Nonstationary Nonseparable Random Fields

Kangrui Wang [1] **Oliver Hamelijnck** [1 2] **Theodoros Damoulas** [1 2] **Mark Steel** [2]

Abstract

We describe a framework for constructing nonstationary nonseparable random fields based on an infinite mixture of convolved stochastic processes. When the mixing process is stationary but the convolution function is nonstationary we arrive at nonseparable kernels with *constant nonseparability* that are available in closed form. When the mixing is nonstationary and the convolution function is stationary we arrive at nonseparable random fields that have *varying nonseparability* and better preserve local structure. These fields have natural interpretations through the spectral representation of stochastic differential equations (SDEs) and are demonstrated on a range of synthetic benchmarks and spatio-temporal applications in geostatistics and machine learning. We show how a single Gaussian process (GP) with these random fields can computationally and statistically outperform both separable and existing nonstationary nonseparable approaches such as treed GPs and deep GP constructions.

1. Introduction

Kernel-based methods (Scholkopf & Smola, 2001) have a long history in both machine learning and spatial statistics (Cressie, 1990) across frequentist and Bayesian paradigms. Standard covariance (kernel) functions, such as the Gaussian and Matérn, are stationary (translation invariant) and separable. Although these covariance functions admit tractable forms they are unrealistic for modelling real world phenomena that are non-stationary and exhibit strong dependencies.

In this work we focus on spatio-temporal random fields in $\mathbb{R}^3$ as our motivation for proposing nonstationary nonseparable covariance functions. However, the methodology is applicable to general $\mathbb{R}^D$ input spaces. Consider a spatio-temporal stochastic process $Z(\mathbf{s}, t)$ that has a stationary and separable structure, where $\mathbf{s} \in \mathbb{R}^2$ indicates the spatial coordinates and $t \in \mathbb{R}$ indicates a temporal dimension. Stationarity implies that the covariance function depends only on the distance of the observations $C(\mathbf{s}, t, \mathbf{s}', t') = C(\mathbf{s}-\mathbf{s}', t-t')$ but not on their specific location. Separability implies independence between input dimensions, for example between space and time as $C(\mathbf{s}, t, \mathbf{s}', t') = C(\mathbf{s}, \mathbf{s}')C(t, t')$. A nonseparable covariance function captures dependencies between the dimensions; when that dependency is constant it can be expressed as $C(\mathbf{s}, t, \mathbf{s}', t') = \rho C(\mathbf{s}, \mathbf{s}')C(t, t')$ we have *constant nonseparability* whereas when the dependency itself is changing across input space, we define it as *varying nonseparability*. Fig. 1 illustrates these different levels of separability and stationarity.

Nonstationary Separable: There is a significant body of work in nonstationary covariance functions either through hierarchical constructions (Paciorek & Schervish, 2004; Remes et al., 2017; Heinonen et al., 2016), compositional (deep) models (Damianou & Lawrence, 2013; Monterrubio-Gómez et al., 2018), input space partitioning approaches (Gramacy & Lee, 2008) or spectral representations (Stein, 2005; Remes et al., 2017).

As shown by Paciorek & Schervish (2004) any stationary covariance function can be used to construct a nonstationary one, where input dependent local-lengthscales are used to define the correlation between two points. This has been extended by Remes et al. (2017); Heinonen et al. (2016) by placing GP priors on the (log) of the lengthscales. These functions are very flexible but suffer from identifiability issues, inefficient inference procedures, and an increased computational burden (Paciorek & Schervish, 2006). Cortes et al. (2009) studies the general problem of kernel learning with a polynomial (potentially non-linear) combination of base kernels that can handle non-stationary data when the combination is location-dependent.

Remes et al. (2017) extend the spectral mixture (SM) kernel (Wilson & Adams, 2013) to a nonstationary one but the spectral representation is unavailable hence it is unclear how it evolves across input space. Similar to the Paciorek & Schervish (2004) construction, the nonstationary SM kernel suffers from identifiability issues. Reece et al. (2015)

[1]Data-centric Engineering, The Alan Turing Institute, London, UK [2]Department of Statistics, University of Warwick, Coventry, UK. Correspondence to: Kangrui Wang <Kwang@turing.ac.uk>, Theodoros Damoulas <T.Damoulas@warwick.ac.uk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

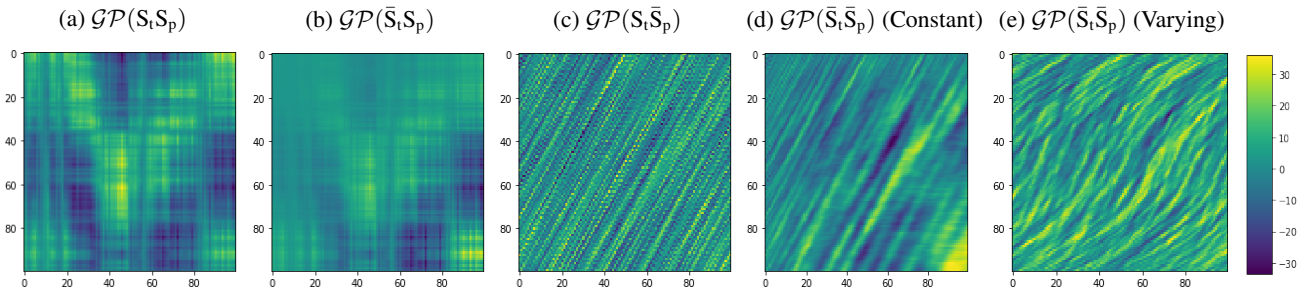

Figure 1. Illustration of samples from 2D Gaussian processes with varying levels of stationarity (a/c,b/d,e) and separability (a/b,c/d,e). Nonstationary fields (b,d) exhibit varying levels of smoothness, changing from top left to bottom right. Higher levels of nonseperability express progressively more complex dependencies: from independence (a) to linear dependency structure (c,d) and varying local correlation structure (e). between the input dimensions. In (e) the smoothness of the process is fixed and the nonstationarity comes from the varying dependency structure.

construct a piece-wise stationary function via the Markov Region Link kernel. The final process is nonstationary while each partition of the process follows a stationary GP. Lewis et al. (2006) combines multiple kernels with a nonstationary warping function and similarly Snoek et al. (2014) introduces nonstationarity into the covariance function by warping the input through another function. While each dimension has its own warping function, the final process is nonstationary but separable.

Stationary Nonseparable: There is some work on stationary nonseparable covariance functions. Gneiting (2002) describes general constructions of stationary nonseparable kernels via Bochner's theorem and Lindgren et al. (2011) show a clear link between the nonseparable Matérn class, stochastic differential equations (SDEs) and Gaussian random fields through the spectral transformation. Rodrigues & Diggle (2010) extend the general class of convolution based nonseparable kernels and Remes et al. (2017); Chen et al. (2019) expand the SM kernel into nonseparable versions, where the stochastic process is constructed using a nonseparable spatial field.

The dominant approach in this class are kernels arising from *Blurring Processes* (Brown et al., 2000) and the closely related *Process Convolutions* (Higdon, 2002; Fonseca & Steel, 2011a; Alvarez et al., 2012) that we introduce in §2 and generalize in this work to hierarchical constructions via infinite mixtures.

Nonstationary Nonseparable: There has been little work in directly constructing nonstationary *and* nonseparable fields beyond the Matérn class (Stein, 2005). Indirect approaches include (deep) compositions (Damianou & Lawrence, 2013), partitioning approaches (Gramacy & Lee, 2008), or random Fourier approximations (Ton et al., 2018).

