 1. Simulated data Example 1: Posterior median and percentiles (2.5% and 97.5%) for the parameters of Model 2 as in (13).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Median</th>
<th>(2.5%, 97.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2 = 0.500$</td>
<td>0.533</td>
<td>(0.400, 0.717)</td>
</tr>
<tr>
<td>$\nu = 2.500$</td>
<td>2.534</td>
<td>(2.162, 3.175)</td>
</tr>
<tr>
<td>$a = 0.200$</td>
<td>0.218</td>
<td>(0.170, 0.271)</td>
</tr>
<tr>
<td>$\lambda_2 = 1.000$</td>
<td>0.619</td>
<td>(0.172, 1.472)</td>
</tr>
<tr>
<td>$b = 0.063$</td>
<td>0.059</td>
<td>(0.050, 0.072)</td>
</tr>
<tr>
<td>$\lambda_0 = 0.300$</td>
<td>0.424</td>
<td>(0.169, 0.910)</td>
</tr>
<tr>
<td>$\tilde{c} = 0.375$</td>
<td>0.526</td>
<td>(0.310, 0.762)</td>
</tr>
</tbody>
</table>

Table 2. Simulated data Example 1: Natural logarithm of the Bayes factor in favor of the nonseparable Model 2 versus its separable version.

<table>
<thead>
<tr>
<th>Method</th>
<th>Bayes Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newton-Raftery</td>
<td>10.8</td>
</tr>
<tr>
<td>Bridge sampling</td>
<td>10.8</td>
</tr>
<tr>
<td>Shifted gamma</td>
<td>9.8</td>
</tr>
</tbody>
</table>

5.1. Example 1 - Matérn covariance in space

We follow the set up suggested in (13) for the covariance function where the parameters in the model are $\sigma^2$ (space-time variance), $\nu$ (spatial smoothness), $a$ (spatial scale), $\lambda_2$, $b$ (time scale) and $\lambda_0$ (separability parameter). Thus, we use Model 2 and the parameter vector is given by $\theta = (\sigma^2, \nu, a, \lambda_2, b, \lambda_0)$. This particular data set was generated using $\theta = (0.5, 2.5, 0.2, 1.0, 0.0625, 0.3)$ implying a space-time dependence of $\tilde{c} = 0.38$ following (14). The prior distributions used were $\lambda_0 \sim \text{Ga}(1, 1)$, $\lambda_2 \sim \text{Ga}(3, 1)$ and $\nu \sim \text{Ga}(2, 1)$. The posterior estimates for the parameters shown in Table 1 are in good agreement with the true values. The prior distributions were quite noninformative. We also conduct inference from the model where we impose $\lambda_0 = 0$ (separable model) and the comparison between these two models is shown in Table 2. As we can see in Table 2 all three methods (Newton-Raftery, Bridge sampling and Shifted gamma) for estimating the marginal likelihood give strong (and very similar) evidence against the separable model in favor of the nonseparable model in this example.

5.2. Example 2 - Generalized Matérn covariance in space

Now we use Model 1, with the parameters $\theta = (\sigma^2, \lambda_1, \delta, a, \lambda_2, b, \lambda_0)$. In this example the data was generated using $\theta = (0.5, -2.0, 0.20, 0.20, 1.0, 0.0625, 0.3)$ implying a space-time de-
pendence of \( c = 0.37 \). The prior distributions used were \( \lambda_0 \sim \text{Ga}(0.5, 1) \), \( \lambda_2 \sim \text{Ga}(2, 1) \), \( \lambda_1 \sim N(0, 4) \) and \( \delta \sim \text{Ga}(2, 1) \). The posterior estimates for the parameters shown in Table 3 are in excellent agreement with the true values. We also conducted inference when we assume various different models. The first considers \( \lambda_0 = 0 \) (separable model) and the comparison is shown in Table 4. As we can see all the three methods (Newton-Raftery, Bridge sampling and Shifted gamma) for estimating the marginal likelihood give strong evidence against the separable model in favor of the nonseparable model in this example. Next we considered the model with mean given by \( m(s, t) = \delta_1 t^2 \). In this comparison, the methods Bridge sampling and Shifted gamma give some evidence against the model with nonzero mean structure but the Newton-Raftery method gives a roughly unitary Bayes factor. To double check we also calculated the Savage-Dickey density ratio estimator (Verdinelli and Wasserman, 1995), which can be computed in this simple case and results in a log Bayes factor in favor of the “correct model” of 1.8 reinforcing the results given by Bridge sampling and Shifted gamma. This Savage-Dickey density ratio can also directly be used to estimate the Bayes factor versus Model 1a, where \( \lambda_1 = -1/2 \). This gives a log Bayes factor of 1.7 in favor of the model that generated the data, which gives us additional evidence that \( \lambda_1 \) is relatively well estimated. Finally, we consider the model with \( \alpha = 1.5 \) and \( \beta = 1 \). In this case, all three methods lead to very strong evidence against this model in favor of the model with \( \alpha = \beta = 2 \). It seems that this deviation from the correct model is much more critical than the misspecification of the mean structure. This might be expected, as the nonzero mean structure can simply be accommodated by having a posterior of \( \delta_1 \) close to zero, whereas changing the smoothness properties of the process can not so easily be compensated for. It is interesting to see that the Bayes factor strongly reacts to imposing the wrong smoothness properties, as it suggests these properties can be successfully estimated from the data (as we indeed do in the real data application in the next section).

