Nonstationary multi-output Gaussian processes via harmonizable spectral mixtures

Matías Altamirano
Departament of Mathematical Engineering
Universidad de Chile
maltamirano@dim.uchile.cl

Felipe Tobar
Initiative for Data & Artificial Intelligence
Universidad de Chile
ftobar@uchile.cl

Abstract

Kernel design for Multi-output Gaussian Processes (MOGP) has received increased attention recently. In particular, the Multi-Output Spectral Mixture kernel (MOSM) approach has been praised as a general model in the sense that it extends other approaches such as Linear Model of Corregionalization, Intrinsic Corregionalization Model and Cross-Spectral Mixture. MOSM relies on Cramér's theorem to parametrise the power spectral densities (PSD) as a Gaussian mixture, thus, having a structural restriction: by assuming the existence of a PSD, the method is only suited for multi-output stationary applications. We develop a nonstationary extension of MOSM by proposing the family of harmonizable kernels for MOGPs, a class of kernels that contains both stationary and a vast majority of non-stationary processes. A main contribution of the proposed harmonizable kernels is that they automatically identify a possible nonstationary behaviour meaning that practitioners do not need to choose between stationary or non-stationary kernels. The proposed method is first validated on synthetic data with the purpose of illustrating the key properties of our approach, and then compared to existing MOGP methods on two real-world settings from finance and electroencephalography.

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

1 INTRODUCTION

Gaussian Processes (GPs) provide a flexible and powerful non-parametric framework for Bayesian inference on time series, and thus are considered in many areas of application [Williams and Rasmussen, 2006]. The main aspect on the design of the GP is the choice of the covariance function (also called kernel), which encapsulates all the properties of the process such as smoothness, periodicity and stationarity to name a few. The extension of the GP to multiple outputs is known as multi-output Gaussian processes (MOGP) [Álvarez et al., 2012], which models the outputs to be jointly Gaussian an thus is able to share information across outputs, potentially improving the estimation. As in the single-output case, designing kernels that successfully model auto- and cross-covariances between channels is a core challenge in MOGPs.

There have been several approaches to design valid MOGP kernels [Álvarez and Lawrence, 2009, Álvarez et al., 2012, Goovaerts et al., 1997], and a number of them are based on linear combinations of latent-factor independent Gaussian processes. These approaches, though they work in practice, avoid the direct parametrisation of multioutput covariances and may result in constrained covariances, especially from a spectral analysis perspective. Recently, [Parra and Tobar, 2017] proposed the multi-output spectral mixture (MOSM) which directly designs the kernel in the spectral domain, using the multivariate version of Bochner's theorem [Bochner et al., 1959], namely Cramér's Theorem [Cramér, 1940].

The MOSM kernel provides a unified perspective of existing MOGP kernels in the literature, however, its principal limitation is that it is restricted to stationary data, i.e., $k(x, x') = k(x - x')$, thus it encodes an identical similarity notion across the input space. This assumption of stationarity is unsuitable for several real world settings, such as vibratory signals [Kim and Kim, 2005, Zhang et al., 2003], free-drifting oceanic instruments

[Lilly and Olhede, 2011], various neuroscience applications [Cranstoun et al., 2002, Ombao et al., 2001], and econometrics [Joyeux, 1980]. Therefore, a flexible non-stationary multioutput kernel becomes necessary and in particular a non-stationary version of the MOSM kernel.

In this article, we propose an expressive and flexible family of MOGP kernels to model non-stationary processes as a natural extension of the MOSM. The proposed kernel relies on the concept of harmonizability, a term introduced in [Loeve, 1978] which generalizes the Fourier spectral representations to non-stationary processes. The harmonizable processes have been widely studied and developed in the statistics community, but there has been a lack of attentiveness among machine learning researchers.

The rest of the paper is organized as follows. We revisit the required concepts supporting our proposal from the GP literature and the concept of harmonizable processes in Section 2. Section 3 presents the proposed MOGP kernel and Section 4 compares it to previous approaches in the literature. We explain practical considerations of the proposed model in Section 5 and Section 6 validates it on synthetic and real-world data. Lastly, we discuss our results and summarise our contribution and future work in Section 7.

2 BACKGROUND

In this section we briefly introduce Gaussian processes, multi-output Gaussian processes, spectral mixture kernels and harmonizable processes; these support the development of the proposed kernel in Section 3.

2.1 Gaussian processes

The GP is a Bayesian nonparametric generative model for functions $f : \mathbb{R}^D \to \mathbb{R}$. The GP is the infinite-dimensional extension of the multivariate normal (MVN) distribution, meaning that it can model second-order relationships among an infinite number of random variables. With mean function m and covariance function k, a GP denoted by

$$f \sim \mathcal{GP}(m, k),$$

has the property that any collection of inputs $\{x_1, \ldots, x_N\} \subset \mathbb{R}^D$, the output $[f(x_1), \ldots, f(x_N)] \in \mathbb{R}^N$ is distributed according to an MVN of mean $m(\mathbf{x})$ and covariance $k(\mathbf{x}, \mathbf{x})$, where $\mathbf{x} = [x_1, \ldots, x_N]$.

The design of the GP, involves choosing the kernel function, which determines key properties in the draws from the GP such as differentiability, periodicity, long-range correlation or stationarity. Furthermore, we say that a kernel is stationary if it can be written as $k(x, x') = k(x - x')$; a GP is said to be stationary if its covariance is stationary. The central result on the characterization of stationary kernels is Bochner's theorem:

Theorem 1 ([Bochner et al., 1959])**.** *A complex-valued function k on $\mathbb{R}^d$ is the covariance function of a weakly-stationary mean-square-continuous stochastic process on $\mathbb{R}^d$ if and only if it admits the following representation*

$$k(x, x') = \int_{\mathbb{R}^d} e^{i\omega^\top (x - x')} S(\omega) d\omega, \tag{1}$$

where $S(\omega)$ is a non-negative bounded function on $\mathbb{R}^d$, called the power spectral density, and i denotes the imaginary unit.

This theorem defines a duality between time and frequency, which allows us to construct new kernels via their parametrization in the (Fourier) frequency domain. Perhaps the main approach to kernel design in this fashion is the spectral mixture (SM) kernel proposed by [Wilson and Adams, 2013].

2.2 Multi-output Gaussian processes

The extension of the GP framework to handle multiple output is referred to as Multi-Output Gaussian Process (MOGP) [Bonilla et al., 2007], which consists in modeling all the outputs as jointly Gaussian where the covariance and cross covariance are ruled by a multi-output kernel. In more detail, if we have M output latent function $\{f_i\}_{i=1}^M$, the element of (i, j) of the covariance kernel $\mathcal{K}$ corresponds to the covariance between outputs f_i and f_j, following the next notation:

$$\text{cov}[f_i(x), f_j(x')] = k_{ij}(x, x') = [\mathcal{K}(x, x')]_{ij}. \tag{2}$$

A kernel function must be symmetric and positive-definite in order to be a valid covariance function, and similar to the single channel case, a multivariate kernel $\mathcal{K}$ is stationary if $\mathcal{K}(x, x') = \mathcal{K}(x - x')$. The design of valid and expressive multi-output kernels is quite challenging because we need to jointly choose functions that model the covariance of each channel and functions that model the cross-covariance between channels [Álvarez et al., 2012]. Several approaches have been proposed to overcome this difficulty, see, e.g., [Álvarez and Lawrence, 2009, Goovaerts et al., 1997, Teh et al., 2005, Ulrich et al., 2015]. However, mostly all of them are based on the idea of model the cross covariances as a linear combination of the covariance of each channel. In the next subsection we review one of the most prominent extensions of the existing methods in expressiveness and interpretation, the multi-output spectral mixture kernel.

2.3 Multi-output spectral mixture kernel

Relying on Cramér's theorem [Cramér, 1940], the multivariate version of the Bochner theorem, [Parra and Tobar, 2017] provided a new approach to design multivariate covariance functions which allows full parametric interpretation of the relationship across channels, in addition to model delays and phase among channels.

Definition 1. *The multi-output spectral mixture (MOSM) kernel between channels i and j has the form:*

$$k_{ij}(\hat{x}) = \sum_{q=i}^{Q} \alpha_{ij}^{(q)} \exp\left(-\frac{1}{2}(\hat{x}+\theta_{ij}^{(q)})^{\top}\Sigma_{ij}^{(q)}(\hat{x}+\theta_{ij}^{(q)})\right) \cos\left((\hat{x}+\theta_{ij}^{(q)})^{T}\mu_{ij}^{(q)} + \phi_{ij}^{(q)}\right), \tag{3}$$

where $\hat{x} = x - x'$ is the difference between input locations, and the hyperparameters $\alpha_{ij}^{(q)}, \Sigma_{ij}^{(q)}, \mu_{ij}^{(q)}, \theta_{ij}^{(q)}, \phi_{ij}^{(q)}$ are the magnitude, covariance, mean, time delay and phase delay (respectively) between channels i and j.

