UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

NONSTATIONARY MULTI-OUTPUT GAUSSIAN PROCESSES VIA HARMONIZABLE SPECTRAL MIXTURES

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN MATEMÁTICAS APLICADAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

MATÍAS IGNACIO ALTAMIRANO MONTERO

PROFESOR GUÍA:
FELIPE TOBAR HENRÍQUEZ

MIEMBROS DE LA COMISIÓN:
ABELINO JIMÉNEZ GAJARDO
JOAQUÍN FONTBONA TORRES

Este trabajo ha sido parcialmente financiado por:
CMM ANID PIA AFB170001 y Fondecyt-Iniciación 11171165

SANTIAGO DE CHILE
2021

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA,
MENCIÓN MATEMÁTICAS APLICADAS
RESUMEN DE LA MEMORIA PARA OPTAR
AL TITULO DE INGENIERO CIVIL MATEMÁTICO
POR: MATÍAS IGNACIO ALTAMIRANO MONTERO
FECHA: 2021
PROF. GUÍA: FELIPE TOBAR HENRÍQUEZ

NONSTATIONARY MULTI-OUTPUT GAUSSIAN PROCESSES VIA HARMONIZABLE SPECTRAL MIXTURES

En varias disciplinas, los procesos Gaussianos (GP) [1] son el referente para modelar series de tiempo o funciones en general, especialmente en los casos en los que se requiere modelar la incertidumbre. La extensión de GP para manejar múltiples salidas se conoce como procesos Gaussianos de múltiples salidas (MOGP) [2], que, al modelar los canales como Gaussianas conjuntas, puede compartir información entre las salidas, mejorando potencialmente la estimación. Tanto los GPs como los MOGP están completamente determinados por una función de covarianza.

El diseño de kernel para MOGP ha recibido mayor atención recientemente, en particular, el enfoque del *Multi-Output Spectral Mixture* (MOSM) [3] kernel ha sido elogiado como un modelo general en el sentido de que se extiende a otros enfoques como el *Linear Model of Corregionalization* [4], el *Intrinsic Corregionalization Model* [4] y el *Cross-SpectralMixture* [5]. El MOSM se basa en el teorema de Cramér [6] para parametrizar las densidades espectrales de potencia (PSD) como una mezcla Gaussiana, por lo que tiene una restricción estructural: asumiendo la existencia de una PSD, el método solo es adecuado para procesos estacionarios.

El objetivo principal de esta tesis es ampliar el MOSM para hacerlo adecuado para procesos no estacionarios. Para abordar esto, proponemos el multi-output harmonizable spectralmixture (MOHSM) kernel, una familia expresiva y flexible de kernels de MOGP para modelar procesos no estacionarios como una extensión natural del MOSM que se basa en el concepto de armonizabilidad, término introducido por Loève [7] que generaliza las representaciones espectrales a procesos no estacionarios. El MOHSM es capaz de modelar procesos tanto estacionarios como no estacionarios mientras mantiene las propiedades deseadas del MOSM: una interpretación clara de los parámetros, desde un punto de vista espectral, y flexibilidad en cada canal. También proponemos una heurística basada en datos para el punto inicial de la optimización, con el fin de mejorar el entrenamiento del modelo. Además, presentamos una variación del modelo, la cual abre la puerta a considerar densidades espectrales más generales. El método propuesto se valida primero en datos sintéticos con el propósito de ilustrar las propiedades clave de nuestro enfoque, y luego se compara con los métodos MOGP existentes en dos escenarios del mundo real, finanzas y electroencefalografía.

RESUMEN DE LA TESIS PARA OPTAR
AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA,
MENCIÓN MATEMÁTICAS APLICADAS
RESUMEN DE LA MEMORIA PARA OPTAR
AL TITULO DE INGENIERO CIVIL MATEMÁTICO
POR: MATÍAS IGNACIO ALTAMIRANO MONTERO
FECHA: 2021
PROF. GUÍA: FELIPE TOBAR HENRÍQUEZ

NONSTATIONARY MULTI-OUTPUT GAUSSIAN PROCESSES VIA HARMONIZABLE SPECTRAL MIXTURES

In several disciplines, Gaussian Processes (GPs)[1] are the gold standard for modelling time series or functions in general, especially in cases where modelling uncertainty is required. The extension of GP to handle multiple outputs is known as multi-output Gaussian processes (MOGP)[2], which, by modelling the outputs as jointly Gaussian, is able to share information across outputs, potentially improving the estimation. Both, the single output and multiple output GP are entirely determined by a covariance function.

Kernel design MOGP has received increased attention recently, in particular, the Multi-Output Spectral Mixture kernel (MOSM) [3] approach has been praised as a general model in the sense that it extends other approaches such as Linear Model of Coregionalization [4], Intrinsic Coregionalization Model [4] and Cross-Spectral Mixture [5]. MOSM relies on Cramér's theorem [6] to parametrize 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 stationary processes.

The main purpose of this thesis is to extend the MOSM model to make it suitable for non-stationary processes. To address this, we propose the multi-output harmonizable spectral mixture (MOHSM) kernel, an expressive and flexible family of MOGP kernels to model non-stationary processes as a natural extension of the MOSM. which relies on the concept of harmonizability, a term introduced by Loève [7] which generalizes the spectral representations to non-stationary processes. The proposed MOHSM is able to model both stationary and non-stationary processes while maintaining the desires properties of the MOSM: a clear interpretation of the parameters, from a spectral viewpoint, and flexibility in each channel. We also propose a data-driven heuristics for the initial point in the optimization, in order to improve the model training. In addition, we also present a variation of the model, which open the door to consider more more general spectral densities. 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.

“Elen síla lúmenn’ omentielvo”
— J.R.R. Tolkien

Agradecimientos

A mi familia, que me han apoyado siempre en todas las decisiones que he tomado, sin ese apoyo no estaría donde estoy ahora. Gracias por motivarme a perseguir mis sueños y siempre apuntar a lo más alto. Gracias por todo en la vida.

A Hernán y Joao, mis amigos de toda la vida. Su amistad ha sido fundamental para convertirme en la persona que soy y lograr las cosas que he conseguido, aún recuerdo cuando eramos chicos y soñabamos con estudiar en la Chile. Gracias por cada consejo, cada desahogo y cada buen momento que hemos compartido.

A Catalina, por estar ahí y hacerme feliz en gran parte de mi vida universitaria. Soy inmensamente feliz por haberme encontrado con una persona como tu. Gracias por tu amor y apoyo en todo momento y sobre todo en los momentos díficiles, sin ti esta tesis hubiese sido imposible.

A Javier y Daniel, sin ustedes la universidad no hubiese sido lo mismo. Gracias por todos los buenos momentos y experiencias que hemos vivimos juntos, sé que vendrán muchos más.

A los cabros: Feña, Freddy, Juan, Pablo, Lete, PL, Quique, Guillaume, Danner, sin ustedes la carrera hubiese sido imposible. Gracias por cada momento vivido, desde las tardes/noches de estudio, los partidos, hasta los carretes.

A cada persona que nombré acá, los quiero mucho.

Table of Contents

Introduction

One of the most frequent problem in the machine learning community is the regression problem, which consist on estimate the relationship between inputs, called the independent variables, and outputs, called the dependent variables. More specifically, regression helps us to understand how the value of the dependent variables is changing corresponding to the independent variables. The input can be a wide range of quantities with arbitrary dimension but usually is the time and/or space. On the other hand, the output can be scalar or vector continuous quantities, which are called univariate or multivariate regression respectively.

From a Bayesian point of view, the regression problem consists in placing a prior distribution over the estimators, then, in conjunction with the likelihood function for the observe data, and using the Bayes' rule of probability, we obtain the posterior distribution which can be use for prediction and forecasting. Within this context, Gaussian processes (GP) [1] provide a flexible and powerful non-parametric framework for Bayesian regression, due to their properties such as a closure of the posterior distribution under a Gaussian data likelihood. The main aspect on the design of the GP is the choice of the covariance function, also called kernel, which encapsulate all the properties of the process such as: smoothness periodicity and stationarity, among others.

Unfortunately, the GP framework is restricted to the univariate case, making unsuitable for problems in which it is necessary to learn multiple tasks simultaneously. The extension of GP to handle multiple outputs is known as multi-output Gaussian processes (MOGP) [2], which, by modelling the outputs as jointly Gaussian, 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 the core aspect of MOGP.

There have been several approaches to design valid kernels to MOGP [8, 9, 4], 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 parametrization of multioutput covariances thus failing to provide model interpretation, specially from a spectral analysis perspective. Recently, Parra et al.[3] proposed the multi-output spectral mixture (MOSM) which directly designs the kernel in the spectral domain, using the multivariate version of Bochner's theorem [10], namely Cramér's Theorem [6].

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 a vast of real world problems, like in vibratory signals [11, 12], free-drifting oceanic instruments [13], various neuroscience applications [14, 15], and econometric [16]. Therefore, a flexible non-stationary multioutput kernel becomes necessary and in particular a non-stationary version of the MOSM kernel.

The main purpose of this thesis is to propose an expressive and flexible family of MOGP kernels to model non-stationary processes as a natural extension of the MOSM, which is achieved through the concept of harmonizability, a term introduced by Loève [7] which generalizes the spectral representations to non-stationary processes. The harmonizable processes have been widely studied and developed in the statistical community, but there has been a lack of attentiveness in the machine learning community.

The remainder of this thesis is organized as follows. We revisit all the required concept to support our proposal in Chapter 1, starting from single output GP, going through MOGP, previous approaches of MOGP, and finalizing with harmonizable processes. In Chapter 2, we present our MOGP kernel, then we compare it to previous approaches, and we explain practical considerations of the proposed model. Then, in Chapter 3 we test our kernel on synthetic and real-world data. Finally, in Conclusion Chapter we discuss our results and summarise our contribution and future work .

Contributions

Furthermore, the main contributions of this thesis are the following:

- To propose a kernel able to model both non-stationary processes as stationary processes while keeping the desires properties of the MOSM, called Multi-Output Harmonizable Spectral Mixture (MOHSM) kernel. Detailed in section 2.1, the proposed kernel base on the concept of harmonizable processes to extend the MOSM kernel to non-stationary processes.

- To suggest an parameter initialization scheme for the MOHSM kernel to overcome one of the MOHSM problem, which is the sensitivity of the optimization. Detailed in section 2.4, the proposed initialization method use the spectral interpretation of the kernel to initialize the parameters.

- To propose possible variations of the MOHSM kernel opening the door for a wider class of processes to model. Detailed in section 2.5, we study possible variation of the proposed framework by using symmetric rectangle function as the global component of the spectral density, instead of squared exponential functions.

- To validate experimentally the proposed MOHSM kernel in real world applications: EEG and financial time series. Detailed in section 3.2 and section 3.3, the proposed kernel is compared against previous MOGP frameworks.

Chapter 1

Background

The objective of this chapter is to introduce all the necessary concepts for the work developed in the thesis. First, we review Gaussian Processes (GP) methods for regression, describing the univariate case as well as the kernel section problem that they pose and the proposed solution using spectral representations. Then, we introduce the multivariate extension of the Gaussian Processes framework called Multi-Output Gaussian Processes (MOGP), revisiting the state-of-art models for MOGP. Finally, we study the harmonizable processes a general class of processes that are the cornerstone of our work.

1.1. Gaussian Processes

A Gaussian Process [1] is a Bayesian non-parametric generative model for functions $f : \mathcal{X} \to \mathbb{R}$. 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 variables. Formally,

Definition 1.1 *A Gaussian process is a stochastic process $\{f(x) : x \in \mathcal{X}\}$ such that for every finite subset $X = \{x_i\}_{i=1}^{N}$ of $\mathcal{X}$, the random vector $f(X) := [f(x_1), ..., f(x_N)]$ is a multivariate Gaussian random variable.*

A GP model, before conditioning on observed data, is completely specified by its mean function $m(x)$ and covariance (also called kernel) function $k(x, x')$, defined as follow for a stochastic process $f(x)$

$$m(x) = \mathbb{E}[f(x)], \quad \forall x \in \mathcal{X} \tag{1.1}$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))], \quad \forall x, x' \in \mathcal{X} \tag{1.2}$$

and we will write the Gaussian process as

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')), \tag{1.3}$$

where the mean function is usually assumed equal to zero, $m(x) = 0$. 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.

Usually the kernel is selected to have a parametric form which depends on a set of parameters θ, that is $k(\cdot,\cdot) = k_\theta(\cdot,\cdot)$, which allow us to control the structure of the GP. However, not all function are a valid kernel, nevertheless a necessary and sufficient condition for a symmetric function $k(x, x')$ to be a kernel is that it be positive-definite, i.e. $\forall\{x_1, ..., x_N\} \subset \mathcal{X}$, $\forall\{c_1, ..., c_N\} \subset \mathbb{R}$ and $\forall N \in \mathbb{N}$:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j k(x_i, x_j) \geq 0, \tag{1.4}$$

which equivalent that for any N points the $N \times N$ matrix given by $k(x, x)$ is positive-semidefinite.

1.1.1. Gaussian Process Regression

The regression problem is perhaps the most frequent in the machine learning community, which can be seen as the estimation of the relationship between an *independent* variable $x \in \mathcal{X}$ and a *dependent* variable $y \in \mathcal{Y}$. The relationship between both variables is represented by a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, thus the goal is to define a family of functions and find the best estimator in this family. This is usually done by minimizing or maximizing some performance indicator, such as some measure of the error between the estimation and the observation. From a Bayesian point of view, this problem can be solved by placing a *prior distribution* over the estimators f, then, in conjunction with the *likelihood* function for the dataset, and using the Bayes' rule of probability, we obtain the *posterior* distribution which can be used for prediction and forecasting.

Gaussian processes are the *gold-standard* Bayesian regression method, since all the appealing properties of the Gaussian distribution, like its marginalization closure and its explicit and tractable calculations. In order to better understand these properties, we now formalize how the Gaussian processes form a robust, non-parametric, non-linear regression framework.