5.3. Example 3 - Cauchy covariance in space

In this example the data was generated from Model 3 using \( \theta = (0.5, 2.5, 0.2, 1.0, 0.0625, 0) \). Notice that we are now generating data from the separable model (\( \lambda_0 = 0 \)) which gives us no space-time dependence (\( c = 0 \)). The prior distribution used were \( \lambda_0 \sim \text{Ga}(1, 1) \), \( \lambda_2 \sim \text{Ga}(3, 1) \) and \( \lambda_1 \sim \text{Ga}(2, 1) \) and the posterior estimates for the parameters are shown in Table 5. All the results shown here refer to the “wrong” model where \( \lambda_0 \) was estimated, rather than set to its correct value of zero. The posterior distribution of \( \lambda_0 \), however, does not assign a lot of mass far away from zero. Formal testing in Table 6 reveals that all three estimators (Newton-Raftery, Bridge sampling and Shifted gamma) indicate that both models are roughly equally well supported by the data. It is reassuring to find that no “spurious” nonseparability is introduced.
Table 3. Simulated data Example 2: Posterior median and percentiles for the parameters of Model 1 in (11).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Median</th>
<th>(2.5%, 97.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2 = 0.500$</td>
<td>0.530</td>
<td>(0.383, 0.814)</td>
</tr>
<tr>
<td>$\lambda_1 = -2.000$</td>
<td>-1.615</td>
<td>(-5.016, -0.904)</td>
</tr>
<tr>
<td>$\delta = 0.200$</td>
<td>0.252</td>
<td>(0.068, 0.818)</td>
</tr>
<tr>
<td>$a = 0.200$</td>
<td>0.210</td>
<td>(0.174, 0.269)</td>
</tr>
<tr>
<td>$\lambda_2 = 1.000$</td>
<td>1.273</td>
<td>(0.534, 2.701)</td>
</tr>
<tr>
<td>$b = 0.063$</td>
<td>0.064</td>
<td>(0.052, 0.082)</td>
</tr>
<tr>
<td>$\lambda_0 = 0.300$</td>
<td>0.290</td>
<td>(0.112, 0.585)</td>
</tr>
<tr>
<td>$c = 0.373$</td>
<td>0.339</td>
<td>(0.163, 0.514)</td>
</tr>
</tbody>
</table>

Table 4. Simulated data Example 2: Natural logarithm of the Bayes factor in favor of the nonseparable Model 1 versus its counterpart that is separable, has nonzero mean structure and assumes $\alpha = 1.5$ and $\beta = 1$.

<table>
<thead>
<tr>
<th>Model</th>
<th>Newton-Raftery</th>
<th>Bridge sampling</th>
<th>Shifted gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>separable</td>
<td>8.4</td>
<td>7.5</td>
<td>8.1</td>
</tr>
<tr>
<td>$m(s, t) = \delta_1 t^2$</td>
<td>0.04</td>
<td>2.8</td>
<td>1.5</td>
</tr>
<tr>
<td>$\alpha = 1.5$ and $\beta = 1$</td>
<td>303</td>
<td>298</td>
<td>305</td>
</tr>
</tbody>
</table>
Table 5. Simulated data Example 3: Posterior median and percentiles for the parameters in Model 3 as in (15) with $\lambda_0$ being estimated.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Median</th>
<th>(2.5%, 97.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2 = 0.500$</td>
<td>0.508</td>
<td>(0.389, 0.692)</td>
</tr>
<tr>
<td>$\lambda_1 = 2.000$</td>
<td>2.492</td>
<td>(1.240, 5.182)</td>
</tr>
<tr>
<td>$a = 0.2000$</td>
<td>0.249</td>
<td>(0.188, 0.338)</td>
</tr>
<tr>
<td>$\lambda_2 = 1.000$</td>
<td>0.833</td>
<td>(0.355, 1.483)</td>
</tr>
<tr>
<td>$b = 0.063$</td>
<td>0.065</td>
<td>(0.057, 0.077)</td>
</tr>
<tr>
<td>$\lambda_0 = 0.000$</td>
<td>0.269</td>
<td>(0.041, 0.724)</td>
</tr>
<tr>
<td>$\tilde{c} = 0.000$</td>
<td>0.155</td>
<td>(0.022, 0.340)</td>
</tr>
</tbody>
</table>

Table 6. Simulated data Example 3: Natural logarithm of the Bayes factor in favor of the nonseparable Model 3 versus its separable counterpart.

<table>
<thead>
<tr>
<th>Newton-Raftery</th>
<th>Bridge sampling</th>
<th>Shifted gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.76</td>
<td>0.42</td>
<td>0.39</td>
</tr>
</tbody>
</table>

through the prior in this case where the data are really generated from a separable model.

6. Application: Irish wind data

The model proposed in this paper can be easily extended to accommodate realistic features of space-time data as decisions regarding time and space can be taken separately and spatial and temporal components are clearly defined. In that context, we illustrate the potential of our proposal by modelling the well known Irish wind data for which nonseparability and asymmetry of the covariance function were identified by several authors. The data was initially described by Haslett and Raftery (1989) where the spatial dependence is used to provide better predictions of the wind power of a potentially new turbine. Several features of this data were not taken into account in this first paper. Gneiting (2002) analyzed this data accounting for nonseparability but the model was not able to accommodate asymmetries which are clearly necessary in this application. Stein (2005) also analyzed the Irish data but considered differences in time of the observations to avoid modelling the spatial means which complicates any comparison with our approach. The asymmetric model with different variances across space proved to be very useful in this application. Gneiting et al. (2007) fitted an asymmetric model for the wind data but the interaction parameter estimate was obtained by fixing all the other parameters at
previous estimates obtained for a separable model, and then estimating the space-time interaction parameter using weighted least squares. Moreover, the latter model assumes constant variance across space which is an unrealistic assumption for this data. Porcu et al. (2007) fit these data (through a composite likelihood method) using an anisotropic and nonseparable covariance structure with different range and nugget effects in both spatial dimensions. They find their model does roughly equally well as that of Gneiting et al. (2007) in terms of predictions in time. In this work, we account for all the features cited in the literature while estimating all the parameters jointly using the Bayesian paradigm. Note that the latter paradigm also immediately allows for predictive distributions that duly take into account parameter uncertainty.

