The Minecraft Kernel: Modelling correlated Gaussian Processes in the Fourier domain

Fergus Simpson Secondmind
Alexis Boukouvalas Secondmind
Vaclav Cadek Secondmind
Elvijs Sarkans Secondmind
Nicolas Durrande Secondmind

Abstract

In the univariate setting, using the kernel spectral representation is an appealing approach for generating stationary covariance functions. However, performing the same task for multiple-output Gaussian processes is substantially more challenging. We demonstrate that current approaches to modelling cross-covariances with a spectral mixture kernel possess a critical blind spot. For a given pair of processes, the cross-covariance is not reproducible across the full range of permitted correlations, aside from the special case where their spectral densities are of identical shape. We present a solution to this issue by replacing the conventional Gaussian components of a spectral mixture with block components of finite bandwidth (i.e. rectangular step functions). The proposed family of kernel represents the first multi-output generalisation of the spectral mixture kernel that can approximate any stationary multi-output kernel to arbitrary precision.

1 INTRODUCTION

Gaussian Processes (GPs) provide a principled and powerful framework for building statistical models (Rasmussen and Williams, 2006). Given some input/output tuples $\{x_i, y_i\}_{i=1}^n$ where $x_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$, a GP model typically consists of two elements: a latent GP f that maps input points $x \in \mathcal{X}$ into a latent space, and an observation model (sometimes referred to as the likelihood) that describes the link between $f(x_i)$ and the observation y_i. The variety of possible choices for the distribution of f and for the observation model make GP models extremely versatile. For example, the choice of the observation model can account for regression problems with outliers by using heavy tail likelihoods, but it can also be leveraged to tackle different classes of problems such as classification (Nickisch and Rasmussen, 2008) or point process modelling (see John and Hensman, 2018, and references therein). On the other hand, a large number of mean and covariance functions are available off-the-shelf and they can be used to encode various prior beliefs or assumptions about the system that generated the data, such as the order of differentiability, (non)-stationarity or (quasi)-periodicity. Using the right covariance function comes with massive benefits in term of model accuracy and uncertainty quantification (see Rasmussen and Williams, 2006, chapter 5).

Three main approaches can be distinguished when it comes to finding a good kernel for the problem at hand. The first one consists of gathering expert knowledge on the phenomenon that generated the data and to define bespoke covariances to encode it (López-Lopera et al., 2019), the second is to search automatically through the space of possible kernel combinations (Bach, 2009; Duvenaud, 2014) and the third consists of choosing a kernel parametrised by a large number of variables such that it is flexible enough to be a good match for a broad range of datasets. This third approach has recently yielded state of the art accuracy on various benchmarks (Wilson and Adams, 2013; Sun et al., 2018) and will be the focus of the present work.

The spectral mixture (SM) kernel (Wilson and Adams, 2013; Remes et al., 2017) is a canonical example of these highly flexible and heavily parametrised covariance functions. It represents a kernel's spectral density as a mixture of Gaussians, which ensures that the kernel itself can be expressed in closed form. One striking advantage of this approach is that any stationary kernel can be well approximated by the model, provided that the mixture of Gaussians is comprised of a sufficient number of components. An alternative spectral representation based upon a mixture of rectangular blocks of constant spectral density has recently been

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

investigated by Tobar (2019). This latter approach and its multi-output generalisation will prove to be of particular interest for the current work.

In many applications, it is desirable to use GPs with multivariate outputs. For example if one wants to predict time series corresponding to the temperatures of various neighbouring cities, it is valuable to build a joint model that can account for the correlations between the observations (Parra and Tobar, 2017). This complicates the choice of the covariance function because one has to select not only a covariance function for each time series, but one must also specify all cross-covariances between the different time-series. For a detailed review of the classic approaches for defining multi-output GPs (Linear Model of Coregionalisation, convolution GPs, etc.), see Alvarez et al. (2011). Generalisations of the SM kernel to a multi-output framework have been proposed by Ulrich et al. (2015) and Parra and Tobar (2017). Their approaches consist of modelling the cross-spectral densities between pairs of processes. As we shall demonstrate in this article, such an approach strips the SM kernel of its key ability: to approximate any stationary covariance to arbitrary precision.

The central contribution of this work is to resolve this limitation. We begin by highlighting the root cause of the limitation: is arises as a result of the inevitable overlap between the spectral densities of Gaussian components. This pathology can be eliminated by utilising component kernels which possess a finite bandwidth, and can therefore be arranged in a non-overlapping manner. We prove that our generalisation of the sinc kernel (Yao, 1967; Tobar, 2019) to the multi-output case is the first kernel in the spectral mixture class that is dense in the space of multi-output stationary covariances. The practical implication is that the correlations available to the model now spans the full theoretical range. This article focuses on a clear exposition of current limitations and how they can be overcome by combining recent work from the literature. Finally we show in the experiments that the inductive bias of the proposed kernels leads to improved performances compared to the conventional form of multi-output spectral mixture kernels.

In § 2 we review the theoretical background behind univariate and multivariate spectral mixture kernels and highlight the limitations of using a mixture of Gaussians to model cross-spectra. A solution based upon components of finite bandwidth is presented in § 3, and we show that the proposed method can accurately approximate any stationary multi-output covariance. Experiments are presented in § 4 where we illustrate the use of the proposed approach for modelling non-stationary time series and for producing colour images. Finally, concluding remarks are presented in § 5.