The MOSM kernel exhibits desirable properties for modeling multivariate time series, in particular, it provides a clear interpretation from a spectral viewpoint through its hyperparameters. However, since this family of kernels stems form Cramér's theorem, the method is only suited for stationary processes.

2.4 Harmonizable processes

To consider a general class of processes beyond stationary ones, we will study the celebrated extension of the stationarity property called harmonizabilty, originally introduced in [Loeve, 1978] for the univariate case, and then in [Kakihara, 1997] for the multidimensional case.

Definition 2. *A stochastic process on $\mathbb{R}^d$ is weakly harmonizable iff its covariance function can be expressed as:*

$$k(x,x') = \iint_{\mathbb{R}^d\times\mathbb{R}^d} e^{i(\omega^{\top}x - \omega'^{\top}x')} F(d\omega, d\omega'), \tag{4}$$

where i is the imaginary unit and $F(d\omega, d\omega')$ is a positive semi-definite bimeasure[1] with finite Fréchet variation, referred to as the spectral bimeasure of the process. A harmonizable process is strongly harmonizable iff its spectral bimeasure is a measure and the integral above coincides with the Lebesgue integral.

Henceforth, we will refer to strongly harmonizable processes simply as *harmonizable processes.*

[1] F is a bimeasure iff $F(A,\cdot)$ and $F(\cdot,B)$ are complex measures $\forall A, B \in \mathcal{B}(\mathbb{R}^d)$, but is not necessarily a measure on $\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{B}(\mathbb{R}^d)$

Remark 1. *When F is absolutely continuous w.r.t. the Lebesgue measure, we denote its Radon-Nikodym derivative as $S = \frac{\partial^2 F}{\partial\omega\partial\omega'}$ and write the covariance as*

$$k(x,x') = \iint_{\mathbb{R}^d\times\mathbb{R}^d} e^{i(\omega^{\top}x - \omega'^{\top}x')} S(w,w') d\omega d\omega', \tag{5}$$

we refer to S as the generalized spectral density of the process in analogy to the spectral density of stationary processes.

Remark 2. *Notice that $S(\omega,\omega')$ is also a covariance function [Hurd, 1973], and measure the interaction between the ω and ω', thus, we can truly interpret the variables ω and ω' as frequencies. This notion of correlation between frequencies is what gives harmonizable processes the property of modeling non-stationary processes.*

Remark 3. *Observe that the harmonizable concept is a consistent extension of stationary processes in the sense that when the measure F concentrates on its diagonal $\omega = \omega'$, eq. (4) collapses to eq. (1) in Bochner Thm.*

It is worth noting that the harmonizable processes as presented above define a much larger class than that of stationary processes. In fact, [Yaglom, 1987] noticed that the only processes with continuous bounded kernels that are not harmonizable are, in his own words:

> *rather complicated and have some unusual, even pathological, properties.*

More recently [Samo, 2017], who studied the universality of the harmonizable kernels, showed that:

Proposition 1. *The family of harmonizable kernels defined on $\mathbb{R}^d\times\mathbb{R}^d$ is pointwise dense in the family of all complex-valued continuous bounded kernels defined on $\mathbb{R}^d\times\mathbb{R}^d$.*

Moreover, the concept of harmonizability can also be extended to the multivariate case as follows.

Theorem 2 ([Kakihara, 1997])**.** *A family $\{k_{ij}(x,x')\}_{i,j=1}^{m}$ of complex-valued functions on $\mathbb{R}^d$ are the covariance functions of a harmonizable multivariate stochastic process on $\mathbb{R}^d$ if and only if they admit the following representation:*

$$k_{ij}(x,x') = \iint_{\mathbb{R}^d\times\mathbb{R}^d} e^{i(\omega^{\top}x - \omega'^{\top}x')} F_{ij}(d\omega, d\omega'), \tag{6}$$

where the matrix spectral measure $F = [F_{ij}(A,B)]$ is such that $\forall i,j, F_{ii}$ is positive semi-definite, and symmetric, that is, $F_{ij}(A,B) = \overline{F}_{ji}(B,A)$.

Remark 4. *In the same manner as in the univariate case, if F is absolutely continuous w.r.t. the Lebesgue measure, we denote its Radon-Nikodym derivative as*

$S = \frac{\partial^2 F}{\partial \omega \partial \omega'}$, the generalized spectral density of the process and we have

$$k_{ij}(x, x') = \iint_{\mathbb{R}^d \times \mathbb{R}^d} e^{i(\omega^\top x - \omega'^\top x')} S_{ij}(w, w') d\omega d\omega'. \tag{7}$$

3 MULTI-OUTPUT HARMONIZABLE SPECTRAL MIXTURE

Following the idea behind the MOSM kernel, we propose a family of Hermitian positive-definite complex-valued functions $\{S_{ij}\}_{i,j=1}^M$ that satisfy the Theorem 2 to use them as building blocks for a generalized cross-spectral densities.

In order to model the relationship among channels, we support the proposed family of Hermitian positive-definite complex-valued functions on its Cholesky decomposition, recalling that every complex-valued positive-definite $m \times m$ matrix S can be decomposed as $S(\omega, \omega') = R^H(\omega, \omega')R(\omega, \omega')$, where $R \in \mathbb{C}^{Q \times m}$, with $Q \in \mathbb{N}$ the rank of the decomposition. For ease of understanding, we choose $Q = 1$, the case for arbitrary Q is shown at the end of the section. Since, the (i, j) entry of $S(\omega, \omega')$ can be expressed as $S_{ij}(\omega, \omega') = \overline{R}_i(\omega, \omega')R_j(\omega, \omega')$, $\forall i, j = 1, \ldots, m$, thus, we choose:

$$R_i(\omega, \omega') = w_i \underbrace{\exp\left(-\frac{1}{4l_i^2}\|\hat{\omega}\|^2\right)}_{(\star)} \times \underbrace{\exp\left(-\frac{1}{4}(\overline{\omega} - \mu_i)^\top \Sigma_i^{-1}(\overline{\omega} - \mu_i) - i(\theta_i^\top \overline{\omega} + \phi_i)\right)}_{(\bullet)}, \tag{8}$$

where $\hat{\omega} = \omega - \omega'$, $\overline{\omega} = (\omega + \omega')/2$, and $w_i, \phi_i \in \mathbb{R}$, $\theta_i, \mu_i \in \mathbb{R}^n$, $\Sigma_i = \text{diag}([\sigma_{i1}^2, ..., \sigma_{in}^2]) \in \mathbb{R}^{n \times n}$ are hyperparameters.

The intuition behind this choice is that the $(\star)$ term controls the correlation between the frequencies: two frequencies that are further away from one another are less correlated. On the other hand, the $(\bullet)$ component models the importance of each frequency. We parametrize both components as square exponential (SE) functions with complex argument, since: i) they are closed under multiplication and anti-Fourier transform, meaning that the resulting kernel is explicit and ii) they can match continuous power spectra to a desired degree of accuracy.

Therefore, selecting this parametrization for $\{R_i\}_{i=1}^m$ we have that $\{S_{ij}\}_{i,j=1}^m$ are given by:

$$\begin{aligned} S_{ij} =& w_{ij} \exp\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2\right) \times \\ & \exp\left(-\frac{1}{2}(\overline{\omega} - \mu_{ij})^\top \Sigma_{ij}^{-1}(\overline{\omega} - \mu_{ij})\right) \times \\ & \exp(-i(\theta_{ij}^\top \overline{\omega} + \phi_{ij})). \end{aligned} \tag{9}$$

Observe that this is a decaying square exponential—due to the factor $\exp\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2\right)$—and the channel parameters obey the following relationships:

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

Observe that S_{ij} is a *locally-stationary kernel*, a concept coined by [Silverman, 1957], which refers to kernels that can be expressed as a product of a stationary kernel and a non-negative function. Therefore we can assure that each S_{ij} is a positive-definite complex-valued function. Additionally, since S was designed through the product $S = R^H R$, it is by construction a positive-definite matrix thus fulfilling Theorem 2. Finally, as we are interested in real-valued GPs, for which the covariance kernel is also required to be real-valued, we make S_{ij} symmetric by reassigning:

$$S_{ij}(\omega, \omega') \longleftarrow \frac{1}{2}(S_{ij}(\omega, \omega') + S_{ij}(-\omega, -\omega')), \tag{10}$$

this guarantees symmetry and real values in the diagonal as the complex terms cancel each other. Therefore, the kernel obtained by taking the integral of the symmetrised spectral density is:

$$\begin{aligned} k_{ij}(x, x') =& \alpha_{ij} \exp\left(-\frac{1}{2}(\hat{x} + \theta_{ij})^\top \Sigma_{ij}(\hat{x} + \theta_{ij})\right) \times \\ & \cos\left((\hat{x} + \theta_{ij})^T \mu_{ij} + \phi_{ij}\right) \exp\left(-\frac{1}{2l_{ij}^2}\|\overline{x}\|^2\right), \end{aligned} \tag{11}$$

where $\overline{x} = \frac{x + x'}{2}$, $\hat{x} = x - x'$ and the magnitude $\alpha_{ij} = w_{ij}(2\pi)^n |\Sigma_{ij}|^{1/2} l_{ij}$.