Fonseca & Steel (2011b) introduced a nonstationary nonseparable kernel through a process convolution approach endowed with scale mixtures whose scale varies across locations. This corresponds to a *constant* nonseparable field, Fig. 1, as the mixing process is constant. We will generalize this construction to more complex mixing processes in order to achieve varying nonseparability.

We offer a Stochastic Process Mixing (SPM) framework that results in closed form nonseparable nonstationary covariance functions when the mixing process is stationary. The SPM is based on an infinite mixture of convolved stochastic processes and when the mixing process is nonstationary this enables us to better capture local correlation structure that changes across the domain. We focus on a Bayesian nonparametric setting, typical in spatial statistics, and demonstrate the capabilities of the resulting covariance functions within a Gaussian process (GP) framework and against other GP based approaches and compositions.

2. Stochastic Process Mixing (SPM)

To ease exposition we define the following subscripts: $Z_{\mathrm{S_t S_p}}$ is a stationary separable process, $Z_{\mathrm{S_t} \bar{\mathrm{S}}_\mathrm{p}}$ is a stationary nonseparable one, $Z_{\bar{\mathrm{S}}_\mathrm{t} \mathrm{S_p}}$ is a nonstationary separable process and $Z_{\bar{\mathrm{S}}_\mathrm{t} \bar{\mathrm{S}}_\mathrm{p}}$ is nonstationary nonseparable one.

We start by describing the general construction of nonstationary nonseparable processes based on the convolution and mixing of base stochastic processes (Higdon, 2002; Fonseca & Steel, 2011a). A stochastic process can be constructed as a kernel convolution over another stochastic process. For example, given a D dimensional white noise process Φ and a valid kernel (convolution) function K then a stochastic process $Z(\mathbf{x})$ can be defined as

$$Z(\mathbf{x}) = \int K(\mathbf{x} - \mathbf{u}) \Phi_\mathbf{u} \, d\mathbf{u} \tag{1}$$

where $\mathbf{x}, \mathbf{u} \in \mathbb{R}^D$ and $\phi_\mathbf{u} \sim \mathcal{N}(0, \mathbf{I})$. The resulting covariance function of Z is then simply given by

$$C(\mathbf{x}, \mathbf{x}') = \int K(\mathbf{x} - \mathbf{u}) K(\mathbf{x}' - \mathbf{u}) \, d\mathbf{u} \,. \tag{2}$$

Nonstationarity: It is often easier to specify the convolving kernel functions rather than the resulting covariance function directly. When K is a stationary function then the resulting $Z(\mathbf{x})$ will also be stationary. When K is nonstationary or ϕ has the form of a nonstationary stochastic process Φ then the resulting field is also nonstationary (Paciorek & Schervish, 2004; Fuentes & Smith, 2001):

$$Z(\mathbf{x}) = \int K_{\mathbf{x}}(\mathbf{x}-\mathbf{u})\Phi_{\mathbf{u}}(\mathbf{x})\, d\mathbf{u} \tag{3}$$

where we have used subscripts to denote dependence on the input location and latent variables. Note that when $\Phi_{\mathbf{u}}(\mathbf{x})$ is a GP then Z will also be a GP. Typically the process Z will depend on some parameters, either from the kernel or the latent $\Phi_{\mathbf{u}}(\mathbf{x})$. Placing a prior over these we can define the marginal process:

$$Z(\mathbf{x}) = \int K_{\mathbf{x}}(\mathbf{x}-\mathbf{u}|\mathbf{a})\Phi(\mathbf{x}|\mathbf{a})p_{\mathbf{u}}(\mathbf{a})\, d\mathbf{a}\, d\mathbf{u} \tag{4}$$

that is difficult to get in closed form. If the mixing process $p(\mathbf{a})$ does not depend on the latent variable $\mathbf{u}$, we have:

$$\begin{aligned} Z(\mathbf{x}) &= \int K_{\mathbf{x}}(\mathbf{x}-\mathbf{u}|\mathbf{a})\Phi(\mathbf{x}|\mathbf{a})p(\mathbf{a})\, d\mathbf{a}\, d\mathbf{u} \\ &= \mathbb{E}_{p(\mathbf{a})}[Z(\mathbf{x}|\mathbf{a})] \end{aligned} \tag{5}$$

Thus, the marginal process $Z(\mathbf{x})$ can be regarded as an infinite mixture of stochastic processes $Z(\mathbf{x}|\mathbf{a})$ with parameters distributed $\mathbf{a} \sim p_{\mathbf{u}}(\mathbf{a})$. If these $Z(\mathbf{x}|\mathbf{a})$ are stationary and $p_{\mathbf{u}}(\mathbf{a}) = p(\mathbf{a})$, the resulting process is also stationary. When the mixing distribution changes with the latent variable $\mathbf{u}$, the resulting process $Z(\mathbf{x})$ will be nonstationary, even if $Z(\mathbf{x}|\mathbf{a})$ is stationary.

Nonseparability: Using the construction of Eq. 4, a nonseparable spatio-temporal process can be written as a convolution of local separable processes. We have:

$$\begin{aligned} Z_{\bar{\mathbb{S}}_t\bar{\mathbb{S}}_p}(\mathbf{s},t) = \int\int\int\int & K_{\mathbf{s}}(\mathbf{s}-\mathbf{u}|\mathbf{a})K_t(t-v|b) \\ & \Phi_{\mathbf{u}}(\mathbf{s}|\mathbf{a})\Phi_v(t|b)p_{\mathbf{u},v}(\mathbf{a},b)\, d\mathbf{a}\, db\, d\mathbf{u}\, dv \end{aligned} \tag{6}$$

When $p(\mathbf{a},b) = p(\mathbf{a})p(b)$ the resulting process $Z(\mathbf{s},t) = Z(\mathbf{s})Z(t)$ will be a separable process. Instead, if $\mathbf{a}$ and b are dependent then the resulting process will be nonseparable. When $p(\mathbf{a},b)$ is not changing with location of $\mathbf{u}$ and v, the process $Z(\mathbf{s},t)$ will be a stationary process.

Thus, the resulting covariance function will be:

$$\begin{aligned} C(\mathbf{s},t,\mathbf{s}',t') = \int\int\int\int & K_{\mathbf{s}}(\mathbf{s}-\mathbf{u}|\mathbf{a})K_t(t-v|b) \\ & C_{\mathbf{u}}(\mathbf{s},\mathbf{s}'|\mathbf{a})C_v(t,t'|b)K_{\mathbf{s}'}(\mathbf{s}'-\mathbf{u}|\mathbf{a}) \\ & K_{t'}(t'-v|b)p_{\mathbf{u},v}(\mathbf{a},b)\, d\mathbf{a}\, db\, d\mathbf{u}\, dv \end{aligned} \tag{7}$$

Where the local covariance depend on the zero-mean latent process $C_{\mathbf{u}}(\mathbf{s},\mathbf{s}'|\mathbf{a}) = \mathbf{E}[\Phi_{\mathbf{u}}(\mathbf{s}|\mathbf{a})\Phi_{\mathbf{u}}(\mathbf{s}'|\mathbf{a})]$ and $C_v(t,t'|b) = \mathbf{E}[\Phi_v(t|b)\Phi_v(t'|b)]$. This approach incorporates both nonstationary convolution and nonstationary mixing for locally stationary processes, which results in increased number of hyper-parameters and raises the computational complexity. Furthermore, the flexibility of learning the dynamics either through the convolution or the mixing function causes identifiability issues. This leads us to consider two different approaches by constraining a different component each time.