The data consist of time series of daily average wind speed in m/s at 11 meteorological stations in Ireland during the period 1961-1978. We use UTM coordinates, so that the scale of the spatial coordinates is in kilometers and consider 10 years (1961-1970) of data. Following the literature we apply some transformations to the data. Firstly, we take the square root transformation in order to obtain data that are approximately normally distributed. Next, we estimate the seasonal effects by calculating the average of the square root of the daily means over years and stations for each day of the year and then regressing the result on the set of annual harmonics \( \sin(\frac{2\pi}{365}rt), \cos(\frac{2\pi}{365}rt) \), \( t = 1, \ldots, 365 \) and \( r = 1, 2, \ldots, 364/2 \). For this subset of the data we used \( r = 1, 2, 3, 6 \). We substract these estimated seasonal effects from the square root data and we work with the deseasonalized data.

The empirical correlation for the transformed wind speed decays fast in time, and for a lag of 4 days the empirical correlation is already very close to 0. Given the dimension of this data set we need to restrict our attention to few lags in time in order to make computation feasible. In this case, we use an approximation of the likelihood function which assumes that observations more than three days apart are uncorrelated. Without such a simplification, we would need to invert matrices of size \( 40, 150 \times 40, 150 \) at each step of the MCMC algorithm in the spatiotemporal modelling. The details about the calculation of the likelihood in this example are presented in Appendix C.

In an initial analysis of the data set, we fit a purely temporal model station by station. We assume nonzero mean \( \mu \) and a Cauchy covariance function given by

\[
C_2(t; \sigma^2, \beta, b) = \frac{\sigma^2}{1 + \frac{|t/b|^\beta}} - 1, \quad |t| \leq 10 \text{ days}
\]  

We consider Gaussian processes here, so for each station with location \( s_i, i = 1, \ldots, 11 \) we have \( z_i \sim N(\mu_i, \Sigma_2) \), where \( z_i \) groups the 3650 observations for station \( i \), \( \mu \) is a vector of ones and \( \Sigma_2_{jj'} = C_2(t_j - t_{j'}, \sigma^2, \beta, b) \). Priors for the relevant parameters are as described in Section 4. Figure 3 shows the posterior medians as well as the 95% credible intervals for the parameters in (19) for each station. For all stations the correlation decays very rapidly in
time. This preliminary analysis shows that the mean of the deseasonalized data varies over sites. This can be seen in Figure 3(d) which shows the posterior mean of the transformed wind speed \( \mu_i, i = 1, \ldots, 11 \) obtained from the purely temporal analysis for each station separately. The means are larger at the coastal sites than inland and a sensible model for the mean should take this into consideration. The earlier papers that analyzed these data subtract the mean by station or consider differences in time in order to avoid modelling this spatial trend. Since our goal is merely to compare models with different covariance structures, we subtract the estimated (through posterior means) station-specific means. The resulting data are often called velocity measures. We also notice from this purely temporal analysis that the variances differ considerably over sites (see Figure 3(a)). We want our model to capture this feature, so we consider a different variance for each site. Stein (2005) comments that allowing for variances to differ with location can improve the fit considerably in this application.

In the full spatiotemporal analysis we first consider the proposed nonseparable model (11) (i.e. Model 1) with a constant \( \mu(s,t) \) in (17). We set \( \lambda_1 = -1 \) and we estimate \( \delta = a_1, \alpha \) and \( a \). As in the purely temporal analysis and as in earlier studies we set \( \lambda_2 = 1 \) for the Cauchy covariance in time and estimate \( b \) and \( \beta \). We also want our model to capture discontinuities at the origin, so we include a purely spatial nugget effect as described in (16). Priors are as described in Section 4.

After a burn in of 10,000 iterations, we record every 18\textsuperscript{th} draw in an MCMC chain of 90,000, resulting in 5000 draws from the posterior distribution. The first two columns of Table 7 show summaries from the posterior distribution for the separable and nonseparable versions of Model 1. The nugget effect is non-negligible and well estimated for both versions, while the posterior distributions of \( \alpha \) and \( \beta \) suggests the process is not infinitely smooth in space and time. Note that the posterior mass for the separability parameter \( \lambda_0 \) is concentrated well away from zero, suggesting nonseparability. Such support for the nonseparable model is also suggested by the posterior distribution of \( c \) in (10), which gives an measure of the degree of dependence in space and time (with \( c = 0 \) indicating complete separability). The 95\% credible interval for \( c \) is \( (0.218, 0.501) \) indicating strong nonseparability in this example. Figure 4 displays the posterior density of \( c \), overlayed with its prior counterpart.

