Nonstationary Nonseparable Random Fields

Kangrui Wang [1 2] **Oliver Hamelijnck** [1 3] **Theodoros Damoulas** [1 2 3] **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 nonstationary 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 the correlation can be expressed as $C(\mathbf{s}, t, \mathbf{s}', t')/C(\mathbf{s}, \mathbf{s}')C(t, t') = \rho$ for $s \neq s'$ and $t \neq t'$ and 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. To ease exposition we define the following subscripts: $Z_{\mathrm{S_t S_p}}$ is a stationary separable process, $Z_{\mathrm{S_t \bar{S}_p}}$ is a stationary nonseparable one, $Z_{\mathrm{\bar{S}_t S_p}}$ is a nonstationary separable process and $Z_{\mathrm{\bar{S}_t \bar{S}_p}}$ is nonstationary nonseparable one.

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 nonstationary data when the combination is location-dependent.

[1]Data-centric Engineering, The Alan Turing Institute, London, UK [2]Department of Statistics, University of Warwick, Coventry, UK [3]Department of Computer Science, University of Warwick. 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, Online, PMLR 119, 2020.

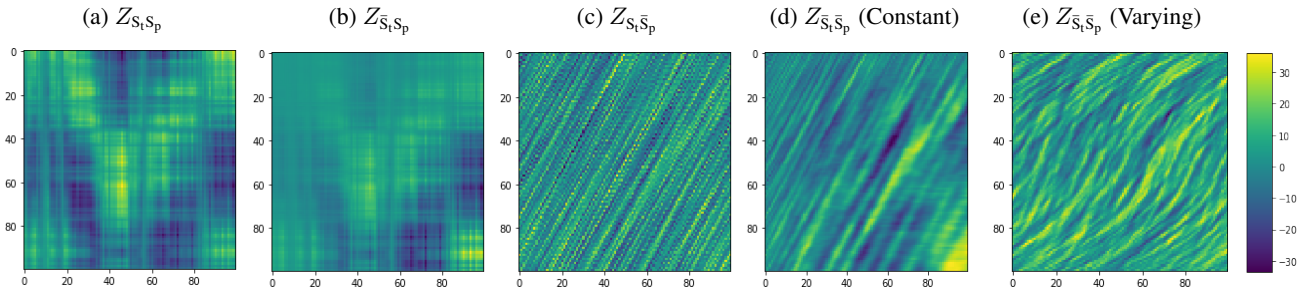

Figure 1. Illustration of samples from 2D Gaussian processes with various combinations of stationarity and separability. The fields (a, c) are stationary, (b, d, e) are nonstationary; fields (a,b) are separable and (c,d,e) are non-separable. Nonstationary fields (b,d) exhibit varying levels of smoothness, changing from top left to bottom right. Higher levels of nonseparability express progressively more complex dependencies between the input dimensions: from independence (a) to linear dependency structure (c,d) and varying local correlation structure (e). In (e) the smoothness of the process is fixed and the nonstationarity arises from the varying dependency structure.

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) 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. When 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). We introduce these in §2 and generalize 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 nonstationary nonseparable 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 non-parametric 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. Code is available at `https://github.com/ohamelijnck/nsns_kernels` and is implemented in GPFlow (Matthews et al., 2017).

2. Stochastic Process Mixing (SPM)

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 white noise process Φ and a valid kernel (convolution) function K then $Z(\mathbf{x})$ can be defined as

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

and is a stochastic process where $\mathbf{x}, \mathbf{u} \in \mathbb{R}^D$ and $\Phi_{\mathbf{u}} \sim \mathcal{N}(0, \mathbf{I})$. We will refer to Φ as the latent process. 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 and Φ is a stationary process then the resulting $Z(\mathbf{x})$ will also be stationary. When K is nonstationary or Φ has the form of a nonstationary stochastic process then the resulting field will be nonstationary (Paciorek & Schervish, 2004; Fuentes & Smith, 2001). A nonstationary process is then given by:

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

where we use subscripts to denote dependence on the input and latent variables. Note that when $\Phi_{\mathbf{u}}(\mathbf{x})$ is a GP then Z will also be a GP, due to convolution being a linear operator. Typically the process Z will depend on some parameters ($\mathbf{a}$), either from the kernel or the latent $\Phi_{\mathbf{u}}(\mathbf{x})$. Placing a prior over these gives 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 (Fonseca, 2010).

Nonseparability: We generalise the construction of Eq. 4 to a nonseparable spatio-temporal process that is a mixture of conditionally separable processes. We have:

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

where we overload $\mathbf{x}$ to denote space-time locations $\{\mathbf{s}, t\}$. 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 (Ma, 2003) and when $p(\mathbf{a}, b)$ depends on $\mathbf{u}, v$ the process $Z(\mathbf{s}, t)$ will be nonstationary (Fuentes & Smith, 2001). The covariance function of Eq.6 is:

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

where the latent covariances 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 convolutions and nonstationary mixings which results in an increased number of hyperparameters and raises the computational complexity. Furthermore, because the nonstationarity may be learnt through both the convolution and the mixing this leads to identifiability issues. Due to this we consider them separately by constructing processes where the nonstationarity is constrained to either the convolution or the mixing.