Our implementation of the proposed method is based on the GPflow framework (De G. Matthews et al., 2017), and the code for generating all of the figures in the paper is available as supplementary material.

2 SPECTRAL KERNELS

Covariance functions of real-valued processes are functions $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that are symmetric $K(x, x') = K(x', x)$ and positive definite: $\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) \geq 0$, $\forall n \in \mathbb{N}, \forall x_i \in \mathcal{X}, \forall a_i \in \mathbb{R}$. Checking that a function is positive definite may seem a daunting prospect but for stationary covariances (kernels that can be represented as univarite functions $K(x, x') = K(x - x')$ using the classic notation overloading) a simpler condition is given by Bochner's Theorem (Wendland, 2004):

Theorem 1 (Bochner's theorem)**.** *An integrable function $K(\cdot)$ is the covariance function of a weakly-stationary real-valued stochastic process if and only if it admits the representation*

$$K(\boldsymbol{r}) = \int_{\mathbb{R}} e^{i\boldsymbol{\nu}^\top \boldsymbol{r}} S(\boldsymbol{\nu}) d\boldsymbol{\nu} \tag{1}$$

where $S: \mathbb{R}^N \to \mathbb{R}$ is integrable, symmetric $S(\boldsymbol{\nu}) = S(-\boldsymbol{\nu})$, and positive $S(\boldsymbol{\nu}) \geq 0$.

The function S is usually referred to as the *power spectrum* or *spectral density*. Bochner's theorem makes it an appealing basis for defining new covariances: the full family of stationary kernels becomes readily available by exploring positive spectral densities, without having to check any positive definiteness condition. This is the approach followed by Wilson and Adams (2013) for defining the spectral mixture kernel.

2.1 Univariate Spectral Mixture kernels

When modelling the spectral density of the kernel, it is advantageous to choose a parameterisation such that its Fourier transform, i.e. the kernel itself, is available in closed form. This increases both the speed and precision with which the marginal likelihood can be evaluated. Wilson and Adams (2013) proposed to use spectral densities given by the sum of Q Gaussian pairs:

$$S(\nu) = \sum_{i=1}^{Q} \frac{A_i}{2} \left[G(\nu, \mu_i, \sigma_i) + G(-\nu, \mu_i, \sigma_i) \right],$$

$$G(\nu, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(\nu - \mu)^2}{2\sigma^2} \Big),$$

where the kernel parameters $\theta = \{A_i, \mu_i, \sigma_i\}_{i=1}^{Q}$ satisfy $A_i \geq 0$ and $\sigma_i > 0$ for all i. This formalism is readily expanded to operate in higher dimensions by replacing the univariate normal distributions by their multivariate counterparts $G(\boldsymbol{\nu}, \boldsymbol{\mu}_i, \Sigma_i) = \mathcal{N}(\boldsymbol{\nu} | \boldsymbol{\mu}_i, \Sigma_i)$ where the

covariance matrix is diagonal. The corresponding kernel, which we will refer to as the *Gaussian-SM*, is

$$K(\boldsymbol{r}) = \sum_{i=1}^{Q} A_i \exp(-2\pi^2 \boldsymbol{r}^\top \Sigma_i \boldsymbol{r}) \cos(2\pi \boldsymbol{r}^\top \boldsymbol{\mu_i}) \,. \quad (2)$$

Gaussian-SM are very flexible and as outlined in Wilson and Adams (2013) one of their key property is that they can approximate (in the L^1 sense) any stationary covariance.

Tobar (2019) proposed to replace the Gaussian components in the spectral mixture kernel by block components, motivated by the potential applications relating to signal processing. As with the Gaussian case, it is straightforward to generalise this 'block-SM' kernel to input spaces of dimension D by taking the product of D one-dimensional blocks:

$$\begin{aligned} S(\boldsymbol{\nu}) &= \sum_{i=1}^{Q} \frac{A_i}{2} [B_{\boldsymbol{\mu}_i, \boldsymbol{w}_i}(\boldsymbol{\nu}) + B_{-\boldsymbol{\mu}_i, \boldsymbol{w}_i}(\boldsymbol{\nu})] \\ B_{\boldsymbol{\mu}, \boldsymbol{w}}(\boldsymbol{\nu}) &= \begin{cases} \prod_k \frac{1}{w_k} & \text{if all } |\nu_k - \mu_k| < \frac{1}{2} w_k \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (3)$$

which results in

$$K(\boldsymbol{r}) = \sum_{i=1}^{Q} A_i \cos(2\pi \boldsymbol{r}^\top \boldsymbol{\mu}_i) \prod_{d=1}^{D} \operatorname{sinc}(r_d \boldsymbol{w}_{id}) \,, \quad (4)$$

where $\operatorname{sinc}(x) \equiv \dfrac{\sin(\pi x)}{\pi x}$.