The fitted (posterior median) correlation function versus the empirical correlation is presented in Figure 5(a-b) for the separable and nonseparable models (for all 55 pairs of locations). Clearly, the nonseparable model fits the empirical correlations better. The relative lack of fit at lag one in time is mainly due to the assumption of symmetry of the covariance function which is not adequate for this data set, as discussed in Li et al. (2007) and Gneiting et al. (2007). Figure 6 shows the difference between the empirical west-to-east (WE, i.e. with the westerly station leading in time) and east-to-west (EW, with the westerly station lagging in time) correlations and
the fitted continuous correlation function using the nonseparable model for stations Valentia and Roche's Point (Figure 6(a)) and Belmullet and Clones (Figure 6(b)). The difference in the empirical correlation at lag one is quite large for both stations. A simple way to address this problem is to consider $C(s - \zeta tw, t)$ where $\zeta$ is a parameter to be estimated and, as the asymmetries in this example are mainly functions of differences in longitude, we take $w = (0, 1)'$ as suggested by Stein (2005). In our framework, this is equivalent to replacing the variogram $\gamma_1(s)$ by $\gamma_1(s - \zeta tw)$. Note that now the resulting covariance function is always nonseparable for nonzero $\zeta$, even if $\lambda_0 = 0$. For a discussion of covariance functions that are not fully symmetric, see Gneiting et al. (2007). The asymmetric version of Model 1 with free $\lambda_0$ leads to an improvement in fit, which is clearly shown in Figure 5(c). In Table 7 we can see that the posterior distribution of $\zeta$ is very far from zero, indicating that asymmetry is essential to this application. The posterior distribution of the smoothness parameters does not change much when compared with the fully symmetric version of this model. Note that in the context of the asymmetric model, $\lambda_0$ is no longer the only parameter controlling separability and $c$ does not have the same interpretation as in the symmetric version.

Table 8 shows the model comparison through Bayes factors in favour of the asymmetric Model 1 with free $\lambda_0$ versus the asymmetric Model 1 with $\lambda_0 = 0$, the nonseparable Model 1, the separable Model 1 and the nonseparable Models 2 and 3. All three methods mentioned in Section 4 for estimating the marginal likelihood give very strong evidence against imposing $\lambda_0 = 0$ in Model 1 in favour of its version with free $\lambda_0$ (with log Bayes factors around 50 for the symmetric version and around 150 under asymmetry). This clearly shows the overwhelming data support for the mixture construction of the covariance, as introduced in Section 3. Comparison of the nonseparable Model 1 with other nonseparable models indicates that Model 1 is also very strongly favoured over the nonseparable exponential model (Model 2 with $\alpha = 2$.
Table 8. Irish wind data: Natural logarithm of the Bayes factor in favour of the asymmetric Model 1 with free $\lambda_0$ versus the asymmetric Model 1 with $\lambda_0 = 0$, the nonseparable Model 1, the separable Model 1, the nonseparable Models 2 and 3 and the model in Gneiting et al. (2007) using Newton-Raftery ($d = 0.01$), Bridge-sampling and Shifted gamma ($\lambda = 0.98$) estimators for the marginal likelihood.

<table>
<thead>
<tr>
<th></th>
<th>Newton-Raftery</th>
<th>Bridge sampling</th>
<th>Shifted gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asym. Model 1 $\lambda_0 = 0$</td>
<td>149</td>
<td>153</td>
<td>148</td>
</tr>
<tr>
<td>Nonseparable Model 1</td>
<td>166</td>
<td>159</td>
<td>162</td>
</tr>
<tr>
<td>Separable Model 1</td>
<td>215</td>
<td>205</td>
<td>212</td>
</tr>
<tr>
<td>Nonseparable Model 2</td>
<td>223</td>
<td>227</td>
<td>205</td>
</tr>
<tr>
<td>Nonseparable Model 3</td>
<td>172</td>
<td>168</td>
<td>169</td>
</tr>
<tr>
<td>Model of Gneiting et al.</td>
<td>206</td>
<td>212</td>
<td>204</td>
</tr>
</tbody>
</table>

and $\nu = 0.5$) and decisively outperforms the nonseparable Cauchy model in space (Model 3). We also estimated the model proposed by (Gneiting et al., 2007, p.167), which is a nonseparable asymmetric model defined as a convex combination of a nonseparable correlation function and a compactly supported Lagrangian correlation function. The Bayes factor in favour of the asymmetric Model 1 with free $\lambda_0$ is shown in Table 8 (using comparable priors). Our proposed model is clearly favoured over the one proposed by Gneiting et al. (2007).

7. Discussion

In this article we have proposed a new covariance model that is nonseparable and includes the separable model as a particular case. The proposed model is obtained through a continuous mixture of separable covariances in space and time. The resulting model has some useful theoretical properties such as different degrees of smoothness across space and time and long-range dependence in time. For practical modelling purposes, we suggest a number of different parameterisations, leading to a variety of special cases with a wide range of spatial behaviour. Under separability, the purely spatial covariance structure can, for example, be a generalized Matérn, a Matérn, a Cauchy or a shifted exponential.

We conduct Bayesian inference with relatively vague priors, using an MCMC sampler. In addition, we implement an approximation to the likelihood which makes it feasible to perform inference for large data sets. Examples with simulated data (for all models) confirm that the prior is not overly informative and that the numerical methods perform well. We present an illustrative example for the Irish wind data of Haslett and Raftery (1989). The results show that use of the nonseparable model introduced in this paper leads to a very substantial improvement
in the fit. In addition, allowing for asymmetries by changing one of the variogram functions in the original construction, leads to a further improvement for these data. Both in symmetric and asymmetric versions of the model, mixing the covariance structure as proposed here is very strongly supported by the data.