3. SPMs through Nonstationary Convolution

In our first construction we limit the nonstationarity of the constructed processes to be expressed in the convolution function. The resulting stochastic functions have constant nonseparability (as seen in Fig. 1 - d) and are a restricted class of those defined in Eq. 6. The general construction of these processes is given by:

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

Because the mixing distribution does not depend on $\mathbf{u}, v$ all of the nonstationarity is expressed in the convolution function. The convolution integral can then be computed before the mixing marginalisation, allowing for nonstationary separable processes to be plugged in (e.g Paciorek & Schervish (2004)) and additionally mixed thus constructing nonstationary nonseparable processes. When the covariance of $Z_{\bar{\mathbf{S}}_t \mathbf{S}_p}(\mathbf{s}|\mathbf{a})$ and $Z_{\bar{\mathbf{S}}_t \mathbf{S}_p}(t|b)$ are available in closed form, then the covariance of the marginalized process can be calculated using a mixture of the conditional covariances:

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

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 nonstationary separable process with scale mixture parameters $\mathbf{a}, b$. We can simplify the form of Eq. 8 by only mixing the parameters of convolution function or the latent process. In any case the resulting process will have a nonseparable covariance function.

Special cases with closed forms

In this section we provide closed form examples as special cases of Eq. 9. This first requires computing the nonstationary separable covariance functions and additionally deriving closed form expression for the mixing of these. In the following examples we mix the parameters of the convolution:

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

such that 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 $\mathbf{u}$ depends on a scalar random variable a. We present more complex mixing in our applications. Defining the convolution function as:

$$K_{\mathbf{s}}(\mathbf{s}-\mathbf{u}|a) = \exp\left(-a(\mathbf{s}-\mathbf{u})\Sigma(\mathbf{s})^{-1}(\mathbf{s}-\mathbf{u})^T\right), \tag{11}$$

the latent covariance as $C_{\mathbf{u}}(\mathbf{s}, \mathbf{s}') = \sigma(\mathbf{s})\sigma(\mathbf{s}')$ and letting

$$Q_{\mathbf{r},\mathbf{r}'} = (\mathbf{r}-\mathbf{r}')\left(\frac{\Sigma(\mathbf{r})+\Sigma(\mathbf{r}')}{2}\right)^{-1}(\mathbf{r}-\mathbf{r}')^T \tag{12}$$

where $\mathbf{r}$ is either $\mathbf{s}$ or t. Substituting into Eq. 10 recovers the nonstationary covariance of Paciorek & Schervish (2006):

$$C_{\bar{\mathbb{S}}_t\mathbb{S}_p}(\mathbf{s}, \mathbf{s}'|a) = A(s, s')\exp(-aQ_{\mathbf{s},\mathbf{s}'})$$

where $A(s, s') = \sigma(\mathbf{s})\sigma(\mathbf{s}')\frac{|\Sigma(\mathbf{s})|^{\frac{1}{4}}|\Sigma(\mathbf{s}')|^{\frac{1}{4}}}{|\Sigma(\mathbf{s})+\Sigma(\mathbf{s}')|^{\frac{1}{2}}}$. Defining $R(\mathbf{s}, \mathbf{s}'|\mathbf{a}) = \exp(-aQ_{\mathbf{s},\mathbf{s}'})$ and $R(t, t'|b) = \exp(-bQ_{t,t'})$ then we can marginalise $\mathbf{a}$ and b:

$$\begin{aligned} R(\mathbf{s}, \mathbf{s}', t, t') &= \mathbb{E}_{a,b}[\exp(-aQ_{\mathbf{s},\mathbf{s}'})\exp(-bQ_{t,t'})] \\ &= M_a(-Q_{\mathbf{s},\mathbf{s}'})M_b(-Q_{t,t'}) \end{aligned} \tag{13}$$

where $M(.)$ is the moment generation function (MGF). Setting a and b as combinations of independent random variables : $a = \lambda_0+\lambda_1$ and $b = \lambda_0+\lambda_2$ and using the properties of MGFs we rewrite Eq. 13 as $M_a(-Q_{\mathbf{s},\mathbf{s}'})M_b(-Q_{t,t'}) =$

$$M_{\lambda_0}(-(Q_{\mathbf{s},\mathbf{s}'}+Q_{t,t'}))M_{\lambda_1}(-Q_{\mathbf{s},\mathbf{s}'})M_{\lambda_2}(-Q_{t,t'}).$$

With specific distributions on $\lambda_0, \lambda_1, \lambda_2$, we arrive at *closed forms* examples of nonstationary nonseparable functions.