From now on, without loss of generality, the input space $\mathcal{X}$ will be assumed to be $\mathbb{R}^d$, and the output space $\mathcal{Y}$ will be assumed to be $\mathbb{R}$. Given the training data $(X, \mathbf{y}) = \{(x_n, y_n)\}_{n=1}^N$ where $x_n \in \mathbb{R}^d$ and $y_n \in \mathbb{R}$. The points y_n correspond to noisy observations of the latent function f:

$$y_n = f(x_n) + \varepsilon_n, \quad \varepsilon_n \sim \mathcal{N}(0, \sigma^2), \sigma \in \mathbb{R}. \tag{1.5}$$

We place a Gaussian prior with zero mean and kernel k with parameters θ, and denote $\mathbf{f} = f(X) = \{f(x_n)\}_{n=1}^N$ the set of training latent values,

$$p(\mathbf{f}|X, \theta) = \mathcal{N}(0, k(X, X)). \tag{1.6}$$

Since we assume that the observations are contaminated with a Gaussian noise, the likelihood of the dataset can be written as:

$$p(\mathbf{y}|X, \mathbf{f}, \theta) = \mathcal{N}(\mathbf{f}, \sigma^2 I_N), \tag{1.7}$$

where I_N denotes the $N \times N$ identity matrix. In order to train the model, we need to find the set of kernel hyperparameters θ that maximize the Gaussian marginal likelihood which takes the following expression,

$$p(\mathbf{y}|X,\theta) = \int p(\mathbf{y}|\mathbf{f},X)p(\mathbf{f}|X,\theta)d\mathbf{f} = \mathcal{N}(\mathbf{y}|0,k(X,X)+\sigma^2 I_N). \tag{1.8}$$

Usually, we minimize the negative log marginal likelihood (NLL), that can be written as:

$$NLL = \log p(\mathbf{y}|X,\theta) = \frac{1}{2}\mathbf{y}^T(k(X,X)+\sigma^2 I_N)^{-1}\mathbf{y} + \frac{1}{2}\log|k(X,X)+\sigma^2 I_N| + \frac{N}{2}\log 2\pi. \tag{1.9}$$

This is known as maximum likelihood estimation (MLE) of the hyperparameters. Once, the hyperparameters of the kernel have been chosen, the posterior by the Bayes' rule:

$$p(\mathbf{f}|X,\mathbf{y},\theta) = \frac{p(\mathbf{f}|X,\theta)p(\mathbf{y}|X,\mathbf{f},\theta)}{p(\mathbf{y}|X,\theta)}. \tag{1.10}$$

Thus, the prediction f^* at a new point x^* have the following expression:

$$p(f^*|\mathbf{y},X,x^*,\theta) = \mathcal{N}(\widehat{m}_*, \widehat{\Sigma}_{**}), \tag{1.11}$$

where

$$\widehat{m}_* = k(x^*,X)(k(X,X)+\sigma^2 I_N)^{-1}\mathbf{y}, \tag{1.12}$$

$$\widehat{\Sigma}_{**} = k(x^*,x^*) - k(x^*,X)(k(X,X)+\sigma^2 I_N)^{-1}k(X,x^*). \tag{1.13}$$

Here, we notice two main challenges. First, the high computational cost of both training and prediction due to the requirement of the inversion of the matrix $k(X,X)+\sigma^2 I_N$ which has a cost $\mathcal{O}(N^3)$, this is an omnipresent issue which practically limits Gaussian Processes to datasets of size $\mathcal{O}(10^4)$. Second, the posterior distribution is entirely determined by the choice of the kernel function k. Design valid new covariance functions is one of the main challenges in the Gaussian processes framework, which in the following section we will cover.

1.2. Spectral Representation of Stationary Kernels

Stationary kernels are one of the most common and well-studied subclasses of kernels. We say that a kernel is stationary if it a function of $\tau = x - x'$, i.e. it is invariant to translation of inputs:

$$k(x+z, x'+z) = k(x,x'), \tag{1.14}$$

and it can be written as:

$$k(x,x') = k(x-x') = k(\tau). \tag{1.15}$$

From Equation (1.15) we notice that stationarity means that the correlation between two input points does not depend on the location itself, but only on the difference (or lag) between them, thus it encodes an identical similarity notion across the input space. Moreover, stationary kernels are of special interest since they can be fully characterized by positive finite measures using Bochner's theorem [10]. This representation can be used to construct new covariance functions in the spectral domain rather than the input domain.

Theorem 1.1 (Bochner's theorem [10]) *A complex-valued 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*

$$k(x, x') = \int_{\mathbb{R}^d} e^{i\omega^\top (x-x')} F(d\omega), \tag{1.16}$$

where F is a positive finite measure and i denotes the imaginary unit.

If the measure F has a density $S(\omega)$, this density is called the spectral density or power spectrum of k, and k and S are Fourier duals:

$$S(\omega) = \mathcal{F}\{k(\tau)\}(\omega) = \int e^{-i\omega\tau} k(\tau) d\tau \tag{1.17}$$

$$k(\tau) = \mathcal{F}^{-1}\{S(\omega)\}(\tau) = \int e^{i\omega\tau} S(\omega) d\omega \tag{1.18}$$

where $\mathcal{F}$ denotes the Fourier transform operator. Bochner's theorem defines a one-to-one mapping from stationary kernels to finite measures via Fourier transform, which allows designing the kernels in the spectral domain rather than the input domain. This is especially useful since strict positivity is much easier to achieve than positive definiteness.

The *spectral kernels* are a family of kernels that are defined in the spectral domain and are able to approximate any integrable stationary covariance function given enough parameters. In this sense, Wilson et al. [17] proposed the Spectral Mixture (SM) kernel, where the spectral density is modeled as a weighted mixture of Q Gaussian functions, with weight w_q, spectral means μ_q and diagonal covariance matrices Σ_q, that is:

$$S(\omega) = \sum_{q=1}^{Q} w_q \frac{1}{(2\pi)^{n(2} |\Sigma_q|^{1/2}} \exp\left(-\frac{1}{2} (\omega - \mu_q)^\top \Sigma_q^{-1} (\omega - \mu_q) \right). \tag{1.19}$$

In order to obtain a real-valued covariance function, the spectral density has to be symmetric with respect to ω. This is achieve by considering the kernel as:

$$k(\tau) = \mathcal{F}^{-1} \left\{ \frac{S(\omega) + S(-\omega)}{2} \right\}, \tag{1.20}$$

obtaining the following kernel:

Definition 1.2 *A spectral-mixture (SM) kernel is a positive-definite stationary covariance function given by*

$$k(\tau) = \sum_{q=1}^{Q} w_q \exp\left(-\frac{1}{2} \tau^\top \Sigma_q \tau \right) \cos\left(\mu_q^\top \tau \right) \tag{1.21}$$

where $\mu_q \in \mathbb{R}^d$, $\Sigma_q = \operatorname{diag}(\sigma_1^{(q)}, ..., \sigma_n^{(q)})$ and $w_q, \sigma_q \in \mathbb{R}_+$

Since the sum of Gaussians are universal approximator of densities, consider here in the frequency domain, the SM kernel is able to recover commonly used stationary kernels, such as squared exponential, Matérn, rational quadratic, and periodic kernels [1]

1.3. Multi-Output Gaussian Processes

Although Gaussian processes are very powerful and flexible models, they are restricted to the univariate case only. This makes them unsuitable for problems in which it is necessary to learn multiple tasks simultaneously. The extension of the GP framework to handle multiple outputs is called Multi-Output Gaussian Processes (MOGP) [2], which consists in model all the outputs jointly as a GP where the covariance and cross-covariance are ruled by a multi-output kernel. Formally,

Definition 1.3 *A Multi-Output Gaussian Process (MOGP) of m-channels is an m-tuple of stochastic processes $\{\mathbf{f}(x) := (f_1(x), ..., f_m(x)) : x \in \mathcal{X}\}$, such that for any family $\{X_i\}_{i=1}^m$ of finite subsets of X, the random vector $[f_1(X_1), ..., f_m(X_m)]$ is a multivariate Gaussian random variable.*

We call a *channel* or an *ouput* at each component of the vector $\mathbf{f}(x)$, $\forall x \in \mathcal{X}$. Similar to the single output GP, the MOGP is completely specified by its mean function $\mathbf{m}(x)$ and covariance function $\mathcal{K}(x, x')$, but in the multivariate case its mean is $\mathbb{R}^m$-valued functions whose i^{th} element denotes the mean function of the i^{th} channel, and its kernel is $\mathbb{R}^m \times \mathbb{R}^m$-valued function whose (i, j) element denotes the covariance between the i^{th} and j^{th} channels, these two functions are defined as follows:

$$\mathbf{m}(x) = \mathbb{E}[f(x)], \quad \forall x \in \mathcal{X} \tag{1.22}$$

$$\mathcal{K}(x, x') = \mathbb{E}[(f(x) - \mathbf{m}(x))(f(x') - \mathbf{m}(x'))], \quad \forall x, x' \in \mathcal{X} \tag{1.23}$$

In the same fashion that in the univariate case, the mean function is usually assumed to be zero, $\mathbf{m}(x) = 0$ $\forall x \in \mathcal{X}$. Moreover, if we have an MOGP with M channels, $\{f_i\}_{i=1}^M$, the $(i, j)^{th}$ element of the covariance kernel $\mathcal{K}$ corresponds to the covariance between outputs f_i and f_j, following the next notation:

$$cov[f_i(x), f_j(x')] = k_{ij}(x, x') = [\mathcal{K}(x, x')]_{ij} \quad \forall i, j = \{1, ..., M\} \tag{1.24}$$

Similar to the single-channel case, in order to be a valid covariance function, a function kernel must be symmetric, i.e.

$$\mathcal{K}(x, x') = \mathcal{K}(x', x) \quad \forall x, x' \in \mathcal{X}, \tag{1.25}$$

and positive-definite, i.e. $\forall n \in \mathbb{N}$, $\forall \{x_i\}_{i=1}^n \subset \mathcal{X}, \forall \{\{c_{ij}\}_{i=1}^n\}_{j=1}^m$

$$\sum_{i,j=1}^{m} \sum_{p,q=1}^{n} c_{ip} c_{jq} k_{ij}(x_p, x_q) \geq 0. \tag{1.26}$$

From the positive-definite condition, we can notice that Eq. (1.26) imposes the diagonal components of a multivariate kernel function $\mathcal{K}(x, x')$ to be positive-definite as in Eq. (1.4), i.e. the functions $k_{ii}(x, x')$ $\forall i \in \{1, ..., m\}$ are univariate covariance functions. On the other hand, the off-diagonal components are not restricted to be positive-definite, i.e. the functions $k_{ij}(x, x')$ $\forall i, j \in \{1, ..., \}$, $i \neq j$ are not restricted to be univariate kernel functions. Futhermore, a multivariate kernel $\mathcal{K}$ is stationary if $\mathcal{K}(x, x') = \mathcal{K}(x - x')$; a MOGP is said to be stationary if its covariance is stationary.

Design valid and expressive multi-output kernels is quite challenging because we need to jointly choose functions that model the covariance of each channel, called auto-covariance, and functions that model the cross-covariance between channels [9]. Thus, it is necessary to select $m(m+1)/2$ covariance functions which are able to model, for example, delays, phase shifts, S negative correlations or to enforce specific spectral content while at the same time maintaining positive definiteness of $\mathcal{K}$

1.3.1. Multi-Output Gaussian Process Regression

Thus far we have considered just the case where the datapoints consist of n dimensional vector-valued inputs and scalar-valued outputs. While this configuration covers the vast majority of regression cases one may well encounter in practice, it is possible to perform regression where both input and output are vector-valued, where each dimension of the output is called a channel. This is often called multiple-output regression, and in this Section, we will cover how to apply the MOGP framework to these kinds of problems.

Let be the training data for channel i, $(X_i, y_i) = \{x_n^{(i)}, y_n^{(i)}\}_{n=1}^{N_i}$ where $x_n^{(i)} \in \mathbb{R}^d$ and $y_n^{(i)} \in \mathbb{R}$. We assume that the points $y_n^{(i)}$ correspond to noisy observations of the latent function f_i:

$$y_n^{(i)} = f_i(x_n^{(i)}) + \varepsilon_n^{(i)}, \quad \varepsilon_n^{(i)} \sim \mathcal{N}(0, \sigma_i^2), \sigma_i \in \mathbb{R} \tag{1.27}$$

Denoting $\hat{N} = \sum_{i=1}^{m} N_i$ the total number of observations, $\mathbf{y} = [y_1, ..., y_m]$ the vector of observations, $X = [X_1, ..., X_m]$ and $\mathbf{F} = [\mathbf{f}_1(X_1), ..., \mathbf{f}_m(X_m)]$ the set of training latent values, where $\mathbf{f}_i(X_i) = \{f_i(x_n^{(i)})\}_{n=1}^{N_1}$. We place a Gaussian prior with zero mean and kernel $\mathcal{K}$ with parameters Θ

$$p(\mathbf{F}|X, \Theta) = \mathcal{N}(0, \mathcal{K}(X, X)) \tag{1.28}$$

Since we assume that the observations are contaminated with a Gaussian noise, the likelihood of the dataset can be written as:

$$p(\mathbf{F}|X, \Theta) = \mathcal{N}(\mathbf{F}, \Sigma) \tag{1.29}$$

where $\Sigma = \text{diag}[I_{N_1}\sigma_1^2, ..., I_{N_m}\sigma_m^2]$, and I_{N_i} denotes the $N_i \times N_i$ identity matrix. In the same fashion that the univariate case, we train the model finding the set of kernel hyperparameters Θ that maximize the Gaussian marginal likelihood which takes the following expression,

$$p(\mathbf{y}|X, \Theta) = \int p(\mathbf{y}|\mathbf{F}, X)p(\mathbf{F}|X, \Theta)d\mathbf{F} = \mathcal{N}(\mathbf{y}|0, \mathcal{K}(X, X) + \Sigma). \tag{1.30}$$

Similar to the univariate case we usually minimize the negative log marginal likelihood (NLL), that can be written as:

$$NLL = \log p(\mathbf{y}|X, \Theta) = \frac{1}{2}\mathbf{y}^T(\mathcal{K}(X, X) + \Sigma)^{-1}\mathbf{y} + \frac{1}{2}\log|\mathcal{K}(X, X) + \Sigma| + \frac{\hat{N}}{2}\log 2\pi. \tag{1.31}$$

Now, finding the posterior by the Bayes' rule we express the prediction f^* at a new point x^* as follow

$$p(f^*|\mathbf{y}, X, x^*, \Theta) = \mathcal{N}(\widehat{m}_*, \widehat{\Sigma}_{**}), \tag{1.32}$$

where

$$\widehat{m}_* = \mathcal{K}(x^*, X)(\mathcal{K}(X, X) + \Sigma)^{-1}\mathbf{y}, \tag{1.33}$$

$$\widehat{\Sigma}_{**} = \mathcal{K}(x^*, x^*) - \mathcal{K}(x^*, X)(\mathcal{K}(X, X) + \Sigma)^{-1}\mathcal{K}(X, x^*). \tag{1.34}$$

In the same way that the single output GP, the MOGP has two main challenges. First, the high computation cost due to the inversion of the matrix $\mathcal{K}(X, X) + \Sigma$, which in this case has a cost $\mathcal{O}(m^3N^3)$. Second, the posterior distribution is entirely determined by the choice of the kernel $\mathcal{K}$. Thus, design expressive covariance functions while satisfying the positive-definiteness condition is one of the main challenges in MOGP.