3. SPMs through Nonstationary Convolution

In our first construction, we assume we have a stochastic function with constant nonseparablity as seen in Fig. 1. In other words, the mixing distribution $p(\mathbf{a},b)$ is not changing with location $\mathbf{u}$. The nonstationarity of the process is handled using the convolution function via Eq. 3. Thus, the general construction of Eq. 6 becomes:

$$\begin{aligned} Z_{\bar{\mathbb{S}}_t\bar{\mathbb{S}}_p}(\mathbf{s},t) &= \mathbb{E}_{p(\mathbf{a},b)}\left[\int K_{\mathbf{s}}(\mathbf{s}-\mathbf{u}|\mathbf{a})\Phi_{\mathbf{u}}(\mathbf{s}|\mathbf{a})\, d\mathbf{u} \right. \\ &\qquad \left. \int K_t(t-v|b)\Phi_v(t|b)\, dv \right] \\ &= \mathbb{E}_{p(\mathbf{a},b)}\left[Z_{\bar{\mathbb{S}}_t\mathbb{S}_p}(\mathbf{s}|\mathbf{a})Z_{\bar{\mathbb{S}}_t\mathbb{S}_p}(t|b)\right] \end{aligned} \tag{8}$$

Note that the mixing does not add any extra nonstationarity. From Eq. 8, we understand that the integral for the mixing is not related to the convolution mixing. When calculating the covariance function, Eq. 7, we could do the convolution first to generate the nonstationary separable process and handle the nonseparability using the process mixing. If the convolution function is separable, we have:

$$\begin{aligned} & C(\mathbf{s},\mathbf{s}',t,t') \\ &= \int\left[\int K_{\mathbf{s}}(\mathbf{s}-\mathbf{u}|\mathbf{a})K_{\mathbf{s}'}(\mathbf{s}'-\mathbf{u}|\mathbf{a})C_{\mathbf{u}}(\mathbf{s},\mathbf{s}')d\mathbf{u}\right] \\ &\quad \left[\int K_t(t-v|b)K_{t'}(t'-v|b)C_v(t,t')dv\right] d\mu(\mathbf{a},b) \\ &= \int C_{\bar{\mathbb{S}}_t}(\mathbf{s},\mathbf{s}'|\mathbf{a})C_{\bar{\mathbb{S}}_t}(t,t'|b)d\mu(\mathbf{a},b) \end{aligned} \tag{9}$$

Where $C_{\bar{\mathbb{S}}_t}(t,t'|b)$ and $C_{\bar{\mathbb{S}}_t}(\mathbf{s},\mathbf{s}'|\mathbf{a})$ are the covariances for the separable nonstationary process with scale mixture parameters $\mathbf{a}, b$. We could simplify the construction of Eq. 8 by conditioning either the convolution function or the latent process w.r.t the mixing parameters. In any case, we always arrive at the nonseparable covariance using the mixture distribution and the conditional separable covariance.

Special cases with closed forms

From Eq. 9, we see that to get the closed form of the random field, we need the closed forms of the component separable nonstationary covariance functions. We now assign the mixing parameter to the convolution function:

$$C_{\bar{\mathbb{S}}_t}(\mathbf{s},\mathbf{s}'|\mathbf{a}) = \int K_{\mathbf{s}}(\mathbf{s}-\mathbf{u}|\mathbf{a}) C_{\mathbf{u}}(\mathbf{s},\mathbf{s}') K_{\mathbf{s}'}(\mathbf{s}'-\mathbf{u}|\mathbf{a}) d\mathbf{u} \tag{10}$$

Any closed form nonstationary construction can be used to generate $C_{\bar{\mathbb{S}}_t}(\mathbf{s},\mathbf{s}'|\mathbf{a})$. For simplicity, we assume the latent variable $\mathbf{u}$ is affected by a scalar random variable a. We present more complex mixing in our applications. When we set the convolution function to:

$$K_{\mathbf{s}}(\mathbf{s}-\mathbf{u}|a) = \exp\left(-a(\mathbf{s}-\mathbf{u})\Sigma(\mathbf{s})^{-1}(\mathbf{s}-\mathbf{u})^T\right)$$

and the latent covariance as $C_{\mathbf{u}}(\mathbf{s},\mathbf{s}') = \sigma(\mathbf{s})\sigma(\mathbf{s}')$ we arrive at Paciorek & Schervish (2006) nonstationary covariance function: $C_{\bar{\mathbb{S}}_t\mathbb{S}_p}(\mathbf{s},\mathbf{s}'|a) =$

$$A(s,s')\exp(-a(\mathbf{s}-\mathbf{s}')\left(\frac{\Sigma(\mathbf{s})+\Sigma(\mathbf{s}')}{2}\right)^{-1}(\mathbf{s}-\mathbf{s}')^T)$$

Where $A(s,s') = \sigma(\mathbf{s})\sigma(\mathbf{s}')\frac{|\Sigma(\mathbf{s})|^{\frac{1}{4}}|\Sigma(\mathbf{s}')|^{\frac{1}{4}}}{|\Sigma(\mathbf{s})+\Sigma(\mathbf{s}')|^{\frac{1}{2}}}$. By setting:

$$\begin{aligned} Q_{\mathbf{s},\mathbf{s}'} &= (\mathbf{s}-\mathbf{s}')\left(\frac{\Sigma(\mathbf{s})+\Sigma(\mathbf{s}')}{2}\right)^{-1}(\mathbf{s}-\mathbf{s}')^T \\ Q_{t,t'} &= (t-t')\left(\frac{\Sigma(t)+\Sigma(t')}{2}\right)^{-1}(t-t')^T \end{aligned} \tag{11}$$

and defining $R(\mathbf{s},\mathbf{s}'|\mathbf{a}) = \exp(-aQ_{\mathbf{s},\mathbf{s}'})$ and $R(t,t'|b) = \exp(-bQ_{t,t'})$ as the exponential part of the covariance function, we can marginalize $\mathbf{a}$ and b as:

$$\begin{aligned} R(\mathbf{s},\mathbf{s}',t,t') &= \mathbb{E}_{a,b}[R(\mathbf{s},\mathbf{s}'|a)R(t,t'|b)] \\ &= \mathbb{E}_{a,b}[\exp(-aQ_{\mathbf{s},\mathbf{s}'}))\exp(-bQ_{t,t'})] \\ &= M_a(-Q_{\mathbf{s},\mathbf{s}'})M_b(-Q_{t,t'}) \end{aligned} \tag{12}$$

Where $M_a(.)$ and $M_b(.)$ are the moment generation functions (MGF). Using the properties of MGFs, if we set the random variable a and b as linear combinations of independent random variables : $a = \lambda_0 + \lambda_1$ and $b = \lambda_0 + \lambda_2$ we can then rewrite the MGF as:

$$\begin{aligned} &M_a(-Q_{\mathbf{s},\mathbf{s}'})M_b(-Q_{t,t'}) \\ =&M_{\lambda_0}(-(Q_{\mathbf{s},\mathbf{s}'}+Q_{t,t'}))M_{\lambda_1}(-Q_{\mathbf{s},\mathbf{s}'})M_{\lambda_2}(-Q_{t,t'}) \end{aligned} \tag{13}$$

With specific distributions on $\lambda_0,\lambda_1,\lambda_2$, we finally arrive at different *closed forms* for nonseparable nonstationary functions. For example, when $\lambda_0 \sim \text{Ga}(\beta_0,1)$, $\lambda_1 \sim \text{Ga}(\beta_1,1)$, $\lambda_2 \sim \text{Ga}(\beta_2,1)$ we have $R(\mathbf{s},\mathbf{s}',t,t') =$

$$(1+Q_{\mathbf{s},\mathbf{s}'}+Q_{t,t'})^{\beta_0}(1+Q_{\mathbf{s},\mathbf{s}'})^{\beta_1}(1+Q_{t,t'})^{\beta_2} \tag{14}$$