Example 1: Let the i.i.d mixing variables follow $\lambda_0 \sim \text{Ga}(\beta_0, 1)$, $\lambda_1 \sim \text{Ga}(\beta_1, 1)$, $\lambda_2 \sim \text{Ga}(\beta_2, 1)$ then $R(\mathbf{s}, \mathbf{s}', t, t') =$

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

Example 2: Let the iid mixing variables follow $\lambda_0 \sim \text{IGa}(\beta_0, 1/4)$, $\lambda_1 \sim \text{IGa}(\beta_1, 1/4)$, $\lambda_2 \sim \text{IGa}(\beta_2, 1/4)$ then $R(\mathbf{s}, \mathbf{s}', t, t') =$:

$$R'_{\beta_0}(Q_{\mathbf{s},\mathbf{s}'}+Q_{t,t'})R'_{\beta_1}(Q_{\mathbf{s},\mathbf{s}'})R'_{\beta_2}(Q_{t,t'})$$

$$R'_{\beta}(Q) = \frac{1}{\gamma(\beta)2^{\beta-1}}\left(\sqrt{2\beta Q}\right)^{\beta}K_{\beta}(\sqrt{2\beta Q})$$

4. SPMs through Nonstationary Mixing

In an alternative approach, we fix the convolution function and let the mixing distribution vary across input space. Fonseca & Steel (2011a) developed the nonstationary convolution based on the spatial dimension: $C(\mathbf{s}, \mathbf{s}', \hat{t}) =$

$$\int K(\mathbf{s}-\mathbf{u})K(\mathbf{s}'-\mathbf{u})\int C_{\mathbf{u}}(\mathbf{s}, \mathbf{s}'|\mathbf{a})C(\hat{t}|b)\, d\mu(a, b)\, d\mathbf{u} \tag{14}$$

where $\hat{t} = \{t, t'\}$ and $C(\hat{t}|b)$ is assumed to be stationary and hence any nonstationarity across time is not modelled. We generalise this to a fully nonstationary nonseparable kernel by constructing an SPM as:

$$\begin{aligned} Z_{\bar{\mathbb{S}}_t\bar{\mathbb{S}}_p}(\mathbf{x}) = \int\int\int\int K(\mathbf{s}-\mathbf{u})K(t-v) \\ \Phi_{\mathbf{u}}(\mathbf{s}|\mathbf{a})\Phi_v(t|b)p_{\mathbf{u},v}(\mathbf{a}, b)\, d\mathbf{a}\, db\, d\mathbf{u}\, dv \end{aligned} \tag{15}$$

If the mixing distribution $p_{\mathbf{u},v}(\mathbf{a}, b)$ is nonstationary across locations $\{\mathbf{u}, v\}$, we cannot, in general, get a closed form for the marginal process $Z_{\bar{\mathbb{S}}_t\bar{\mathbb{S}}_p}(\mathbf{s}, t)$. However, we can write down the conditional covariances when the latent process $\Phi_{\mathbf{u}}(\mathbf{s}|\mathbf{a}), \Phi_v(t|b)$ is separable across locations $\{\mathbf{u}, v\}$:

$$\begin{aligned} C(\mathbf{x}, \mathbf{x}') = \int\int K(\mathbf{s}-\mathbf{u})K(\mathbf{s}'-\mathbf{u})K(t-v)K'(t'-v) \\ \int\int C_{\mathbf{u}}(\mathbf{s}, \mathbf{s}'|\mathbf{a})C_v(t, t'|b)\, d\mu_{\mathbf{u},v}(\mathbf{a}, b)\, d\mathbf{u}\, dv \end{aligned} \tag{16}$$

where the mixing integral is now performed before the convolution. By defining the convolution function to be stationary and separable the above covariance function can be seen as a special case of Eq. 7. In this construction the convolution function is acting as a smoother on a nonstationary nonseparable process and so in general it is difficult to derive closed form expressions for the full covariance. But, in some cases, it is possible to derive closed form expressions for the latent conditional covariance:

$$C_{\mathbf{u},v}(\mathbf{s}, \mathbf{s}', t, t') = \mathbb{E}_{p_{\mathbf{u},v}(\mathbf{a},b)}[C(\mathbf{s}, \mathbf{s}'|\mathbf{a})C(t, t'|b)] \tag{17}$$

This covariance function defines all the nonstationarity and nonseparability of the resulting processes. Because the covariance depends on the convolution parameters we say that

it defines the local nonseparable structure. Any nonseparable covariance function may be used here including the closed-form special cases presented in §3.