The proposed model easily allows for extensions e.g. to include a nugget effect (see Subsection 3.2) and asymmetries (as we have illustrated in the modeling of the Irish wind data). Another direction in which the model could be extended is the modeling of nonstationarities in space which is an empirically important issue. Here we could combine our methodology with some of the models proposed in the spatial literature. For example, consider the convolution approach proposed in Fuentes and Smith (2001). Combining this with the nonseparable model presented here we obtain a nonstationary and nonseparable spatiotemporal covariance function given by

\[ C(s,s',t) = \int_D K(s-w)K(s'-w)\sigma^2(w)M_0(-\gamma_1(s-s';w)-\gamma_2(t))M_1(-\gamma_1(s-s';w))M_2(-\gamma_2(t))dw, \]  

(20)

where \( K(\cdot) \) is a convolution kernel and we mix through \( w \in D \). If the kernel \( K(\cdot) \) decreases rapidly and the parameters \( \sigma^2(w) \) and \( \gamma_1(\cdot; w) \) vary slowly then we have local stationarity. However, these parameters are allowed to vary across the whole spatial domain resulting in a non-stationary process. For instance we could define \( \gamma_1(d; w) = ||d/a(w)||^{\alpha(w)} \). Note that the simplicity of this extension derives from the mixture construction that allows us to choose appropriate models for the spatial and the temporal components.

In this paper we have assumed Gaussianity. However, this is an assumption that is not appropriate for all applications. The models proposed here could also be extended in order to account for non-Gaussian behaviour, for example by considering heavy tailed processes. This could be easily done by following the approach in Palacios and Steel (2006) and mixing over the processes \( Z_1(s; U) \) and \( Z_2(t; V) \). For instance we could define the process \( \tilde{Z}(s,t) = \tilde{Z}_1(s; U)\tilde{Z}_2(t; V) \) where

\[ \tilde{Z}_1(s; U) = \sqrt{1 - \tau^2} \frac{Z_1(s; U)}{\sqrt{\lambda_1(s)}} + \tau \frac{\epsilon(s)}{\sqrt{h(s)}}, \]  

(21)

with \( \{\lambda_1(s); s \in D\} \) a positively valued mixing process which is independent of \( \epsilon(s) \) and \( Z_1(s; U) \). As in Subsection 3.2, the process \( \{\epsilon(s); s \in D\} \) denotes an uncorrelated Gaussian process with zero mean and unitary variance and introduces a nugget effect parameterized by \( \tau \). Finally, \( \{h(s); s \in D\} \) is an uncorrelated process in \( \mathbb{R}_+ \) with distribution \( P_h \). The mixing process \( \lambda_1(s) \) is spatially correlated and allows for regions in space with larger variance while the process \( h(s) \) can create traditional outliers, i.e. observations with unusually large nugget effects. The same approach could be taken in time. The unconditional finite dimensional distributions induced by (21) have heavier tails than the normal distribution. Computation is relatively
easy since the finite dimensional distributions are Gaussian, conditionally upon \( \lambda_1(s) \) and \( h(s) \). This approach is implemented in detail in Fonseca and Steel (2009b).

**Appendix A. GIG Distribution**

A random variable \( X \) has a generalized inverse Gaussian (GIG) distribution if the density function of \( X \) is given by

\[
f_{GIG}(x; \lambda, \delta, a) = \left( \frac{a}{\delta} \right)^{\lambda/2} x^{\lambda-1} \frac{1}{2K_{\lambda}(2\sqrt{a\delta})} \exp\{-ax + \delta x^{-1}\}, \quad x > 0.
\]  

(22)

Permitted parameter values are \( a > 0, \delta \geq 0 \) if \( \lambda > 0 \), while \( a > 0, \delta > 0 \) if \( \lambda = 0 \) and \( \delta > 0 \) if \( \lambda < 0 \). A standard reference for the GIG distribution is Jørgensen (1982). We use the notation \( X \sim GIG(\lambda, \delta, a) \). An important aspect is that this class covers many special cases such as the gamma distribution (\( \lambda > 0 \) and \( \delta = 0 \)), the inverse gamma distribution (\( \lambda < 0 \) and \( a = 0 \)), the inverse Gaussian distribution (\( \lambda = -1/2 \)) and the reciprocal inverse Gaussian distribution (\( \lambda = 1/2 \)). For \( a, \delta > 0 \), the mean of \( X \) is

\[
E(X) = \sqrt{\delta/a} \left\{ \frac{K_{\lambda+1}(2\sqrt{a\delta})}{K_{\lambda}(2\sqrt{a\delta})} \right\}.
\]

(23)

**Appendix B. Proofs**

**Proof of Proposition 2.1**

Consider the covariance model given in (5) and the specification

\[
C_1(s; U) = \exp\{-\gamma_1 U\},
\]

\[
C_2(t; V) = \exp\{-\gamma_2 V\}.
\]

If the random vector \((U, V)\) has cumulative function \( \mu(u, v) \) the joint moment generating function \( M(r_1, r_2) = \int \exp\{r_1 u + r_2 v\} d\mu(u, v) \). Then,

\[
C(s, t) = \int \exp\{-\gamma_1 u - \gamma_2 v\} d\mu(u, v) = M(-\gamma_1, -\gamma_2).
\]

