Spectral Mixture Kernels for Multi-Output Gaussian Processes

Gabriel Parra
Department of Mathematical Engineering
Universidad de Chile
gparra@dim.uchile.cl

Felipe Tobar
Center for Mathematical Modeling
Universidad de Chile
ftobar@dim.uchile.cl

Abstract

Initially, multiple-output Gaussian processes models (MOGPs) were constructed as linear combinations of independent, latent, single-output Gaussian processes (GPs). This resulted in cross-covariance functions with limited parametric interpretation, thus conflicting with single-output GPs and their intuitive understanding of lengthscales, frequencies and magnitudes to name but a few. On the contrary, current approaches to MOGP are able to better interpret the relationship between different channels by directly modelling the cross-covariances as a spectral mixture kernel with a phase shift. We propose a parametric family of complex-valued cross-spectral densities and then build on Cramer’s Theorem, the multivariate version of Bochner’s Theorem, to provide a principled approach to design multivariate covariance functions. The so-constructed kernels are able to model delays among channels in addition to phase differences and are thus more expressive than previous methods, while also providing full parametric interpretation of the relationship across channels. The proposed method is first validated on synthetic data and then compared to existing MOGP methods on two real-world examples.

1 Introduction

The extension of Gaussian processes (GPs [1]) to multiple outputs is referred to as multi-output Gaussian processes (MOGPs). MOGPs model temporal or spatial relationships (as scalar GPs) but also account for the statistical dependence across different sources of data (or channels); this is crucial in a number of real-world applications such as fault detection, data imputation and denoising. For any two input points \( x, x' \) the covariance function of an \( m \)-channel MOGP \( k(x, x') \) is a symmetric positive definite \( m \times m \) matrix of functions, the design of this matrix-valued kernel is challenging since we have to deal with the trade off between (i) choosing a broad class of \( m(m - 1)/2 \) cross-covariances and \( m \) auto-covariances, while at the same time (ii) ensuring positive definiteness of the symmetric matrix containing these \( m(m + 1)/2 \) covariance functions. In particular, unlike widely available families of auto-covariance functions (e.g., [2]), cross-covariances are not bound to positive definite and therefore can be designed freely; the construction of these functions with interpretable functional form is the main focus of this article.

A classical approach to define cross-covariances for a MOGP is by a linear operation over independent latents GPs, this is the case of the Linear Model of Coregionalization (LMC [3]) and the Convolution Model (CONV, [4]), in these cases, the resulting kernel is a function of both the covariance functions of the latent GPs and the linear operator used which usually results in symmetric and centered cross-covariances, while these approaches are simple, they lack interpretability of the dependencies learnt and force the auto-covariances to have similar behaviour. The LMC method has also inspired the Cross-Spectral Mixture (CSM) kernel [5] which uses the scalar-valued spectral mixture kernel in [6] within LMC to model phase differences across channels by manually introducing a shift between...
the cosine and exponential factors of the spectral mixture kernel. The reader is referred to [7] [8] for a review of MOGP models.

We take a different approach to extend the spectral mixture concept to multiple outputs: Recall that a GP [1] defines a distribution over functions. An integrable function $w$ is called an $n$-component stationary mean-square-continuous stochastic process if and only if it admits the following representation

$$k(\tau) = \int_{\mathbb{R}^n} e^{i\omega \cdot \tau} S(\omega) d\omega \tag{1}$$

where $S(\omega)$ is a non-negative bounded function on $\mathbb{R}^n$ and $i$ denotes the imaginary unit.

For a proof see [11]. The above theorem gives an explicit relationship between the spectral density $S$ and the covariance function $k$ of the stochastic process $f$, in this sense [6] proposed to model the spectral density $S$ as a weighted mixture of $Q$ square-exponential functions, with weights $w_q$, centres $\mu_q$ and diagonal covariance matrices $\Sigma_q$, that is,

$$S(\omega) = \sum_{q=1}^{Q} w_q \left( \frac{1}{2\pi} \right)^{n/2} |\Sigma_q|^{1/2} \exp \left( -\frac{1}{2} (\omega - \mu_q)^T \Sigma_q^{-1} (\omega - \mu_q) \right). \tag{2}$$

Relying on Theorem [11] the kernel associated to the spectral density $S(\omega)$ in eq. (2) is given the spectral mixture kernel defined as follows.