In Eq. 15 the nonseparability of the process is obtained by mixing the latent processes. Alternatively the convolution kernel may be mixed:

$$Z_{\bar{\mathbf{S}}_t\bar{\mathbf{S}}_p}(\mathbf{s},t) = \int\int\int\int K(\mathbf{s}-\mathbf{u}|\mathbf{a})K(t-v|b) \Phi(\mathbf{s})\Phi(t)p_{\mathbf{u},v}(\mathbf{a},b)\,d\mathbf{a}\,db\,d\mathbf{u}\,dv \tag{18}$$

thus defining a nonseparable convolution: $C(\mathbf{s},\mathbf{s}',t,t') =$

$$\int\int \mathbb{E}_{p_{\mathbf{u},v}(\mathbf{a},b)}\big[K(\mathbf{s}-\mathbf{u}|\mathbf{a})K(t-v|b) K(\mathbf{s}'-\mathbf{u}|\mathbf{a})K(t'-v|b)\big]C(\mathbf{s},\mathbf{s}')C(t,t')\,d\mathbf{u}\,dv \tag{19}$$

that also defines local non separability because the mixing function still depends on the convolution parameters. Both mixing approaches in Eq. 17 and Eq. 19 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. 17 rather than Eq. 19 and vice versa.

Example 3: SPM via Nonstationary Mixing
We derive a covariance function following the nonstationary mixing construction in Eq. 16 but we restrict the latent covariance to be stationary. Additionally we define the convolution function $K(\cdot)$ to be a stationary Gaussian convolution. Let the locally stationary processes follow the covariance $C'(\mathbf{s},\mathbf{s}'|a(\mathbf{u}))C'(t,t'|b(v)) =$

$$\exp\left(-a(\mathbf{u})\sum_{i=1}^{D}\frac{(s_i-s_i')^2}{\ell_{s_i}}\right)\exp\left(-b(v)\frac{(t-t')^2}{\ell_t}\right)$$

then we can construct the nonstationary mixing via a linear function of independent variables.
Let $\lambda_0 \sim \text{Ga}(\beta_0,1)$, $\lambda_1 \sim \text{Ga}(\beta_1(\mathbf{u}),1)$ 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 . Then:

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

where $Q_s = (s-s')^2/\ell_s$ and $C'_\beta(Q) = (1+Q)^{-\beta}$. We demonstrate this SPM experimentally in §6.

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):

$$a_pZ^{(p)}(t) + a_{p-1}Z^{(p-1)}(t) + ... + a_0Z(t) = \phi(t) \tag{20}$$

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.

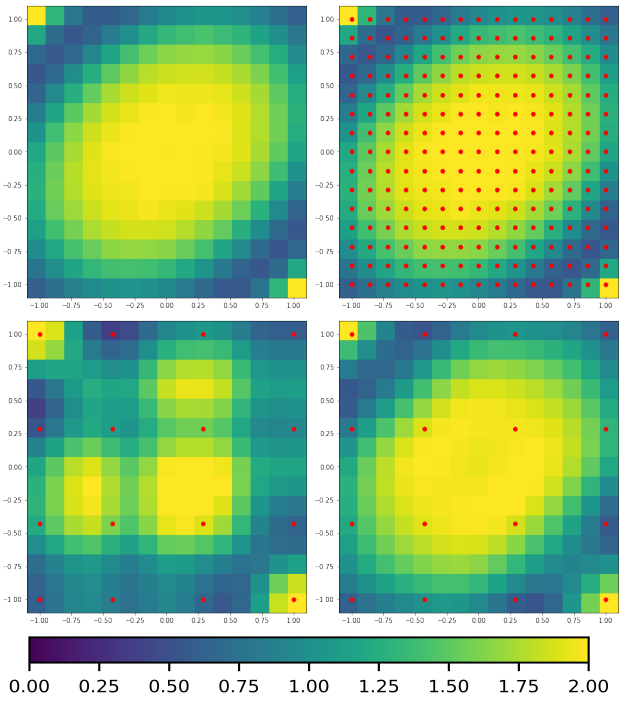

Figure 2. Illustrations of GPs modelling a solution to the 2D heat equation. The true surface is plotted in the top left and red points denote observation locations. Top right and bottom left are GP predictions using a squared exponential kernel across varying sample sizes. Bottom right shows predictions from our SDE informed SPM. Because the SPM encodes the physical behaviour of the process it achieves superior performance in recovering the true surface, even on a small sample size.

Instead of solving Eq. 20 directly, we can find the corresponding Green's function and rewrite the process of interest as a convolution against this:

$$Z(t) = \int G(t-u)\phi_u(t)\,du \tag{21}$$

where $G(t-u)$ is the Green's function for the SDE in Eq. 20 (Duffy, 2015). 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 solutions of the SDE as arising from a convolution we can cast it into both our nonstationary convolution and nonstationary mixing frameworks.

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

$$\begin{aligned} Z(t) &= \int Z(t|b)d\mu(b) \\ &= \int\int G(t-u|b)\phi_u(t|b)\,d\mu(b)\,du, \end{aligned} \tag{22}$$

where $\mu(b)$ is the probability measure for random variable b. In Eq.22 we do not directly express $Z(t)$ as a convolution of a Green's function, because in general it is hard to find the corresponding SDE. Instead we find the SDE for the conditional $Z(t|b)$ where b is a mixing variable that varies across input space. When the input space is multi-dimensional, the correlation of the mixing variables captures the dependency between the input dimensions.