When $\lambda_0 \sim \text{IGa}(\beta_0,1/4)$, $\lambda_1 \sim \text{IGa}(\beta_1,1/4)$, $\lambda_2 \sim \text{IGa}(\beta_2,1/4)$ we arrive at:

$$\begin{aligned} R(\mathbf{s},\mathbf{s}',t,t') &= R'_{\beta_0}(Q_{\mathbf{s},\mathbf{s}'}+Q_{t,t'})R'_{\beta_1}(Q_{\mathbf{s},\mathbf{s}'})R'_{\beta_2}(Q_{t,t'}) \\ R'_{\beta}(Q) &= \frac{1}{\gamma(\beta)2^{\beta-1}}\left(\sqrt{2\beta Q}\right)^{\beta}K_{\beta}(\sqrt{2\beta Q}) \end{aligned} \tag{15}$$

4. SPMs through Nonstationary Mixing

In an alternative approach, we fix the convolution function and make the mixing function vary amongst locations. Fonseca & Steel (2011a) developed the nonstationary convolution based on the spatial dimension:

$$C(\mathbf{s},\mathbf{s}',t,t') = \int K(\mathbf{s}-\mathbf{u})K(\mathbf{s}'-\mathbf{u}) \int\int C_{\mathbf{u}}(\mathbf{s},\mathbf{s}'|\mathbf{a})C(t,t'|b)\,d\mathbf{a}\,db\,d\mathbf{u} \tag{16}$$

However, the covariance amongst time is not handled by this function. It is possible to develop a fully space-time mixture kernel. Construct an SPM as:

$$Z_{\bar{\mathbb{S}}_t\bar{\mathbb{S}}_p}(\mathbf{s},t) = \int\int\int\int K(\mathbf{s}-\mathbf{u})K(t-v) \Phi_{\mathbf{u}}(\mathbf{s}|\mathbf{a})\Phi_v(t|b)p_{\mathbf{u},v}(\mathbf{a},b)\,d\mathbf{a}\,db\,d\mathbf{u}\,dv \tag{17}$$

If the mixing distribution $p_{\mathbf{u},v}(\mathbf{a},b)$ is nonstationary between locations $\{\mathbf{u},v\}$, we cannot get a general closed form for the marginal process $Z_{\bar{\mathbb{S}}_t\bar{\mathbb{S}}_p}(\mathbf{s},t)$. However, we can write down the covariance function when $\Phi_{\mathbf{u}}(\mathbf{s}|\mathbf{a}),\Phi_v(t|b)$ are independent for location $\{\mathbf{u},v\}$:

$$C(\mathbf{s},\mathbf{s}',t,t') = \int\int K(\mathbf{s}-\mathbf{u})K(\mathbf{s}'-\mathbf{u})K(t-v)K'(t'-v) \int\int C_{\mathbf{u}}(\mathbf{s},\mathbf{s}'|\mathbf{a})C_v(t,t'|b)\,d\mu(\mathbf{a},b)\,d\mathbf{u}\,dv \tag{18}$$

As we can see, this is a special case of the general construction of Eq. 7. This construction ensures the smoothing of the convolution will be in each of the dimension while the mixing is nonstationary in each of the location amongst the space-time location. This makes the full covariance structure difficult to derive in closed form. However, we are still able to get the closed form of $C_{\mathbf{u},v}(\mathbf{s},\mathbf{s}',t,t')$ under each location:

$$C_{\mathbf{u},v}(\mathbf{s},\mathbf{s}',t,t') = \mathbb{E}_{p(\mathbf{a},b)}[C(\mathbf{s},\mathbf{s}'|\mathbf{a})C(t,t'|b)] \tag{19}$$

The covariance function $C_{\mathbf{u},v}(\mathbf{s},\mathbf{s}',t,t')$ presents the local nonseparable structure. Any nonseparable covariance function can be used here. For a closed form local covariance, we could use the special cases generated in §3.

As shown in the general construction in Eq. 6, the mixing parameter can be included in the convolution function:

$$Z_{\bar{\mathbf{S}}_t\bar{\mathbf{S}}_p}(\mathbf{s},t) = \int\int\int\int K(\mathbf{s}-\mathbf{u}|\mathbf{a})K(t-v|b) \Phi(\mathbf{s})\Phi(t)p_{\mathbf{u},v}(\mathbf{a},b)\,d\mathbf{a}\,db\,d\mathbf{u}\,dv \tag{20}$$

Thus, the covariance function can be written using a nonseparable convolution: $C(\mathbf{s},\mathbf{s}',t,t') =$

$$\int\int \mathbb{E}_{p_{\mathbf{u},v}(\mathbf{a},b)}\big[K(\mathbf{s}-\mathbf{u}|\mathbf{a})K(t-v|b) K(\mathbf{s}'-\mathbf{u}|\mathbf{a})K(t'-v|b)\big]C(\mathbf{s},\mathbf{s}')C(t,t')\,d\mathbf{u}\,dv \tag{21}$$

Both mixing approaches in Eq. 19 and Eq. 21 handle nonstationarity and nonseparability through the mixing distribution. There is no significant qualitative difference between the two constructions but in some situations, it is easier to calculate using Eq. 19 rather than Eq. 21 and vice versa.

Example: SPM through Nonstationary Mixing We generate the covariance function using nonstationary mixing:

$$\begin{aligned} &C(\mathbf{s},\mathbf{s}',t,t') \\ &\quad= \int\int K(\mathbf{s}-\mathbf{u})K(\mathbf{s}'-\mathbf{u})K(t-v)K'(t'-v) \\ &\qquad \int\int C'(\mathbf{s},\mathbf{s}'|a)C'(t,t'|b)\,d\mu(a,b)\,d\mathbf{u}\,dv \end{aligned} \tag{22}$$

For the convolution functions $K(\mathbf{s}-\mathbf{u}), K(t-v)$, we assume they are stationary and use an exponential convolution. We assume the local stationary processes follow the covariance:

$$C'(\mathbf{s},\mathbf{s}'|a(\mathbf{u}))C'(t,t'|b(v)) = \exp\left(-a(\mathbf{u})\frac{(\mathbf{s}-\mathbf{s}')^2}{\ell_s}\right)\exp\left(-b(v)\frac{(t-t')^2}{\ell_t}\right) \tag{23}$$

Thus, we can construct the nonstationary mixing via a linear function and independent variables. Assume $\lambda_0 \sim \text{Ga}(\beta_0,1)$, $\lambda_0 \sim \text{Ga}(\beta_1(\mathbf{u}),1)$ and $\lambda_2 \sim \text{Ga}(\beta_2(v),1)$, we have $a(\mathbf{u}) = \lambda_0+\lambda_1$ and $b(v) = \lambda_0+\lambda_2$, where $\beta_1(\mathbf{u})$ and $\beta_2(v)$ are the polynomial functions related to the location $\mathbf{u}$ and v . Thus, we have:

$$\begin{aligned} C_{\mathbf{u},v}(\mathbf{s},\mathbf{s}',t,t') =& \mathbb{E}_{a,b}(C(\mathbf{s},\mathbf{s}'|a(\mathbf{u}))C(t,t'|b(v)) \\ =& C'_{\beta_0}(Q_s+Q_t)C'_{\beta_1(\mathbf{u})}(Q_s)C'_{\beta_2(v)}(Q_t) \end{aligned} \tag{24}$$

Where $Q_s = (s-s')^2/\ell_s$ and $C'_\beta(Q) = (1+Q)^{-\beta}$. We demonstrate this SPM in a synthetic setting at §6.1.