1.4. Existing work on Multi-Output Gaussian Processes

As we mention in Section 1.3, design multivariate covariance functions are quite challenging since we need to jointly choose functions that model auto-covariance, and functions that model the cross-covariance while maintaining the positive-definite and symmetric condition. Several approaches have been proposed to overcome this difficulty where most 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 subsections, we review one by one the previous proposals of multivariate covariance functions, finalizing with one of the most prominent extensions of the existing methods in expressiveness and interpretation, the multi-output spectral mixture kernel.

1.4.1. Linear Model of Coregionalization

The central and most basic idea of MOGP, which comes from Geostatistics, is known as the Linear Coregionalization Model. (LMC) [4]. The LMC models each output as a linear combination of independent latent processes $\{u_q\}_{q=1}^Q$, where each latent process is a Gaussian process with zero mean and kernel $k_q(x, x')$, namely

$$f_i(x) = \sum_{q=1}^{Q} a_{iq}u_q(x) \quad i = 1, ..., m. \tag{1.35}$$

Since a linear combination of GPs is also a GP, each f_i is a GP. The m channels are naturally correlated and the covariance between them can be calculated explicitly, leading to the following multivariate covariance function

$$\mathcal{K}(x, x') = \sum_{q=1}^{Q} A_q A_q^\top k_q(x, x'), \tag{1.36}$$

where $A_q = [a_{1q}, ..., a_{mq}]^\top \in \mathbb{R}^m$. Furthermore, considering that some latent processes may have the same covariance function, while remaining independent, we grouped them together leading to the following expression of the ouput functions f_i:

$$f_i(x) = \sum_{q=1}^{Q}\sum_{r=1}^{R_q} a_{iq}^{(r)}u_q^{(r)}(x) \quad i = 1, ..., m. \tag{1.37}$$

by defining $A_q^{(r)} = [a_{1q}^{(r)}, ..., a_{mq}^{(r)}]^\top \in \mathbb{R}^m$, we obtain the LMC multivariate covariance function:

$$\mathcal{K}(x, x') = \sum_{q=1}^{Q} \left(\sum_{r=1}^{R_q} A_q^{(r)} A_q^{(r)\top} \right) k_q(x, x') \tag{1.38}$$

$$= \sum_{q=1}^{Q} B_q k_q(x, x') \tag{1.39}$$

where $B_q = \sum_{r=1}^{R_q} A_q^{(r)} A_q^{(r)\top} \in \mathbb{R}^{m \times m}$ is the matrix known as *coregionalization matrix*, and has entries $[B_q]_{ij} = b_{ij}^{(q)} = \sum_{r=1}^{R_q} a_{iq}^{(r)} a_{jq}^{(r)}$.

We can note that the LMC kernel is essentially a multivariate covariance function obtained by multiplying a univariate covariance function with a positive-definite matrix. This model is a simple, economical, and effective way to build valid multivariate covariance functions, it even allows the construction of non-stationary multivariate covariance functions by considering non-stationary latent processes. The main drawback of this formulation is that it is limited to model just symmetric and centered cross-covariances functions, leaving out an entire spectrum of problems that cannot be modeled properly, since the cross-covariances are simply linear combinations of univariate covariances functions.

1.4.2. Intrinsic Coregionalization Model

The intrinsic coregionalization model (ICM) [4] is a particular case of the LMC, where it is assumed that the coregionalization matrix can be decoupled in an output component and a latent processes component, that is, the elements of the coregionalization matrix B_q can be written as $b_{ij}^{(q)} = v_{ij} b_q$, thus the multivariate kernel proposed by the ICM takes the form

$$k_{ij}(x, x') = v_{ij} \sum_{q=1}^{Q} b_q k_q(x, x') \tag{1.40}$$

$$= v_{ij} k(x, x'), \tag{1.41}$$

where $k = \sum_{q=1}^{Q} b_q k_q$. Furthermore, the proposed multivariate covariances function is

$$\mathcal{K}(x, x') = B k(x, x'), \tag{1.42}$$

where $B \in \mathbb{R}^{m \times m}$ is a positive definite matrix with entries $[B]_{ij} = v_{ij}$. From Eq. 1.42, we observe that the ICM is equivalent to a LMC with $Q = 1$, thus is a valid multivariate kernel. Futhermore, we notice that each component k_q contribute equally to covariance and cross-covariance between outputs, hence is a more restrictive model than LMC.

1.4.3. Semi-Parametric Latent Factor Model

The semi-parametric latent factor model (SLFM) [18] is a particular case of LMC where $R_q = 1 \ \forall q \in \{1, ..., Q\}$, that is,

$$\mathcal{K}(x,x') = \sum_{q=1}^{Q} A_q A_q^\top k_q(x,x') \tag{1.43}$$

$$= \sum_{q=1}^{Q} B_q k_q(x,x'), \tag{1.44}$$

where $B_q \in \mathbb{R}^{m\times m}$ is a rank 1 matrix with entries $[B_q]_{ij} = a_{iq}a_{jq}$. The semi-parametric name comes from the linear parametric combination of non parametric latent processes.

1.4.4. Convolution Model

The convolution model (CONV) [8] is a generative model for multivariate covariance function based on the idea of using convolution rather than instantaneous mixing, like the linear combination. Formally, let $\{u_q\}_{q=1}^{Q}$ a family of Q independent latent processes, where each latent process is a Gaussian process with zero mean and kernel $k_q(x,x')$, and let $\{\{k_{jq}\}_{j=1}^{m}\}_{q=1}^{Q}$ a family of stationary kernels, called *smoothing kernels*. The base idea of the convolution model is to model each channel as the sum of those Q latent processes convolved with its corresponding smoothing kernel, that is,

$$f_i(x) = \sum_{q=1}^{Q} (k_{iq} \star u_q)(x) \quad i = 1,...,m \tag{1.45}$$

$$= \sum_{q=1}^{Q} \int_{\mathbb{R}^d} k_{iq}(x-z)u_q(z)dz \quad i = 1,...,m, \tag{1.46}$$

where $\star$ denotes the convolution operator. We notice that each channel is a Gaussian process, since the convolution of a Gaussian process with a smoothing kernel is also a Gaussian processes. The multivariate covariance function of the above MOGP takes the following form,

$$k_{ij}(x,x') = \sum_{q=1}^{Q} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} k_{iq}(x-z)k_{jq}(x'-z)k_q(z,z')dzdz'. \tag{1.47}$$

The integral in Eq. 1.47 does not always have an explicit and tractable form, so choosing the smoothing kernels and covariance functions for the latent processes wisely becomes imperative. Since the Gaussian functions are closed under convolution, the previous problem is avoided considering both the smoothing kernels and the latent covariances as Gaussian functions, that is to choose the smoothing kernels and latent covariances functions as follow:

$$k_{iq}(\tau) = \frac{w_{iq}|\Sigma_{iq}|^{1/2}}{(2\pi)^{n/2}} \exp\left(\frac{1}{2}\tau^\top \Sigma_{iq}\tau\right), \tag{1.48}$$

$$k_q(\tau) = \exp\left(\frac{1}{2}\tau^\top \Sigma_q \tau\right), \tag{1.49}$$

we obtain the Gaussian convolution model, which takes the following form

$$k_{ij}(\tau) = \sum_{q=1}^{Q} \frac{w_{iq} w_{jq} |\sigma_q^{-1}|^{1/2}}{|\Sigma_{ijq}|^{1/2}} \exp\left(\frac{1}{2}\tau^\top \Sigma_{ijq}^{-1} \tau\right), \tag{1.50}$$

where $\Sigma_{ijq} = \Sigma_{iq}^{-1} + \Sigma_{jq}^{-1} + \Sigma_q^{-1}$. We observe that the convolution model is a generalization of the LMC, in which each channel has different kernel parameters, in contrast to the LMC where the kernel parameters of k_q are shared across all channels. Although CONV allows a broader family of multivariate processes to be modeled, they are still limited to linear combinations of univariates covariances functions.

1.4.5. Cross-Spectral Mixture kernel

The cross-spectral mixture (CSM) kernel [5] generalized the LMC framework allowing the coregionalization matrix to be a complex matrix which leads to non-symmetric cross-covariances. The main idea of this model is to consider the LMC with latent processes as GPs with spectral mixture kernels, that is

$$\mathcal{K}(\tau) = \sum_{q=1}^{Q} \left(\sum_{r=1}^{R_q} A_q^{(r)} A_q^{(r)H} \right) k_q(\tau) \tag{1.51}$$

where $(\cdot)^H$ is the Hermitian operator and $k_q(\tau) = \mathcal{F}^{-1}(S_q(\omega))(\tau) = \exp\left(-\frac{1}{2}\tau^\top \Sigma_q \tau\right) \cos(\mu_q \tau)$ is the spectral mixture kernel. The principal innovation of this model is that it can model the phase cross-channel, this is done by considering the coefficient of the coregionalization matrices to the form $a_{iq}^{(r)} = b_{iq}^{(r)} \exp\left(i\varphi_{iq}^{(r)}\right)$, where $b_{iq}^{(r)}, \varphi_{iq}^{(r)} \in \mathbb{R}$ and i is the imaginary unit, which leads to

$$\begin{aligned} k_{ij}(\tau) &= \sum_{q=1}^{Q} \sum_{r=1}^{R_q} b_{iq}^{(r)} b_{jq}^{(r)} e^{i(\varphi_{iq}^{(r)} - \varphi_{jq}^{(r)})} k_q(\tau) && (1.52) \\ &= \sum_{q=1}^{Q} \sum_{r=1}^{R_q} b_{iq}^{(r)} b_{jq}^{(r)} e^{i(\varphi_{iq}^{(r)} - \varphi_{jq}^{(r)})} \int_{\mathbb{R}^d} e^{i\omega^\top \tau} S_q(\omega) d\omega && (1.53) \\ &= \sum_{q=1}^{Q} \sum_{r=1}^{R_q} b_{iq}^{(r)} b_{jq}^{(r)} \int_{\mathbb{R}^d} e^{i(\omega^\top \tau + \varphi_{iq}^{(r)} - \varphi_{jq}^{(r)})} S_q(\omega) d\omega && (1.54) \end{aligned}$$

Solving the above integral, in conjunction with the reparameterization $\varphi_{iq}^{(r)} = \mu_q^\top \phi_{iq}^{(r)}$ where μ_q is the frequency of the SM kernel and $\phi_{iq}^{(r)} \in \mathbb{R}$, we obtain the CSM kernel which takes the following form:

$$k_{ij}(\tau) = \sum_{q=1}^{Q} \sum_{r=1}^{R_q} b_{iq}^{(r)} b_{jq}^{(r)} \exp\left(\tau^\top \Sigma_q \tau\right) \cos\left(\mu_q^\top \left(\tau + \phi_{iq}^{(r)} - \phi_{jq}^{(r)}\right)\right) \tag{1.55}$$

The CSM kernel is, to the best of our knowledge, the first multivariate covariance function that allows non-symmetric cross-covariances through cross-phases $\phi_{iq}^{(r)}$. This hints at the construction of more flexible and expressive multivariate covariances functions, where we consider, for example, phase shifts and time delay. Since the cornerstone of this model is the spectral mixture kernel, the CSM kernel is only suited for stationary processes.

1.4.6. Multi-Output Spectral Mixture Kernel

The Multi-Output Spectral Mixture (MOSM) Kernel, proposed by Parra et al. [3], provide 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. This approach rely on Cramér's theorem, the multivariate version of the Bochner's theorem,

Theorem 1.2 (Cramér's Theorem [6]) *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 admit the following representation*

$$k_{ij}(\tau) = \int_{\mathbb{R}^d} e^{i\omega^\top \tau} S_{ij}(\omega) d\omega \quad \forall i,j \in \{1,...,M\} \tag{1.56}$$

where each S_{ij} is an integrable complex-valued function $S_{ij} : \mathbb{R}^d \to \mathbb{C}$ known as the spectral density associated to the covariance function k_{ij}, and fulfil the positive definiteness condition

$$\sum_{i,j=1}^{m} \overline{z_i} z_j S_{ij}(\omega) \geq 0 \qquad \forall \{z_1,...,z_m\} \subset \mathbb{C}, \omega \in \mathbb{R}^d \tag{1.57}$$

Following the idea behind Spectral Mixture kernel, they proposed a family of Hermitian positive-definite complex-valued functions $\{S_{ij}\}_{i,j=1}^{m}$ which fulfil the Theorem 1.2. Parra et al. proposed to construct the family of $\{S_{ij}\}_{i,j=1}^{m}$ through the Cholesky decomposition, recalling that any complex-valued positive definite matrix $S(\omega)$ can be decomposed in the form $S(\omega) = R^H(\omega)R(\omega)$, meaning that the $(i,j)^{th}$ entry of the $S(\omega)$ can be expressed as

$$S_{ij} = R^H_{:i}(\omega) R_{:j}(\omega) \tag{1.58}$$

where $R(\omega) \in \mathbb{C}^{Q\times m}$ and Q is the rank of the decomposition. They start considering $Q = 1$, where the columns of $R(\omega)$ are complex-valued functions $R_i(\omega)_{i=1}^{m}$, and $S(\omega)$ is modeled as a rank-one matrix according to $S_{ij}(\omega) = \overline{R_i}(\omega)R_j(\omega)$, $\forall i,j = 1,...,m$. Therefore, they proposed to model each R_i as a squared exponential function, namely

$$R_i(\omega) = w_i \exp\left(-\frac{1}{4}(\omega-\mu_i)^\top \Sigma_i^{-1}(\omega-\mu_i)\right) \exp\left(-i(\theta_i^\top \omega + \phi_i)\right), \quad i = 1,...,m \tag{1.59}$$

where $w_i, \phi_i \in \mathbb{R}$, $\mu_i, \theta_i \in \mathbb{R}^d$ and $\Sigma_i = \text{diag}([\sigma_{i1}^2,...,\sigma_{in}^2])^\top \in \mathbb{R}^{d\times d}$. This choice of the function $R_{i i,j=1}^{m}$ yields the spectral densities $S_{ij}{}_{i,j=1}^{m}$ given by

$$S_{ij}(\omega) = w_{ij} \exp\left(-\frac{1}{2}(\omega-\mu_{ij})^\top \Sigma_{ij}^{-1}(\omega-\mu_{ij}) + i(\theta_{ij}^\top \omega + \phi_{ij})\right), \quad i = 1,...,m. \tag{1.60}$$

where each cross-spectral density is a complex-valued Squared exponential function with the following parameters