Let the separable process $Z(\mathbf{s},t|\mathbf{a},b) = Z_s(\mathbf{s}|\mathbf{a})Z_t(t|b)$ be written as the solution to the following SDE:

$$\begin{aligned} \sum_i a_i \frac{\partial^i Z_s(\mathbf{s}|\mathbf{a})}{\partial s^i} + a_0 Z_s(\mathbf{s}|\mathbf{a}) &= \phi(\mathbf{s}|\mathbf{a}), \\ \sum_j b_j \frac{\partial^j Z_t(t|\mathbf{b})}{\partial t^j} + b_0 Z_t(t|\mathbf{b}) &= \phi(t|\mathbf{b}). \end{aligned} \tag{23}$$

then as we have seen in our nonstationary mixing framework (§4) we can induce correlation between the input dimensions by mixing $\mathbf{a}$ and $\mathbf{b}$. Let $\mathbf{a} = \{a_0, ...a_I\}, \mathbf{b} = \{b_0, ...b_J\}$ (where I, J are the order of SDEs for Z_s, Z_t respectively) be random variables from the joint distribution $p(\mathbf{a}, b)$ that mix the above SDE solutions. As in §4 we can induce a nonstationary nonseparable process, even when the latent SDEs are stationary, i.e. $Z_{\bar{\mathrm{S}}_\mathrm{t}\bar{\mathrm{S}}_\mathrm{p}}(\mathbf{x})$

$$\begin{aligned} &= \int\int \mathbb{E}_{p_{\mathbf{u},v}(\mathbf{a},b)}\left[G(\tau_s)G'(\tau_t)\Phi(\mathbf{s}|\mathbf{a})\Phi(t|b)\right]d\mathbf{u}dv \\ &= \int\int G(\tau_s)G'(\tau_t)\mathbb{E}_{p_{\mathbf{u},v}(\mathbf{a},b)}[\Phi(\mathbf{s}|\mathbf{a})\Phi(t|b)]d\mathbf{u}dv \end{aligned}$$

where $\tau_s = \mathbf{s} - \mathbf{u}$, $\tau_t = t - v$. We can also handle a nonstationary mixture using a nonstationary Green's function; the process will then be constructed as $Z_{\bar{\mathrm{S}}_\mathrm{t}\bar{\mathrm{S}}_\mathrm{p}}(\mathbf{s},t)$

$$\begin{aligned} &= \int\int \mathbb{E}_{p_{\mathbf{u},v}}\left[G(\tau_s|\mathbf{a})G'(\tau_t|b)\Phi(\mathbf{s})\Phi(t)p_{\mathbf{u},v}(\mathbf{a},b)\right]d\mathbf{u}dv \\ &= \int\int \mathbb{E}_{p_{\mathbf{u},v}(\mathbf{a},b)}\left[G\tau_s|\mathbf{a})G'(\tau_t|b)\right]\Phi(\mathbf{s})\Phi(t)d\mathbf{u}dv \end{aligned}$$

Example 4: Spatio-temporal Heat Equation
The spatio-temporal heat equation:

$$\frac{df(\mathbf{x})}{dt} - D\cdot\left[\frac{d^2f(\mathbf{x})}{ds_1^2} + \frac{d^2f(\mathbf{x})}{ds_2^2}\right] = \phi(\mathbf{x})$$

is an SDE that defines the dispersion of heat through an object (Duffy, 2015). The stochastic factor $\phi(\mathbf{x})$ of the system is related to the input space $\mathbf{x} = \{s_1, s_2, t\}$ and the fundamental solution is:

$$G(\mathbf{x}) = \frac{1}{(4\pi Dt)}\exp\left(-\frac{s_1^2+s_2^2}{4Dt}\right).$$

When the stochastic factor depends on the input location then the covariance function of $f(\mathbf{x})$ is given by:

$$\begin{aligned} C(\mathbf{x},\mathbf{x}') = \int\int &G(\tau_{s_1},\tau_{s_2},\tau_t)G(\tau_{s_1'}\tau_{s_2'},\tau_{t'}) \\ &C_{u_{s_1},u_{s_2},v}(\mathbf{x},\mathbf{x}')du_{s_2}du_{s_1}dv \end{aligned}$$

where $\tau_{s_1} = s_1 - u_{s_1}, \tau_{s_2} = s_2 - u_{s_2}, \tau_t = t - v$. We will instantiate this in §6.2 as a benchmark problem. For computational simplicity, we consider that the stochastic factor is only nonseparable in the spatial dimension. The latent covariance is then: $C_{u_{s_1},u_{s_2},v}(\mathbf{x},\mathbf{x}') = C_{u_{s_1},u_{s_2}}(\mathbf{s},\mathbf{s}')C_v(t,t')$ where the spatial covariance is given by Example. 1 and the temporal covariance is an squared exponential kernel (SQE).
Although the latent covariance is separable across space-time the covariance of the resulting process we be fully nonseparable. This is because $G(\tau_{s_1},\tau_{s_2},\tau_t)$ is informed by a corresponding SDE that is itself nonseparable across space-time.