**Definition 2.** A spectral-mixture (SM) kernel is a positive-definite stationary kernel given by

$$k(\tau) = \sum_{q=1}^{Q} w_q \exp \left( -\frac{1}{2} \tau^T \Sigma_q \tau \right) \cos(\mu_q^T \tau) \tag{3}$$

A function $g(x)$ is said to be integrable if $\int_{\mathbb{R}^n} |g(x)| dx < +\infty$.

2 Background

**Definition 1.** A Gaussian process (GP) over the input set $X$ is a real-valued stochastic process $(f(x)_{x \in X})$ such that for any finite subset $\{x_1, x_2, \ldots, x_k\} \subset X$, the vector $[f(x_1), f(x_2), \ldots, f(x_k)]^T \in \mathbb{R}^k$ is a multivariate Gaussian random variable. Without loss of generality we will choose $X = \mathbb{R}^n$.

A GP [11] defines a distribution over functions $f(x)$ that is uniquely determined by its mean function $m(x) := \mathbb{E}(f(x))$, typically assumed $m(x) = 0$, and its covariance function (also known as kernel) $k(x, x') := \text{cov}(f(x), f(x'))$, $x, x' \in X$. We now equip the reader with the necessary background to follow our proposal: we first review a spectral-based approach to the design of scalar-valued covariance kernels and then present the definition of a multioutput GP.

2.1 The spectral mixture kernel

To bypass the explicit construction of positive-definite functions within the design of stationary covariance kernels, it is possible to design the power spectral density instead [6], which only requires to be positive, and then transform it into a covariance function using the inverse Fourier transform. This is motivated by the fact that strict positivity is much easier to achieve than positive definiteness and the theoretical support of this construction is given by the following theorem:

**Theorem 1.** (Bochner’s theorem) An integrable function $k : \mathbb{R}^n \rightarrow \mathbb{C}$ is the covariance function of a weakly-stationary mean-square-continuous stochastic process $f : \mathbb{R}^n \rightarrow \mathbb{C}$ if and only if it admits the following representation

$$k(\tau) = \int_{\mathbb{R}^n} e^{i\omega \cdot \tau} S(\omega) d\omega \tag{1}$$

where $S(\omega)$ is a non-negative bounded function on $\mathbb{R}^n$ and $i$ denotes the imaginary unit.

For a proof see [11]. The above theorem gives an explicit relationship between the spectral density $S$ and the covariance function $k$ of the stochastic process $f$, in this sense [6] proposed to model the spectral density $S$ as a weighted mixture of $Q$ square-exponential functions, with weights $w_q$, centres $\mu_q$ and diagonal covariance matrices $\Sigma_q$, that is,
where \( \mu_q \in \mathbb{R}^n \), \( \Sigma_q = \text{diag}(\sigma_1^{(q)}, \ldots, \sigma_n^{(q)}) \) and \( w_q, \sigma_q \in \mathbb{R}_+ \).

Due to the universal function approximation of the sum-of-Gaussians (considered here in the frequency domain) and the relationship given by Theorem 1, the SM kernel is able to approximate continuous stationary kernels at an arbitrary precision given enough spectral components. This concept points in the direction of sidestepping the kernel selection problem in GPs and it will be extended to cater for multivariate GPs in the next section.

### 2.2 Multi-Output Gaussian Processes

A multivariate-output extension of GPs can be constructed by considering an ensemble of scalar-valued stochastic processes where any finite collection of values across all such processes are jointly Gaussian. We formalise this definition as follows.

**Definition 3.** An \( m \)-channel multi-output Gaussian process \( f(x) := (f_1(x), f_2(x), \ldots, f_m(x)) \), \( x \in \mathcal{X} \), is an \( m \)-tuple of stochastic processes \( f_p : \mathcal{X} \to \mathbb{R} \) such that for any (finite) subset of inputs \( \{x_i\}_{i=1}^n \subset \mathcal{X} \), the random variables \( \{f_{c(i)}(x_i)\}_{i=1}^n \) are jointly Gaussian for any choice of indices \( c(i) \in \{1, \ldots, m\} \).

Recall that the construction of a scalar-valued GPs requires choosing a scalar-valued mean function and a scalar-valued covariance function. Conversely, an \( m \)-channel MOGP is defined by an \( \mathbb{R}^m \)-valued mean function whose \( i \)-th element denotes the mean function of the \( i \)-th channel, and an \( \mathbb{R}^m \times \mathbb{R}^m \)-valued covariance function whose \( (i, j) \)-th element denotes the covariance between the \( i \)-th and \( j \)-th channels. The symmetry and positive-definiteness conditions of the MOGP kernel are defined as follows.