2.2 Multi-Output Spectral Kernels

Modelling N outputs (or channels) demands the construction of a model capable of describing not only N spectral densities, but also $\mathcal{O}(N^2)$ distinct cross-spectra. The first multi-output version of SM kernels consisted of using the Linear Model of Coregionalisation (LMC) approach, where each channel is defined as a linear combination of R independent processes with SM kernels (Wilson, 2014). This has been generalised by Ulrich et al. (2015) who included phase shifts between the various channels. Finally, Parra and Tobar (2017) further refined the previous proposals by using the generalisation of Bochner's theorem for multi-output processes:

Theorem 2 (Cramér's Theorem)**.** *A family $\{K_{ij}\}_{i,j=1}^{N}$ of integrable functions is the kernel of a weakly-stationary multivariate stochastic process if and only if they admit the representation*

$$K_{ij}(\boldsymbol{r}) = \int_{\mathbb{R}} e^{\mathrm{i}\boldsymbol{\nu}^\top \boldsymbol{r}} S_{ij}(\boldsymbol{\nu}) d\boldsymbol{\nu} \qquad \forall i, j \in \{1, \ \ldots, \ N\} \quad (5)$$

where $S_{ij} : \mathbb{R}^N \to \mathbb{C}$ are integrable functions that fulfil the positive definiteness condition pointwise

$$\sum_{i,j=1}^{N} \bar{z}_i z_j S_{ij}(\boldsymbol{\nu}) \geq 0 \qquad \forall z_1, \ \ldots, \ z_N \in \mathbb{C}, \forall \boldsymbol{\nu} \in \mathbb{R}^N. \quad (6)$$

The kernel obtained by Parra and Tobar (2017) is more general than the previous proposals since it can account for both phase shifts and delays between channels. In all three cases, the formalism ensures that each channel takes the form of a stationary GP with a conventional Gaussian-SM kernel.

Despite these increasing levels of refinement in the modelling of multi-output kernels in the spectral domain, the following section will show that none of the aforementioned methods can represent the full range of cross-correlations which a given pair of channels is capable of exhibiting.

2.3 The case of missing cross-covariance

Arguably the most valuable characteristic of the Spectral Mixture kernel lies in its ability to mimic any stationary kernel. We will now show that this property has been lost in the previous attempts to generalise the SM formalism to multivariate processes. To highlight the crux of the problem at hand, it is instructive to walk through a minimal worked example. We shall explore the case of a GP where the two channels, say a and b, have a Gaussian-SM kernel with two components: $S_a = C_{a_1} + C_{a_2}$ and $S_b = C_{b_1} + C_{b_2}$, with the notations $C_{a_1} = \frac{1}{2} A_1 (G^+_{a_1} + G^-_{a_1})$ and $G^{\pm}_{a_1} = G(\cdot, \pm\mu_{a_1}, \sigma_{a_1})$. There is no time delay or phase shift to be concerned about in this example, so the three approaches reviewed in § 2.2 are now equivalent. The resulting cross-spectrum, as given in Parra and Tobar (2017), is

$$S_{ab} = \sqrt{G^+_{a_1} G^+_{b_1}} + \sqrt{G^-_{a_1} G^-_{b_1}} + \sqrt{G^+_{a_2} G^+_{b_2}} + \sqrt{G^-_{a_2} G^-_{b_2}} \,. \quad (7)$$

Plugging these expressions into the definition of the correlation coefficient $\rho(\nu)$ which measures if the Fourier modes of a and b tend to be in phase with each other yields

$$\begin{aligned} \rho(\nu) &= \frac{S_{ab}(\nu)}{\sqrt{S_a(\nu) S_b(\nu)}} \\ &= \frac{\sqrt{G^+_{a_1} G^+_{b_1}} + \sqrt{G^-_{a_1} G^-_{b_1}} + \sqrt{G^+_{a_2} G^+_{b_2}} + \sqrt{G^-_{a_2} G^-_{b_2}}}{\sqrt{(C_{a_1} + C_{a_2})(C_{b_1} + C_{b_2})}} \,. \end{aligned} \quad (8)$$

At any given frequency ν, the Cauchy-Schwarz inequality tells us that this coefficient is 1 if the modes of the two processes are maximally correlated. However, given the specific shape that is assumed for S_{ab}

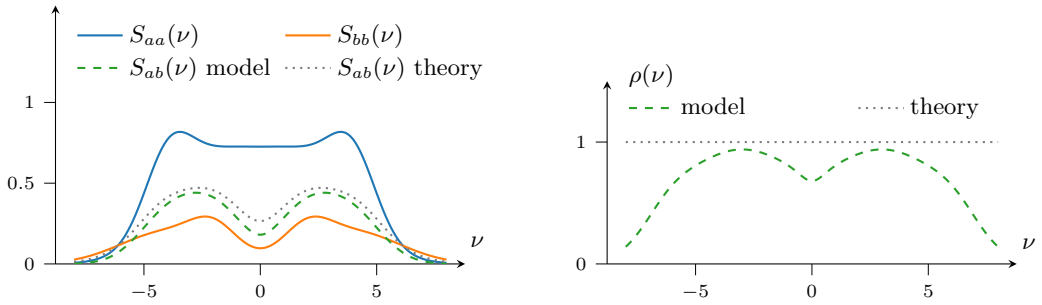

Figure 1: An illustration of the limitation of Gaussian-SM multi-output covariances. Left: The kernel's maximal cross spectrum (dashed line) falls short of the theoretical limit (dotted line) across the whole frequency spectrum. Right: This shortcoming means that possible correlations between the spectral modes are truncated to a reduced range, only permitted to reach the dashed line, and far below the theoretical limit.

in Eq. 7 this can only arise if the relative contributions from the two components are identical, such that $C_{a_1}/C_{a_2} = C_{b_1}/C_{b_2}$. For this condition to hold across all frequencies suggests that the two spectra must be of exactly the same shape. Aside from that special case of matching spectra, the Cauchy-Schwarz bound cannot be saturated by the model, and so we find that some loss in cross-covariance is inevitable. Naturally, the same limitation also arises for the less commonly encountered case where the pair of processes are anti-correlated.