From the kernel and spectral expressions we can interpret the parameters of the constructed kernel as follows:

- The spectral mean μ_i represents the main frequency
- The spectral covariance Σ_i represents the uncertainty of the distribution in the spectrum
- The cross spectral delay θ_{ij} serves as the time delay between channels
- The cross spectral phase ϕ_{ij} provides the difference in phase between channels
- The spectral length-scale l_i controls the correlation between the frequencies.

One drawback of the presented formulation is that, due the last exponential term in eq. (11), the proposed kernel vanishes outside the origin with length-scale l_i. We address this limitation by placing input shifts, which will allow us to control in which parts of the input domain the kernel is activated. We can include the input shifts in our formulation by multiplying our initial matrix S by $\exp(-x_p(w-w'))$, which comply with the properties needed for Theorem 2 to hold. Putting together all the aforementioned components, the kernel is defined as follows:

$$\begin{aligned} k_{ij}(x,x') =& \alpha_{ij} \exp\left(-\frac{1}{2}(\hat{x}+\theta_{ij})^\top \Sigma_{ij}(\hat{x}+\theta_{ij})\right)\times \\ & \cos\left((\hat{x}+\theta_{ij})^T \mu_{ij} + \phi_{ij}\right)\times \\ & \exp\left(-\frac{1}{2l_{ij}^2}\|\overline{x}-x_p\|^2\right). \end{aligned}$$

Lastly, for the general case we can expand the kernel to an arbitrary rank matrix Q by taking S as a sum of Q of these matrices of rank 1, and considering P input shifts. This yields the final expression for the proposed kernel, termed MOHSM:

Definition 3. *The Multi-Output Harmonizable Spectral Mixture (MOHSM) kernel has the form:*

$$\begin{aligned} k_{ij}(\hat{x},\overline{x}) = \sum_{p=1}^{P}\sum_{q=1}^{Q_p} & \alpha_{ij}^{(q)} \exp\left(-\frac{1}{2}(\hat{x}+\theta_{ij}^{(q)})^\top \Sigma_{ij}^{(q)}(\hat{x}+\theta_{ij}^{(q)})\right)\times \\ & \cos\left((\hat{x}+\theta_{ij}^{(q)})^\top \mu_{ij}^{(q)} + \phi_{ij}^{(q)}\right)\times \\ & \exp\left(-\frac{l_{ij}^{2(p)}}{2}\|\overline{x}-x_p\|^2\right), \end{aligned} \tag{12}$$

where recall that $\hat{x} = x - x'$ and $\overline{x} = \frac{x+x'}{2}$, P is the number of input shifts, Q_p is the number of spectral components for the p^{th} input shift with location x_p, $\alpha_{ij}^{(q)} = w_{ij}^{(q)}(2\pi)^n|\Sigma_{ij}^{(q)}|^{1/2} l_{ij}^{(q)}$, the super index $(\cdot)^{(q)}$ denotes the parameter of the q^{th} component of the spectral mixture and the super index $(\cdot)^{(p)}$ denotes the parameters for the p^{th} input shift.

Using $Q_p > 1$ can also be justified by the latent-factor construction of MOGPs. There, Q_p denotes the number of latent GPs, and since each of these latent signals has different kernels, the "capacity" of the model is not dominated by the number of channels m but by Q_p, which is (proportional) to the amount of the kernel's hyperparameters.

4 RELATIONSHIP TO PREVIOUS WORK

Even though the idea of considering a non-stationary kernel derived from the harmonizable processes is not new, all previous attempts are restricted to the single output case. For example, [Samo and Roberts, 2015] proposed a family of spectral kernels that they prove can approximate any continuous bounded nonstationary kernel which they called Generalized Spectral Kernels. In the same line, [Shen et al., 2019] proposed the harmonizable spectral mixture (HSM) kernel which is also a family derived from mixture models on the generalized spectral representation. [Remes et al., 2017] presented a non-stationary kernel based on the idea of harmonizable processes and parametrized the frequencies, length scales, and mixture weights as Gaussian processes. Our work can be seen as a multivariate extension of the family proposed by [Shen et al., 2019] with expressive cross-covariance functions.

In general, classical MOGP approaches (such as the Linear Model of Corregionalization and the Intrinsic Corregionalization Model) can represent non-stationary processes, since they model the cross-correlation functions as a linear combination of the auto-correlation functions, thus choosing non-stationary auto-correlations leads to a non-stationary multivariate process. The issue with these formulations is they force the auto-covariance and the cross-covariance to have similar behavior. Also, these methods based on linear mixtures are not able to introduce temporal correlation other than those of the latent GP component. Furthermore, by modeling the cross-correlation in this fashion, the interpretability of the learned dependence is almost null. Though MOSM solves these previous problems, adding interpretability of the dependencies and not imposing similar behaviors through different channels, is restricted to stationary cases by construction.

In this way, our work is a generalization of MOSM since the proposed MOHSM can model the correlation between the channels like the MOSM do, but allowing changes in the regimes across time, leading to a more general model. A natural question that arises in our context is whether the MOHSM can recover its stationary counterpart MOSM. We can notice that when $x_p = 0$ and $l_{ij} \to 0$, $\forall i,j$ we recover the MOSM kernel,

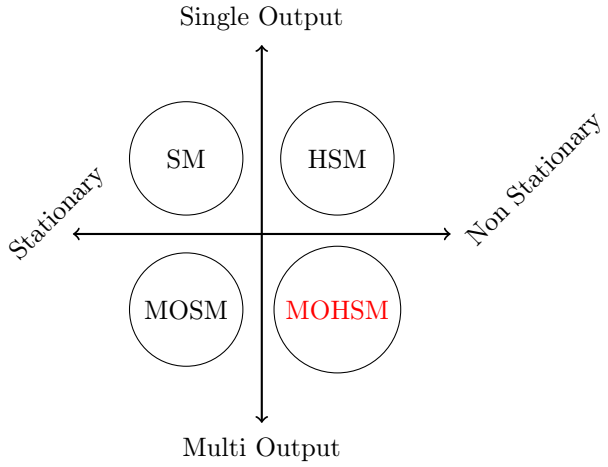

Figure 1: Relationship between the different spectral kernels

successfully extending the stationary model (more details in the supplementary material). In the frequency domain, this can be seen as the absence of correlation between frequencies, which is the assumption of stationarity. Figure 1 shows a summary of the relationships of the MOHSM with the different spectral kernels.

It is worth discussing what kind of non-stationarity the proposed MOHSM model can capture. Observe, from Def. 3, that each mixture of the proposed kernel is a locally stationary kernel, which means that the proposed kernel is a combination of a component of global structure multiplied by components that describe the local structure of the data. The chosen global structure is an exponential window operating over the regions described by the local structure component. Thus, the kernel can be seen as a union of different regimes that can be disjoint or overlapped, depending on the locations of the input shifts x_p and their lengthscales $l^{(p)}$.

5 PRACTICAL CONSIDERATIONS

5.1 Training

Training an MOGP with the MOHSM kernel follows the same procedure as the single output GP, by closed-form maximum likelihood. Since the model is jointly Gaussian, we can concatenate the locations $x \in \mathbb{R}^n$, the channel identifier $i \in \{1, ..., M\}$ and the observed value $y \in \mathbb{R}$ into three vectors and express the negative log-likelihood (NLL) as considering the multi-output covariance matrix of all the observed samples.

The NLL is minimised with respect to all the hyperparameters of the MOHSM including the noise hyperparameter, that is

$$\Theta = \{w_i^{(q)}, \mu_i^{(q)}, \Sigma_i^{(q)}, \theta_i^{(q)}, \phi_i^{(q)}, \sigma_{i,\mathrm{n}}^{(q)}, l_i^{(p)}, x_p\}_{i=1,q=1,p=1}^{M,Q_p,P}.$$

Once the optimization is concluded, computing the predictive posterior in the proposed model follows the standard GP procedure as well.

In order to make the optimization more stable we reparametrized the MOHSM kernel using $\gamma_{ij} = l_{ij}^{-1}$. This will allow us to express the global component of the MOHSM in a more amenable form to find models where γ_{ij} approaches zero rather than l_{ij} going to infinity.

5.2 Parameter initialization

The MOHSM kernel, akin to the MOSM and other spectral mixture kernels, is challenging to train as its likelihood is particularly susceptible to local maxima, and sensible to initial conditions: a poor choice of initial condition will lead to suboptimal solutions. In this context, we suggest a way of initializing the hyperparameters based on the interpretation of each parameter from a spectral viewpoint.