**Definition 4.** A two-input matrix-valued function \( K(x, x') : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m \times m} \) defined element-wise by \( [K(x, x')]_{ij} = k_{ij}(x, x') \) is a multioutput covariance kernel if it is:

(i) Symmetric, i.e., \( K(x, x') = K(x', x) \), \( \forall x, x' \in \mathcal{X} \), and

(ii) Positive definite, i.e., \( \forall N \in \mathbb{N}, c \in \mathbb{R}^{N \times m}, x \in \mathcal{X} \) such that, \( [c]_{pi} = c_{pi}, [x]_p = x_p \), we have

\[
\sum_{i,j=1}^{m} \sum_{p,q=1}^{N} c_{pi} c_{qj} k_{ij}(x_p, x_q) \geq 0 \quad (4)
\]

Furthermore, we say that a multivariate kernel \( K(x, x') \) is stationary if \( K(x, x') = K(x - x') \) or equivalently \( k_{ij}(x, x') = k_{ij}(x - x') \) \( \forall i, j \in \{1, \ldots, m\} \), in this case, we denote \( \tau = x - x' \).

The design of the MOGP covariance kernel involves jointly choosing functions that model the covariance of each channel (diagonal elements in \( K \)) and functions that model the cross-covariance between different channels at different input locations (off-diagonal elements in \( K \)). Choosing these \( m(m + 1)/2 \) covariance functions is challenging when we want to be as expressive as possible and include, for instance, delays, phase shifts, negative correlations or to enforce specific spectral content while at the same time maintaining positive definiteness of \( K \). For a comprehensive review of MOGPs the reader is referred to [7].

### 3 Designing Multi-Output Gaussian Processes in the Fourier Domain

We extend the spectral-mixture approach [6] to multioutput Gaussian processes relying on the multivariate version of Theorem 1 first proved by Cramer and thus referred to as Cramer’s Theorem [10][11] given by

**Theorem 2.** (Cramer’s Theorem) A family \( \{k_{ij}(\tau)\}_{i,j=1}^{m} \) of integrable functions are the covariance functions of a weakly-stationary multivariate stochastic process if and only if they (i) admit the representation

\[
k_{ij}(\tau) = \int_{\mathbb{R}^n} e^{i\omega^\top \tau} S_{ij}(\omega) d\omega \quad \forall i, j \in \{1, \ldots, m\} \quad (5)
\]

where each \( S_{ij} \) is an integrable complex-valued function \( S_{ij} : \mathbb{R}^n \to \mathbb{C} \) known as the spectral density associated to the covariance function \( k_{ij}(\tau) \), and (ii) fulfill the positive definiteness condition

\[
\sum_{i,j=1}^{m} z_{i} z_{j} S_{ij}(\omega) \geq 0 \quad \forall \{z_1, \ldots, z_m\} \subset \mathbb{C}, \, \omega \in \mathbb{R}^n \quad (6)
\]
where $\overline{z}$ is the complex conjugate of $z \in \mathbb{C}$. Note that eq. (5) states that each covariance function $k_{ij}$ is the inverse Fourier transform of a spectral density $S_{ij}$, therefore, we will say that these functions are *Fourier pairs* and refer to the set of arguments of the covariance function $\tau \in \mathbb{R}^n$ as *time* or *space domain* depending of the application considered, and to the set of arguments of the spectral densities $\omega \in \mathbb{R}^n$ as *Fourier* or *spectral domain*. Furthermore, a direct consequence of the above theorem is that for any element $\omega$ in the Fourier domain, the matrix defined by $S(\omega) = [S_{ij}(\omega)]_{i,j=1}^{m} \in \mathbb{R}^{m \times m}$ is Hermitian, i.e., $S_{ij}(\omega) = \overline{S}_{ji}(\omega) \quad \forall \ i,j,\omega$.