The cross-spectrum between these two processes is shown in the left hand panel of Figure 1. Note the gap between the maximum attainable value of the model cross-spectrum and the theoretical maximal cross-spectrum. The missing power (or equivalently, the lost correlation) in the right panel is due to the summation of the Gaussian components where they are treated *as if they were independent*, when in reality they need not be. There is inevitably some degree of overlap between the tails of any two Gaussian components, and yet correlations which exist between these tails cannot be fully captured by the conventional model. In the event that the two processes are highly correlated, this can prove to be an extremely poor approximation.

The cross-spectral density S_{AB} between two processes A and B at a given frequency ν can be interpreted as the expectation of the product of their Fourier modes $\langle f_A(\nu), f_B^*(\nu)\rangle$. The Cauchy-Schwarz inequality tells us that $|\langle u, v\rangle|^2 \leq \langle u, u\rangle\langle v, v\rangle$. The theoretical range of a cross spectral density is therefore given by $S_{AB}^2 \leq S_{AA}S_{BB}$, and the dotted line in the figure represents the regime where this inequality is saturated. This corresponds to the Fourier components being perfectly in phase.

As the overlap between a given pair of components grows, so too does the potential loss in cross-covariance experienced by a Gaussian-SM kernel. This highlights the danger of using broadband components, those which span large frequencies due to a high bandwidth σ, in a multi-output setting.

Note that introducing more than two components does not resolve the problem - it exacerbates it. This is because a process described by N spectral components generates $\mathcal{O}(N^2)$ pairs of overlapping Gaussians, yielding greater opportunities for the cross-covariance to be lost.

3 A UNIVERSAL MODEL FOR CROSS SPECTRAL DENSITIES

In this section we shall specify a GP kernel which permits a full exploration of the space of possible cross-covariances, while ensuring that the number of model parameters grows linearly with the number of components. This constitutes the central contribution of this work.

3.1 The Minecraft kernel

Since the limitation of the multi-output Gaussian-SM comes from the overlap between the tails of the components, the issue can be resolved by selecting components with disjoint support. By replacing the pairs of Gaussians with pairs of 'blocks' (i.e. rectangular functions, as defined in § 2.1) we can reap two major benefits. First of all, it will allow a complete description of the cross-covariance for highly correlated processes, resolving the issue highlighted in the previous section, thereby allowing any multi-output stationary kernel to be well approximated. Secondly, it becomes easier to evaluate the coefficients which determine the cross-covariances, using a technique which is outlined in the following subsection. This motivates the introduction of the *Minecraft kernel* whose spectral representation is a sum over Q symmetrised block components as defined

in (3):

$$S_{ij}(\boldsymbol{\nu}) = \sum_{q=1}^{Q} \frac{1}{2} A_{ij}^q (B_{\boldsymbol{\mu}^q, \boldsymbol{w}^q}(\nu) + B_{-\boldsymbol{\mu}^q, \boldsymbol{w}^q}(\nu)) \quad (9)$$

$$K_{ij}(\boldsymbol{r}) = \sum_{q=1}^{Q} A_{ij}^q \cos(\boldsymbol{r}^\top \boldsymbol{\mu}^q) \prod_{d=1}^{D} \operatorname{sinc}(r_d w_d^q) \,. \quad (10)$$

Provided that the Q amplitude matrices $\{A_{ij}^q\}_{i,j=1}^N$ are positive definite, Cramér's theorem guarantees that the Minecraft kernel is a valid covariance function. Note that the Minecraft Kernel can either be interpreted as the multi-output generalisation of the block-SM kernel (Tobar, 2019), or as a Riemann approximation of the integral in Cramér's theorem if the S_{ij} are seen as a constant per block functions. This remark suggests that the Minecraft kernel can recover the universal approximation property that had been lost in previous multi-output SM kernels. This is guaranteed by the following result:

Theorem 3. *Minecraft kernels are dense in the space of multi-output stationary real-valued covariance functions for the L^1 norm.*

Proof. Let $K = \{K_{ij}\}_{i,j=1}^N$ be the covariance of weakly-stationary real-valued multivariate stochastic process, and let $\{S_{ij}\}_{i,j=1}^N$ be the associated spectral density given by Cramér's theorem. Finally, let $R_{\boldsymbol{\mu},\boldsymbol{w}}$ denote the hyper-rectangle in $\mathbb{R}^D$ centred on $\boldsymbol{\mu}$ with width w_i along the ith axis. Since simple functions are dense in L^1 (Bogachev, 2007), and since $S_{ij}(\boldsymbol{\nu}) = S_{ij}(-\boldsymbol{\nu})$, we can find a sequence of sets of non-overlapping rectangles $(\{R_{\boldsymbol{\mu}_{kl},\boldsymbol{w}_{kl}}\}_{l=1}^{L_k} \cup \{R_{-\boldsymbol{\mu}_{kl},\boldsymbol{w}_{kl}}\}_{l=1}^{L_k})_{k\geq 0}$ such that for all $i, j \in \{1, \ldots, N\}$