6. Experiments

To demonstrate our SPMs we apply them on two synthetic datasets, on the well-studied Irish wind dataset and on the challenging setting of forecasting NO_2 across London. We compare against nonstationary separable kernels (Paciorek & Schervish, 2004) denoted as $GP(\bar{\mathrm{S}}_\mathrm{t}\mathrm{S}_\mathrm{p})$, and stationary nonseparable kernels (Fonseca & Steel, 2011a) denoted as $GP(\mathrm{S}_\mathrm{t}\bar{\mathrm{S}}_\mathrm{p})$, Treed GPs (Gramacy & Lee, 2008) and a two-layer Deep Gaussian process (DGP) (Damianou & Lawrence, 2013) with the doubly stochastic framework (Salimbeni & Deisenroth, 2017). We denote SPM:nonstationary convolutions (SPM:NC) as $GP(\bar{\mathrm{S}}_\mathrm{t}\bar{\mathrm{S}}_\mathrm{p})$:NC and SPM:nonstationary mixings (SPM:NM) as $GP(\bar{\mathrm{S}}_\mathrm{t}\bar{\mathrm{S}}_\mathrm{p})$:NM. A summary of results is provided in Table. 1.

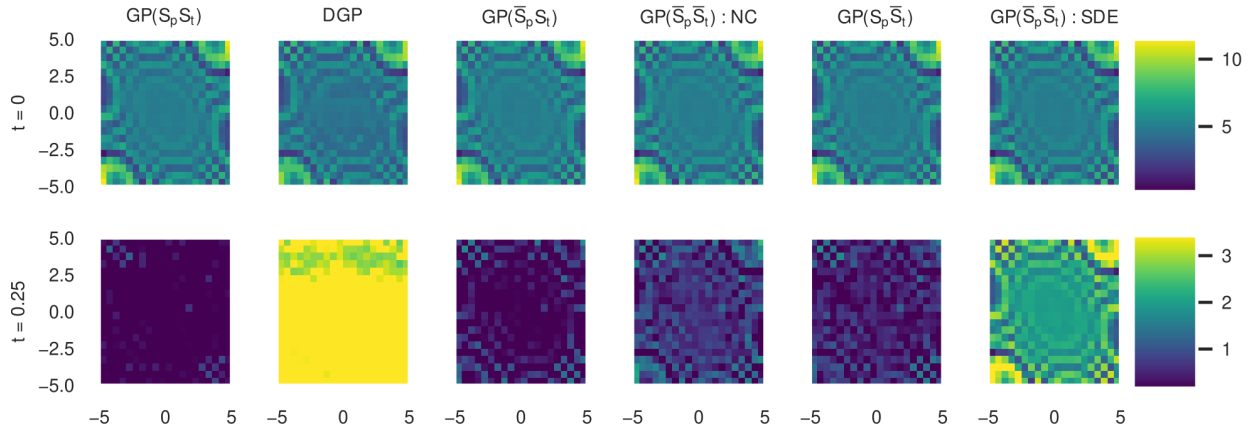

Figure 3. Predictive mean surfaces from GPs with different kernel functions for two time points of the spatio-temporal heat equation. The training data is a solution to the spatio-temporal heat equation and is only generated from $t = 0$ to $t = 0.1$. At $t = 0$ all the covariance functions capture the structure of the data because observations are available. At $t = 0.25$, only the nonstationary mixing with the SDE convolution ($GP(\bar{S}_t\bar{S}_p)$:SDE) maintains this structure, because the kernel is informed via the SDE, whereas all other models start to return to the mean. Note that the nonstationary convolution ($GP(\bar{S}_t\bar{S}_p)$:NC) still captures some of the structure through the hierarchical nonstationary nonseparable structure and therefore has second-best performance.

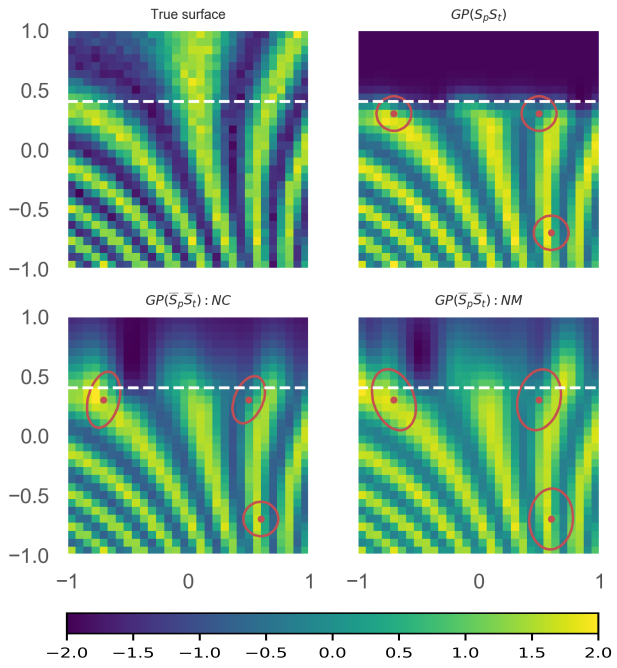

Figure 4. Illustration of GP correlations across multiple input locations. Top left is the true surface. The red ellipses denote the 0.1 correlation contour line for the corresponding centre point (red dot). We train a GP using observations below the white dashed line and predict on the region above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