- Covariance: $\Sigma_{ij} = 2\Sigma_i(\Sigma_i + \Sigma_j)^{-1}\Sigma_j$

- Mean: $\mu_{ij} = (\Sigma_i + \Sigma_j)^{-1}(\Sigma_i \mu_j + \Sigma_j \mu_i)$
- 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: $\phi_{ij} = \phi_i - \phi_j$

In the same fashion that in the SM kernel, in order to restrict to this generative model to a real-valued GPs, the power spectral densities has to be symmetric with respect ω. This is achieve by considering the kernel as:

$$k_{ij}(\tau) = \mathcal{F}^{-1}\left\{\frac{S_{ij}(\omega) + S_{ij}(-\omega)}{2}\right\}(\tau) \tag{1.61}$$

$$= \alpha_{ij} \exp\left(-\frac{1}{2}(\tau + \theta_{ij}^\top \Sigma_{ij}(\tau + \theta_{ij})\right) \cos\left((\tau + \theta_{ij})^T \mu_{ij} + \phi_{ij}\right), \tag{1.62}$$

where $\alpha_{ij} = w_{ij}(2\pi)^{\frac{n}{2}}|\Sigma_{ij}|^{1/2}$. Finally, for the general case, where $Q \in \mathbb{N}$, S is written as the sum of these matrices of rank 1 . Thus, the multi-Output Spectral Mixture (MOSM) Kernel has the form:

$$k_{ij} = \sum_{q=i}^{Q} \alpha_{ij}^{(q)} \exp\left(-\frac{1}{2}(\tau + \theta_{ij}^{(q)})^\top \Sigma_{ij}^{(q)}(\tau + \theta_{ij}^{(q)})\right) \cos\left((\tau + \theta_{ij}^{(q)})^T \mu_{ij}^{(q)} + \phi_{ij}^{(q)}\right) \tag{1.63}$$

where $\alpha_{ij}^{(q)} = w_{ij}^{(q)}(2\pi)^n|\Sigma_{ij}^{(q)}|^{1/2}$ and the super index $(\cdot)^{(q)}$ denotes the parameter of the q^{th} component of the spectral mixture.

The MOSM kernel has shown desirable properties like 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 MOSM kernel is the ability of joint model delays and phase differences, this is possible due to the complex-valued model for the cross-spectral density.

All the useful properties of the MOSM kernel are restricted for stationary data, due the fact that the cornerstone of this family of kernels is Cramér's theorem. This assumption of stationarity is unsuitable for a vast of real world problems, like in vibratory signals [11, 12], free-drifting oceanic instruments [13], various neuroscience applications [14, 15], and econometric [16]. In the next section we will study harmonizable processes, a concept created by Loéve that will help us extend the MOSM kernel to non-stationary data while keeping all its desirable properties.

1.5. Harmonizable Processes

In order to consider a broad and general class of processes beyond stationary ones, we will study the celebrated extension of the stationarity property called harmonizability, originally introduced by Loève [7] for the univariate case, and then by Kahihara [19] for the multidimensional case. We now recall the definition and relevant properties of the harmonizable processes.

Definition 1.4 *A stochastic process on $\mathbb{R}^d$ is weakly harmonizable if its covariance function can be expressed as:*

$$k(x, x') = \iint_{\mathbb{R}^d \times \mathbb{R}^d} e^{i(\omega^\top x - \omega'^\top x')} F(d\omega, d\omega'), \tag{1.64}$$

where F is a positive definite bimeasure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \times (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ of bounded semivariation (or Fréchet variation), and the integral is in the Morse-Transue sense [20]. Thus, F is called the spectral bimeasure and satisfies the following conditions:

- *(Bimeasure) $\forall A, B \in \mathcal{B}(\mathbb{R}^d)$, $F(A, \cdot)$ and $F(\cdot, B)$ are complex measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$*
- *(Positive definiteness) $\forall n \in \mathbb{N}$, $\forall \{\alpha_i\}_{i=1}^n \subset \mathbb{C}, \forall \{A_i\}_{i=1}^n \subset \mathcal{B}(\mathbb{R}^d)$ it holds that*

$$\sum_{j=1}^{n} \sum_{i=1}^{n} \alpha_j \overline{\alpha_j} F(A_j, A_j) \geq 0 \tag{1.65}$$

- *(Fréchet variation boundedness)$\|F\|(\mathbb{R}^d, \mathbb{R}^d) < \infty$, where the Fréchet variation is given by*

$$\forall A, B \in \mathcal{B}(\mathbb{R}^d), \|F\|(A, B) = \sup \sum_{i=1}^{n} \sum_{j=1}^{m} |\overline{\alpha_i} F(A_i, B_j) \beta_j|, \tag{1.66}$$

Definition 1.5 *A stochastic process on $\mathbb{R}^d$ is strongly harmonizable if is weakly harmonizable and its spectral bimeasure F is of finite total (Vitali) variation, that is $|F|(\mathbb{R}^d, \mathbb{R}^d) < \infty$, where the Vitali variation is given by*

$$\forall A, B \in \mathcal{B}(\mathbb{R}^d), |F|(A, B) = \sup \sum_{i=1}^{n} \sum_{j=1}^{m} |F(A_i, B_j)| \tag{1.67}$$

where the supremum is taken over all finite measurable partitions $\{A_i\}_{i=1}^n$ of A, and $\{B_j\}_{j=1}^m$ of B respectively.

In the strong harmonizability case, due to the finite total variation, the spectral bimeasure F can be extended to a measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^d))$, and in this case the Morse-Transue integral in Eq. (1.64) coincides with the Lebesgue integral on $\mathbb{R}^d \times \mathbb{R}^d$ [19].

Remark 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 we can write the covariance as

$$k(x, x') = \iint_{\mathbb{R}^d \times \mathbb{R}^d} e^{i(\omega^\top x - \omega'^\top x')} S(\omega, \omega') d\omega d\omega' \tag{1.68}$$

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

It is important to point out that $S(\omega, \omega')$ is also a covariance function [21], and measure the interaction between the ω, ω', where we can interpret the variables ω and ω' as frequencies. This correlation between frequencies is what gives harmonizable processes the property of modeling non-stationary processes.

Notice that the strongly harmonizable concept is a consistent extension of stationary processes in the sense that when the measure F concentrates on its diagonal $\omega = \omega'$, eq. (1.64) collapses to Bochner Thm eq. (1.16). Nonetheless, harmonizable processes as presented above define a much larger class than that of stationary processes, actually, is well known that all continuous bounded kernels of practical interest are indeed strongly harmonizable. In fact, Yaglom [22] noticed that the only processes with continuous bounded kernels that are not strongly harmonizable are, in his own words, 'rather complicated and have some unusual, even pathological properties'. Therefore, we will restrict ourself to the strongly harmonizable case and from now on we will refer to strongly harmonizable processes simply as *harmonizable processes.*

The definitions presented above are only suitable for univariate processes, and since we are interested in model multivariate processes, we need a multidimensional extension for the definition of harmonizable processes. Kakihara extended the harmonizable property to the multidimensional case, which will be the cornerstone of our proposal, as follow:

Theorem 1.3 (Kakihara's theorem [19]) *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 it admits the following representation:*

$$k_{ij}(x,x') = \iint_{\mathbb{R}^d\times\mathbb{R}^d} e^{i(\omega^\top x-\omega'^\top x')} F_{ij}(d\omega,d\omega') \quad \forall i,j\in\{1,...,m\}, \tag{1.69}$$

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

Remark 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 can write the covariance as

$$k_{ij}(x,x') = \iint_{\mathbb{R}^d\times\mathbb{R}^d} e^{i(\omega^\top x-\omega'^\top x')} S_{ij}(\omega,\omega')d\omega d\omega' \quad \forall i,j\in\{1,...,m\}, \tag{1.70}$$

where the matrix spectral densities $S=[S_{ij}(A,B)]$ is such that $\forall i,j,S_{ii}$ is positive semi-definite, while $S_{ij}(A,B)=\overline{S}_{ji}(B,A)$

Theorem 1.3 gives the guidelines to construct multivariate covariance functions for non-stationary MOGP by designing their corresponding spectral densities instead, i.e., the design is performed in the frequency rather than the space domain

Chapter 2

Multi-Output Harmonizable Spectral Mixture Kernel

The objective of this chapter is to present the main contribution of this thesis, the Multi-Output Harmonizable Spectral Mixture (MOHSM) kernel, which generalized the MOSM kernel, a multi-output kernel built in the frequency domain, to non-stationary processes. First, we present the MOHSM kernel, deriving our multi-output covariance function using the Harmonizable processes seen in the chapter before. Then, we study the relationship of our model with other non-stationary kernels and the MOSM kernel. After that, we show the expressiveness of the MOHSM, analyzing what kind of non-stationarity our covariance function is able to model. Then, we present a parameter initialization scheme, to overcome one of the MOHSM problems, which is the sensitivity of the optimization. Finally, we study possible variations of the proposed framework.

2.1. Derivation of the Proposed Kernel

Following the idea behind the Multi-output Spectral Mixture (MOSM) kernel proposed by Parra et al. [3], we aim to propose a family of Hermitian positive-definite complex-valued functions $\{S_{ij}\}_{i,j=1}^{m}$ that satisfy the Theorem 1.3 to use them as building blocks for a cross-spectral densities. In contrast with what was done by Parra et al., we rely on the spectral representation provided by the harmonizable processes, this will allow us to model a broader family of processes such as non-stationary processes as well as stationary processes. The proposed family of functions is designed to extend the MOSM kernel to non-stationary processes while keeping its desirable properties, like its physical parametric interpretation and covariance functions with closed-form.

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 matrix S can be decomposed in

$$S(\omega, \omega') = R^{H}(\omega, \omega')R(\omega, \omega'), \tag{2.1}$$

where $R \in \mathbb{C}^{Q\times m}$, Q is the rank of the decomposition, and $(\cdot)^{H}$ denotes the Hermitian, transpose and conjugate, operator. Thus, rather than design directly the family of cross-spectral functions, we choose the family of the Cholesky decomposition. For ease of understanding we suppose that $Q = 1$, where the case for arbitrary Q is shown at the end of the section. Now, the (i, j) entry of $S(\omega, \omega')$ can be expressed as

$$S_{ij}(\omega, \omega') = \overline{R}_i(\omega, \omega')R_j(\omega, \omega') \quad \forall i, j = 1, \dots, m, \tag{2.2}$$

where $R \in \mathbb{C}^{Q\times 1}$ and $\overline{(\cdot)}$ denotes the conjugate operator. Now, we consider the family $\{R_i\}_{i=1}^m$ as follow:

$$R_i(\omega, \omega') = w_i \underbrace{\exp\left(-\frac{1}{4l_i^2}\|\hat{\omega}\|^2\right)}_{(\star)} \underbrace{\exp\left(-\frac{1}{4}(\overline{\omega} - \mu_i)^\top \Sigma_i^{-1}(\overline{\omega} - \mu_i)\right) \exp\Big(-i(\theta_i^\top \overline{\omega} + \phi_i)\Big)}_{(\bullet)}, \tag{2.3}$$

where $\hat{\omega} = \omega - \omega', \overline{\omega} = \frac{\omega+\omega'}{2}$, $w_i, l_i, \phi_i \in \mathbb{R}$, $\theta_i, \mu_i \in \mathbb{R}^n$ and $\Sigma_i = \mathrm{diag}([\sigma_{i1}^2, \dots, \sigma_{in}^2]) \in \mathbb{R}^{n\times n}$. The intuition behind this choice is that the $(\star)$ component 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 model the importance of each frequency. We modeled both components as square exponential (SE) functions since they have closed form for multiplications and Fourier transforms. Therefore, selecting this $\{R_i\}_{i=1}^m$ in conjunction with the following property of SE functions

Proposition 2.1 *The product of SE function is closed up to a constant, that is*

$$e^{\left(-\frac{1}{2}(x-\mu_i)^\top \Sigma_i (x-\mu_i)\right)} e^{\left(-\frac{1}{2}(x-\mu_j)^\top \Sigma_j (x-\mu_j)\right)} = \alpha_{ij} e^{\left(-\frac{1}{2}(x-\mu_{ij})^\top \Sigma_{ij} (x-\mu_{ij})\right)} \tag{2.4}$$

where

$$\alpha_{ij} = e^{\left(-\frac{1}{2}(\mu_i-\mu_j)^\top (\Sigma_i+\Sigma_j)^{-1} (\mu_i-\mu_j)\right)} \tag{2.5}$$

$$\mu_{ij} = (\Sigma_i + \Sigma_j)^{-1}(\Sigma_i\mu_j + \Sigma_j\mu_i) \tag{2.6}$$

$$\Sigma_{ij} = \Sigma_i(\Sigma_i + \Sigma_j)^{-1}\Sigma_j \tag{2.7}$$

We have that $\{S_{ij}\}_{i,j=1}^m$ are given by:

$$S_{ij} = w_{ij} \exp\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2\right) \exp\left(-\frac{1}{2}(\overline{\omega} - \mu_{ij})^\top \Sigma_{ij}^{-1}(\overline{\omega} - \mu_{ij})\right) \exp\Big(-i(\theta_{ij}^\top \overline{\omega} + \phi_{ij})\Big), \tag{2.8}$$

where we can observe that this is a complex decaying square exponential, due to the decaying factor at the left, 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: $w_{ij} = w_i w_j \exp\left(\frac{1}{4}(\mu_i - \mu_j)^\top(\Sigma_i + \Sigma_j)^{-1}(\mu_i - \mu_j)\right)$
- delay: $\theta_{ij} = \theta_i - \theta_j$
- phase: $\phi_{ij} = \phi_i - \phi_j$
- length-scale: $l_{ij}^2 = 2l_i^2 l_j^2(l_i^2 + l_j^2)^{-1}$

Observe that each S_{ii} is a locally-stationary kernel, a concept coined by Silverman [23], which refers to kernels that can be expressed as a product of a stationary kernel and a non-negative function. In facts,

$$S_{ii} = \underbrace{w_{ij} \exp\left(-\frac{1}{2l_{ii}^2}\|\hat{\omega}\|^2\right)}_{(1)} \underbrace{\exp\left(-\frac{1}{2}(\overline{\omega} - \mu_{ii})^\top \Sigma_{ii}^{-1} (\overline{\omega} - \mu_{ii})\right)}_{(2)}, \tag{2.9}$$

where we first notice that the auto-spectral densities are real-valued functions, since $\phi_{ii} = \theta_{ii} = 0$. Second, we observe that the (1) term is a stationary kernel and the (2) term is a non-negative function, therefore we can assure that each S_{ii} is a kernel function, hence a positive-definite function. On the other hand, since S can be decomposed in the form $S = R^H R$, it is clearly a positive-definite matrix thus fulfilling Theorem 1.3.