Theorem 2 gives the guidelines to construct covariance functions for MOGP by designing their corresponding spectral densities instead, i.e., the design is performed in the Fourier rather than the space domain. The simplicity of design in the Fourier domain stems from the positive-definiteness condition of the spectral densities in eq. (6) which is much easier to achieve than that of the covariance functions in eq. (4). This can be understood through an analogy with the univariate model: in the single-output case the positive-definiteness condition of the kernel only requires positivity of the spectral density, whereas in the multioutput case the positive-definiteness condition of the multivariate kernel only requires that the matrix $S(\omega), \quad \forall \omega \in \mathbb{R}^n$, is positive definite but there are no constraints on each function $S_{ij} : \omega \mapsto S_{ij}(\omega)$.

### 3.1 The multivariate spectral mixture kernel

We now propose a family of Hermitian positive-definite complex-valued functions $\{S_{ij}(\cdot)\}_{i,j=1}^{m}$, thus fulfilling the requirements of Theorem 2, eq. (6), to use them as cross-spectral densities within MOGP. This family of functions is designed with the aim of providing physical parametric interpretation and closed-form covariance functions after applying the inverse Fourier transform.

Recall that complex-valued positive-definite matrices be decomposed in the form $S(\omega) = R(\omega)R(\omega)^H$, meaning that the $(i,j)^{th}$ entry of $S(\omega)$ can be expressed as $S_{ij}(\omega) = R_{ij}(\omega)R_{ji}(\omega)$; where $R(\omega) \in \mathbb{C}^{Q \times m}$, $R_{i}(\omega)$ is the $i^{th}$ column of $R(\omega)$, and $(\cdot)^H$ denotes the Hermitian (transpose and conjugate) operator. Note that this factor decomposition fulfills eq. (6):

$$
\sum_{i,j=1}^{m} \sum_{\omega} |R_{ij}(\omega)R_{ji}(\omega)|^2 = ||R(\omega)z||^2 \geq 0 \quad \forall z = [z_1, \ldots, z_m]^\top \in \mathbb{C}^m, \omega \in \mathbb{R}^n
$$

(7)

We refer to $Q$ as the rank of the decomposition, since by choosing $Q < m$ the rank of $S(\omega) = R(\omega)R(\omega)^H$ can be at most $Q$. For ease of notation we choose $Q = 1$, where the columns of $R(\omega)$ are complex-valued functions $\{R_i\}_{i=1}^{m}$, and $S(\omega)$ is modeled as a rank-one matrix according to $S_{ij}(\omega) = \overline{R_i}(\omega)R_{ij}(\omega)$. Since Fourier transforms and multiplications of square exponential (SE) functions are also a SE, we model $R_i(\omega)$ to as a complex-valued SE function so as to ensure closed-form expression of its corresponding covariance kernel, that is,

$$
R_{i}(\omega) = w_i \exp \left( -\frac{1}{4}(\omega - \mu_i)^\top \Sigma_i^{-1}(\omega - \mu_i) \right) \exp \left( -i(\theta_i^\top \omega + \phi_i) \right), \quad i = 1, \ldots, m
$$

(8)

where $w_i, \phi_i \in \mathbb{R}, \mu_i, \theta_i \in \mathbb{R}^n$ and $\Sigma_i = \text{diag}([\sigma_i^2, \ldots, \sigma_i^2])^\top \in \mathbb{R}^{n \times n}$. With this choice of the functions $\{R_i\}_{i=1}^{m}$, the spectral densities $\{S_{ij}\}_{i,j=1}^{m}$ are given by

$$
S_{ij}(\omega) = w_{ij} \exp \left( -\frac{1}{2}(\omega - \mu_{ij})^\top \Sigma_{ij}^{-1}(\omega - \mu_{ij}) + i(\theta_{ij}^\top \omega + \phi_{ij}) \right), \quad i, j = 1, \ldots, m
$$

(9)

meaning that the cross-spectral density between channels $i$ and $j$ is modeled as a complex-valued SE function with the following parameters:

- covariance: $\Sigma_{ij} = 2\Sigma_i(\Sigma_i + \Sigma_j)^{-1}\Sigma_j$
- mean: $\mu_{ij} = (\Sigma_i + \Sigma_j)^{-1}(\Sigma_i \mu_i + \Sigma_j \mu_j)$
- magnitude: $w_{ij} = w_i w_j \exp \left( -\frac{1}{4}(\mu_i - \mu_j)^\top (\Sigma_i + \Sigma_j)^{-1}(\mu_i - \mu_j) \right)$
- delay: $\theta_{ij} = \theta_i - \theta_j$
- phase: $\phi_{ij} = \phi_i - \phi_j$

\footnote{The extension to arbitrary $Q$ will be presented at the end of this section.}
where the so-constructed magnitudes \( w_{ij} \) ensure positive definiteness and, in particular, the auto-
spectral densities \( S_{ii} \) are real-valued SE functions (since \( \theta_{ii} = \phi_{ii} = 0 \)) as in the standard (scalar-
valued) spectral mixture approach \([6]\).