$$\Bigg(\sum_{l=1}^{L_k} \int_{R_{\boldsymbol{\mu}_{kl},\boldsymbol{w}_{kl}}} S_{ij}(\boldsymbol{\nu}) d\boldsymbol{\nu} \left(B(\cdot, \boldsymbol{\mu}_{kl}, \boldsymbol{w}_{kl}) + B(\cdot, -\boldsymbol{\mu}_{kl}, \boldsymbol{w}_{kl}) \right) \Bigg)_{k\geq 0} \quad (11)$$

is a sequence of constant per block functions that converges to S_{ij} in L^1. Let S_{ij}^k denote the elements of these sequences, we will now prove that they satisfy the two conditions of Cramér's theorem. First, these functions are clearly integrable. Second, for some given k and $\boldsymbol{\nu}$, we have either $\boldsymbol{\nu} \notin \bigcup_l R_{\pm\boldsymbol{\mu}_{kl},\boldsymbol{w}_{kl}}$ which implies $S_{ij}^k(\boldsymbol{\nu}) = 0$ for all i, j, or there exists a unique l such that $\boldsymbol{\nu} \in R_{\pm\boldsymbol{\mu}_{kl},\boldsymbol{w}_{kl}}$ and $S_{ij}^k(\boldsymbol{\nu}) = \frac{1}{|R_{\boldsymbol{\mu}_{kl},\boldsymbol{w}_{kl}}|} \int_{R_{\boldsymbol{\mu}_{kl},\boldsymbol{w}_{kl}}} S_{ij}(\boldsymbol{\nu}) d\boldsymbol{\nu}$ where $|\cdot|$ is the area of the rectangle. In both cases, $(S_{ij}^k(\boldsymbol{\nu}))_{i,j=1}^N$ is positive definite: either because a matrix of zero is trivially positive definite, or because the set of positive definite matrices is a convex cone and $S_{ij}^k(\boldsymbol{\nu})$, is the rescaled integral of positive definite matrices. Cramér's theorem thus applies and tells us that the Fourier transform of S_{ij}^k is the covariance function of a weakly-stationary multivariate process which by definition is a Minecraft kernel. By continuity of the Fourier transform, this sequence of Minecraft kernels converges to K_{ij} for the L^1 norm. □

Since the Minecraft kernel can be interpreted as a Linear Model of Coregionalisation, one corollary of Theorem 3 is that Linear Models of Coregionalisation can approximate any stationary multi-output covariance to arbitrary precision.

The ability of the Minecraft kernel to model highly correlated channels is illustrated in Figure 2. Contrary to the case where Gaussian components are used, components can be arranged in a non-overlapping manner, This configuration enables the cross-covariances to now reach the theoretical limit. This can be seen explicitly in the right hand panel, as indicated by a correlation coefficient which saturates at unity across the full range of frequencies.

If we wish to generalise the model to incorporate a possible delay between the different processes, we can adopt the prescription advocated by Parra and Tobar (2017) to yield

$$\begin{aligned} K_{ij}(\boldsymbol{r}) &= \sum_{q=1}^{Q} A_{ij}^q C_{ij}^q S_{ij}^q \,, \\ C_{ij}^q &= \cos\left[(\boldsymbol{r} + \theta_{ij}^q)^\top \boldsymbol{\mu}^q + \phi_{ij}^q)\right] \,, \\ S_{ij}^q &= \prod_{d=1}^{D} \operatorname{sinc}\left[(r_d + \theta_{ij}^q) w_d\right] \,. \end{aligned} \quad (12)$$

3.2 Block shapes

In principle, the blocks we use to model the spectral density need not be rectangular. Indeed we note that, when working in higher input dimensions, it may prove beneficial for each spectral component to take the form of a pair of ellipsoids. This would allow the potentially long product of sinc functions in equation (10), one for each input dimension, to be replaced by a single Bessel function:

$$K_e(\boldsymbol{r}) = \sum_{q=1}^{Q} A_{ij}^q ||\boldsymbol{r} \odot \boldsymbol{w}||^{n/2} \cos(2\pi \boldsymbol{r}^\top \boldsymbol{\mu}) J_{n/2}(||\boldsymbol{r} \odot \boldsymbol{w}||) \,, \quad (13)$$

where $||\cdot||$ is the Euclidean norm and $\odot$ denotes an element-wise product. This kernel retains the desired property of the minecraft kernel, that the components are of finite bandwidth.

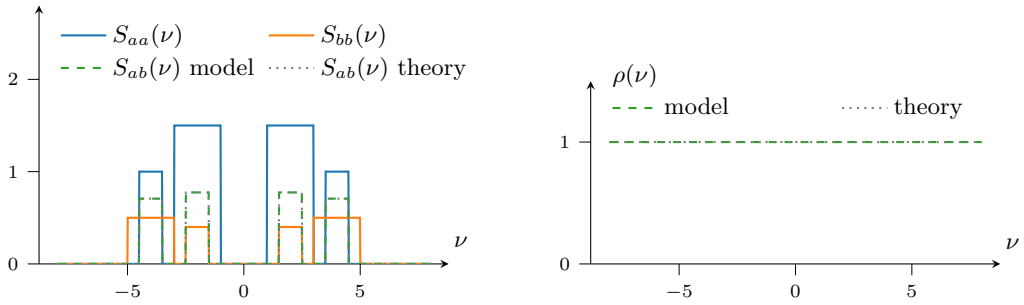

Figure 2: The introduction of spectral blocks resolves the issue highlighted in Figure 1. The model cross-spectrum can now be constructed from non-overlapping components, allowing the cross-spectrum to replicate the target. All Fourier modes can now be fully correlated, so there is no loss in cross-covariance.