Now, we calculate the inverse generalized Fourier transform of the spectral densities S_{ij} 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'^\top x')} S_{ij}(\omega, \omega') d\omega d\omega' && (2.10)\\
&= \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' && (2.11)\\
&= \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) && (2.12)\\
&= \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}))} d\omega d\omega' && (2.13)\\
&= 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}\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' && (2.14)\\
&= 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}\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} && (2.15)\\
&= w_{ij} \int_{\mathbb{R}^n} e^{\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2 + i\hat{\omega}^\top \overline{x}\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} && (2.16)\\
&= w_{ij} \int_{\mathbb{R}^n} e^{\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2 + i\hat{\omega}^\top \overline{x}\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} && (2.17)\\
&= \alpha_{ij} e^{-\frac{l_{ij}^2}{2}\|\overline{x}\|^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)} && (2.18)\\
&= \alpha_{ij} e^{-\frac{l_{ij}^2}{2}\|\overline{x}\|^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)}, && (2.19)
\end{aligned}$$

where $\alpha_{ij} = w_{ij}(2\pi)^n |\Sigma_{ij}|^{1/2} l_{ij}^n$. In line (2.15) we applied the change of variables Theorem with $\hat{\omega} = \omega - \omega'$ and $\overline{\omega} = \frac{\omega+\omega'}{2}$, and in line (2.18) we solved the integral following the Gaussian integral, which state as follow:

Proposition 2.2 *For any diagonal matrix* $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ *and for any* $b \in \mathbb{C}^n$, $c \in \mathbb{C}$*:*

$$\int_{R^n} \exp\Big(-x^\top \Lambda x - 2b^\top x + c\Big) dx = \frac{\pi^{\frac{n}{2}}}{|\Lambda|^{\frac{1}{2}}} \exp\Big(b^\top \Lambda^{-1} b + c\Big). \tag{2.20}$$

The kernel obtained above is a complex-valued covariance function which yields to a complex-valued MOGP. In order to restrict our model to real-valued MOGP the covariance functions must be real-valued functions. This can be done by two equivalent procedures : taking the real part of the covariance function will leads to a real-valued kernel functions, and symetrizing the spectral densities $S_{ij}(\omega, \omega')$ by reassigning:

$$S_{ij}(\omega, \omega') \longleftarrow \frac{1}{2}(S_{ij}(\omega, \omega') + S_{ij}(-\omega, -\omega')), \tag{2.21}$$

this guarantees symmetry and real values in the diagonal as the complex terms cancel each other. Therefore, the kernel obtained by performing either of the two previous procedures is:

$$k_{ij}(x, x') = \alpha_{ij} \exp\Big(-\frac{1}{2}(\tau + \theta_{ij})^\top \Sigma_{ij}(\tau + \theta_{ij})\Big) \cos\left((\tau + \theta_{ij})^\top \mu_{ij} + \phi_{ij}\right) \exp\Big(-\frac{l_{ij}^2}{2}\|\overline{x}\|^2\Big), \tag{2.22}$$

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

One drawback of this formulation is that, due the last exponential term in eq. (2.22), the proposed kernel vanishes outside the origin with length-scale $\frac{1}{l_i}$. We address this limitation by placing input shifts, which will allow us to control where the kernel contribute. We can add this input shifts on our formulation by multiplying our initial matrix S by $\exp\Big(-i(\omega - \omega')^\top x_p\Big)$, which will keep all of properties needed for the Theorem 1.3. This shift can be seen from the point of view of Fourier transform, since the integral in eq. (1.70) is equivalent to a Fourier transform of the concatenated vector $\begin{pmatrix} x \\ -x \end{pmatrix}$. The translation in the input leads to closed form Fourier transforms:

$$k_{ij}(x - x_p, x' - x_p) = \iint_{\mathbb{R}^n \times \mathbb{R}^n} e^{i(\omega^\top x - \omega'^\top x')} S_{ij}(\omega, \omega') e^{-i(\omega - \omega')^\top x_p} d\omega d\omega'. \tag{2.23}$$

Putting together all the aforementioned, the kernel is defined as follows:

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

Lastly, for the general case we can expand the kernel to higher rank matrix by taking S as a sum of these matrix of rank 1, which yields the expression for the proposed kernel:

Definition 2.1 *The Multi-Output Harmonizable Spectral Mixture (MOHSM) kernel has the form:*

$$
\begin{aligned}
k_{ij}(x,x') = \sum_{q=1}^{Q} \alpha_{ij}^{(q)} \exp\Big(-\frac{1}{2}(\tau+\theta_{ij}^{(q)})^\top \Sigma_{ij}^{(q)}(\tau+\theta_{ij}^{(q)})\Big) \\
\cos\big((\tau+\theta_{ij}^{(q)})^\top \mu_{ij}^{(q)} + \phi_{ij}^{(q)}\big) \exp\left(-\frac{l_{ij}^{2(q)}}{2}\|\overline{x}-x_p^{(q)}\|^2\right),
\end{aligned}
\tag{2.24}
$$

where $\alpha_{ij}^{(q)} = w_{ij}^{(q)}(2\pi)^n|\Sigma_{ij}^{(q)}|^{1/2} l_{ij}^{(q)}$ and the super index $(\cdot)^{(q)}$ denotes the parameter of the q^{th} component of the spectral mixture.

From the kernel and spectral expressions we can interpret the kernel parameters 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 which control the correlation between the frequencies.

Finally, a useful particular case of MOHSM is restricting that certain components have the same center at all times, this is because in certain places of the input space more expressiveness of the model is needed than in others. This can be done by defining the MOHSM as follows:

Proposition 2.3 *The particular case where some components of the MOHSM share the same center can be written as follow:*

$$
\begin{aligned}
k_{ij}(x,x') = \sum_{p=1}^{P}\sum_{q=1}^{Q_p} \alpha_{ij}^{(q)} \exp\Big(-\frac{1}{2}(\tau+\theta_{ij}^{(q)})^\top \Sigma_{ij}^{(q)}(\tau+\theta_{ij}^{(q)})\Big) \\
\cos\big((\tau+\theta_{ij}^{(q)})^\top \mu_{ij}^{(q)} + \phi_{ij}^{(q)}\big) \exp\left(-\frac{l_{ij}^{2(p)}}{2}\|\overline{x}-x_p\|^2\right),
\end{aligned}
\tag{2.25}
$$

where $\sum_{p=1}^{P} Q_p = Q$, $\alpha_{ij}^{(q)} = w_{ij}^{(q)}(2\pi)^n|\Sigma_{ij}^{(q)}|^{1/2} l_{ij}^{(q)}$, 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 parameter of the p^{th} center of the spectral mixture.

This expression of the MOHSM will be extremely useful in the following sections where we will study the properties of the kernel and propose an initialization scheme for the hyperparameters.

2.2. Relationship with Other Models

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 et al. [24] proposed a family of spectral kernels that they prove can approximate any continuous bounded nonstationary kernel which they called Generalized Spectral Kernels (GSK). In the same line, Shen et al. [25] proposed the harmonizable mixture kernel (HMK) which also is a family derived from mixture models on the generalized spectral representation. Remes et al.[26] presented a non-stationary kernel based on the idea of harmonizable processes and parameterized the frequencies, length scales, and mixture weights as Gaussian processes. While all these previous works of non-stationary kernel derived from the harmonizable processes can be extended easily extended to multivariate cases using the LCM or ICM framework, the obtained cross-covariance functions lack expressiveness and interpretability. Our work can be seen as a multivariate extension of the family proposed by Shen et al. [25] 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 problem with these formulations is they force the auto-covariance to have similar behavior across channels. Furthermore, modeling the cross-correlation in this fashion, the interpretability of the dependence learned is almost null. The MOSM solves all the previous problems, adding interpretability of the dependencies and not imposing similar behaviors through different channels, but restricting itself to stationary cases. Our work extend the MOSM kernel since the MOHSM can model the correlation between the channels like the MOSM do, but allowing changes in the regimes within 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 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 frequencies domain we notice that when $\omega \neq \omega'$:

$$\exp\left(-\frac{1}{2l_{ij}^2}\|\omega-\omega'\|^2\right) \overset{l_{ij}\to 0}{\longrightarrow} 0. \tag{2.26}$$

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

$$\exp\left(-\frac{1}{2l_{ij}^2}\|\omega-\omega'\|^2\right) \overset{l_{ij}\to 0}{\longrightarrow} 1, \tag{2.27}$$

Combining equations (2.26) and (2.27) we obtain:

$$S_{ij}(\omega.\omega') \overset{l_{ij}\to 0}{\longrightarrow} \delta(\omega-\omega')\exp\left(-\frac{1}{2}(\overline{\varpi}-\mu_{ij})^\top\Sigma_{ij}^{-1}(\overline{\varpi}-\mu_{ij})\right)\exp\Big(-i(\theta_{ij}^\top\overline{\varpi}+\phi_{ij})\Big) = \hat{S}(\omega,\omega'), \tag{2.28}$$

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 generalized 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, and consequently extend the CSM kernel.

2.3. Expressiveness of the Model

It is important to remark what kind of non-stationarity the proposed MOHSM model can capture. This is relevant because it gives us a clue as to what kinds of problems we can face with our model and which ones it will work best on.

First, we observe that each mixture of the proposed kernel is a locally stationary kernel, which means that have a component that describes the global structure multiply by a component that describes the local structure of the data. More in details, a locally stationary kernel has the following form:

$$k(x,x') = k_1\left(\frac{x+x'}{2}\right)k_2(x-x') \tag{2.29}$$

where k_1 is a non-negative function, which describe the global structure, and k_2 is a stationary kernel, which describe the local structure. The variable $\frac{x+x'}{2}$ represent the centroid of the example x and x'. Genton [27] showed that a locally stationary kernel k is completely determined by its values on the diagonal $x = x'$ and antidiagional $x = -x'$. Fig 2.1 shows an example of a locally-stationary kernel.

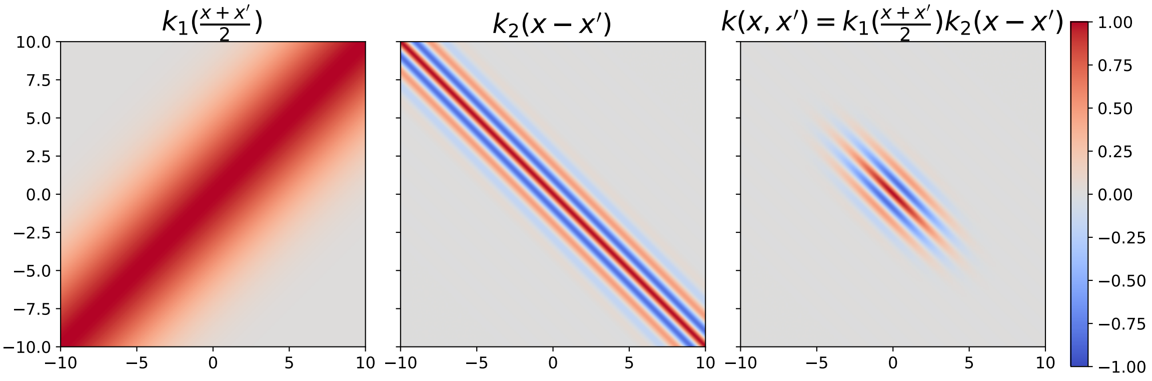

Figure 2.1: Example of a locally stationary kernel, where the global component k_1 is an SE function and the local component k_2 is a SM kernel.

In our model the global structure was selected as an exponential window operating over of the regime described by the local structure component, which is our case is a SM kernel. Hence, each component is an SM kernel windowed by an exponential function and thus, the kernel can be seen as a union of different regimen that can be disjoint or overlapped. In the places where the regimes overlap, the model is more expressive since it can be seen as an SM with more components which have the property that they can approximate any stationary kernel with the desired precision. Fig 2.2 shows an example of an auto-covariance of our model with 2 different regimes, which are disjoint at the edges but overlap in the center.