The power spectral density in eq. (9) corresponds to a complex-valued kernel and therefore to a
complex-valued GP. In order to restrict this generative model only to real-valued GPs, the proposed
power spectral density has to be symmetric with respect to \( \omega \) \([12]\), we then make \( S_{ij}(\omega) \) symmetric
simply by reassigning \( S_{ij}(\omega) \rightarrow \frac{1}{2}(S_{ij}(\omega) + S_{ij}(-\omega)) \), this is equivalent to choosing \( R_i(\omega) \) to be a
vector of two mirrored complex SE functions.

The resulting (symmetric with respect to \( \omega \)) cross-spectral density between the \( i \)-th and \( j \)-th channels
\( S_{ij}(\omega) \) and its corresponding real-valued kernel \( k_{ij}(\tau) = F^{-1}\{S_{ij}(\omega)\}(\tau) \) are the Fourier pairs
given by

\[
S_{ij}(\omega) = \frac{w_{ij}}{2} \left( e^{\frac{i}{2}((\omega - \mu_{ij})^T \Sigma_{ij}^{-1}(\omega - \mu_{ij}) + \epsilon(\theta_{ij}^T \omega + \phi_{ij}))} + e^{\frac{i}{2}((\omega + \mu_{ij})^T \Sigma_{ij}^{-1}(\omega + \mu_{ij}) + \epsilon(-\theta_{ij}^T \omega + \phi_{ij}))} \right)
\]

\[
k_{ij}(\tau) = \alpha_{ij} \exp\left(-\frac{1}{2}(\tau + \theta_{ij})^T \Sigma_{ij}(\tau + \theta_{ij})\right) \cos\left((\tau + \theta_{ij})^T \mu_{ij} + \phi_{ij}\right)
\]

where the magnitude parameter \( \alpha_{ij} = w_{ij}(2\pi)^{\frac{Q}{2}} |\Sigma_{ij}| \) absorbs the constant resulting from the inverse
Fourier transform.

We can again confirm that the autocovariances \((i = j)\) are real-valued and contain square-exponential
and cosine factors as in the scalar SM approach since \( \alpha_{ii} \geq 0 \) and \( \theta_{ii} = \phi_{ii} = 0 \). Conversely,
the proposed model for the cross-covariance between different channels \((i \neq j)\) allows for (i) both
negatively- and positively-correlated signals \((\alpha_{ij} \in \mathbb{R})\), (ii) delayed channels through the delay
parameter \( \theta_{ij} \neq 0 \) and (iii) out-of-phase channels where the covariance is not symmetric with respect
to the delay for \( \phi_{ij} \neq 0 \). Fig. 1 shows cross-spectral densities and their corresponding kernel for a
choice of different delay and phase parameters.

![Cross-spectral densities](image)

Figure 1: Cross-spectral densities (bottom, real part in blue and imaginary part in green) and cross-
covariance function (top) generated by the proposed model in eq. (10). From left to right: zero delay
and zero phase; zero delay and non-zero phase; non-zero delay and zero phase; and non-zero delay
and non-zero phase. The dashed lines denote the SE envelopes.

The kernel in eq. (10) resulted from a low rank choice for the PSD matrix \( S_{ij} \), therefore, increasing
the rank in the proposed model for \( S_{ij} \) is equivalent to consider several kernel components. Arbitrarily
choosing \( Q \) of these components yields the expression for proposed multioutput kernel:

**Definition 5.** The Multi-Output Spectral Mixture Kernel (MOSM) has the form:

\[
k_{ij}(\tau) = \sum_{q=1}^{Q} \alpha_{ij}^{(q)} \exp\left(-\frac{1}{2}(\tau + \theta_{ij}^{(q)})^T \Sigma_{ij}^{(q)}(\tau + \theta_{ij}^{(q)})\right) \cos\left((\tau + \theta_{ij}^{(q)})^T \mu_{ij}^{(q)} + \phi_{ij}^{(q)}\right)
\]

where \( \alpha_{ij}^{(q)} = w_{ij}^{(q)}(2\pi)^{\frac{Q}{2}} |\Sigma_{ij}^{(q)}| \) and the superindex \((\cdot)^{(q)}\) denotes the parameter of the \( q \)-th component
of the spectral mixture.
This multivariate covariance function has spectral-mixture positive-definite kernels as autocovariances, while the cross-covariances are non-symmetric, delayed, and not necessarily positive-definite, spectral mixture kernels with different parameters for different pairs of outputs. Therefore, the MOSM kernel is a multi-output generalisation of the spectral mixture approach [6] where the positive definiteness is guaranteed by the factor decomposition of $S_{ij}$ as shown in eq. (7).

3.2 Training the model and computing the predictive posterior

Fitting the model to observed data follows the same rationale of standard GP, that is, maximising log-probability of the data. Recall that the observations in the multivariate case consist of (i) a location $x \in X$, (ii) a channel identifier $i \in \{1, \ldots, m\}$, and (iii) an observed value $y \in \mathbb{R}$, therefore, we denote $N$ observations as the set of 3-tuples $D = \{(x_i, i, y_i)\}_{i=1}^N$. As all observations are jointly Gaussian, we concatenate the observations into the three vectors $x = [x_1, \ldots, x_N]^T \in \mathcal{X}$, $\mathbf{i} = [i_1, \ldots, i_N]^T \in \{1, \ldots, m\}^m$, and $y = [y_1, \ldots, y_N]^T \in \mathbb{R}^{N \times m}$, to express the negative log-likelihood (NLL) by

$$-\log p(y|x, \Theta) = \frac{k}{2} \log 2\pi + \frac{1}{2} \log |K_{xi}| + \frac{1}{2} y^T K_{xi}^{-1} y$$  (12)

where all model parameters are denoted by $\Theta$, and $K_{xi}$ is the covariance matrix of all observed samples, that is, the $(r, s)^{th}$ element $[K_{xi}]_{rs}$ is the covariance between the process at $x_r$, channel: $i_r$ and the process at $x_s$, channel: $i_s$, which under the proposed MOSM model is given by eq. (11) plus a diagonal term to cater for uncorrelated observation noise. The NLL is then minimised with respect to $\Theta = \{w_i^{(q)}, \mu_i^{(q)}, \Sigma_i^{(q)}, \omega, \phi_i^{(q)}\}_{i=1,q=1}$, that is, the original parameters chosen to construct $R(\omega)$ in Section 3.1.

Once the hyper-parameters are optimised, computing the predictive posterior in the proposed MOSM follows the standard GP procedure with the joint covariances given by eq. (11).

3.3 Related work

Generalising the scalar spectral mixture kernel to MOGPs can be achieved from the LMC framework as pointed out in [5] (which we denote SM-LMC). As this formulation only considers real-valued cross spectral densities, the authors propose a multivariate covariance function by including a complex component to the cross spectral densities to cater for phase differences across channels, which they call the Cross Spectral Mixture kernel (denoted CSM). This multivariate covariance function can be seen as the proposed MOSM model with $\mu_i = \mu_j$, $\Sigma_i = \Sigma_j$, $\theta_i = \theta_j$ $\forall i, j \in \{1, \ldots, m\}$ and $\phi_i = \phi_j$ for $\psi_i \in \mathbb{R}^n$, thus the SM-LMC is a particular case of the MOSM model where the parameters $\mu_i, \Sigma_i, \theta_i$ are restricted to be same for all channels and therefore only phase shifts and no delays are allowed unlike the example in Fig. 1.

4 Experiments

We first validated the ability of the proposed MOSM model in the identification of known auto- and cross-covariances of synthetic data, and then compared it against the spectral-mixture linear model of coregionalization (SM-LMC, [3, 6, 5]), the Gaussian convolution model (CONV, [4]), and the cross-spectral mixture model (CSM, [5]) in the estimation of missing real-world data in two different distributed settings: climate signals and metal concentrations. All models were implemented in Tensorflow [13] using the package GPflow [14].