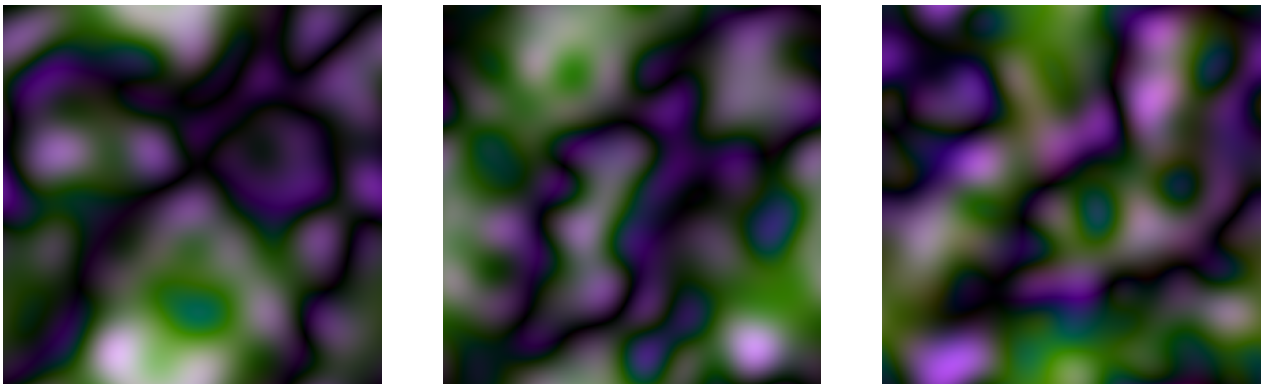

Figure 3: A comparison of images where the RGB channels are determined by drawing samples from a multi-output Gaussian process. Left: A sample from the target kernel we wish to model, comprising three Matern kernels, which possess correlations given by Table (1). Middle: The GP has a Minecraft kernel which can ensure a high correlation between the three channels. Right: This GP has a Gaussian-SM covariance, and the limited cross-correlation range results in a substantial loss of coherence between the channels.

3.3 Parameterisation

Two remarks can be made regarding the parameterisation of the Minecraft kernel. As with the multi-output Gaussian-SM, the parameters $\boldsymbol{w}^q$ and $\boldsymbol{\mu}^q$ do not depend on i, j (which means that all spectral densities share the same blocks as basis functions), and that we want the blocks' support to be non-overlapping.

These choices to not imply any loss of generality (they may require using a larger number of blocks) but they confer two benefits. The first one is to guarantee that no cross-covariance is lost and that the maximum correlation between the signal can be reached. The second is to ensure that each component cannot interact with more than one block per channel, which implies that the number of free parameters in the model grows only as $\mathcal{O}(Q)$.

A final remark on parametrisation is each A_{ij}^q matrix contains $\mathcal{O}(D^2)$ model parameters that typically need to be optimised, but it must at the same time satisfy the positive definiteness constraint. A convenient method is to reparametrise them by their Cholesky factors before exposing the latter to the optimiser. If this $\mathcal{O}(D^2)$ scaling in the number of parameters cannot be afforded, a low rank decomposition of A_{ij}^q may be used instead.

4 EXPERIMENTS

In this section we illustrate the advantages of the proposed approach on two case studies that require the cross-covariances to be accurately accounted for. The first one involves modelling color channels in an image, and the second revisits the use of change points for modelling non-stationary time series.

4.1 Rendering of images

Colour images are a form of multi-channel data in which there tends to be a significant degree of correlation between the three RGB channels. While some distinct information is invariably carried within each channel, a simple modulation in brightness across an image will

be shared across all three of them.

In this section, we aim to generate images which possess a significant degree of correlation between the three channels, so as to generate coherent fluctuations in brightness. As a baseline we select a different covariance for each R, G and B channel corresponding to isotropic 2D Matérn kernels, with regularity 1/2, 3/2, and 5/2 respectively.

Table 1: Correlations between the RGB channels presented in Figure 3, and the relative errors compared to the target correlations. The values obtained with Minecraft are more than an order of magnitude closer to the target correlations than those from the conventional Gaussian model.

Channels		R, G	R, B	G, B
Correl.	Target	0.897	0.960	0.744
	Gaussian	0.849	0.908	0.566
	Minecraft	0.901	0.962	0.754
Error	Target	-	-	-
	Gaussian	-5.4%	-5.4%	-23.9%
	Minecraft	+0.4%	+0.2%	+1.2%

Given this fiducial model, we shall attempt to approximate the covariance using both variants of the multi-output SM kernels: the conventional Gaussian components and the Minecraft kernel. In each case, we tile the area $[-2, 2]^2$ of the spectral domain with 64 basis functions, which means that we use 32 components. The amplitude of each component is obtained by minimising the L^1 distance between the original Matérn spectrum and the approximations.