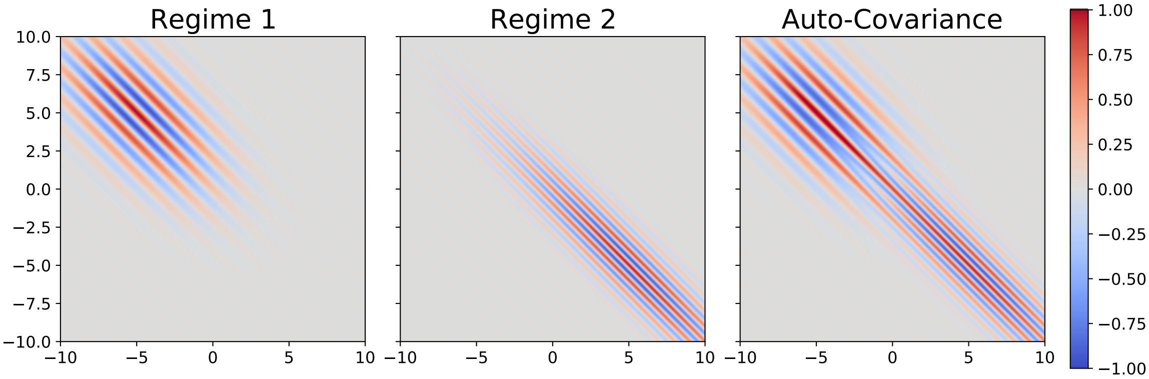

Figure 2.2: Example of an auto-covariance of the MOHSM. **Left and center** show the different regime of the model separately, and **right** shows the union of both regimes.

Therefore, our framework allows us to model and discover different regimes on the data, adding more expressiveness where needed. These properties of the MOHSM make it suitable for a wide range of problems, both in tasks where the data is known to have changes and in those where there is no knowledge of the data patterns. In addition to this, the proposed kernel is able to jointly model delays and phase differences between channels, which gives the model more expressiveness, which will be experimentally validated in the following sections.

2.4. Practical Considerations

2.4.1. Training and Prediction

Training the proposed model follows the same line as we saw in Section 1.3.1, by closed-form maximum likelihood. For the general case of the MOHSM, the NLL is minimized with respect to the parameters of the MOHSM plus the noise of the observations $\Theta = \{w_i^{(q)}, \mu_i^{(q)}, \Sigma_i^{(q)}, \theta_i^{(q)}, \phi_i^{(q)}, \sigma_{i,\text{noise}}^{(q)}, l_i^{(q)}, x_p^{(q)}\}_{i=1,q=1}^{M,Q}$. On the other hand, in the particular case where some components of the MOHSM share the same center, we add the restriction that all the centers $x_p^{(q)}$ and lenghtscales $\{l_{ij}^{(q)}\}_{i,j=1}^{m}$ are equals for $q = 1, ..., Q_p$. Once the optimization is concluded, computing the predictive posterior in the proposed model follow the standard MOGP procedure.

2.4.2. Parameter Initialization

The MOHSM kernel, like the MOSM and any other spectral mixture kernel, has a problem regarding the extreme sensibility in the optimization with respect to the initial points: a bad choice of initial condition will lead to poor optimization. This problem arises due to the large number of parameters of the spectral mixture kernels, which increase with the number of components. Therefore, exists a correlation between the expressiveness of the model, related to the number of components, and the complexity by increasing the dimension of the optimization problem. In this context, to avoid falling to local minima, a fundamental aspect of the optimization process is the initialization of the hyperparameters.

For spectral kernels, the problem of initializing the hyperparameters can be addressed relying on the spectral interpretation of the kernel parameters. In recent work, Cuevas [28] has proposed an initialization scheme for the SM kernel and the MOSM kernel, which leads to consistent and better results. Both rely on the idea of estimate the power spectral density (PSD) of the available data to then use it to obtain initial values of the parameters, considering the spectral interpretation of it. Since the MOHSM is a spectral kernel, we can rely on the idea proposed by Cuevas to initialize the hyperparameters. In this section, we propose a hyperparameters initialization scheme for the MOHSM kernel based on the ideas proposed by Cuevas.

As we discussed in Section 2.2, the MOHSM can be seen as a union of different regimens that can be disjoint or overlapped, where each regimen can be seen like a MOSM. Based on this, the problem of initializing the hyperparameters can be split into two parts: first, we have to placed the regimens, choosing the centers x_p and the lengthscales l_{ij}, to then initialize the hyperparameters of each regimen. For the initialization scheme, we will start from the model representation given by eq. (2.25), which will be useful for the second step of the method.

Following the above idea, we start placing the P different regimens by choosing the centers and the lengthscales. This proposed two methods to select the centers and the lengthscales: the first consists on place them equidistant and cover the input space with windows of the same lengthscale. This method, despite to be fast and easy to implement, it may not be the best for all the dataset, for example in datasets that are not uniformly distributed. The second method consists on initialize the centers x_p using some clustering methods, like K-means, and the choose the lengthscale to cover the input space.

After defining the different regime windows, it is time to initialize the hyperparameters of the spectral components of each regime. Extending the idea of Cuevas, in each regime we estimate the PSD for each channel. If the input has multiple dimension, we estimate the PSD for each input dimension and each channel. Note that for each output dimension and input dimension, a different estimation of the Nyquist is required. Since the regime are defined trough an exponential windows, it is not clear where to estimate the PSD. To address this problem we suggest to place a rectangular windows center on x_p and widths equal to $2l_i^{(p)}$. Therefore, this window let us to define $(X_i^{(p)}, y_i^{(p)})$, the data of the channel i and the regimen p, and use it to estimate the PSD of the regimen.

Let Q_p be the number of components with center p in the MOHSM kernel. Once the PSD of the regime p is estimated for channel i, we select the greater Q_p peaks of the PSD for each input dimension. The magnitudes a_i, taken proportional to the magnitude of the peaks, normalized so the sum of squared weights equals the channel sample variance of the observations of said channel,

$$\sum_{q=1}^{Q_p}(a_i^{(q)})^2 \propto \operatorname{Var}\left(\{y_n^{(i)(p)}\}_{n=1}^{N_i^{(p)}}\right), \tag{2.30}$$

where $\{y_n^{(i)(p)}\}_{n=1}^{N_i^{(p)}}$ correspond to the observation at channel i on the regimen p. The spectral means $\mu_i^{(q)}$ are initialized as the position of the peaks; the spectral variance $\Sigma_i^{(q)}$ are set proportional to the width of each peak. In order to prevent overfit, the variances are multiplied by 2 so that the uncertainty on a given frequency starts lower.

In our experience, the initial conditions of the delays and phases 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 discover the non-zero delay and phase if the data suggest so.

The base of the proposed initialization methods is the estimation of the PSD of the process. There are several techniques in the literature to estimate the PSD, such as the periodogram [29], Welch periodogram [30], Lomb Scargle (LS) [31] or the Bayesian non-parametric spectral estimation (BNSE) [32], among others. We recommend to obtained the PSD employing the BNSE or LS method, due that in most GP regression problems the observations are not uniformly sampled. For both methods, a grid of frequencies is required. We suggest to use a uniform grid of frequencies up to the estimated Nyquist frequency, with the Nyquist frequency estimated as half of the inverse of the smallest interval between input points.

A summary of the proposed method is shown below, where the algorithm 1 summarizes the procedure of the second part of the initialization method, and algorithm 2 outlines the approach.

Algorithm 1: Regime/Channel Spectral initialization

```
Input: $(X_i^{(p)}, y_i^{(p)})$ the observations of the regimen $p$ in channel $i$, integer $Q_p$ the
  number of components with center $x_p$
psd ← $EstimatePSD(X_i^{(p)}, y_i^{(p)})$;
peaks ← $HighestPeaks$(psd,$Q_p$);
for $q$ ← 1 to $Q_p$ do
    $\mu_i^{(q)}$ ← position of $q$th peak;
    $\Sigma_i^{(q)}$ ← proportional to the width of $q$th peak;
    $a_i^{(q)}$ ← proportional to the magnitudes of $q$th peak, following eq. (2.30) ;
end
Output: Regime/channel parameters $\mu_i^{(q)}$, $\Sigma_i^{(q)}$, $a_i^{(q)}$ for $q = 1, \ldots, Q_p$,
```

Algorithm 2: MOHSM Spectral initialization

```
Input: $(X, y)$ the observations, integer $P$ the number of different centers, integer $Q_p$
  the number of components with center $p$, for $p = 1, \ldots, P$
for $p \leftarrow 1$ to $P$ do
    $x_p \leftarrow$ PlaceCenter($P$, X);
    for $i \leftarrow 1$ to $m$ do
        $l_i^{(p)} \leftarrow$ ChooseLenghtscale($x_p$, $X_i$);
        $(X_i^{(p)}, y_i^{(p)}) \leftarrow$ Regime($X_i$, $y_i$, $x_p$, $l_i^{(p)}$);
        $(\mu_i^{(p)}, \Sigma_i^{(q)}, a_i^{(q)}) \leftarrow$ RegimeChannelInitialization($X_i^{(p)}$, $y_i^{(p)}$, $Q_p$);
    end
end
Output: MOHSM parameters $\mu_i^{(q)}$, $\Sigma_i^{(q)}$, $a_i^{(q)}$ for $q = 1, \ldots, Q_p$, $i = 1, \ldots, M$ and
  $p = 1, \ldots, P$
```

2.5. Varying the Global Component of the Spectral Densities

An interesting variation of the MOHSM model is, instead of modeling the global component of the spectral density with an exponential, consider a symmetric rectangle function. Considering symmetric rectangle function as spectral densities is not arbitrary, Tobar [33], motivated by the potential applications relating to signal processing, used it to craft the *sinc kernel*. In this Section, we investigate the possibility of using our MOHSM framework but with a different global component. To do so, we first review the definition of the *sinc kernel* generalized to input spaces of dimension d, which is done by taking the product of d one-dimensional blocks:

Definition 2.2 *The symmetric rectangle function with center μ, width σ and power α has the form:*

$$\text{simrect}_{\mu,\sigma,\alpha}(\omega) = \frac{\alpha}{2}(\text{rect}_{\mu,\sigma}(\omega) + \text{rect}_{-\mu,\sigma}(\omega)) \tag{2.31}$$

where $\sigma > 0$, $\alpha, \mu \geq 0$ and $\text{rect}_{\mu,\sigma}$ *denote the rectangular function with center μ and width σ, given by*

$$\text{rect}_{\mu,\sigma}(\omega) = \begin{cases} \prod_{k=1}^{d} \frac{1}{\sigma_k} & \textit{if all } |\omega_k - \mu_k| < \frac{1}{2}\sigma_k \\ 0 & \textit{otherwise} \end{cases} \tag{2.32}$$

Using the Bochner's theorem with the symmetric rectangle function as spectral density yield to the well-known *Sinc kernel*, which has the following form:

Definition 2.3 *The sinc kernel has the form:*

$$k(\tau) = \alpha \cos\Big(\tau^\top \mu\Big) \prod_{k=1}^{d} \operatorname{sinc}(\tau_k \sigma_k), \tag{2.33}$$

where $\sigma > 0$, $\alpha, \mu \geq 0$ and $\operatorname{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$ is known as the the normalized sinc function.

In order to integrate the *sinc kernel* in our framework, we combine the approach of Simpson et al. [34], where they extend the idea of the *sinc kernel* to MOGP, and our MOHSM. Therefore, we consider the following S_{ij}, where the local component is modelled by a exponential and the global component is modelled by a symmetric rectangle function:

$$S_{ij}(\omega, \omega') = \exp\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2\right) \operatorname{simrect}_{\alpha_{ij},\mu,\sigma}(\overline{\omega}), \tag{2.34}$$

where l_{ij} are defined alike the MOHSM model. Hence, in virtue of Kakihara's theorem, the covariance functions are given by

$$k_{ij}(x, x') = \exp\left(-\frac{l_{ij}^2}{2}\|\overline{x}\|^2\right) \cos\left(\tau^\top \mu\right) \prod_{k=1}^{d} \operatorname{sinc}(\tau_k \sigma_k), \tag{2.35}$$

This multivariate kernel is a non-stationary extension of the approach of Simpson et al. [34] which they proved that can approximate any stationary multi-output kernel to arbitrary precision. To generalize the model to allow phase and delay modeling, we extend these real-valued spectral densities by adding an imaginary component to them, as we did in the MOHSM kernel, to yield:

$$S_{ij}(\omega, \omega') = \exp\left(-\frac{1}{2l_{ij}^2}\|\hat{\omega}\|^2\right) \operatorname{simrect}_{\alpha_{ij},\mu,\sigma}(\overline{\omega}) \exp\Big(-i(\theta_{ij}^\top \overline{\omega} + \phi_{ij})\Big), \tag{2.36}$$

where θ_{ij}, ϕ_{ij} are defined alike the MOHSM model. We calculate the inverse generalized Fourier transform of the spectral densities S_{ij} to obtain the following multivariate covariance function

$$k_{ij}(x, x') = \alpha_{ij} \exp\left(-\frac{l_{ij}^2}{2}\|\overline{x}\|^2\right) \cos\left((\tau + \theta_{ij})^\top \mu + \phi_{ij}\right) \prod_{k=1}^{d} \operatorname{sinc}((\tau_k + (\theta_{ij})_k)\sigma_k), \tag{2.37}$$

Similarly to the MOHSM kernel, we increase the rank of the S by considering a mixture of spectral densities defined in eq. (2.36), which yields the expression for the proposed variation:

Definition 2.4 *The variation of the MOHSM considering the sinc kernel for the local component has the following form:*

$$k_{ij}(x, x') = \sum_{q=1}^{Q} \alpha_{ij}^{(q)} \exp\left(-\frac{l_{ij}^{2(q)}}{2}\|\overline{x}\|^2\right) \cos\left((\tau + \theta_{ij}^{(q)})^\top \mu^{(q)} + \phi_{ij}^{(q)}\right) \prod_{k=1}^{d} \operatorname{sinc}((\tau_k + (\theta_{ij}^{(q)})_k)\sigma_k^{(q)}), \tag{2.38}$$

where the Q amplitude matrices $\{\alpha_{ij}^{(q)}\}_{i,j=1}^{m}$ are assumed to be positive definite.

Like the MOHSM kernel, this multivariate covariance function allows for different delays and phases across channels. However, it does not allows to model different center μ and width σ across channels, as the MOHSM does. This opens the door to question whether it is possible to consider a more general family with this type spectral densities fulfilling the Theorem 1.3.

Chapter 3

Experiments

We tested the proposed MOHSM kernel in different settings, first we learnt a synthetic multi-output GP with three component, a reference GP, its derivative and a delayed version of the GP. Then, we applied our proposal kernel to two real-world datasets: a dataset comprising series of gold and oil prices, the NASDAQ and the USD index (henceforth referred to as GONU); and a dataset of a electroencephalography (EEG), which are known to have dependencies between frequencies.