$$f(\mathbf{s}, t) = \sin\left(3 \cdot (s_1^2 + s_1 \cdot (2 - s_2)^2 + t)\right) + 2$$
$$y(\mathbf{s}, t) = f(g(\mathbf{s}), t) + \epsilon \tag{24}$$

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 nonseperability, $\Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$ and $\epsilon \sim \mathcal{N}(0, 0.1)$.

To measure the amount of nonseparability we calculate the empirical nonseparability 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 both nonstationary and nonseparable.

We generate 7 data sets with increasing 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 use single GPs for all covariance functions and to make comparisons with the DGP fair we optimize all models w.r.t to their variational lower bounds and 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.

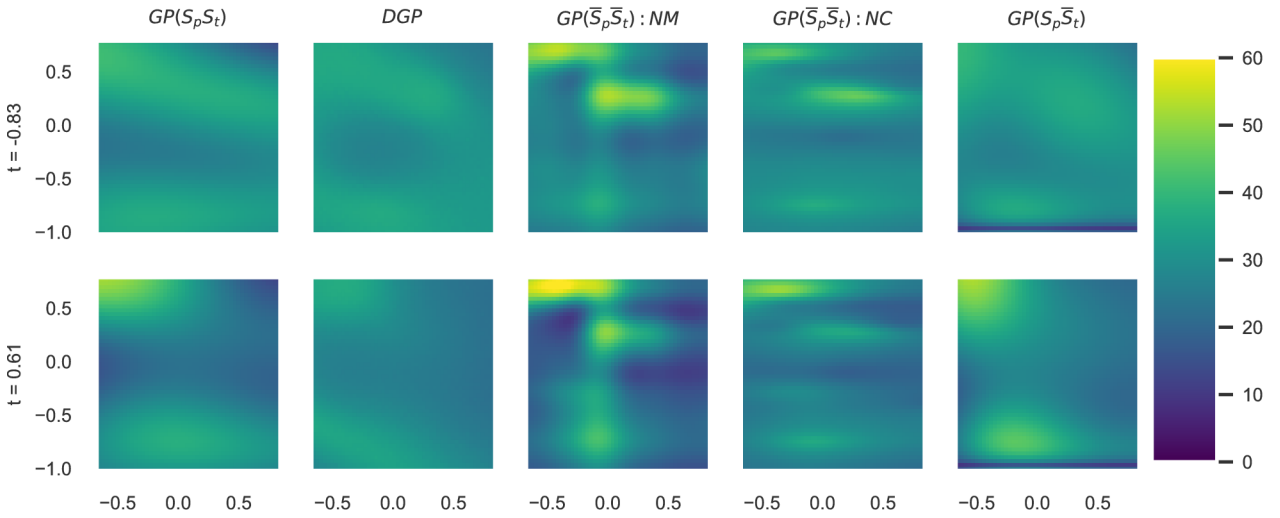

Figure 5. We plot the predictive mean of GP models with multiple covariance functions estimating NO_2 across London for two time slices. Due to the complex urban environment, e.g road layout, the covariance of NO_2 varies across input locations. Due to this we see that the separable kernels over smooth, whereas both SPM convolutions ($GP(\bar{S}_t\bar{S}_p)$:NC and $GP(\bar{S}_t\bar{S}_p)$:NM) recover more structure. Treed GP results in Appendix

We plot the results of this experiment in Fig. 6. As expected with only 10 observations we find that all models achieve a high MSE but as the number of observations increase we find that the nonseparable kernels converge to the lowest MSE the quickest. Across all number of observations the SPM kernels achieves the lowest MSEs because they are able to learn the correlations between inputs.

6.2. Spatio-temporal Heat equation

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

$$f(\mathbf{s},t) = 0.1 \cdot [50 - (x) \cdot \sin(\pi \cdot (x)/3)] \cdot \exp(-5t)$$
$$y(\mathbf{s},t) = f(g(\mathbf{s}),t) + \epsilon \tag{25}$$

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 three. 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 accurately forecast .