First, we set the centers and the length-scale of each component. We suggest, after choosing how many different centers consider, to place them equidistant and cover the input space with windows of the same length-scale. Now, to initialize the parameters of each regime we recommend a scheme based on the estimated power spectrum of the (windowed) data. Recently, [Cuevas, 2020] proposed an initialization scheme for the SM kernel and the MOSM kernel, which leads to consistent and better results. This concept estimates the power spectral density (PSD) of the available data to then use it to obtain initial values of the parameters, which have a spectral interpretation. Since the MOHSM is a spectral-inspired kernel, we can rely on the above idea to initialize its hyperparameters too. The MOGPTK toolkit [de Wolff et al., 2020] has implemented this initialization strategy for spectral mixture kernels, in particular, using the Bayesian non parametric spectral estimation method by [Tobar, 2018]. In our experience, the initial conditions of the time delay and phase parameters are best set to zero, thus making a initial assumption that there is no input-delay or phase-delay between channels, leaving to the optimization process to find the non-zero delay and phase if the data reveals so.

6 EXPERIMENTS

We tested the proposed MOHSM kernel in different settings. First we learned a synthetic three-output GP, then we applied MOHSM to two real-world scenarios:

a dataset comprising series of gold and oil prices, the NASDAQ and the USD index (henceforth referred to as GONU), and electroencephalography (EEG) data which are known to have dependencies among frequencies.

Since the MOSM kernel has shown better results than the others stationary MOGP kernels, such as the CONV and the CSM, and it is conceptually more general than them, we only considered MOSM as our stationary benchmark. Additionally, we compared MOHSM against 2 other non-stationary MOGP kernels: an independent non-stationary kernel per channel, and non-stationary linear model of coregionalization. The selected non-stationary kernel for both non-stationary MOGP kernels is defined as follow:

$$k(x,x') = \sum_{q=1}^{Q} w_q \exp\left(-\frac{1}{2l_q^2}\left\|\frac{x+x'}{2} - c_q\right\|^2\right) \exp\left(-\frac{1}{2}\tau^\top \Sigma_q \tau\right)\cos(\mu_q^\top \tau), \tag{13}$$

where $\tau = x - x'$.

The above defined kernel is from the family of harmonizable mixture kernels proposed by Shen et al. [Shen et al., 2019], which can be seen as a SM kernel where each component is windowed and centered in c_q. Therefore, we refer to the independent harmonizable HM (HSM) kernel per channel simply as HSM, and the non-stationary linear model of coregionalization using the HSM kernel is called HSM-LMC.

All models were implemented for our experiments based on the architecture of MOGPTK for GPU-accelerated ML-training of GPs [de Wolff et al., 2020]. Moreover, the MOHSM kernel is now part of MOGPTK, and the code for the experiments can be found in its Github repository.[2] The simulations were executed on a Intel Core i7 - 7500U 2.7 GHz CPU with 8 GB of RAM and a 940MX GPU.

6.1 Learning derivatives and delayed signals

This experiment demonstrates the expressiveness of the MOHSM by using it to recover the auto and cross covariance of a nonstationary MOGP. We simulated the following three signals: a sample f from a GP with a non stationary kernel and zero mean, its derivative and a delayed version of the GP. The relevance of this experiment stems from the fact that the true covariance and cross covariance of the mentioned process are known explicitly, thus we can test the performance of the model.

[2] https://github.com/GAMES-UChile/mogptk

We produced $N = 500$ samples in the interval $[-25, 25]$ for each channel. For the experiment, the derivative was computed numerically and we removed observations between $[-5, 5]$ for the derivative and between $[-5, 5]$ for the delayed signal. We randomly split the dataset into 70% for training and 30% for testing. We used a HSM kernel with 2 mixtures, one centered in -25, and the other centered in 25. In order to measure the performance of the models, we will use the distance between the correlation matrix, which is a metric defined by [Herdin et al., 2005] that measures the similarity between two covariance matrices, and is defined as follows:

Definition 4. *The correlation matrix distance (CMD) [Herdin et al., 2005] is the distance between two correlation matrices K_1 and K_2 as defined by*

$$CMD(K_1, K_2) = 1 - \frac{\mathrm{Tr}(K_1 \cdot K_2)}{\|K_1\| \cdot \|K_2\|}, \tag{14}$$

where the norm is the Frobenius norm and Tr *denotes the trace.*

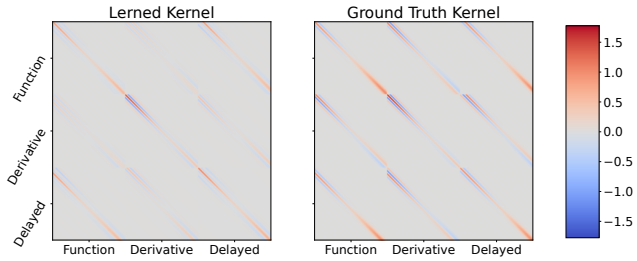

Figure 2: **Left:** learned covariance function by MOHSM kernel. **Right:** ground truth covariance function of the synthetic dataset.

Table 1: Performance indices for the synthetic dataset using the correlation matrix distance (CMD) over 5 realizations

Method	CMD
MOSM	0.85 ± 0.01
HSM	0.86 ± 0.00
HSM-LMC	0.80 ± 0.00
MOHSM	**0.48 ± 0.12**

Figure 2 shows the learned covariance function by MOHSM kernel compared to the ground truth covariance for the synthetic dataset. In this figure we observe that the MOHSM kernel is able to recover almost perfectly the ground truth kernel, learning both regimes and the delay of the third channel. Figure 3 shows that the MOHSM model was able to accurately recover the auto and cross covariance of the reference GP and the delayed version, yet it exhibits slightly larger error bars

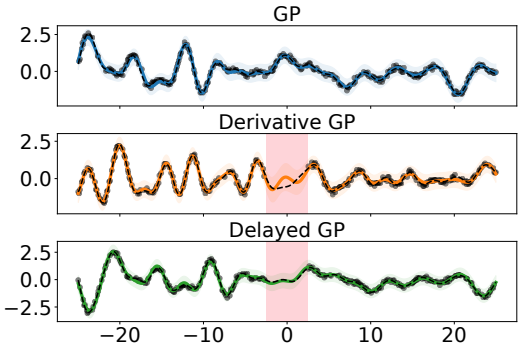

Figure 3: Synthetic data set with the trained **MOHSM** kernel. Training points are shown in black, dashed lines are the ground truth and the colour coded lines are the posterior means. The coloured bands show the 95% confidence intervals. The red shaded areas mark the data imputation ranges.

for the derivative. Among the 4 models considered, Table 1 confirms that the MOHSM largely outperforms the benchmarks in terms of the correlation matrix distance. Notably, in this example MOHSM achieved satisfactory results learning the cross-covariance between the GP and the delayed version, and autocovariances without prior information about the delays.

6.2 GONU dataset

In previous work, [de Wolff et al., 2020] proved the capability of MOSM to impute and predict financial observation by learning the relationships among financial time series. Although MOSM has performed successfully in these applications, the underlying assumption of stationarity is too strong, since financial series are known to be nonstationary. For example, [Joyeux, 1980] have found that, with new housing starts, the high and low frequency components were inter-correlated, and thus the series might be mismodeled by assuming stationarity.

In order to validate hypothesis on nonstationarity, we applied MOHSM to one of the [de Wolff et al., 2020] experiments, a dataset comprising series of gold and oil prices, the NASDAQ and the USD index, between January 2017 and December 2018 with a weekly frequency, so as to account for the known stylized facts of financial series[3]. To simulate missing data we removed regions in each channel. We trained a MOHSM kernel with 4 mixtures on 385 points, and tested on 446 points.

Figure 4 shows that the MOHSM gives an accurate fit to the dataset, where practically all the data are within the predicted confidence interval. Moreover, in the regions where the data were removed, the model was able to predict within an acceptable precision. The performance of the other methods can be found in the supplementary material.Table 2 shows a quantitative comparison of different models against the MOHSM. We performed 5 trials per trained model and reported the mean and the standard deviation of the mean absolute percentage error (MAPE) and root mean square error (RMSE)

[3]The dataset can be found in the websites: www.eia.gov/dnav/pet/hist/RBRTEd.htm, finance.yahoo.com & fred.stlouisfed.org/series/TWEXB

Table 2: Performance indices for the GONU dataset using the mean absolute percentage error (MAPE), root mean square error (RMSE) and negative log likelihood (NLL) over 5 realizations

Method	MAPE	RMSE	NLL
MOSM	3.05 ± 0.27	49.55 ± 5.79	-155.147 ± 9.09
HSM	3.20 ± 0.37	43.54 ± 2.67	-99.97 ± 3.58
HSM-LMC	2.49 ± 0.40	64.20 ± 12.11	-111.43 ± 20.84
MOHSM	**1.67 ± 0.16**	**40.44 ± 4.74**	**-164.83 ± 10.15**

6.3 EEG dataset

In the field of neuroscience, practitioners need to accurately model encephalography (EEG) data so as to correctly study brain states, in particular, the frequency-based perspective is the standard in multivariate EEG analysis [Gorrostieta et al., 2019]. Since the MOHSM is able to learn the interaction between frequencies, the proposed model is well suited to these challenge.