Since the MOSM kernel have shown better results than the others stationary MOGP kernels, like the CONV and the CSM, we only compared our MOHSM with the MOSM. In addition, we compared the MOHSM with 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 kernel is defined as follow:

$$k(x,x') = \sum_{q=1}^{Q} w_q \exp\left(-\frac{1}{2l_q^2}\left\|\frac{x+x'}{2} - c_q\right\|^2\right) \exp\left(-\frac{1}{2}\tau^\top \Sigma_q \tau\right) \cos\!\left(\mu_q^\top \tau\right). \tag{3.1}$$

The above defined kernel is from the family of harmonizable mixture kernels proposed by Shen et al. [25], which can be seen as a SM kernel where each component is windowed and centered in c_q, thus we will called Windowed Spectral Mixture (WSM) kernel.

All models were implemented using the toolkit MOGPTK for GPU-accelerated ML-training of GPs [35], and were executed on a laptop with 8 GB of RAM and a 940MX GPU running Manjaro 5.12.

3.1. Learning Derivatives and Delayed Signals

First, 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 3.1 *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 [1] with the following multivariate covariance function*

$$\mathcal{K}(x,x') = \begin{pmatrix} k(x,x') & \frac{\partial k(x,x')}{\partial x'} \\ \frac{\partial k(x,x')}{\partial x} & \frac{\partial^2 k(x,x')}{\partial x \partial x'} \end{pmatrix} \tag{3.2}$$

The experiment consisted in the reconstruction of the GP, and the interpolation of the derivative and delayed signals over different intervals. We considered $N = 500$ samples in the interval $[-20, 20]$ for each channel. For the experiment, the derivative was computed numerically and we removed observations between $[-10, 5]$ for the derivative and between $[-5, 5]$ for the delayed signal. We randomly splited the dataset70% for training and 30% for testing. The selected non-stationary kernel was the WSM with 2 mixtures, one centered in -20, and the other centered in 20.

In order to measure the performance of the models, we will use the distance of the correlation matrix, which is a metric defined by Herdin et al. [36] that measures the similarity between two covariance matrices, and is defined as follows:

Definition 3.1 *The correlation matrix distance (CMD)[36] is the distance between two correlation matrices K_1 and K_2 as defined by*

$$CMD(K_1, K_2) = 1 - \frac{\mathrm{Tr}(K_1 \cdot K_2)}{\|K_1\| \cdot \|K_2\|} \tag{3.3}$$

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

Fig 3.1 and 3.2 shows that the MOHSM model was able to accurately recovering the auto and cross covariance of the reference GP and the delayed version, but did not achieve the same results with the derivative. This can happen because the derivative of the selected non-stationary kernel is too complex to be modeled with the selected number of mixtures. Despite the above, the model achieves satisfactory results learning the cross-covariance between the GP and the delayed version, and autocovariances without prior information about the delays.

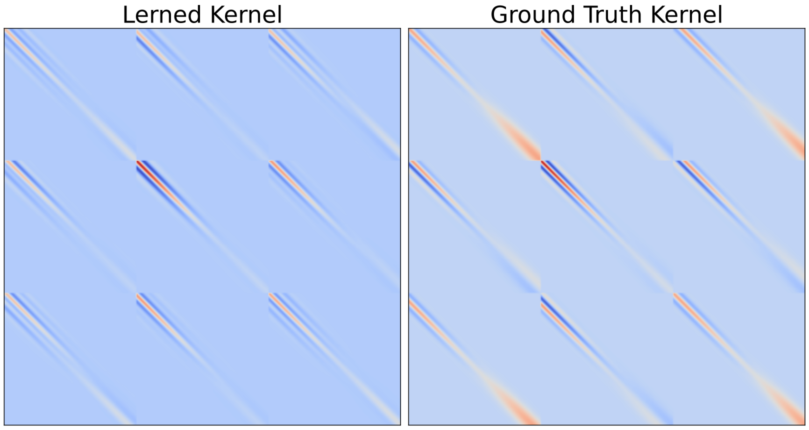

Figure 3.1: **Left:** learned covariance function by MOHSM kernel.
Right: ground truth covariance function of the synthetic dataset.

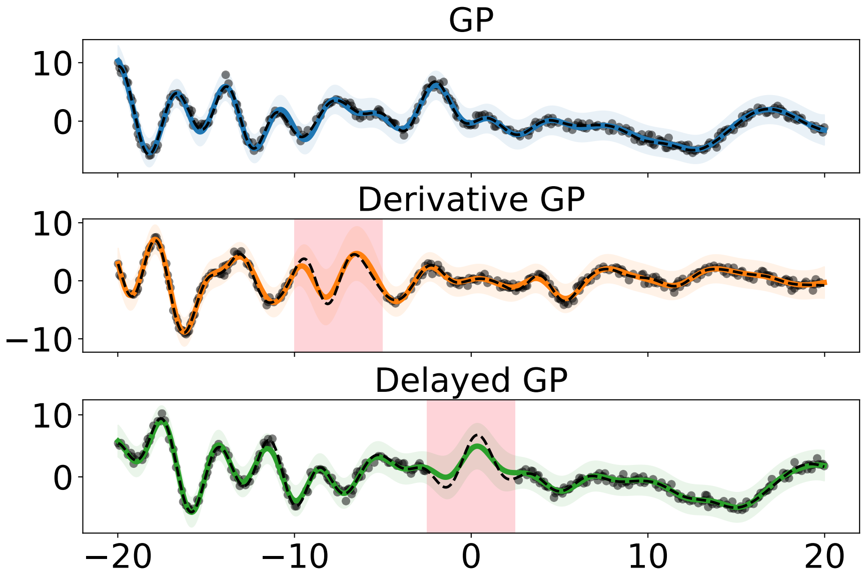

Figure 3.2: Synthetic data set with the trained **MOHSM** kernel. Training points are shown in black, dashed lines are the ground truth and the colour coded lines are the posterior means. The coloured bands show the 95% confidence intervals. The red shaded areas mark the data imputation ranges.

Table 3.1 shows the results of comparing different models against the MOHSM. We performed 5 trials per trained model and reported the mean and he correlation matrix distance (CMD).

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

Table 3.1: Performance indices for the synthetic dataset using the correlation matrix distance (CMD) over 5 realizations

3.2. GONU Data

In previous work, de Wolf et al. [37] proved the capability of the MOSM to impute and predict financial observation by learning the relationships among financial time series. Although the MOSM has shown great performance in these type of problems, the underlying assumption of stationarity is too strong for financial observation. For example, Joyeux [16] have found that, with new housing starts, the high and low frequency components were inter-correlated, and thus the series was nonstationary.

In order to validate our hypothesis that stationarity is a strong requirement, we applied the MOHSM to one of the de Wolf et al. [37] experiments, a dataset comprising series of gold[1] and oil prices[2], the NASDAQ[3] and the USD index[4], between January 2017 and December 2018 with a weekly granularity. The signals were detrented and log-transformed, moreover, to simulate missing data we removed regions in each channel. For gold we removed observations between 2017-03-15 and 2017-05-01; for NASDAQ we removed observations between 2018-04-01 and 2018-07-01; and for oil we removed observations between 2018-10-05 and 2018-12-13. Finally, for each channel, we removed 30% of all observations randomly. Altogether, the experiment included 275 training points and 118 test points.

Finally, we considered the mean absolute percentage error (MAPE), and root-mean-square error (RMSE) as our error metrics, which are defined as follows:

$$\text{MAPE} = \frac{100}{N}\sum_{n=1}^{N}\frac{|y_n - \hat{y}_n|}{y_n} \tag{3.4}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2} \tag{3.5}$$

where N is the number of training points, $\{y_n\}_{n=1}^{N}$ the observations and $\{\hat{y}_n\}_{n=1}^{N}$ the predictive mean of the model at observation locations.

Fig 3.3 shows that the MOHSM model is close to a perfect fit in 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 almost faultless. Table 3.2 shows the results of comparing 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).

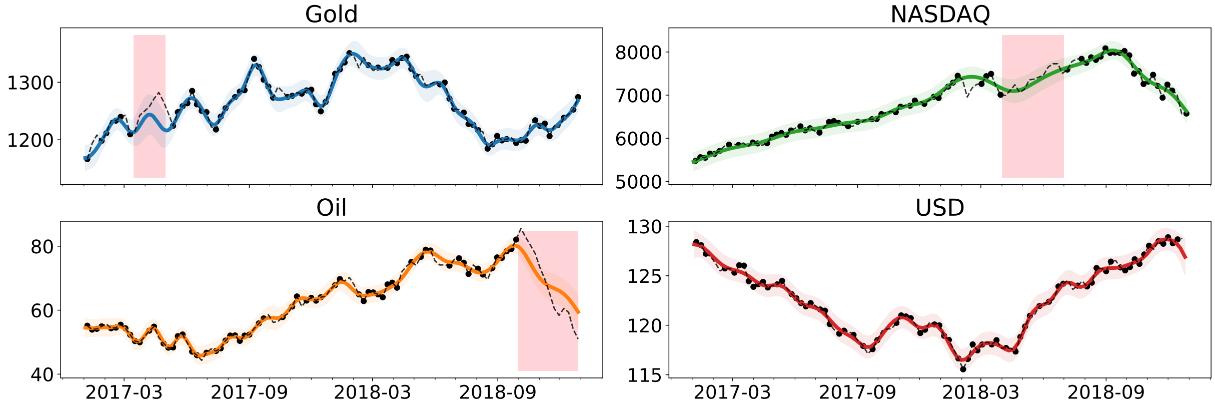

Figure 3.3: GONU data set with the trained **MOHSM** kernel. Training points are shown in black, dashed lines are the ground truth and the colour coded lines are the posterior means. The coloured bands show the 95% confidence intervals. The red shaded areas mark the data imputation ranges.

[1] https://finance.yahoo.com/quote/GC%3DF/history?p=GC%3DF
[2] https://www.eia.gov/dnav/pet/hist/RBRTEd.htm
[3] https://finance.yahoo.com/quote/%5EIXIC/history?p=%5EIXIC
[4] https://fred.stlouisfed.org/series/TWEXB

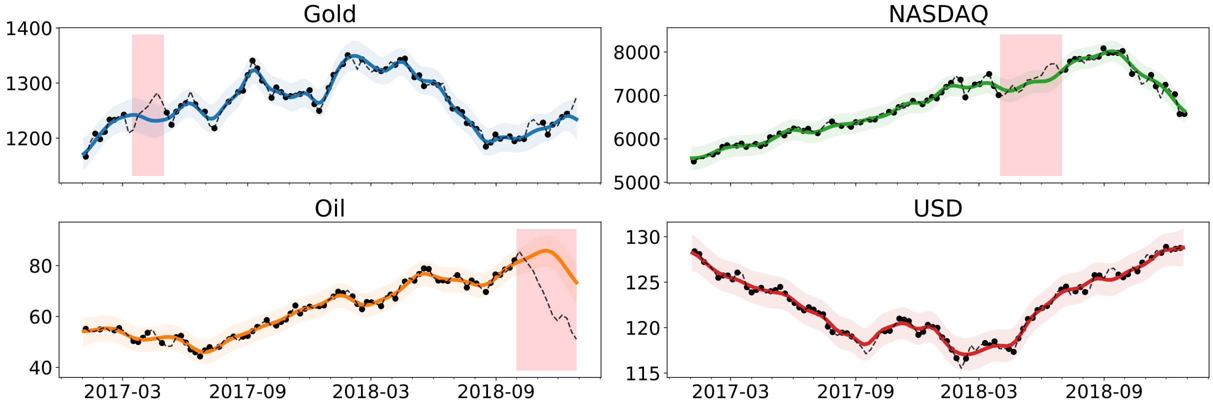

Figure 3.4: GONU data set with the trained **MOSM** kernel. Training points are shown in black, dashed lines are the ground truth and the colour coded lines are the posterior means. The coloured bands show the 95% confidence intervals. The red shaded areas mark the data imputation ranges.

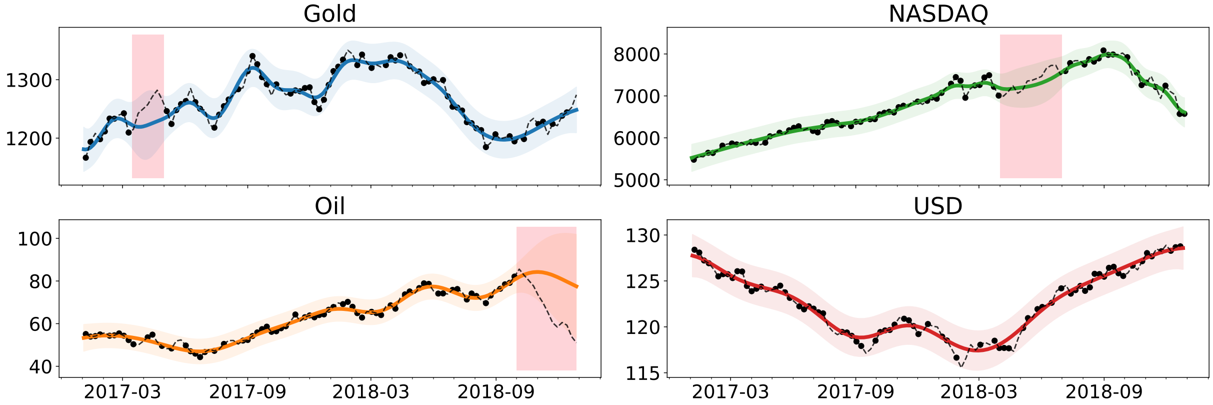

Figure 3.5: GONU data set with the trained **WSM** kernel. Training points are shown in black, dashed lines are the ground truth and the colour coded lines are the posterior means. The coloured bands show the 95% confidence intervals. The red shaded areas mark the data imputation ranges.