5. SDE informed SPMs

For most random fields the optimal form of the convolution is generally unknown. Hence practitioners typically fall back on the Gaussian convolution. Although this provides appealing properties, unbiased estimates and closed form kernel functions, the Gaussian kernel does not provide any additional information and simply acts as a smoother. However, in many cases the observed process can be described as the solution to a stochastic differential equation (**SDE**):

$$a_pZ^{(p)}(t) + a_{p-1}Z^{(p-1)}(t) + ... + a_0Z(t) = \phi(t) \tag{25}$$

where $Z^{(p)}(t)$ is p-th derivative of $Z(t)$ and $\phi(t)$ is the forcing term that brings uncertainty into the process. In general the forcing term can be any stochastic process but is typically assumed to be a white noise process.

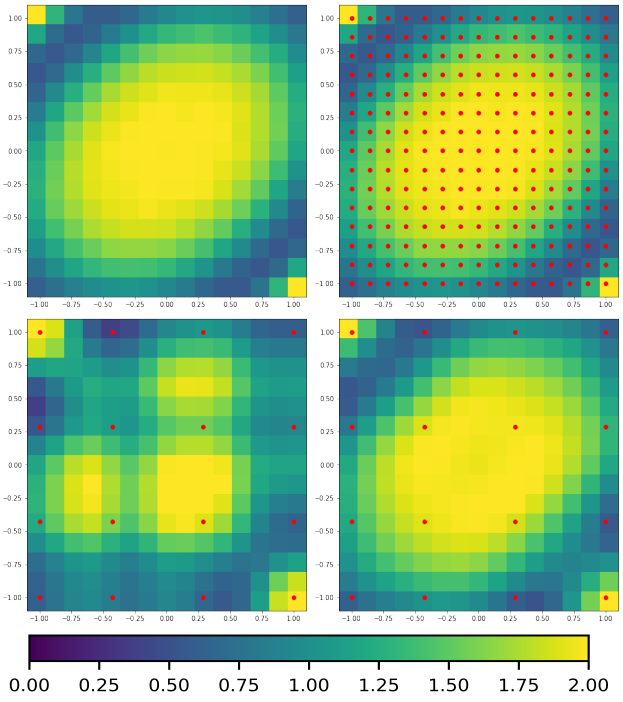

Figure 2. The true 2D heat kernel surface is plotted in the top left and red points denote observation locations. Top right and bottom left show a GP prediction using a squared exponential kernel across varying sample sizes. Bottom right shows that our SDE informed SPM better recovers the true surface, even on a small sample size, because it can encode the physical behaviour of the process.

Instead of solving Eq. 25 directly, we can find the corresponding Green's function and rewrite the process of interest as a convolution against this:

$$Z(t) = \int G(t-u)\phi_u(t)\,du \tag{26}$$

where $G(t-u)$ is the Green's function for the SDE in Eq. 25. Through this form we are injecting physical/mechanistic structure into our prior that will allow us to learn the process $Z(\cdot)$ more effectively. By viewing the SDE as arising from a convolution we can cast it into both our nonstationary convolution and nonstationary mixing framework.

Nonstationary SDEs We can simply create a nonstationary process by mixing the SDE with a nonstationary distribution. Following Sec. 4 we have:

$$\begin{aligned} Z(t) &= \int Z(t|b)d\mu(b) \\ &= \int\int G(t-u|b)\phi_u(t|b)\, d\mu(b)\, du, \end{aligned} \tag{27}$$

where $\mu(b)$ is the probability measure for random variable b. For the marginal process $Z(t)$, it is hard to find a corresponding SDE. However, we can find the SDE representation on the conditional process $Z(t|b)$. When b is the random variable varying on the input space, we can define the involving structure for each location of the input space via the Green's function of the SDE. When the input space is high dimensional, the correlation of the random variables captures the dependency structure of the input space.

For a separable SDE, $Z(s,t|\mathbf{a},b) = Z'(s|\mathbf{a})Z^*(t|b)$, the process is the solution to the following SDE:

$$\begin{aligned} \sum_i a_i \frac{\partial^i Z'(s|\mathbf{a})}{\partial s^i} + a_0 Z'(s|\mathbf{a}) &= \phi(s|\mathbf{a}), \\ \sum_j b_j \frac{\partial^j Z^*(t|b)}{\partial t^j} + b_0 Z^*(t|b) &= \phi(t|b). \end{aligned} \tag{28}$$

We have seen in our Nonstationary mixing framework (§4), that we can induce correlation between the input dimensions. Let $\mathbf{a} = \{\mathbf{a}_0, ...\mathbf{a}_I\}, b = \{b_0, ...b_J\}$ be random variables from the joint distribution $p(\mathbf{a}, b)$ that mix the above SDE. We can induce a nonstationary, nonseparable process, even when the latent SDEs are stationary:

$$\begin{aligned} Z_{\bar{\mathrm{S}}_\mathrm{t}\bar{\mathrm{S}}_\mathrm{p}}(\mathbf{s},t) &= \int\int\int\int G(\mathbf{s}-\mathbf{u})G'(t-v) \\ &\quad \Phi(\mathbf{s}|\mathbf{a})\Phi(t|b)p_{\mathbf{u},v}(\mathbf{a},b)d\mathbf{a}dbd\mathbf{u}dv \\ &= \int\int G(\mathbf{s}-\mathbf{u})G'(t-v) \\ &\quad \mathbb{E}_{p_{\mathbf{u},v}(\mathbf{a},b)}[\Phi(\mathbf{s}|\mathbf{a})\Phi(t|b)]d\mathbf{u}dv \end{aligned} \tag{29}$$

We can also handle a nonstationary mixture using Green's function; the process will then be constructed as:

$$\begin{aligned} Z_{\bar{\mathrm{S}}_\mathrm{t}\bar{\mathrm{S}}_\mathrm{p}}(\mathbf{s},t) &= \int\int\int\int G(\mathbf{s}-\mathbf{u}|\mathbf{a})G'(t-v|b) \\ &\quad \Phi(\mathbf{s})\Phi(t)p_{\mathbf{u},v}(\mathbf{a},b)d\mathbf{a}dbd\mathbf{u}dv \\ &= \int\int \mathbb{E}_{p_{\mathbf{u},v}(\mathbf{a},b)}[G(\mathbf{s}-\mathbf{u}|\mathbf{a})G'(t-v|b)] \\ &\quad \Phi(\mathbf{s})\Phi(t)d\mathbf{u}dv \end{aligned} \tag{30}$$

Example: Spatio-temporal Heat Equation

The spatio-temporal heat equation:

$$\begin{aligned} \frac{df(x_1,x_2,t)}{dt} - D(&\frac{d^2f(x_1,x_2,t)}{dx_1^2} \\ &+ \frac{d^2f(x_1,x_2,t)}{dx_2^2}) = \phi(x_1,x_2,t) \end{aligned} \tag{31}$$

is an SDE that defines the dispersion of heat through an object. The fundamental solution is given by:

$$G(x_1,x_2,t) = \frac{1}{(4\pi Dt)} exp\left(-\frac{x_1^2+x_2^2}{4Dt}\right) \tag{32}$$

.

We can construct the covariance function for $f(x_1,x_2,t)$ as:

$$\begin{aligned} &C(x_1,x_2,t,x_1',x_2',t') \\ &= \int\int G(x_1-u_{x_1},x_2-u_{x_2},t-v) \\ &G(x_1'-u_{x_1},x_2'-u_{x_2},t'-v) \\ &C_{u_{x_1},u_{x_2},v}(x_1,x_2,t,x_1',x_2',t')du_{x_2}du_{x_1}dv \end{aligned} \tag{33}$$

We will instantiate this in §6.2 as a benchmark problem. For computational simplicity , we assume only space dependency exists in the local structure, i.e. $C_{u_{x_1},u_{x_2},v}(x_1,x_2,t,x_1',x_2',t') = C_{u_{x_1},u_{x_2}}(x_1,x_2,x_1',x_2')C(t,t')$. We use the construction of Eq. 24 as our space mixture. Although the variance of time is assumed to be separable in the local covariance, the final covariance of $f(x_1,x_2,t)$ is still purely nonseparable since the convolution operator $K(x_1-u_{x_1},x_2-u_{x_2},t-v)$ is nonseparable across space-time.