4.1 Synthetic example: Learning derivatives and delayed signals

MOSM was implemented to recover the auto- and cross-covariances of a three-output GP with the following components: (i) a reference signal sampled from a GP $f(x) \sim \mathcal{GP}(0, K_{SM})$ with spectral mixture covariance kernel $K_{SM}$ and zero mean, (ii) its derivative $f'(x)$, and (iii) a noiseless delayed version $f_d(x) = f(x-\delta)$. The motivation for this illustrative example is that the covariances and cross covariances of the aforementioned processes are known explicitly (see [1], sec. 9.4) and we can therefore compare our estimates to the true model. The reference signal was sampled on the domain $[-10, 10]$ and the derivative was computed numerically (first order) in the same range. Fig.
shows the ground truth and MOSM estimates for all three synthetic signals and the covariances (normalised), where the proposed model successfully learnt all cross-covariances $\text{cov}(f(\cdot), f'(x))$ and $\text{cov}(f(x), f(x - \delta))$, and autocovariances without prior information about the delay or the derivative relationship between the two channels.

Figure 2: MOSM learning of the covariance functions of a synthetic reference signal, its derivative and a delayed version. **Left**: synthetic signals, **middle**: autocovariances, **right**: cross-covariances. The dashed line represents the ground truth and the solid colour lines the MOSM estimates.

### 4.2 Climate data

The first real-world dataset contained measurements\(^3\) from a sensor network of four climate stations in the south on England: Cambermet, Chimet, Sotonmet and Bramblemet. We considered the normalised air temperature signal from 12 March, 2017 to 16 March, 2017, in 5-minute intervals (5692 samples) from where we randomly chose $N = 1000$ samples for training. Following [4], we simulated a sensor failure by removing the second half of the measurements for one sensor and leaving the remaining three sensors operating correctly, we reproduced the same setup across all four sensors thus having four independent experiments. All models considered had five latent signals/spectral components.

For all four models considered, Fig. 3 shows the estimates of missing data for the case where Cambermet was the simulated faulty sensor. Table 1 presents the mean absolute error for all models over the missing data region given by

$$\text{MAE} : \frac{1}{M} \sum_{i=1}^{M} |y_i - \hat{y}_i|$$  \hspace{1cm} (13)

where $y_i$ denotes the true value and $\hat{y}_i$ the MOGP estimate. Observe how all models were able to capture the behaviour of the signal in the missing range, this is because the considered climate signals are very similar to one another. This shows that the MOSM can also collapse to models that share parameters across pairs of outputs when required.

### 4.3 Heavy metal concentration

The Jura dataset contains, in addition to other geological data, the concentration of seven heavy metals in a region of 14.5 km\(^2\) of the Swiss Jura, this dataset [2] is divided into a training set (259 locations) and a validation set (100 locations). We followed [3, 4] where the motivation was to aid the prediction of a variable that is expensive to measure by using abundant measurements of correlated variables which are less expensive to acquire. Specifically, we estimated Cadmium and Copper at the validation locations using measurements of related variables (Nickel and Zinc for Cadmium, and

\(^3\)Data can be obtained from [www.cambermet.co.uk](http://www.cambermet.co.uk) and the sites therein.
Figure 3: Climate data: Prediction of the Cambermet sensor in a simulated missing-data scenario. The red points denote the observations, the dashed black line the true signal, and the solid colour lines the considered models: From left to right: MOSM, CONV, SM-LMC and CSM.

Table 1: Mean absolute error for all four experiments with one-standard-deviation error bars over ten realisations.

| Model    | Cambermet | Chimet     | Sotonmet | Bramblemet |
|----------|-----------|------------|----------|------------|
| CONV     | 0.098 ± 0.008 | 0.192 ± 0.015 | 0.211 ± 0.038 | 0.163 ± 0.009 |
| SM-LMC   | 0.084 ± 0.004 | 0.176 ± 0.003 | 0.273 ± 0.001 | 0.134 ± 0.002 |
| CSM      | 0.094 ± 0.003 | 0.129 ± 0.004 | 0.195 ± 0.011 | 0.130 ± 0.004 |
| MOSM     | 0.097 ± 0.006 | 0.137 ± 0.007 | 0.162 ± 0.011 | 0.129 ± 0.003 |

Lead, Nickel and Zinc for Copper) at the training and test locations. The MAE—see eq. (13)—is shown in Table 2, where the results for the CONV model were obtained from [4] and all models considered five latent signals/spectral components except the independent Gaussian process (denoted IGP). Observe how the proposed MOSM model outperforms all other models over the Cadmium data, while in the Copper case, the higher variability of the data causes that the simplest MOGP model (SM-LMC) achieves the best results.