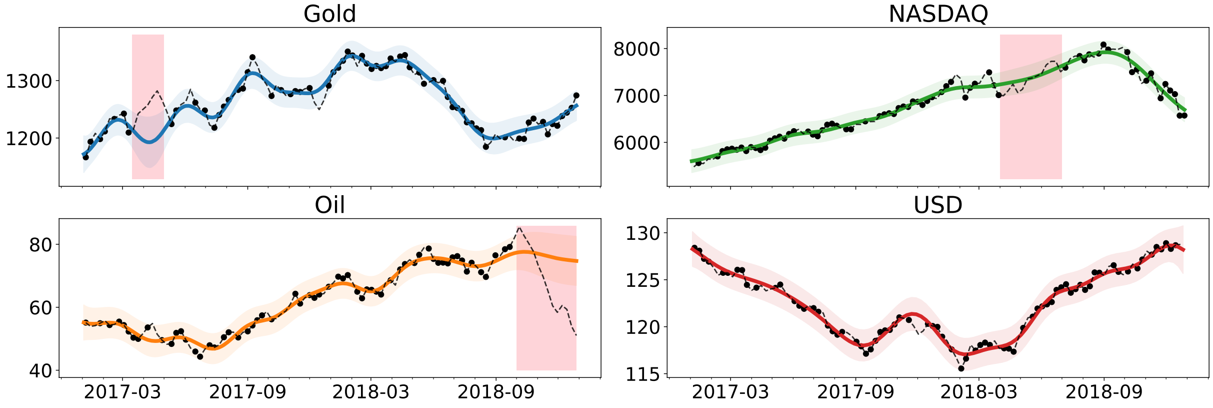

Figure 3.6: GONU data set with the trained **WSM-LMC** kernel. Training points are shown in black, dashed lines are the ground truth and the colour coded lines are the posterior means. The coloured bands show the 95% confidence intervals. The red shaded areas mark the data imputation ranges.

Method	MAPE	RMSE
MOSM	3.05 ± 0.27	49.55 ± 5.79
WSM	3.20 ± 0.37	43.54 ± 2.67
WSM-LMC	2.49 ± 0.40	64.20 ± 12.11
MOHSM	**1.67 ± 0.16**	**40.44 ± 4.74**

Table 3.2: Performance indices for the GONU dataset using the mean absolute percentage error (MAPE) and root mean square error (RMSE) over 5 realizations

3.3. EEG Data

In the field of neuroscience, one of the main challenges is to correctly model encephalography (EEG) data so as to correctly study brain states, in particular, the frequency-based perspective is the standard in multivariate EEG analysis [38]. 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, fig 3.7 shows the position of the electrodes used. We selected a 60-second window resampled at 2 [Hz]. Similar to the GONU experiment, the signals was detrented and log-transformed. We randomly splited the data set 70% for training and 30% for testing. Overall, we trained on 735 points a MOHSM kernel with 4 mixture, and tested on 315 points. In this case, the error metric that we used was the normalized mean absolute error, which are defined as follows:

$$nMAE = \frac{1}{N}\sum_{n=1}^{N}\frac{|y_n - \hat{y}_n|}{\overline{y}} \tag{3.6}$$

where N is the number of training points, $\{y_n\}_{n=1}^{N}$ the observations, $\{\hat{y}_n\}_{n=1}^{N}$ the predictive mean of the model at observation locations, and $\overline{y}$ the mean of the observations.

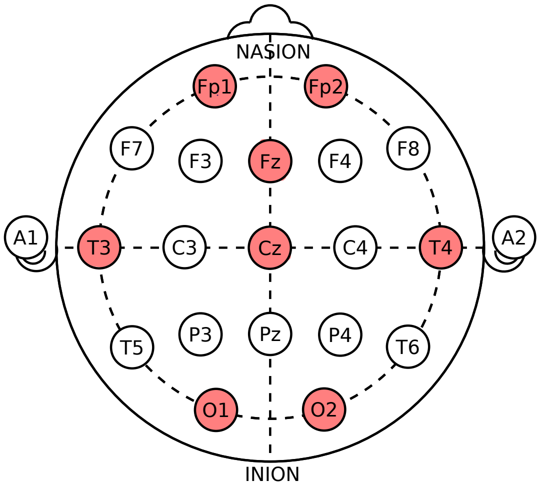

Figure 3.7: Electrode locations for a standard EEG, the red ones are the channels selected for the experiment

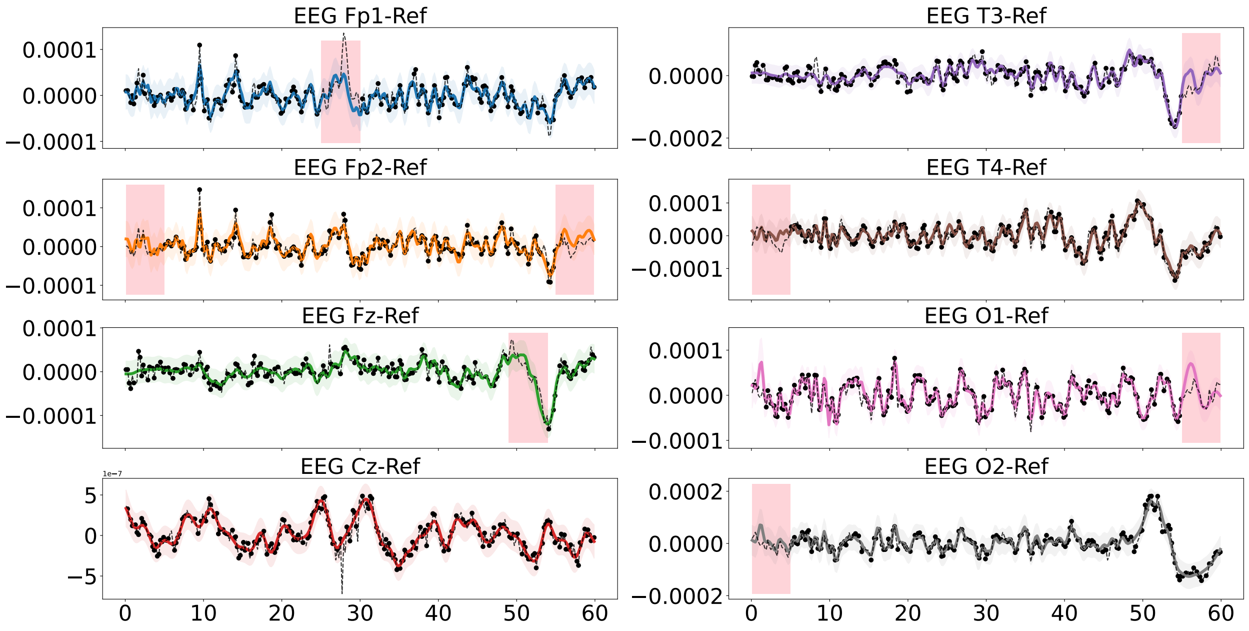

Figure 3.8: EEG data set with the trained **MOHSM** kernel. Training points are shown in black, dashed lines are the ground truth and the colour coded lines are the posterior means. The coloured bands show the 95% confidence intervals. The red shaded areas mark the data imputation ranges.

Table 3.3 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 mean absolute percentage error (MAE), for each channel. Notice that MOHSM performed either better than MOSM or within its error bars for most electrodes.

Model	Fp1	Fp2	Fz	Cz	T3	T4	O1	O2
MOSM	0.16 ± 0.01	0.12 ± 0.01	0.16 ± 0.01	**0.12 ± 0.00**	**0.13 ± 0.02**	0.16 ± 0.02	**0.17 ± 0.00**	0.12 ± 0.02
WSM	0.16 ± 0.01	0.17 ± 0.01	0.17 ± 0.00	0.12 ± 0.01	0.38 ± 0.01	0.24 ± 0.01	0.27 ± 0.00	0.29 ± 0.01
WSM-LMC	0.16 ± 0.01	**0.11 ± 0.01**	0.14 ± 0.01	0.13 ± 0.01	0.16 ± 0.02	0.15 ± 0.01	0.33 ± 0.06	0.17 ± 0.03
MOHSM	**0.14 ± 0.01**	0.13 ± 0.02	**0.13 ± 0.01**	**0.12 ± 0.00**	0.14 ± 0.01	**0.13 ± 0.00**	0.18 ± 0.02	**0.09 ± 0.01**

Table 3.3: Performance for the EEG dataset of each channel using the normalized mean absolute error (nMAE) over 5 realisations

Conclusion

In this thesis, we have presented the multi-output harmonizable spectral mixture (MOHSM) kernel which is a generalization to non-stationary processes of the well-known multi-output spectral mixture (MOSM) kernel. 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 the desires properties of the MOSM: a clear interpretation of the parameters from a spectral viewpoint, and flexibility in each channel. We also presented a parameter initialization scheme, to overcome one of the MOHSM problems, which is the sensitivity of the optimization. Furthermore, we studied possible variation of the proposed framework by using symmetric rectangle function as the global component of the spectral density, instead of squared exponential functions.

We compared the MOHSM kernel against the MOSM and existing MOGP models on two real-world dataset, getting good results in terms of the different error metrics. Therefore, we have showed that our method can effectively model non-stationary data and is a sound extension of the MOSM kernel.

MOHSM is motivated by the intuition that we can model a non-stationary process by considering different regimes within the process itself. This assumption is useful as it allows us to model a wide variety of processes, but it also leads to certain limitations. The first limitation is the number of regimes to consider, which in our model must be previously defined. This requires prior knowledge of the data to be worked on and may lead to overestimating or underestimating the number of components required. The second limitation is that, since the regimes in our model are defined by an exponential, the model is not capable of long-term forecasting.

Future work include considering more complex spectral densities instead of Gaussian functions, this allows to prescind of the infinite differentiability of sampled functions given by spectral mixture kernels. On the other hand, due to the high computational cost of the model, a sparse implementation becomes necessary.

Finally, we hope our proposal contribute towards the goal of developing more expressive kernels to address for challenging problems involving multivariate data, and also to catalyze interest in the harmonizable processes which we believe can contribute a great deal to the machine learning community.

Bibliography

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

[2] C. Williams, E. V. Bonilla, and K. M. Chai, “Multi-task Gaussian process prediction,” *Advances in neural information processing systems*, pp. 153–160, 2007.

[3] G. Parra and F. Tobar, “Spectral mixture kernels for multi-output gaussian processes,” in *Advances in Neural Information Processing Systems*, pp. 6684–6693, 2017.

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

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

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

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

[8] M. Alvarez and N. Lawrence, “Sparse convolved Gaussian processes for multi-output regression,” in *Advances in Neural Information Processing Systems*, pp. 57–64, 2009.

[9] M. A. Alvarez, L. Rosasco, and N. D. Lawrence, “Kernels for vector-valued functions: a review,” *stat*, vol. 1050, p. 16, 2012.

[10] S. Bochner *et al.*, *Lectures on Fourier Integrals*, vol. 42. Princeton University Press, 1959.

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

[12] Z. Zhang, Z. Ren, and W. Huang, “A novel detection method of motor broken rotor bars based on wavelet ridge,” *IEEE Transactions on Energy Conversion*, pp. 417–423, 2003.

[13] J. M. Lilly and S. C. Olhede, “Analysis of modulated multivariate oscillations,” *IEEE Transactions on Signal Processing*, pp. 600–612, 2011.

[14] S. D. Cranstoun, H. C. Ombao, R. Von Sachs, W. Guo, and B. Litt, “Time-frequency spectral estimation of multichannel eeg using the auto-slex method,” *IEEE transactions on Biomedical Engineering*, pp. 988–996, 2002.

[15] H. C. Ombao, J. A. Raz, R. von Sachs, and B. A. Malow, “Automatic statistical analysis of bivariate nonstationary time series,” *Journal of the American Statistical Association*, pp. 543–560, 2001.

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

[17] A. Wilson and R. Adams, "Gaussian process kernels for pattern discovery and extrapolation," in *International conference on machine learning*, pp. 1067–1075, PMLR, 2013.

[18] Y. W. Teh, M. Seeger, and M. I. Jordan, "Semiparametric latent factor models," in *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, pp. 333–340, PMLR, 2005.

[19] Y. Kakihara, *Multidimensional Second Order Stochastic Processes*, vol. 2. World Scientific, 1997.

[20] M. Morse, "Bimeasures and their integral extensions," *Annali di Matematica Pura ed Applicata*, pp. 345–356, 1955.

[21] H. Hurd, "Testing for harmonizability," *IEEE Transactions on Information Theory*, vol. 19, pp. 316–320, 1973.

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

[23] R. Silverman, "Locally stationary random processes," *IRE Transactions on Information Theory*, vol. 3, pp. 182–187, 1957.

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

[25] Z. Shen, M. Heinonen, and S. Kaski, "Harmonizable mixture kernels with variational Fourier features," in *Proceedings of International Conference on Artificial Intelligence and Statistics*, pp. 3273–3282, 2019.

[26] S. Remes, M. Heinonen, and S. Kaski, "Non-stationary spectral kernels," in *Advances in Neural Information Processing Systems*, pp. 4645–4654, 2017.

[27] M. G. Genton, "Classes of kernels for machine learning: a statistics perspective," *Journal of machine learning research*, no. Dec, pp. 299–312, 2001.

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

[29] A. Schuster, "On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena," *Terrestrial Magnetism*, pp. 13–41, 1898.

[30] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, pp. 70–73, 1967.

[31] N. R. Lomb, "Least-squares frequency analysis of unequally spaced data," *Astrophysics and space science*, pp. 447–462, 1976.

[32] F. Tobar, "Bayesian nonparametric spectral estimation," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 10148–10158, 2018.

[33] F. Tobar, "Band-limited Gaussian processes: The sinc kernel," *Advances in Neural In-*

formation Processing Systems, pp. 12749–12759, 2019.

[34] F. Simpson, A. Boukouvalas, V. Cadek, E. Sarkans, and N. Durrande, “The minecraft kernel: Modelling correlated gaussian processes in the Fourier domain,” in *International Conference on Artificial Intelligence and Statistics*, pp. 1945–1953, PMLR, 2021.

[35] T. de Wolff, A. Cuevas, and F. Tobar, “MOGPTK: The Multi-Output Gaussian Process Toolkit,” *Neurocomputing*, 2020.

[36] M. Herdin, N. Czink, H. Ozcelik, and E. Bonek, “Correlation matrix distance, a meaningful measure for evaluation of non-stationary mimo channels,” in *IEEE 61st Vehicular Technology Conference*, pp. 136–140, 2005.

[37] T. de Wolff, A. Cuevas, and F. Tobar, “Gaussian process imputation of multiple financial series,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8444–8448, 2020.

[38] C. Gorrostieta, H. Ombao, and R. Von Sachs, “Time-dependent dual-frequency coherence in multivariate non-stationary time series,” *Journal of Time Series Analysis*, pp. 3–22, 2019.