6. Experiments

To demonstrate our SPMs we apply them on two synthetic datasets (§6.1 Compound function, §6.2 Heat equation), on the well-studied Irish wind dataset that is nonseparable and on a challenging setting of forecasting NO_2 across London. We compare against nonstationary separable (Paciorek & Schervish, 2004) kernels denoted as $GP(\bar{\mathrm{S}}_\mathrm{t}\mathrm{S}_\mathrm{p})$, and stationary nonseparable (Fonseca & Steel, 2011a) kernels denoted as $GP(\mathrm{S}_\mathrm{t}\bar{\mathrm{S}}_\mathrm{p})$ as well as a Treed GPs (Gramacy & Lee, 2008) and a two-layer Deep Gaussian process (DGP) (Damianou & Lawrence, 2013) with the doubly stochastic framework introduced by Salimbeni & Deisenroth (2017). We denote the SPM:nonstationary convolution (SPM:NC) as $GP(\bar{\mathrm{S}}_\mathrm{t}\bar{\mathrm{S}}_\mathrm{p}) : NC$ and the SPM:nonstationry mixing (SPM:NM) as $GP(\bar{\mathrm{S}}_\mathrm{t}\bar{\mathrm{S}}_\mathrm{p}) : NM$

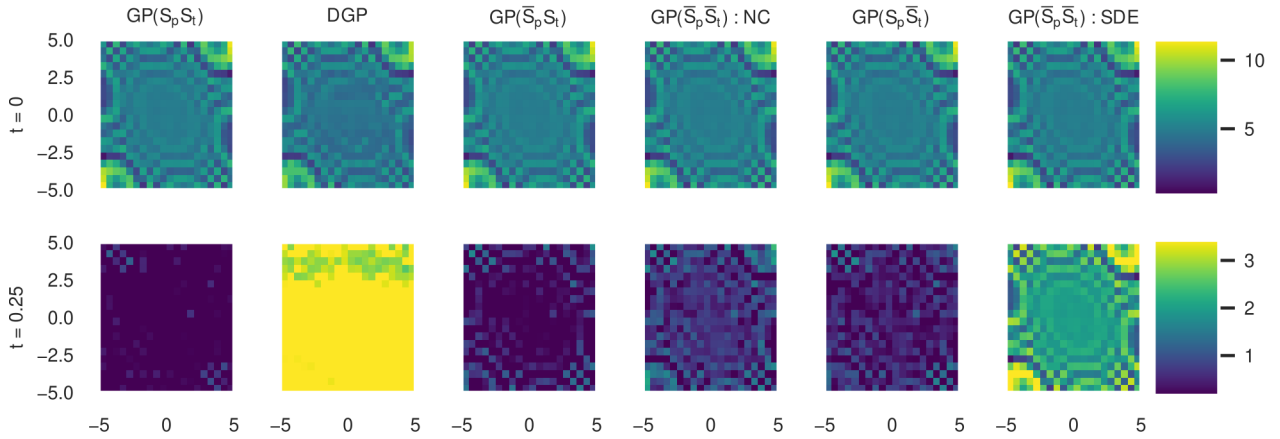

Figure 3. Predictive mean surface of spatio-temporal heat equation for different time points. The training data is only available at $t = 0$ to $t = 0.1$. With enough training data at $t = 0$, all the covariance functions predict well. For $t = 0.25$, only the nonstationary mixing with SDE convolution keeps all the structure. However, the SMP:NC still captures the change of the function using the hierarchical nonstationary nonseparable structure and predicts better than other competing approaches.

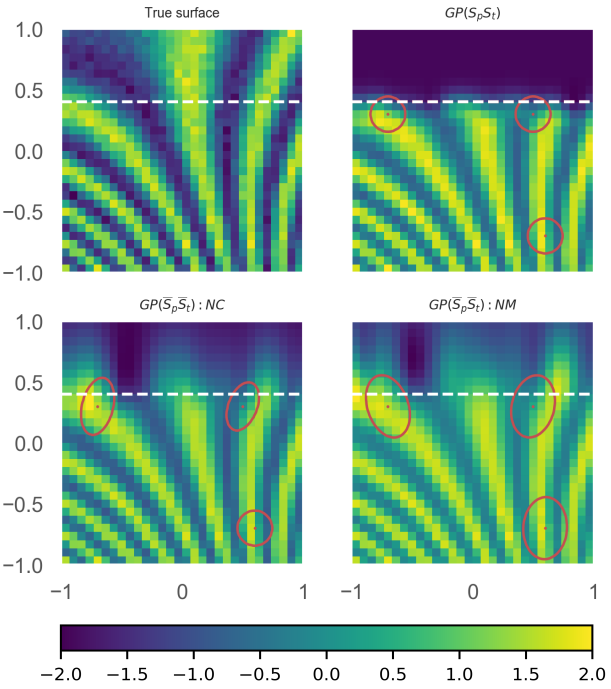

Figure 4. Illustration of kernels in different locations. Top left is the true surface. The red ellipses denote the 0.1 correlation contour line for the corresponding centre point (red dot). We train a GP using observations below the white dashed line and predict on the region $[0.4, 1.0]$. The RBF kernel (top right) is unable the capture the changing correlation structure and so learns a small length-scale, therefore is unable to predict well in the testing region. Whereas both the SPM kernels can capture the information between the input dimensions, allowing them to better predict. From the contour lines, we see that the RBF kernel has a constant shape, the SPM:NC can only model model a global dependency (all ellipses in the same direction) whereas the SPM:NM has varying correlation structure.

.

6.1. Nonseparable compound function

In our first toy example we are interested in recovering the following non-stationary non-separable surface:

$$f(\mathbf{s}, t) = \sin\left(3 \cdot (s_1^2 + s_1 \cdot (2 - s_2)^2 + t)\right) + 2$$
$$y(\mathbf{s}, t) = f(g(\mathbf{s}), t) + \epsilon \tag{34}$$

where s_1, s_2 are the first and second input dimensions of $\mathbf{s}$ respectively, $g(\mathbf{s}) = \Sigma^{\frac{1}{2}}\mathbf{s}$ is the input warping function that provides additional non-seperability, $\Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$ and $\epsilon \sim \mathcal{N}(0, 0.1)$.

To measure the amount of non-separability we calculate the empirical non-separability index ratio (0.58)(De Iaco & Posa, 2013) and run the augmented Dickey–Fuller test (-2.27) for stationarity (Lobato & Velasco, 2007), which indicates that the dataset is nonstationary and nonseparable. We generate 7 data sets with varying sample sizes using 10, 20, 30, 50, 100, 200 and 500 randomly selected observations. We expect our proposed constructions to have pronounced improvements when the sample size is small relative to the complexity of the field. For each dataset we repeat the comparison 3 times using a different random seed. We fit GP for all covariance functions. For all models, except the DGP, we optimise the hyper parameters through the marginal likelihood. To make the DGP experiment fair we use as many inducing points as input observations. We found that the single GP models were easy to fit and robust to initialization whereas the DGP has a tendency to explain the observations as noise; this required us to first hold the noise variance constant and release it half way through optimisation.