**Proof of Proposition 3.1**

From the conditions of Proposition 2.1 it follows that \( C(s, t) = \sigma^2 M(-\gamma_1, -\gamma_2) \), where \( M(., .) \) is the joint moment generating function of \((U, V)\). Define \( U = X_0 + X_1 \) and \( V = X_0 + X_2 \) with \( X_0, X_1 \) and \( X_2 \) independent nonnegative random variables with finite moment generating function \( M_0, M_1 \) and \( M_2 \), respectively. Then,

\[
M(-\gamma_1, -\gamma_2) = E[\exp\{-\gamma_1 U - \gamma_2 V\}] = E[\exp\{-(\gamma_1 + \gamma_2)X_0 - \gamma_1 X_1 - \gamma_2 X_2\}] = M_0(-\gamma_1 - \gamma_2)M_1(-\gamma_1)M_2(-\gamma_2).
\]
Proof of Theorem 3.1
Consider conditions of Proposition 3.1. Let $X_i \sim Ga(\lambda_i, a_i)$, $i = 0, 2$, then

\[ M_i(r) = \mathbb{E}[\exp\{ rX_i \}] = \int_0^\infty a_i^{\lambda_i} \frac{x^{\lambda_i-1} \exp\{-a_i - r\}x}{\Gamma(\lambda_i)} dx = \left( \frac{a_i}{a_i - r} \right)^{\lambda_i}, \quad r < a_i. \]

Let $X_1 \sim GIG(\lambda_1, \delta, a_1)$, if $a_1 \neq 0$ then

\[
M_1(r) = \mathbb{E}[\exp\{ rX_1 \}] = \int_0^\infty \frac{(\delta/a_1)^{-\lambda_1/2} x^{\lambda_1-1} \exp\{-[(a_1 - r)x + \delta x^{-1}]\} dx}{2K_1(2\sqrt{\delta a_1})} = \left( \frac{a_1}{a_1 - r} \right)^{\lambda_1/2} \frac{K_1(2\sqrt{(a_1 - r)\delta})}{K_1(2\sqrt{\delta a_1})}, \quad r < a_1,
\]

if $a_1 = 0$ we use the asymptotic formula $K_1(2\sqrt{\delta a_1}) = 2^{\lambda_1-1} \Gamma(\lambda_1)(2\sqrt{\delta a_1})^{-\lambda_1}$ implying

\[
M_1(r) = \left( \frac{a_1}{a_1 - r} \right)^{\lambda_1/2} \frac{K_1(2\sqrt{-r\delta})}{2^{\lambda_1-1} \Gamma(\lambda_1)(2\sqrt{\delta a_1})^{-\lambda_1}} = \frac{(2\sqrt{-r\delta})^{\lambda_1}}{\Gamma(\lambda_1)2^{\lambda_1-1}K_1(2\sqrt{-r\delta})}, \quad r < 0,
\] 

(24)
Theorem 3.1 follows.

Proof of Proposition 3.2
\( (a) \) The covariance function for the process \( \{Z(s_0, t) : t \in T\} \) at a fixed location \( s_0 \in D \) is given by

\[ C(0, t) = \sigma^2 M_0(-\gamma_2(t))M_2(-\gamma_2(t)) \] 

(25)
where $M_i(r) = \{1 - \frac{r}{a_i}\}^{-\lambda_i}$, $r < a_i$, $i = 0, 2$ and $a_0 = 1$. A (weakly) stationary process with covariance function $K(t)$ is $m$ times mean square differentiable if and only if $K^{(2m)}(0)$ exists and is finite (see Stein (1999) pp 20-22). By Faà di Bruno’s formula, termwise differentiation of (25) results in

\[
C^{(2m)}(0, t) = \sigma^2 \sum_A \frac{m!}{k_1!k_2!...k_{2m}!} y^{(k)}(-\gamma_2(t)) \prod_{j \neq 0} \left\{ -\frac{\gamma_2^{(j)}(t)}{j!} \right\}^{k_j} \]

(26)
where $A = \{k_1, k_2, ..., k_{2m} : k_1 + 2k_2 + ... + 2mk_{2m} = 2m\}$, $k = k_1 + k_2 + ... + k_{2m}$, $k_i \geq 0$, $i = 1, 2, ..., 2m$ and

\[
y^{(k)}(r) = \sum_{i=0}^k \begin{pmatrix} k \cr i \end{pmatrix} M_0^{(k-i)}(r)M_2^{(i)}(r).
\]

The terms $y^{(1)}(-\gamma_2(t)), ..., y^{(2m)}(-\gamma_2(t))$ exist and are finite as $t \to 0$ since

\[
y^{(k)}(0) = \sum_{i=0}^k \begin{pmatrix} k \cr i \end{pmatrix} \mathbb{E}(X_0^{k-i})\mathbb{E}(X_2) = \sum_{i=0}^k \begin{pmatrix} k \cr i \end{pmatrix} \frac{\Gamma(\lambda_0 + k - i)}{\Gamma(\lambda_0)} \frac{\Gamma(\lambda_2 + i)}{\Gamma(\lambda_2)} \left( \frac{1}{a_2} \right)^i,
\]

where $k = 1, 2, ..., 2m$. In the expression (26), the highest order derivative of $\gamma_2(t)$ is $2m$ obtained when $k_{2m} = 1$ and $k_1 = \ldots = k_{2m-1} = 0$. Thus the behaviour of $C^{(2m)}(0, t)$ as $t \to 0$
depends only on the local behaviour of \( \gamma_2^{(2m)}(t) \) as \( t \to 0 \), that is, if \( \gamma_2^{(2m)}(0) \) exists and is finite then the purely temporal process is \( m \) times mean square differentiable.