We tested MOHSM on an EEG dataset with 8 channels. We selected a 60-second window resampled at 2 [Hz] and randomly split the data set into 70% for training and 30% for testing. We trained on 735 points a MOHSM kernel with 4 mixture, and tested on 315 points.

Table 3 shows the results for the different models considered the EEG experiment, where we performed 5 trials per trained model and reported the mean and the standard deviation of the normalized mean absolute error (nMAE), for each channel.

7 CONCLUSION

We have presented the multi-output harmonizable spectral mixture (MOHSM) kernel as a generalization of the well-known multi-output spectral mixture (MOSM) kernel to the non-stationary case. The proposed family of kernels relies upon the concept of harmonizable processes, a rather general class of processes in the sense that it contains stationary processes and a large portion of the non-stationary processes. The resulting kernel, termed MOHSM, provides flexibility to model both stationary and non-stationary processes while maintaining

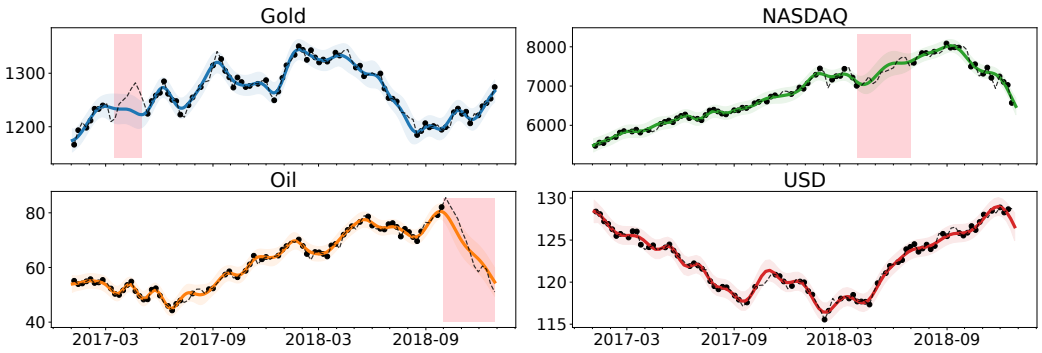

Figure 4: GONU data set with the trained MOHSM kernel. Training points are shown in black, dashed lines are the ground truth and the colour coded lines are the posterior means. The coloured bands show the 95% confidence intervals. The red shaded areas mark the data imputation ranges

Model	Fp1	Fp2	Fz	Cz	T3	T4	O1	O2	**Overall**
MOSM	0.16 ± 0.01	0.12 ± 0.01	0.13 ± 0.01	0.13 ± 0.01	0.25 ± 0.04	0.17 ± 0.02	**0.25 ± 0.04**	**0.12 ± 0.01**	0.17 ± 0.02
HSM	0.18 ± 0.00	0.13 ± 0.01	**0.12 ± 0.01**	**0.12 ± 0.01**	0.43 ± 0.01	0.23 ± 0.01	0.30 ± 0.00	0.26 ± 0.01	0.25 ± 0.01
HSM-LMC	0.15 ± 0.01	**0.10 ± 0.01**	0.14 ± 0.01	0.13 ± 0.01	0.16 ± 0.02	0.17 ± 0.01	0.39 ± 0.15	0.15 ± 0.05	0.17± 0.01
MOHSM	**0.15 ± 0.01**	0.11 ± 0.02	**0.12 ± 0.00**	**0.12 ± 0.00**	**0.23 ± 0.04**	**0.16 ± 0.01**	0.30 ± 0.02	**0.12 ± 0.01**	**0.16 ± 0.01**

Table 3: Performance for the EEG dataset of each channel using the normalized mean absolute error (nMAE) over 5 realisations

the desired properties of the MOSM: a clear interpretation of the parameters from a spectral viewpoint, and flexibility in each channel. Furthermore, we have implemented a parameter initialization scheme to overcome the sensibility of the MOHSM to initial conditions. We have showed that our method can effectively model non-stationary data and is a sound extension of the MOSM kernel. Future work includes considering more complex spectral densities instead of Gaussian functions, this would allow us to prescind of the infinite differentiability requirement of sampled functions assumed by spectral mixture kernels. A sparse implementation of MOSHM is also part of the future work, yet that is a challenge in its own since the relationship between the locations of inducing inputs in the multichannel case has not been thoroughly studied so far. Finally, we hope our work catalyzes interest in the harmonizable processes and their role on multichannel models.

Acknowledgments

We thank Taco de Wolff for his advice using the MOGPTK toolbox and Jou-Hui Ho for insightful discussions about the EEG experiment. This work was funded by Google, Fondecyt-Regular 1210606, ANID-FB210005 (CMM) and ANID-FB0008 (AC3E).

References

[Álvarez and Lawrence, 2009] Álvarez, M. and Lawrence, N. (2009). Sparse convolved Gaussian processes for multi-output regression. In *Advances in Neural Information Processing Systems*, pages 57–64.

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

[Bochner et al., 1959] Bochner, S. et al. (1959). *Lectures on Fourier Integrals*, volume 42. Princeton University Press.

[Bonilla et al., 2007] Bonilla, E. V., Chai, K., and Williams, C. (2007). Multi-task Gaussian process prediction. *Advances in neural information processing systems*, pages 153–160.

[Cramér, 1940] Cramér, H. (1940). On the theory of stationary random processes. *Annals of Mathematics*, pages 215–230.

[Cranstoun et al., 2002] Cranstoun, S. D., Ombao, H. C., Von Sachs, R., Guo, W., and Litt, B. (2002). Time-frequency spectral estimation of multichannel eeg using the auto-slex method. *IEEE Transactions on Biomedical Engineering*, pages 988–996.

[Cuevas, 2020] Cuevas, A. (2020). Multi-output gaussian process toolkit with sparse formulation for spectral kernels. Master's thesis, Universidad de Chile.

[de Wolff et al., 2020] de Wolff, T., Cuevas, A., and Tobar, F. (2020). Gaussian process imputation of

multiple financial series. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8444–8448.

[de Wolff et al., 2020] de Wolff, T., Cuevas, A., and Tobar, F. (2020). MOGPTK: The Multi-Output Gaussian Process Toolkit. *Neurocomputing*.

[Goovaerts et al., 1997] Goovaerts, P. et al. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press on Demand.

[Gorrostieta et al., 2019] Gorrostieta, C., Ombao, H., and Von Sachs, R. (2019). Time-dependent dual-frequency coherence in multivariate non-stationary time series. *Journal of Time Series Analysis*, pages 3–22.

[Herdin et al., 2005] Herdin, M., Czink, N., Ozcelik, H., and Bonek, E. (2005). Correlation matrix distance, a meaningful measure for evaluation of non-stationary mimo channels. In *IEEE 61st Vehicular Technology Conference*, pages 136–140.

[Hurd, 1973] Hurd, H. (1973). Testing for harmonizability. *IEEE Transactions on Information Theory*, 19:316–320.

[Joyeux, 1980] Joyeux, R. (1980). *Harmonizable Processes and Their Applications to the Estimation of Interactions Between Frequences for Non-stationary Economic Processes*. Cornell University Press.

[Kakihara, 1997] Kakihara, Y. (1997). *Multidimensional Second Order Stochastic Processes*, volume 2. World Scientific.

[Kim and Kim, 2005] Kim, I. K. and Kim, Y. Y. (2005). Damage size estimation by the continuous wavelet ridge analysis of dispersive bending waves in a beam. *Journal of Sound and Vibration*, pages 707–722.

[Lilly and Olhede, 2011] Lilly, J. M. and Olhede, S. C. (2011). Analysis of modulated multivariate oscillations. *IEEE Transactions on Signal Processing*, pages 600–612.

[Loeve, 1978] Loeve, M. (1978). *Probability Theory II*. Springer.

[Ombao et al., 2001] Ombao, H. C., Raz, J. A., von Sachs, R., and Malow, B. A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association*, pages 543–560.

[Parra and Tobar, 2017] Parra, G. and Tobar, F. (2017). Spectral mixture kernels for multi-output gaussian processes. In *Advances in Neural Information Processing Systems*, pages 6684–6693.

[Remes et al., 2017] Remes, S., Heinonen, M., and Kaski, S. (2017). Non-stationary spectral kernels. In *Advances in Neural Information Processing Systems*, pages 4645–4654.

[Samo, 2017] Samo, Y.-L. K. (2017). *Advances in kernel methods: towards general-purpose and scalable models*. PhD thesis, University of Oxford.