6.3. London Air Quality data

We model NO_2 across London using observations from 34 sensors from the London air quality network (LAQN) that

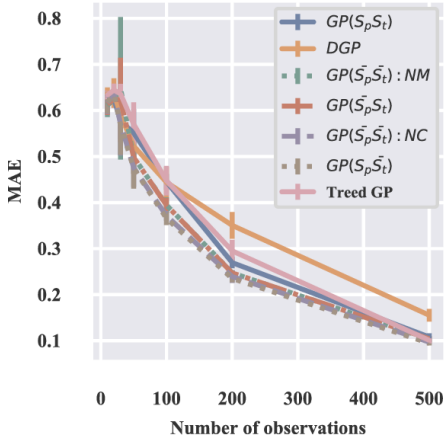

Figure 6. We plot the MAE while varying the number of observations in the nonseparable compound function experiment. With low number of observations the models that capture the complex covariance structure achieve lower MAE values. With high number of observations most models achieve similar values.

.

records data every hour. The levels of NO_2 are impacted by global factors, such as 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 an SQE covariance function. We use the construction of Example. 3 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 $v_t = \lambda_0 + \lambda_4$.

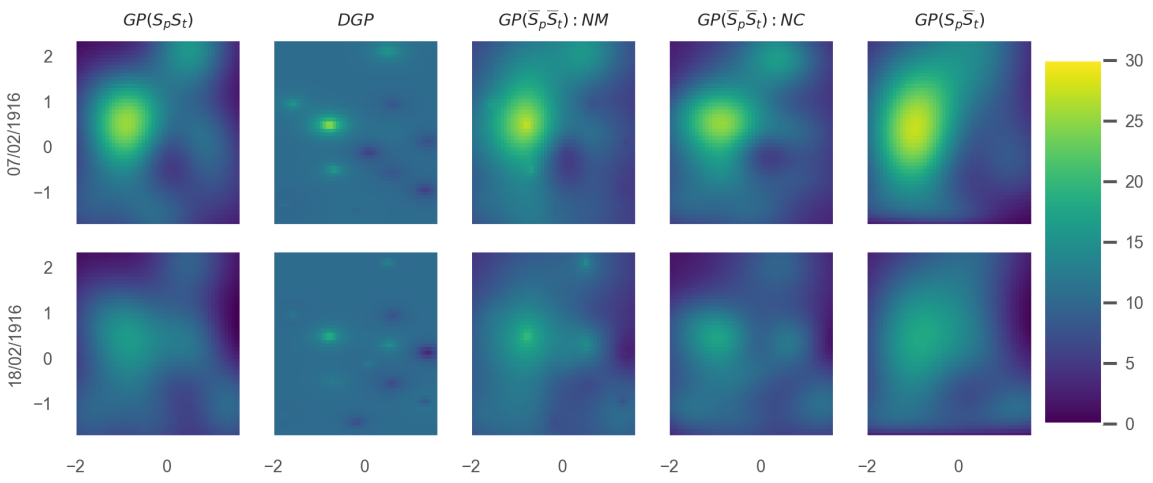

Figure 7. Prediction surfaces on Irish Wind data for two time points. The SQE kernel over smooths and therefore fails to capture the local structure. Whereas both SPM constructions are able to capture this structure. We can see that for both SPM constructions, the reconstructed surfaces are similar and this is because the 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 with nonstationary convolution ($GP(\bar{S}_t\bar{S}_p)$:NC) and SPM with nonstationary mixing ($GP(\bar{S}_t\bar{S}_p)$: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. MSE across datasets and competing approaches

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

Thus, we can use the nonseparable construction in Eq. 13 and the exponential convolution kernel (Eq. 11). For the nonstationary convolution, we assume all parameters in the convolution are a linear function of the input space.

The results of this experiment are shown in in Fig 5. The SQE kernel cannot learn the correlation structure between spatio-temporal location and so it over smooths resulting in a large noise variance. The SPM:NC kernel assumes that the structure for the location parameters are fixed and hence learns the same structure across time slices. However the SPM:NM kernel learns the varying nonseparability successfully and so 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 because the nonstationarity smoothly changes due to the convolution.

6.4. Irish Wind

The Irish wind data consists of average daily wind speeds across 12 different locations in Ireland and is well known that it exhibits nonseparability. After standardizing the data, we run a separability test (De Iaco & Posa, 2013) that results in a score of 0.38, whilst the separability ratio over two individual stations is also around 0.38. This implies that the nonseperability is approximately constant across the sensors. This is reinforced from our experiments (Fig. 7) where we see that SPM:NM and SPM:NC perform similarly and the learnt covariances exhibit 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 in using GPs with these SPM kernels. Because 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 convolutions provide superior predictions even 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 EPSRC grant EP/T004134/1 and 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 non-separability for space–time covariance models. *Journal of Statistical Planning and Inference*, 143(2):378–391, 2013.

Duffy, D. G. *Green's functions with applications*. CRC Press, 2015.

Fonseca, T. C. *On flexible modelling of spatiotemporal processes*. PhD thesis, University of Warwick,Department of Statistics, 3 2010.

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.

Ma, C. Families of spatio-temporal stationary covariance models. *Journal of statistical planning and inference*, 116(2):489–501, 2003.

Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18(40):1–6, apr 2017. URL http://jmlr.org/papers/v18/16-537.html.

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.