(b) The covariance function for the process \( \{ Z(s, t_0) : s \in D \} \) at a fixed time \( t_0 \in T \) is given by

\[
C(s, 0) = \sigma^2 M_0(-\gamma_1(s)) M_1(-\gamma_1(s)),
\]

where

\[
M_0(r) = \{1 - r\}^{-\lambda_0}, \quad r < 1 \quad \text{and} \quad M_1(r) = \{1 - \frac{r}{a_1}\}^{-\lambda_2} \frac{K_{\lambda_1}(2\sqrt{(a_1 - r)\delta})}{K_{\lambda_1}(2\sqrt{a_1\delta})}, \quad r < a_1.
\]

By Faà di Bruno’s formula, termwise differentiation of (27) results in

\[
C^{(2m)}(s, 0) = \sigma^2 \sum_{A} \frac{m!}{k_1!k_2!...k_{2m}!} y^{(k)}(-\gamma_1(s)) \prod_{k_j \neq 0} \left\{ \frac{-\gamma_1^{(j)}(s)}{j!} \right\}^{k_j}
\]

where \( A = \{k_1, k_2, ..., k_{2m} : k_1 + 2k_2 + ... + 2mk_{2m} = 2m\}, k = k_1 + k_2 + ... + k_{2m}, \ k_i \geq 0, \ i = 1, 2, ..., 2m \) and \( y^{(k)}(x) = \sum_{i=0}^{k} \binom{k}{i} M_0^{(k-i)}(x) M_1^{(i)}(x) \).

(i) Consider \( a_1 \neq 0 \). The terms \( y^{(1)}(-\gamma_2(t)), ..., y^{(2m)}(-\gamma_2(t)) \) exist and is finite for all integer \( m \) as \( s \to 0 \) since

\[
y^{(k)}(0) = \sum_{i=0}^{k} \binom{k}{i} \mathbb{E}(X_0^{k-i}) \mathbb{E}(X_1^{i}) = \begin{cases} \sum_{i=0}^{k} \binom{k}{i} \frac{\Gamma(\lambda_0 + k-i)}{\Gamma(\lambda_0)} \frac{\Gamma(\lambda_1 + i)}{\Gamma(\lambda_1)} \left( \frac{1}{a_1} \right)^i \text{ if } \delta = 0 \\ \sum_{i=0}^{k} \binom{k}{i} \frac{\Gamma(\lambda_0 + k-i)}{\Gamma(\lambda_0)} \frac{K_{\lambda_1+i}(2\sqrt{a_1\delta})}{K_{\lambda_1}(2\sqrt{a_1\delta})} \left( \frac{\delta}{a_1} \right)^{i/2} \text{ if } \delta \neq 0 \end{cases}
\]

\[ k = 1, 2, ..., 2m. \] In the expression (28), the highest order derivative of \( \gamma_1(s) \) is \( 2m \) obtained when \( k_{2m} = 1 \) and \( k_1 = ... = k_{2m-1} = 0 \). Thus the behaviour of \( C^{(2m)}(s, 0) \) as \( s \to 0 \) depends only on the local behaviour of \( \gamma_1^{(2m)}(s) \) as \( s \to 0 \).

(ii) Consider \( a_1 \neq 0 \) (which implies \( \lambda_1 < 0 \)). We need to study the behaviour of

\[
y^{(1)}(-\gamma_1(s))[-\gamma_1^{(2m)}(s)]
\]

obtained when \( k_{2m} = 1 \) and \( k_1 = ... = k_{2m-1} = 0 \) and

\[
y^{(2m)}(-\gamma_1(s))[-\gamma_1^{(1)}(s)]^{2m}
\]

obtained when \( k_1 = 2m \) and \( k_2 = ... = k_{2m} = 0 \) as the other terms in (28) require lower order derivatives.

Consider the Taylor expansion of \( \gamma_1(s) \) about \( 0 \) given by

\[
\gamma_1(s) = \sum_{j=0}^{\infty} c_j ||s||^j.
\]

Thus the behaviour of \( \gamma_1(s) \) as \( s \to 0 \) is determined by \( ||s||^l \) where \( l \) is the smallest power such that \( c_l \neq 0 \).

The expression (30) is given by

\[
y^{(2m)}(-\gamma_1(s))[-\gamma_1^{(1)}(s)]^{2m} = \sum_{i=1}^{2m} \binom{2m}{i} M_0^{(2m-i)}(-\gamma_1(s)) M_1^{(i)}(-\gamma_1(s))[-\gamma_1^{(1)}(s)]^{2m}
\]
and the highest order derivative of $M_1(-\gamma_1(s))$ is $2m$ obtained when $i = 2m$. Thus (30) will exist and be finite if $||s||^{l(-\lambda_1-2m)} \times ||s||^{(l-1)2m} = ||s||^{l(-\lambda_1-2m)}$ exists and is finite, implying that $m$ need to satisfy $2m < -l\lambda_1$.

The expression (29) is given by

$$M_0^{(1)}(-\gamma_1(s))M_1(-\gamma_1(s))[-\gamma_1^{(2m)}(s)] + M_0(-\gamma_1(s))M_1^{(1)}(-\gamma_1(s))[-\gamma_1^{(2m)}(s)]$$

and as $s \to 0$ we obtain $-\mathbb{E}[X_0](-\gamma_1^{(2m)}(s)) + ||s||^{l(-\lambda_1-1)}||s||^{l-2m}$, where $l^* \geq l$ implying that $m$ is sufficient for $m$ to satisfy $2m < -l\lambda_1$ and $\gamma_1^{(2m)}(0)$ needs to exist and be finite. Thus, the purely spatial process is $m$ times mean square differentiable if and only if $\gamma_1^{(2m)}(s)$ as $s \to 0$ exists and is finite and $2m < -l\lambda_1$.