We find that all models start off with a high MSE (as to be

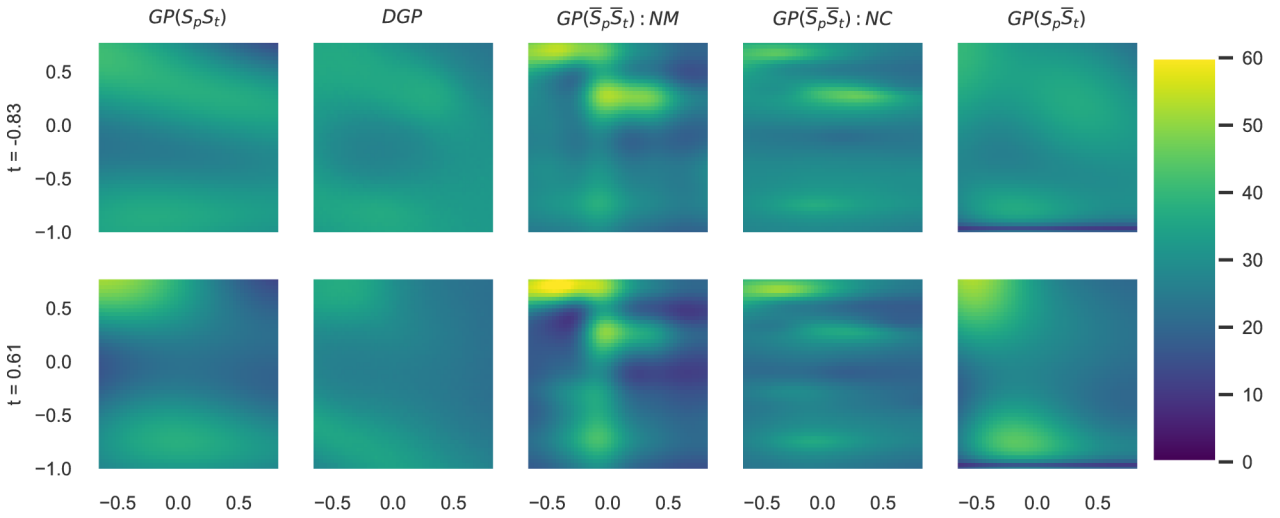

Figure 5. Predictive mean of GP models with different covariance functions estimating NO2 across London for two time slices. We see that the separable kernels oversmoothes, while both SPM:NC and SPM:NM recover more structure. Treed GP results in Appendix

expected with only 10 observations) and find that all nonseparable kernels converge to the lowest MSE the quickest. For all the covariance function trained in the experiments, when we have enough observations, the corresponding GP can fit the data quite well. However, when we have less evidence, the predictive accuracy of the separable kernel drops faster than the nonstationary nonseparable kernel as the covariance function fails to learn the correlations amongst inputs as shown in Fig. 6.

6.2. Spatio-temporal Heat equation

We are now interested in recovering a specific solution to the spatio-temporal heat equation, Eq. 31:

$$f(\mathbf{s},t) = 0.1 \cdot [50 - (x) \cdot \sin(\pi \cdot (x)/3)] \cdot \exp(-5t)$$
$$y(\mathbf{s},t) = f(g(\mathbf{s}),t) + \epsilon \tag{35}$$

where $x = s_1^2 + s_2^2$, $\epsilon \sim \mathcal{N}(0, 0.1)$ and $g(\mathbf{s}) = \Sigma^{\frac{1}{2}}\mathbf{s}$. We generate a $20 \times 20 \times 5$ uniform grid between $[-5, -5, 0]$ and $[5, 5, 0.3]$ for input s_1, s_2, t. We take the first two time slices as our training set and then predict on the remaining slices. For all models we follow the same training regime as described in Sec. 6.1. The results are shown in Fig. 3 and Table 1. In the first time step all models are able to fit to the data well but in the final slice all models apart from our SDE kernel have quickly returned to the prior mean (note that DGP returns to the mean of the data). By encoding the SDE into our prior and mixing over the parameters of the convolution we are able to forecast accurately.

6.3. Air Pollution in London

We model NO_2 across London using observations from the London air quality network and 34 sensors recording every

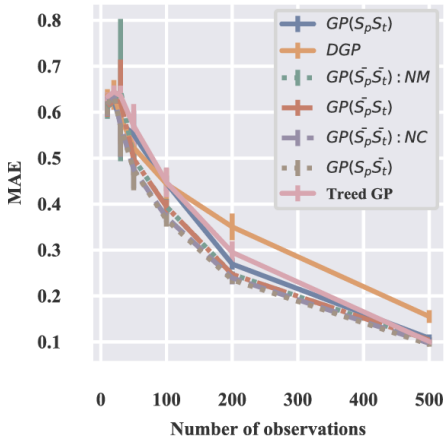

Figure 6. MAE while varying the number of observations.

.

hour. The levels of NO_2 are impacted by both global factors like weather and external air pollution sources as well as local factors such as industry and traffic. Hence we expect the correlations between sensors to be very dynamic, depending on both local and global factors. For comparison we fit the data with a single GP with separable squared exponential (SQE) covariance functions. We use the construction of Eq. 23 to handle nonseparability. To construct the mixture, we simplify the spatiotemporal random mixture as: $u_{s_1} = \lambda_0 + \lambda_1 + \lambda_2$, $u_{s_2} = \lambda_0 + \lambda_1 + \lambda_3$ and $v_t = \lambda_0 + \lambda_4$.

Thus, we can use the nonseparable construction in Eq. 23 and the exponential convolution kernel (Eq. 3). For the nonstationary convolution approach, we assume all parameters in the convolution are a linear function related to the location. For the nonstationary mixing, we assume $\lambda_0, ..., \lambda_4$

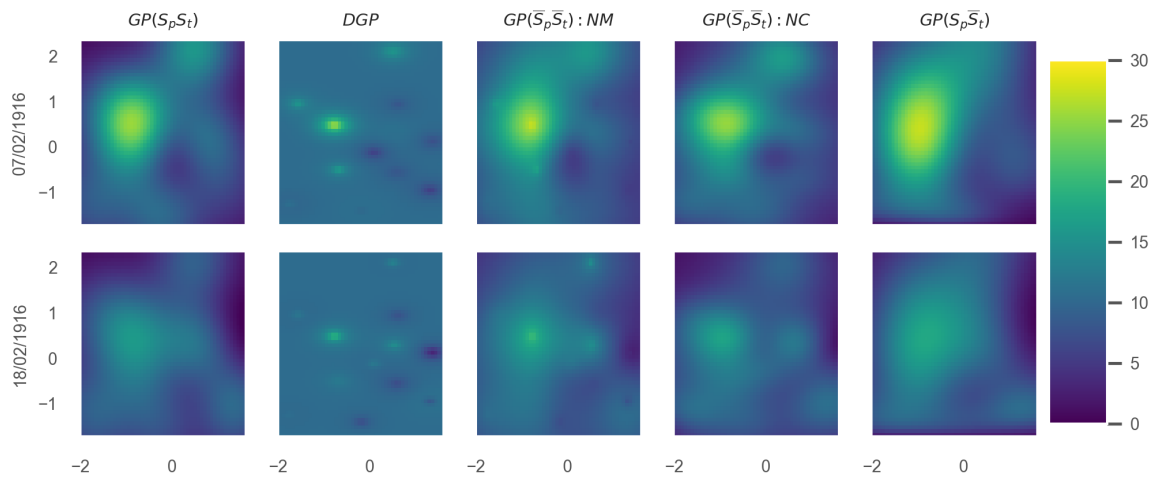

Figure 7. Prediction surface of Irish Wind data for different time points. The RBF kernel over-smooths the surface and therefore fails to capture the local structure. Whereas both SPM constructions are able to capture this structure. We can see that, for both SPM constructions, the reconstructed surfaces are similar and this is because dependency structure in the dataset is close to constant. The Deep GP also captures localized structure but achieves a predictive log likelihood of -3133.60 compared to SPM:NC and SPM:NM that achieve -2991.63 and -2707.07 respectively. This indicates that the DGP is slightly overfitting the data and is not capturing sufficient uncertainty.