[Samo and Roberts, 2015] Samo, Y.-L. K. and Roberts, S. (2015). Generalized spectral kernels. *arXiv preprint arXiv:1506.02236*.

[Shen et al., 2019] Shen, Z., Heinonen, M., and Kaski, S. (2019). Harmonizable mixture kernels with variational Fourier features. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 3273–3282.

[Silverman, 1957] Silverman, R. (1957). Locally stationary random processes. *IRE Transactions on Information Theory*, 3:182–187.

[Teh et al., 2005] Teh, Y. W., Seeger, M., and Jordan, M. I. (2005). Semiparametric latent factor models. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, pages 333–340. PMLR.

[Tobar, 2018] Tobar, F. (2018). Bayesian nonparametric spectral estimation. In *Advances in Neural Information Processing Systems*, pages 10148–10158.

[Ulrich et al., 2015] Ulrich, K., Carlson, D. E., Dzirasa, K., and Carin, L. (2015). GP kernels for cross-spectrum analysis. In *Advances in Neural Information Processing Systems*, pages 1999–2007.

[Williams and Rasmussen, 2006] Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA.

[Wilson and Adams, 2013] Wilson, A. and Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *International conference on machine learning*, pages 1067–1075. PMLR.

[Yaglom, 1987] Yaglom, A. M. (1987). Correlation theory of stationary and related random functions. *Volume I: Basic Results.*, 526.

[Zhang et al., 2003] Zhang, Z., Ren, Z., and Huang, W. (2003). A novel detection method of motor broken rotor bars based on wavelet ridge. *IEEE Transactions on Energy Conversion*, pages 417–423.

Supplementary Material: Nonstationary multi-output Gaussian processes via harmonizable spectral mixtures

A DERIVATION OF THE MOHSM KERNEL

Consider the following cross spectral density:

$$S_{ij} = w_{ij} e^{\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2\right)} e^{\left(-\frac{1}{2}(\overline{\omega}-\mu_{ij})^\top \Sigma_{ij}^{-1}(\overline{\omega}-\mu_{ij})\right)} e^{(-i(\theta_{ij}^\top \overline{\omega}+\phi_{ij}))} \underbrace{e^{(-i\hat{\omega}^\top x_p)}}_{\text{(input shift)}}, \tag{15}$$

where $\hat{\omega} = \omega - \omega', \overline{\omega} = \frac{\omega+\omega'}{2}$, $w_{ij}, \phi_{ij} \in \mathbb{R}$, $\theta_{ij}, \mu_{ij}, x_p \in \mathbb{R}^n$ and $\Sigma_{ij} = \text{diag}([\sigma_{ij1}^2, ..., \sigma_{ijn}^2]) \in \mathbb{R}^{n\times n}$. We calculate the inverse generalized Fourier transform of the spectral densities $S_{ij}(\omega, \omega')$ above to obtain the multivariate covariance function:

$$\begin{aligned}
k_{ij}(x,x') &= \iint_{\mathbb{R}^n\times\mathbb{R}^n} e^{i(\omega^\top x - \omega'^T)} S_{ij}(\omega,\omega') d\omega d\omega' \\
&= \iint_{\mathbb{R}^n\times\mathbb{R}^n} e^{i\left(\left(\frac{\omega+\omega'}{2}\right)^\top (x-x') + (\omega-\omega')^\top \left(\frac{x+x'}{2}\right)\right)} S_{ij}(\omega,\omega') d\omega d\omega' \\
&= \iint_{\mathbb{R}^n\times\mathbb{R}^n} e^{i\left(\overline{\omega}^\top \tau + \hat{\omega}^\top \overline{x}\right)} S_{ij}(\omega,\omega') d\omega d\omega' \quad \left(\text{defining } \tau = x - x' \text{ and } \overline{x} = \frac{x+x'}{2}\right) \\
&= \iint_{\mathbb{R}^n\times\mathbb{R}^n} e^{i\left(\overline{\omega}^\top \tau + \hat{\omega}^\top \overline{x}\right)} w_{ij} e^{\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2\right)} e^{\left(-\frac{1}{2}(\overline{\omega}-\mu_{ij})^\top \Sigma_{ij}^{-1}(\overline{\omega}-\mu_{ij})\right)} e^{(-i(\theta_{ij}^\top \overline{\omega}+\phi_{ij}))} e^{(-i\hat{\omega}^\top x_p)} d\omega d\omega' \\
&= w_{ij} \iint_{\mathbb{R}^n\times\mathbb{R}^n} e^{\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2 + i\hat{\omega}^\top(\overline{x}-x_p)\right)} e^{\left(-\frac{1}{2}(\overline{\omega}-\mu_{ij})^\top \Sigma_{ij}^{-1}(\overline{\omega}-\mu_{ij}) - i(\theta_{ij}^\top \overline{\omega}+\phi_{ij}) + \overline{\omega}^\top \tau\right)} d\omega d\omega' \\
&= w_{ij} \iint_{\mathbb{R}^n\times\mathbb{R}^n} e^{\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2 + i\hat{\omega}^\top(\overline{x}-x_p)\right)} e^{\left(-\frac{1}{2}(\overline{\omega}-\mu_{ij})^\top \Sigma_{ij}^{-1}(\overline{\omega}-\mu_{ij}) - i(\theta_{ij}^\top \overline{\omega}+\phi_{ij}) + \overline{\omega}^\top \tau\right)} d\overline{\omega} d\hat{\omega} \\
&= w_{ij} \int_{\mathbb{R}^n} e^{\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2 + i\hat{\omega}^\top(\overline{x}-x_p)\right)} d\hat{\omega} \int_{\mathbb{R}^n} e^{\left(-\frac{1}{2}(\overline{\omega}-\mu_{ij})^\top \Sigma_{ij}^{-1}(\overline{\omega}-\mu_{ij}) - i(\theta_{ij}^\top \overline{\omega}+\phi_{ij}) + \overline{\omega}^\top \tau\right)} d\overline{\omega} \\
&= w_{ij} \int_{\mathbb{R}^n} e^{\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2 + i\hat{\omega}^\top(\overline{x}-x_p)\right)} d\hat{\omega} \int_{\mathbb{R}^n} e^{\left(-\frac{1}{2}\overline{\omega}^\top \Sigma_{ij}^{-1}\overline{\omega} - (\Sigma_{ij}^{-1}\mu_{ij} + i(\tau+\theta_{ij}))^\top \overline{w} - \frac{1}{2}\mu_{ij}^\top \Sigma_{ij}^{-1}\mu_{ij} + i\phi_{ij}\right)} d\overline{\omega} \\
&= \alpha_{ij} e^{-\frac{l_{ij}^2}{2}\|\overline{x}-x_p\|^2} e^{\left(\frac{1}{2}(\Sigma_{ij}^{-1}\mu_{ij} + i(\tau+\theta_{ij}))^\top \Sigma_{ij} (\Sigma_{ij}^{-1}\mu_{ij} + i(\tau+\theta_{ij})) - \frac{1}{2}\mu_{ij}^\top \Sigma_{ij}^{-1}\mu_{ij} + i\phi_{ij}\right)} \\
&= \alpha_{ij} e^{-\frac{l_{ij}^2}{2}\|\overline{x}-x_p\|^2} e^{\left(-\frac{1}{2}(\tau+\theta_{ij})^\top \Sigma_{ij}(\tau+\theta_{ij})\right)} e^{\left(i(\tau+\theta_{ij})^\top \mu_{ij} + \phi_{ij}\right)}.
\end{aligned}$$

Now, in order to obtain a real-valued covariance function $k_{ij}(x,x')$ we take the real part of the above covariance, which is equivalent to *symmetrize* the spectral densities $S_{ij}(\omega,\omega')$. Thus, we set:

$$k_{ij}(x,x') = \alpha_{ij} e^{-\frac{l_{ij}^2}{2}\|\overline{x}-x_p\|^2} e^{\left(-\frac{1}{2}(\tau+\theta_{ij})^\top \Sigma_{ij}(\tau+\theta_{ij})\right)} \cos\left(i(\tau+\theta_{ij})^\top \mu_{ij} + \phi_{ij}\right).$$

B RECOVERING THE MOSM KERNEL FROM THE MOHSM KERNEL

Since our goal is to extend the multi-output spectral mixture (MOSM) to the non-stationary case, we study in more detail how to recover the MOSM from the proposed multi-output harmonizable spectral mixture (MOHSM).