**Appendix C. Computational issues**

Consider the matrix $Z$ with elements $Z_{ij} = Z(s_i, t_j)$, $i = 1, \ldots, I$ and $j = 1, \ldots, J$. We split the data into $Z_1, Z_2, \ldots, Z_{J/L}$ each of size $I \times L$ and $L$ is the lag in time (we assume, for simplicity, that $J/L$ is integer, but the code can deal with the general case by simply adapting the dimension of the last block). We will approximate the likelihood by assuming that the temporal correlation is positive only between observations that are not more than $L$ time periods apart and that it is zero otherwise. This is justified by the very small empirical correlation at time lags larger than $L$. For the Irish wind data we take $L = 3$ and taking larger values of $L$ did not affect the results appreciably. Consider $y_k = \text{Vec}(Z_k)$, where the first $I$ elements correspond to observations for all locations at time $(k-1)L + 1$, etc., then the covariance matrix of $(y_1, \ldots, y_{J/L})$ has a block Toeplitz structure given by

$$\Sigma = \begin{pmatrix}
T_{11} & T_{12} & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\
T_{21} & T_{11} & T_{12} & 0 & \ldots & 0 & 0 & 0 & 0 \\
0 & T_{21} & T_{11} & T_{12} & \ldots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \ldots & 0 & T_{21} & T_{11} & T_{12} \\
0 & 0 & 0 & 0 & \ldots & 0 & 0 & T_{21} & T_{11}
\end{pmatrix}$$

where

$$T_{11} = \begin{pmatrix}
\Sigma^{(0)} & \Sigma^{(1)} & \ldots & \Sigma^{(L-2)} & \Sigma^{(L-1)} \\
\Sigma^{(1)} & \Sigma^{(0)} & \ldots & \Sigma^{(L-3)} & \Sigma^{(L-2)} \\
\Sigma^{(L-2)} & \Sigma^{(L-3)} & \ldots & \Sigma^{(0)} & \Sigma^{(1)} \\
\Sigma^{(L-1)} & \Sigma^{(L-2)} & \ldots & \Sigma^{(1)} & \Sigma^{(0)}
\end{pmatrix}$$
and

\[
T_{12} = \begin{pmatrix}
\Sigma^{(L)} & 0 & \ldots & 0 & 0 \\
\Sigma^{(L-1)} & \Sigma^{(L)} & \ldots & 0 & 0 \\
\Sigma^{(2)} & \Sigma^{(3)} & \ldots & \Sigma^{(L)} & 0 \\
\Sigma^{(1)} & \Sigma^{(2)} & \ldots & \Sigma^{(L-1)} & \Sigma^{(L)}
\end{pmatrix}
\]

where \( \Sigma_{ij}^{(t)} = C(s_i - s_j, t) \), \( i, j = 1, \ldots, I \). In our example, time is equally spaced with intervals of one day. Then, we calculate the likelihood using the factorization

\[
p(y|\theta) = p(y_1|\theta)p(y_2|y_1, \theta) \ldots p(y_J|y_1, \ldots, y_{J-1}, \theta)
\]

where \( y_1|\theta \sim N(\mu_1 = 0, V_1 = T_{11}) \) and \( y_k|y_1, \ldots, y_{k-1}, \theta \sim N(\mu_k, V_k) \). \( \mu_k = T_{21}V_{k-1}^{-1}(y_{k-1} - \mu_{k-1}) \) and \( V_k = T_{11} - T_{21}V_{k-1}^{-1}T_{12} \) for \( k = 2, \ldots, J \).

For the Irish wind data we have relatively few stations and many replications in time. Therefore the critical issue was how to treat the time dimension. Setting the correlations to be zero from a certain lag in time implies we only need the inversion of small matrices with dimension \( IL \times IL \) for the evaluation of the likelihood. The strategy chosen was particularly advantageous in this situation but if we had many observations in space and not as many in time, other approximations could be considered. For instance, Stein et al. (2004) proposed the use of conditional distributions and considered only a subset of neighbors in the calculation of the likelihood for large spatial data sets. Another proposal for the spatial dimension is tapering (Furrer et al., 2006) where the basic idea is to taper the covariance function to zero beyond a certain range.

Acknowledgements

We are grateful to two Referees and an Associate Editor for insightful and constructive comments, which have substantially improved this paper. Thaís Fonseca acknowledges financial support from the Center for Research in Statistical Methodology (CRiSM).

References


Fig. 3. Irish wind data: (a-c) Posterior 95% credible intervals and median for the parameters in the temporal model for the 11 stations. (d) Posterior mean of $\mu$ for the 11 stations (the circle radius is proportional to the level).
Fig. 4. Irish wind data: Nonseparable model 1: Posterior (solid line) and prior (dotted line) densities for \( \epsilon \) in (10).

Fig. 5. Irish wind data: Empirical versus fitted (posterior median) correlations at temporal lags zero until five, with the higher lags corresponding to lighter shades, for the separable (a), nonseparable (b) and asymmetric (c) versions of Model 1.
Fig. 6. Irish wind data: Empirical EW and WE compared with fitted correlations (nonseparable Model 1). The drawn line is the posterior median and the dashed lines are the 95% credible intervals.