Table 1. Mean square errors across datasets and competing approaches

MODEL	COMPOUND	3D HEAT EQUATION	LONDON AIR QUALITY	IRISH WIND
SINGLE GP	0.415 ± 0.05	0.93	51.24 ± 1.32	12.79 ± 1.27
TREED GP	0.483 ± 0.04	4.34	70.32	13.02 ± 1.02
DEEP GP	0.430 ± 0.03	2.36	50.33 ± 4.37	2.61 ± 0.12
SPM:NM	0.366 ± 0.04	0.14	18.31 ± 2.26	2.92 ± 0.32
SPM:NC	0.360 ± 0.03	0.26	29.80 ± 1.74	9.65 ± 0.15

follows independent Gamma distributions.

As shown in Fig 5, the SQE kernel does not learn the correlation structure between location variables and it oversmooths resulting in large noise variance. The SPM:NC kernel assumes that the structure for the location parameters are fixed. This shows in the predictive surface at different times. The SPM:NM kernel learns the varying nonseparability successfully. Since the varying nonseparability relies on the local structure of the surface, the kernel infers more local structure than all the other kernels. In contrast with the treed GP, which assumes that different partitions are independent, the SMP:NM kernel is still able to learn long term correlations between different local structures through the convolution.

6.4. Irish Wind

The Irish wind data is well known for exhibiting nonseparability. It measures daily average wind speed for 12 different location in Ireland. After standardizing the data, we run a separability test (De Iaco & Posa, 2013) on the data that results in a separability of 0.38, while the separability ratio over two individual stations is also around 0.38. This indicates shared structure amongst all stations. We observe this in the results; the SPM:NM and SPM:NC are similar to each other (Fig. 7), as the final covariance functions all converge into a surface with constant nonseparability.

7. Conclusion

We have generalized process convolution kernels using stochastic mixing to handle both nonstationarity and nonseparability in the data. We demonstrate improved estimates and forecasts as the underlying GP model gains from both the nonstationarity and nonseparability properties of the kernels. As the form of the convolution kernel is generally unknown, we can motivate our convolution function from stochastic differential equations. Thus, any additional physical information can be brought into the covariance function. We illustrate in §6.2 that the SDE informed convolution can reach better prediction with less observations. Finally, we show that our SMP:NM captures local varying structure which is crucial in real world spatio-temporal problems.

Acknowledgements

K. W., O. H. and T. D. are funded by the Lloyd's Register Foundation programme on Data Centric Engineering through the London Air Quality project. O. H. is funded through The Alan Turing Institute PhD fellowship programme. This work is supported by The Alan Turing Institute for Data Science and AI under EPSRC grant EP/N510129/1 in collaboration with the Greater London Authority. In addition we acknowledge support and funding from Microsoft. We would like to thank the anonymous reviewers for their feedback and Daniel Tait, Jeremias Knoblauch and Patrick O'Hara for their help on multiple aspects of this work.

References

Alvarez, M. A., Rosasco, L., Lawrence, N. D., et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Brown, P. E., Roberts, G. O., Kåresen, K. F., and Tonellato, S. Blur-generated non-separable space–time models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):847–860, 2000.

Chen, K., van Laarhoven, T., Chen, J., and Marchiori, E. Incorporating dependencies in spectral kernels for gaussian processes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 565–581. Springer, 2019.

Cortes, C., Mohri, M., and Rostamizadeh, A. Learning nonlinear combinations of kernels. In *Advances in neural information processing systems*, pp. 396–404, 2009.

Cressie, N. The origins of kriging. *Mathematical geology*, 22(3):239–252, 1990.

Damianou, A. and Lawrence, N. Deep gaussian processes. In *Artificial Intelligence and Statistics*, pp. 207–215, 2013.

De Iaco, S. and Posa, D. Positive and negative nonseparability for space–time covariance models. *Journal of Statistical Planning and Inference*, 143(2):378–391, 2013.

Fonseca, T. C. and Steel, M. F. A general class of nonseparable space–time covariance models. *Environmetrics*, 22 (2):224–242, 2011a.

Fonseca, T. C. and Steel, M. F. Non-gaussian spatiotemporal modelling through scale mixing. *Biometrika*, 98(4):761–774, 2011b.

Fuentes, M. and Smith, R. L. A new class of nonstationary spatial models. Technical report, Technical report, North Carolina State University, Raleigh, NC, 2001.

Gneiting, T. Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458):590–600, 2002.

Gramacy, R. B. and Lee, H. K. H. Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483): 1119–1130, 2008.

Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., and Lähdesmäki, H. Non-stationary gaussian process regression with hamiltonian monte carlo. In *Artificial Intelligence and Statistics*, pp. 732–740, 2016.

Higdon, D. Space and space-time modeling using process convolutions. In Anderson, C. W., Barnett, V., Chatwin, P. C., and El-Shaarawi, A. H. (eds.), *Quantitative Methods for Current Environmental Issues*, pp. 37–56, London, 2002. Springer London.

Lewis, D. P., Jebara, T., and Noble, W. S. Nonstationary kernel combination. In *Proceedings of the 23rd international conference on Machine learning*, pp. 553–560, 2006.

Lindgren, F., Rue, H., and Lindström, J. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

Lobato, I. N. and Velasco, C. Efficient wald tests for fractional unit roots. *Econometrica*, 75(2):575–589, 2007.

Monterrubio-Gómez, K., Roininen, L., Wade, S., Damoulas, T., and Girolami, M. Posterior inference for sparse hierarchical non-stationary models. *arXiv preprint arXiv:1804.01431*, 2018.

Paciorek, C. J. and Schervish, M. J. Nonstationary covariance functions for gaussian process regression. In *Advances in neural information processing systems*, pp. 273–280, 2004.

Paciorek, C. J. and Schervish, M. J. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics: The official journal of the International Environmetrics Society*, 17(5):483–506, 2006.

Reece, S., Garnett, R., Osborne, M., and Roberts, S. Anomaly detection and removal using non-stationary gaussian processes. *arXiv preprint arXiv:1507.00566*, 2015.

Remes, S., Heinonen, M., and Kaski, S. Non-stationary spectral kernels. In *Advances in Neural Information Processing Systems*, pp. 4642–4651, 2017.

Rodrigues, A. and Diggle, P. J. A class of convolution-based models for spatio-temporal processes with non-separable covariance structure. *Scandinavian Journal of Statistics*, 37(4):553–567, 2010.

Salimbeni, H. and Deisenroth, M. P. Deeply non-stationary gaussian processes. In *Proc. NIPS Workshop Bayesian Deep Learn*, 2017.

Scholkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

Snoek, J., Swersky, K., Zemel, R., and Adams, R. Input warping for bayesian optimization of non-stationary functions. In *International Conference on Machine Learning*, pp. 1674–1682, 2014.

Stein, M. L. Nonstationary spatial covariance functions. *Unpublished technical report*, 2005.

Ton, J.-F., Flaxman, S., Sejdinovic, D., and Bhatt, S. Spatial mapping with gaussian processes and nonstationary fourier features. *Spatial Statistics*, 28:59 – 78, 2018. ISSN 2211-6753. doi: https://doi.org/10.1016/j.spasta.2018.02.002. URL http://www.sciencedirect.com/science/article/pii/S2211675317302890. One world, one health.

Wilson, A. and Adams, R. Gaussian process kernels for pattern discovery and extrapolation. In *International conference on machine learning*, pp. 1067–1075, 2013.