The images shown in Figure 3 are obtained by sampling jointly from the channels of the three GPs. In the left hand panel we see a sample drawn from the correlated Matérn kernels, in the centre panel is a sample drawn from a Minecraft kernel, and in the right hand panel is a sample is drawn from a conventional Gaussian-SM kernel. The brightness of a given channel - red, green, or blue - is determined by the absolute magnitude of the sample at the given pixel location. Black contours therefore correspond to the regime where the sample crosses zero. Changes in colour correspond to uncorrelated, decoherent fluctuations across the three channels. Meanwhile changes in brightness correspond to correlated, coherent fluctuations across the three channels. More correlated processes will therefore tend to produce fewer changes in colour but more pronounced changes in brightness.

In order to gain a more quantitative measure of performance, beyond this visual illustration, we provide a numerical comparison of the correlations between the channels for the three different kernels. Since the kernels are all stationary, the cross-correlations do not depend on the input location. We can therefore pick an arbitrary input point and compute the correlations between the three channels. As shown in Table 1, the Minecraft formalism offers a much more accurate account of the cross-channel dependencies.

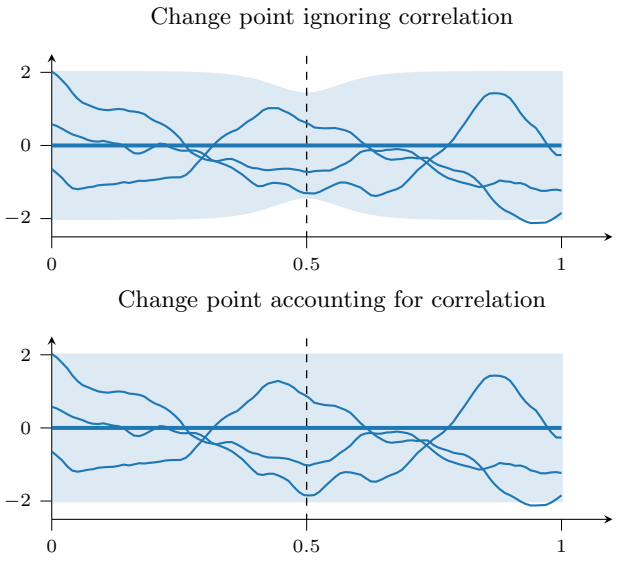

Figure 4: Samples drawn from a Gaussian process with a change point located at 0.5 (vertical dashed line). Top: If f_1 and f_2 are taken to be independent, some variance is lost (the shaded confidence intervals can be seen to contract around the change point location. Bottom: There is no loss of variance when the cross-correlation between f_1 and f_2 is accounted for.

We also note that the sign of the error is consistently different in the two cases, and this is due to their fundamentally different origins. The Gaussian model under-represents the magnitude of all the correlations, due to its inability to reproduce the full range of possible cross-covariance, as illustrated in Figure 1. In the case of the Minecraft model, it does not possess this limitation, so the leading source of modelling error stems from the accuracy of reproducing the shape of the target spectral density. This is limited by the finite number of components chosen ($Q = 32$ in this case). Within the bandwidth of a single component, the spectra of two outputs in the minecraft kernel share the same functional form, and this leads to a slight overestimation of the cross-correlation compared with the target case.

4.2 Change points in non-stationary time series

A useful technique for generating a non-stationary process is via a linear combination of stationary processes. For example, given two Gaussian processes of the form $f_1 \sim \mathcal{GP}(0, k_1)$, $f_2 \sim \mathcal{GP}(0, k_2)$ and a sigmoid function s, one can define a GP that smoothly transitions from f_1 to f_2:

$$f(x) = s(x)f_1(x) + (1 - s(x))f_2(x)\,. \qquad (14)$$

This approach is fairly common in the GP community, it is implemented within the GPflow package, and it serves as one of the building blocks within the automatic statistician (Duvenaud, 2014; Lloyd et al., 2014). However, it is typically assumed that the processes f_1 and f_2 are independent, which results in some of the variance of f vanishing at the transition point. This can be illustrated with a simple experiment where the two kernels we wish to connect are identical, such that $k_1 = k_2$. Since f_1 and f_2 have the same distribution, one could expect that the global distribution remains unchanged when applying a change point. As seen in the upper panel of Figure 4 this is however not the case, and a significant proportion of the variance is lost around the transition point.

This unwanted behaviour can be addressed by relaxing the assumption that f_1 and f_2 are independent. On the example detailed above, choosing $f_1 = f_2$ results in the expected behaviour where the distribution of f is exactly the same as the distribution of f_1 and f_2 (see the lower panel of Figure 4). This pedagogical example highlights the importance of adequately modelling cross-covariances.

Practitioners may be concerned whether blocks are as efficient as Gaussians at replicating real-world spectral densities. To assess their capabilities, we study the thirteen benchmark time series used in Lloyd et al. (2014). For each of these time series, we fit models with one change point (as defined in equation 14) and compare the accuracy of using Gaussian components versus Minecraft's block components. As discussed earlier, the Gaussian-SM cannot fully account for the correlation between the components whereas the block-SM can. In both models, we make use of $Q = 10$ components. Hyperparameters are initialised by selecting the highest marginal likelihood from 1,000 random starting points, before being optimised via SciPy's implementation of the conjugate gradients algorithm for 2,000 iterations. Finally, the accuracy of the models are evaluated by computing for each time series the Standard Mean Square Error (SMSE). Following Lloyd et al. (2014), we adopt a 90/10 train/test split ratio.