We notice that considering $x_p = 0$ and taking $l_{ij} \to 0$ in MOHSM we recover the MOSM kernel, successfully extending the stationary model. Indeed, in the frequency domain we notice that when $\omega \neq \omega'$:

$$\exp\left(-\frac{1}{2l_{ij}^2}\|\omega - \omega'\|^2\right) \overset{l_{ij}\to 0}{\longrightarrow} 0. \tag{16}$$

On the other hand in the case $\omega = \omega'$ we observe that:

$$\exp\left(-\frac{1}{2l_{ij}^2}\|\omega - \omega'\|^2\right) \overset{l_{ij}\to 0}{\longrightarrow} 1, \tag{17}$$

Combining equations (16) and (17) we obtain:

$$S_{ij}(\omega.\omega') \overset{l_{ij}\to 0}{\longrightarrow} \delta(\omega - \omega')\exp\left(-\frac{1}{2}(\overline{\omega} - \mu_{ij})^\top \Sigma_{ij}^{-1}(\overline{\omega} - \mu_{ij})\right)\exp(-i(\theta_{ij}^\top \overline{\omega} + \phi_{ij})) = \hat{S}(\omega, \omega'), \tag{18}$$

where $\delta(\cdot)$ is the Kronecker delta. This can be seen as no correlation between frequencies, which is the supposition of stationary. Moreover, the above equation is equivalent to the cross-spectral densities of the MOSM, and calculating the inverse generalised Fourier transform of these cross-spectral densities we obtain:

$$\begin{aligned}
\hat{k}(x, x') &= \iint_{\mathbb{R}^n \times \mathbb{R}^n} e^{i(\omega^\top x - \omega'^{Top})}\hat{S}_{ij}(\omega, \omega')d\omega d\omega' \\
&= \iint_{\mathbb{R}^n \times \mathbb{R}^n} e^{i(\omega^\top x - \omega'^{Top})}\delta(\omega - \omega')e^{\left(-\frac{1}{2}(\overline{\omega} - \mu_{ij})^\top \Sigma_{ij}^{-1}(\overline{\omega} - \mu_{ij})\right)}e^{(-i(\theta_{ij}^\top \overline{\omega} + \phi_{ij}))}d\omega d\omega' \\
&= \int_{\mathbb{R}^n} e^{i\omega^\top (x - x')}e^{\left(-\frac{1}{2}(\omega - \mu_{ij})^\top \Sigma_{ij}^{-1}(\omega - \mu_{ij})\right)}e^{(-i(\theta_{ij}^\top \omega + \phi_{ij}))}d\omega \\
&= \alpha_{ij}e^{\left(-\frac{1}{2}(\tau + \theta_{ij})^\top \Sigma_{ij}(\tau + \theta_{ij})\right)}e^{\left(i(\tau + \theta_{ij})^\top \mu_{ij} + \phi_{ij}\right)}.
\end{aligned}$$

Taking the real part of the above expression we recover the MOSM kernel. Thus, the MOHSM kernel successfully extend the MOSM kernel.

C EXPERIMENT DETAILS AND ADDITIONAL FIGURES

C.1 Learning derivatives and delayed signals

we demonstrate the expressiveness of the MOHSM by using it to recover the auto and cross covariance of a MOGP. We considered an MOGP with the following three components: a sample f from a GP with a non stationary kernel and zero mean, its derivative and a delayed version of the GP. This experiment is very illustrative since the covariance and cross covariance of the mentioned process are known explicitly, thus we can test the expressiveness of the model, namely

Proposition 2. *Let f be a Gaussian process with covariance function k, then the derivative stochastic process f' is also a Gaussian process and its covariance function is $\frac{\partial^2 k(x,x')}{\partial x \partial x'}$. Furthermore, $(f(x), f'(x))$ form a two-channel Multi-Output Gaussian process [Williams and Rasmussen, 2006] with the following multivariate covariance function*

$$\mathcal{K}(x, x') = \begin{pmatrix} k(x, x') & \frac{\partial k(x,x')}{\partial x'} \\ \frac{\partial k(x,x')}{\partial x} & \frac{\partial^2 k(x,x')}{\partial x \partial x'} \end{pmatrix} \tag{19}$$

C.2 GONU dataset

In this section we present extra figures for the GONU experiment. We show the performance of each method with which we compare the MOHSM.

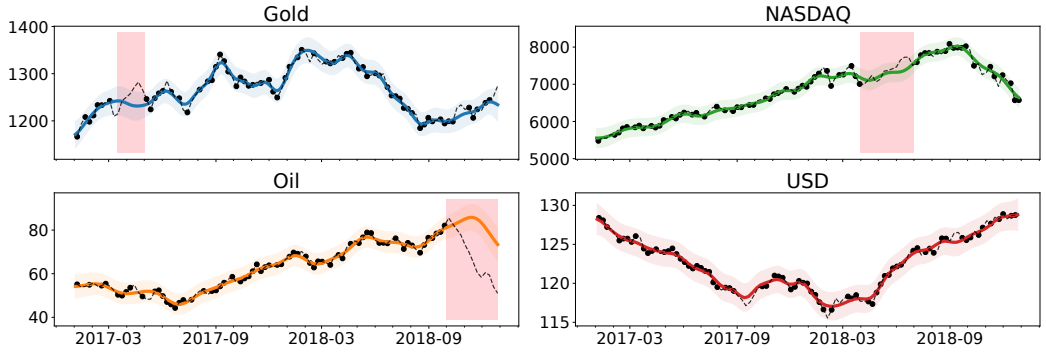

Figure 5: GONU data set with the trained **MOSM** kernel. Training points are shown in black, dashed lines are the ground truth and the colour coded lines are the posterior means. The coloured bands show the 95% confidence intervals. The red shaded areas mark the data imputation ranges

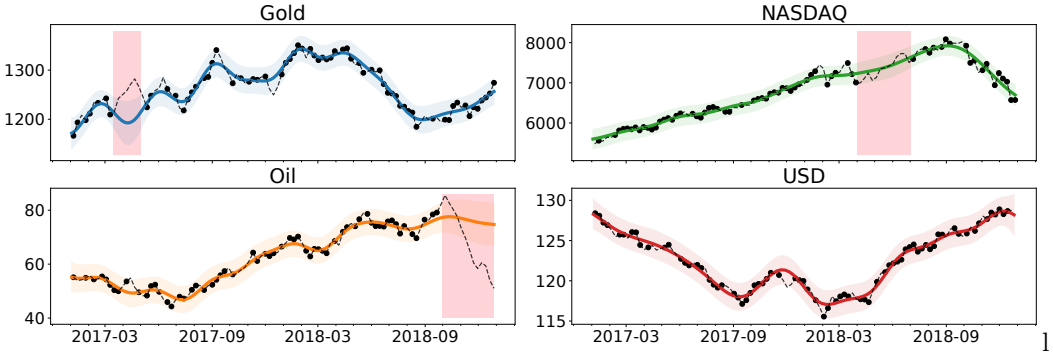

Figure 6: GONU data set with the trained **HSM-LMC** kernel. Training points are shown in black, dashed lines are the ground truth and the colour coded lines are the posterior means. The coloured bands show the 95% confidence intervals. The red shaded areas mark the data imputation ranges

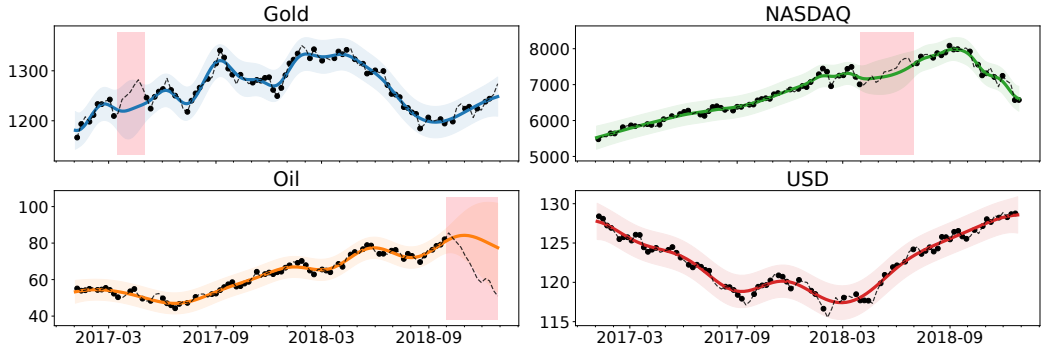

Figure 7: GONU data set with the trained **HSM** kernel. Training points are shown in black, dashed lines are the ground truth and the colour coded lines are the posterior means. The coloured bands show the 95% confidence intervals. The red shaded areas mark the data imputation ranges