Table 2: Mean absolute error for the estimation of Cadmium and Copper concentrations with one-standard-deviation error bars over ten repetitions of the experiment.

| Model | Cadmium | Copper |
|-------|---------|--------|
| IGP   | 0.56 ± 0.005 | 16.5 ± 0.1 |
| CONV  | 0.443 ± 0.006 | 7.45 ± 0.2 |
| SM-LMC| 0.46 ± 0.01  | 7.0 ± 0.1  |
| CSM   | 0.47 ± 0.02  | 7.4 ± 0.3  |
| MOSM  | 0.43 ± 0.01  | 7.3 ± 0.1  |

5 Discussion

We have proposed the multioutput spectral mixture (MOSM) kernel to model rich relationships across multiple outputs within Gaussian processes regression models. This has been achieved by constructing a positive-definite matrix of complex-valued spectral densities and then transforming them using the inverse Fourier transform according to Cramer’s Theorem. The resulting kernel provides clear interpretation from a spectral viewpoint, where each of its parameters can be identified with frequency, magnitude, phase and delay for a pair of outputs; a key feature that is unique to the proposed kernel is the ability joint model delays and phase differences, this is possible due to the complex-valued model for the cross-spectral density considered and validated experimentally using a synthetic example—see fig. 2. The MOSM kernel has also been compared against existing MOGP models on two real-world datasets, where the proposed model performed competitively in terms of the the mean absolute error. Further research should point towards a sparse implementation of the proposed MOGP which can build on [4, 15] to design inducing variables that exploit the spectral content of the processes as in [16, 17].
References

[1] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[2] D. Duvendau, “Automatic model construction with Gaussian processes,” Ph.D. dissertation, University of Cambridge, 2014.

[3] P. Goovaerts, *Geostatistics for natural resources evaluation*. Oxford University Press on Demand, 1997.

[4] M. Alvarez and N. D. Lawrence, “Sparse convolved Gaussian processes for multi-output regression,” in *Proceedings of the 21st International Conference on Neural Information Processing Systems*, 2008, pp. 57–64.

[5] K. R. Ulrich, D. E. Carlson, K. Dzirasa, and L. Carin, “GP kernels for cross-spectrum analysis,” in *Advances in Neural Information Processing Systems 28*, 2015, pp. 1999–2007.

[6] A. G. Wilson and R. P. Adams, “Gaussian process kernels for pattern discovery and extrapolation,” in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1067–1075.

[7] M. A. Álvarez, L. Rosasco, and N. D. Lawrence, “Kernels for vector-valued functions: A review,” *Found. Trends Mach. Learn.*, vol. 4, no. 3, pp. 195–266, Mar. 2012. [Online]. Available: http://dx.doi.org/10.1561/2200000036

[8] M. G. Genton and W. Kleiber, “Cross-covariance functions for multivariate geostatistics,” *Institute of Mathematical Statistics*, vol. 30, no. 2, 2015.

[9] S. Bochner, M. Tenenbaum, and H. Pollard, *Lectures on Fourier Integrals*, ser. Annals of mathematics studies. Princeton University Press, 1959.

[10] H. Cramér, “On the theory of stationary random processes,” *Annals of Mathematics*, pp. 215–230, 1940.

[11] A. Yaglom, *Correlation Theory of Stationary and Related Random Functions*, ser. Correlation Theory of Stationary and Related Random Functions. Springer, 1987, no. v. 1.

[12] S. M. Kay, *Modern spectral estimation: Theory and application*. Englewood Cliffs, N.J.: Prentice Hall, 1988.

[13] M. Abadi et al., “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[14] A. G. d. G. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, and J. Hensman, “GPflow: A Gaussian process library using TensorFlow,” 2016.

[15] M. A. Alvarez, D. Luengo, M. K. Titsias, and N. D. Lawrence, “Efficient multioutput Gaussian processes through variational inducing kernels.” in *AISTATS*, vol. 9, 2010, pp. 25–32.

[16] J. Hensman, N. Durrande, and A. Solin, “Variational fourier features for Gaussian processes,” *arXiv preprint arXiv:1611.06740*, 2016.

[17] F. Tobar, T. D. Bui, and R. E. Turner, “Learning stationary time series using Gaussian processes with nonparametric kernels,” in *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015, pp. 3501–3509.