As seen in Table 2, the blocks compare as well as, and in many cases better than, their Gaussian counterparts. Quoted uncertainties are estimated by repeating the experiments with ten different random seeds. It should however be noted that we do not necessarily recommend using change points for these datasets since in some cases better performances are achievable with stationary covariances.

Table 2: Comparing the SMSE values for benchmark time series when adopting two different choices of spectral component.

Dataset	Block	Gaussian
airline	**0.39** (±0.07)	0.55 (±0.06)
births	0.99 (±0.01)	**0.81** (±0.4)
call centre	**7.28** (±3.8).	7.83. (±4.5)
gas production	1.46 (±0.24)	**0.57** (±0.16)
internet	**1.00** (±0.04)	0.89 (±0.13)
mauna	0.52 (±0.1)	**0.45** (±0.05)
radio	**1.12** (±0.08).	1.46 (±0.09)
solar	**1.35** (±0.4)	1.50 (±0.16)
sulphuric	**5.35** (±0.13)	5.94 (±0.1)
temperature	**0.77** (±0.14)	0.93 (±0.1)
unemployment	**1.21** (±0.05)	1.22 (±0.12)
wages	2.38 (±0.09)	**2.08** (±0.12)
wheat	**1.19** (±0.7)	1.41 (±0.1)
mean	**1.92** (±0.1)	1.97 (±0.1)
best performance	**9/13**	4/13

5 CONCLUSIONS

Modelling the cross-covariances between Gaussian Processes is a challenging but important task in machine learning. We have highlighted a significant blind spot in the conventional approach of modeling cross-covariances in the spectral domain: the important regime where a pair of processes are significantly correlated (either in the positive or negative sense) cannot be adequately modelled via a mixture of Gaussians. Aside from the trivial case where the spectral densities of each process are of the same functional form, overlapping components lead to a loss of cross-covariance. In general, we advise that multi-output spectral kernels adopt different initialisation strategies to the standard single-output case - one which suppresses the generation of very broad components.

We present a new multi-output spectral kernel which offers a resolution to this problem. By utilising a basis kernel of finite bandwidth, we can avoid the loss of cross-covariance caused by overlapping components. By replacing a conventional mixture of Gaussians with a mixture of blocks of constant spectral density, this kernel opens up access to the full range of cross-covariances

associated with stationary processes. A further key advantage of this approach is that, without loss of generality, all spectral and cross spectral densities can share the same base parameterisation, which helps restrict the number of free parameters in the model. Finally, we have also presented and demonstrated a method for combining these correlated processes in order to model non-stationary time series.

Acknowledgements

We would like to thanks James Hensman for several helpful discussions, and the anonymous reviewers for their constructive feedback.

References

Carl E Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9(Oct):2035–2078, 2008.

ST John and James Hensman. Large-scale cox process inference using variational fourier features. *arXiv preprint arXiv:1804.01016*, 2018.

Andrés Felipe López-Lopera, Nicolas Durrande, and Mauricio Alexander Alvarez. Physically-inspired gaussian process models for post-transcriptional regulation in drosophila. *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

Francis Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. *arXiv preprint arXiv:0909.0844*, 2009.

David Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pages 1067–1075, 2013.

Shengyang Sun, Guodong Zhang, Chaoqi Wang, Wenyuan Zeng, Jiaman Li, and Roger Grosse. Differentiable compositional kernel learning for gaussian processes. *arXiv preprint arXiv:1806.04326*, 2018.

Sami Remes, Markus Heinonen, and Samuel Kaski. Non-stationary spectral kernels. *arXiv preprint arXiv:1705.08736*, 2017.

Felipe Tobar. Band-limited gaussian processes: The sinc kernel. In *Advances in Neural Information Processing Systems*, pages 12728–12738, 2019.

Gabriel Parra and Felipe Tobar. Spectral mixture kernels for multi-output gaussian processes. In *Advances in Neural Information Processing Systems*, pages 6681–6690, 2017.

Mauricio A Alvarez, Lorenzo Rosasco, and Neil D Lawrence. Kernels for vector-valued functions: A review. *arXiv preprint arXiv:1106.6251*, 2011.

Kyle R Ulrich, David E Carlson, Kafui Dzirasa, and Lawrence Carin. Gp kernels for cross-spectrum analysis. In *Advances in neural information processing systems*, pages 1999–2007, 2015.

Kung Yao. Applications of reproducing kernel hilbert spaces–bandlimited signal models. *Information and Control*, 11(4):429–444, 1967.

Alexander G De G. Matthews, Mark Van Der Wilk, Tom Nickson, Keisuke Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. Gpflow: A gaussian process library using tensorflow. *The Journal of Machine Learning Research*, 18(1):1299–1304, 2017.

Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Andrew Gordon Wilson. *Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

Vladimir I Bogachev. *Measure theory*, volume 1. Springer Science & Business Media, 2007.

James Robert Lloyd, David Duvenaud, Roger Grosse, Joshua Tenenbaum, and Zoubin Ghahramani. Automatic construction and natural-language description of nonparametric regression models. In *Twenty-eighth AAAI conference on artificial intelligence*, 2014.