C.3 EEG dataset

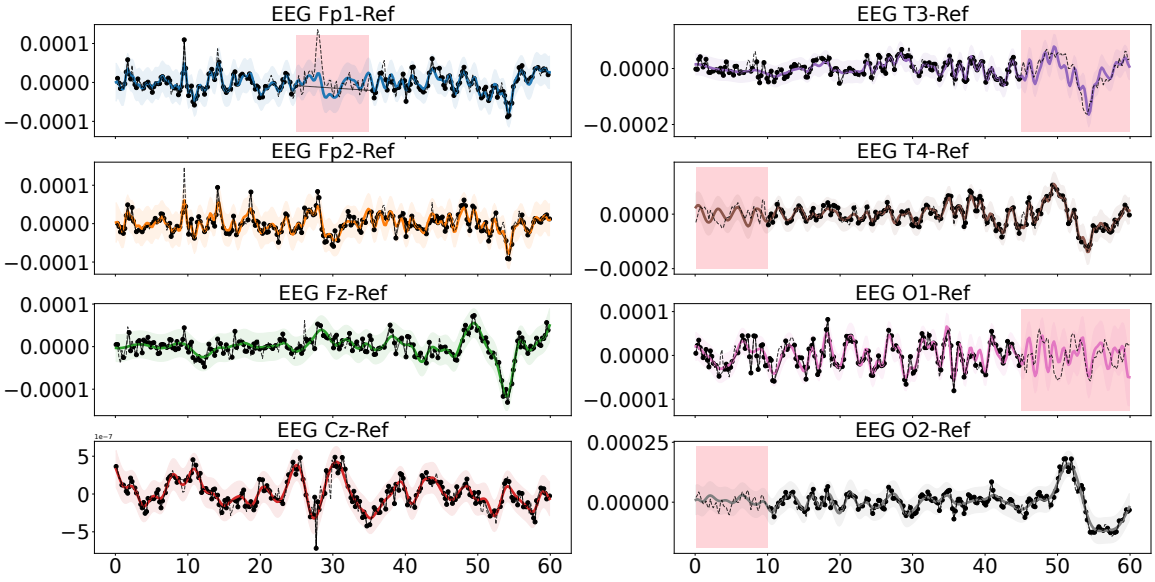

Figure 8: Extrapolation example using **MOHSM** kernel on EEG: training points (black), ground truth (dashed lines), and posterior means with 95% error bars (color). The red shades denote data imputation ranges.

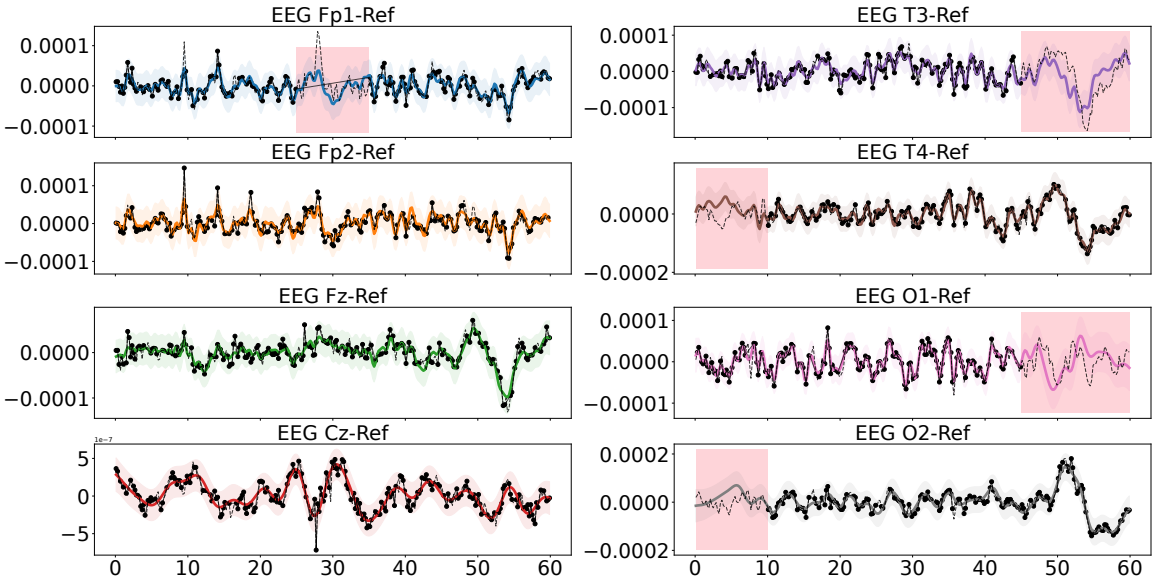

Figure 9: Extrapolation example using **MOSM** kernel on EEG: training points (black), ground truth (dashed lines), and posterior means with 95% error bars (color). The red shades denote data imputation ranges.

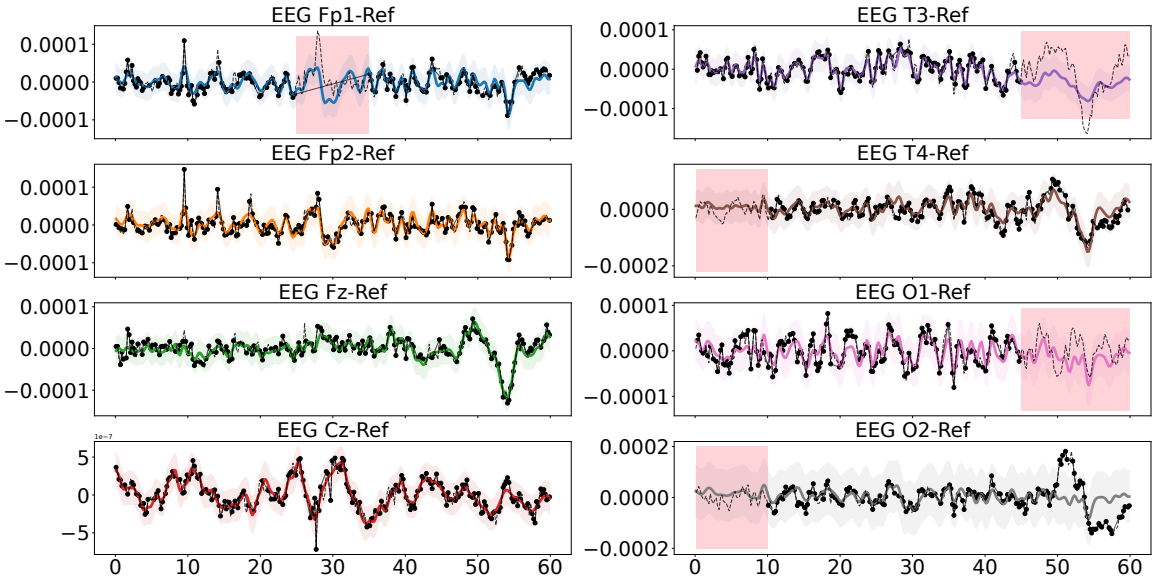

Figure 10: Extrapolation example using **HSM-LMC** kernel on EEG: training points (black), ground truth (dashed lines), and posterior means with 95% error bars (color). The red shades denote data imputation ranges.

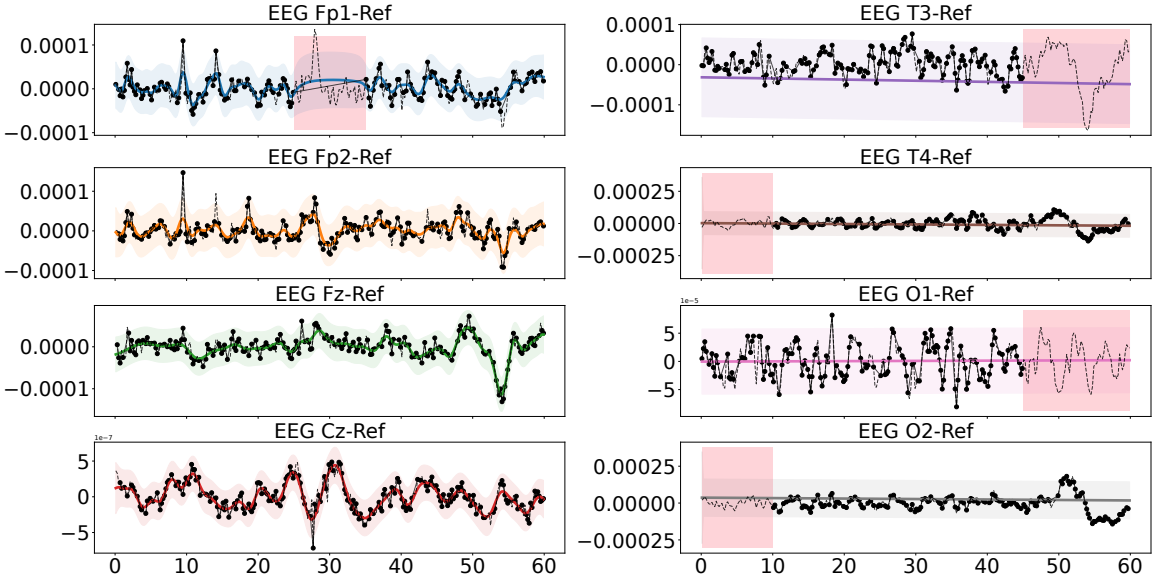

Figure 11: Extrapolation example using **HSM** kernel on EEG: training points (black), ground truth (dashed lines), and posterior means with 95% error bars (color). The red shades denote data imputation ranges.