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

SPECTRAL MIXTURE KERNELS FOR MULTI-OUTPUT GAUSSIAN PROCESSES

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

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

GABRIEL ENRIQUE PARRA VÁSQUEZ

PROFESOR GUÍA:
FELIPE TOBAR HENRÍQUEZ

MIEMBROS DE LA COMISIÓN:
KARIM PICHARA BAKSAI
JAIME SAN MARTÍN ARISTEGUI
JORGE SILVA SÁNCHEZ

SANTIAGO DE CHILE
2017

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: GABRIEL ENRIQUE PARRA VÁSQUEZ
FECHA: 2017
PROF. GUÍA: SR. FELIPE TOBAR HENRÍQUEZ

SPECTRAL MIXTURE KERNELS FOR MULTI-OUTPUT GAUSSIAN PROCESSES

Multi-Output Gaussian Processes (MOGPs) are the multivariate extension of Gaussian processes (GPs [1]), a Bayesian nonparametric method for univariate regression. MOGPs address the multi-channel regression problem by modeling the correlation in time and/or space (as scalar GPs do), but also across channels and thus revealing statistical dependencies among different sources of data. This is crucial in a number of real-world applications such as fault detection, data imputation and financial time-series analysis.

Analogously to the univariate case, MOGPs are entirely determined by a multivariate covariance function, which in this case is matrix valued. The design of this matrix-valued covariance function is challenging, since we have to deal with the trade off between (i) choosing a broad class of cross-covariances and auto-covariances, while at the same time (ii) ensuring positive definiteness of the symmetric matrix containing these scalar-valued covariance functions. In the stationary univariate case, these difficulties can be bypassed by virtue of Bochner's theorem, that is, by building the covariance function in the spectral (Fourier) domain to then transform it to the time and/or space domain, thus yielding the (single-output) Spectral Mixture kernel [2].

A classical approach to define multivariate covariance functions for MOGPs is through linear combinations of independent (latent) GPs; this is the case of the Linear Model of Coregionalization (LMC [3]) and the Convolution Model [4]. In these cases, the resulting multivariate covariance function is a function of both the latent-GP covariances and the linear operator considered, which usually results in symmetric cross-covariances that do not admit lags across channels. Due to their simplicity, these approaches fail to provide interpretability of the dependencies learnt and force the auto-covariances to have similar structure.

The main purpose of this work is to extend the spectral mixture concept to MOGPs: We rely on Cramér's theorem [5, 6], the multivariate version of Bochner's theorem, to propose an expressive family of complex-valued square-exponential cross-spectral densities, which, through the Fourier transform yields the Multi-Output Spectral Mixture kernel (MOSM). The proposed MOSM model provides clear interpretation of all the parameters in spectral terms. Besides the theoretical presentation and interpretation of the proposed multi-output covariance kernel based on square-exponential spectral densities, we inquiry the plausibility of complex-valued t-Student cross-spectral densities. We validate our contribution experimentally through an illustrative example using a tri-variate synthetic signal, and then compare it against all the aforementioned methods on two real-world datasets.

Agradecimientos

Agradezco al Centro de Modelamiento Matemático por el financiamiento mientras trabajaba en esta tesis mediante el proyecto: BASAL PFB 03.

Quiero agradecer al profesor Felipe Tobar H. por haber dirigido este trabajo, por su apoyo tanto en recursos computacionales como económicos, por su ayuda y empuje en que esta tesis desembocara en una publicación, pero además, por su constante trabajo en el establecimiento de un área de *Machine Learning* dentro de la facultad.

Quiero agradecer al Departamento de Ingeniería Matemática, a sus académicos, funcionarios y estudiantes, que me enseñaron que el rigor y la excelencia no es un acto, sino un habito.

Agradezco a mis padres por darme la libertad de ser el arquitecto de mi propio destino y por su confianza y apoyo incondicional de inicio a fin durante esta carrera.

Agradezco a mis hermanos: Roberto, Erick y Javier, que pavimentaron este camino pedregoso, sin lugar a dudas no hubiera podido realizar todo esto sin ellos.

Finalmente, agradezco a la vida, por darme la oportunidad de estudiar y comprender este hermoso lenguaje que son las Matemáticas.

Contents

Introduction

One of the most elementary problems in Statistics is that of regression, which can be defined as the estimation of the relationship between independent variables (also known as input variables) and dependent variables (also known as output variables), the first ones being usually *time* and/or *space* and the second ones being, scalar or vectorial representing a continuous quantity.

The Bayesian approach to the regression problem consists in the choice of a *prior distribution* over the relationship between input an output variables, which in conjunction with a *likelihood function* for observed data, yields a posterior distribution through Bayes' theorem, that can be used for prediction and forecasting. Within this context, in particular in the Machine Learning field, Gaussian processes (GPs [1]) play the role of prior distributions over continuous functions. This prior is fully characterized by a *covariance function or positive-definite kernel* that represents to what degree two similar input variables produce similar or *close* output variables. This generates a full non-parametric Bayesian method for univariate regression, i.e. the output variable is scalar, which excels at flexibility, tractability and interpretability due to its explicit matrix calculations and since it is a probabilistic method allows to define confidence intervals for variance estimation ranges.

Unfortunately, all these useful advantages are restricted to univariate regression. Gaussian processes can be generalized to tackle multivariate regression, i.e. the output variable is now a vector and each component of the vector output variable is called an *output* or a *channel*. This generalization it is known as Multi-Output Gaussian Processes (MOGPs [7]) which, analogously to the univariate case, yields a prior distribution over multivariate continuos functions that is fully characterized by a *multivariate covariance function* or *multivariate kernel*. This multivariate covariance function breaks down into two kind of functions: *(i) auto-covariance functions* that represent to what degree two similar input variables produce similar output variables at each channel and *(ii) cross-covariance functions* that describes the relationship across the different channels. This last peculiarity is the principal attribute of MOGPs as it allows us to share *statistical strength* across channels, having applications such as fault detection, data imputation and denoising, among others.

While, for the univariate case, there is an entire family of covariance functions to choose, the principal obstacle to the proper use of MOGPs in multivariate regression is that there is no clear way for constructing multivariate covariance functions, this is due the fact that *cross-covariance functions* are difficult to define. An arbitrary choice for these functions does not guarantee positive definiteness of the matrices needed for inference. Most of the

previous work [7, 3, 8, 4] goes in the direction of bypassing this problem by applying linear operations over latent processes which is equivalent to linear parametrizations of the *cross-covariance functions*, which, while simple and handy, are limited by its linear parametric nature. The main purpose of this thesis is to propose a parametric generative model for stationary multivariate covariance functions that allows a fully interpretative understanding of its parameters and that allows flexible cross-covariance while maintaining the autonomy of auto-covariances, which is achieved through the spectral representation theorem of stationary multivariate covariance functions known as Cramér's theorem [5, 6].

This work is an extended version of the paper "Spectral Mixture Kernels for Multi-Output Gaussian Processes" to appear in the Proceeding of Advances in Neural Processing Systems (NIPS 2017) and it is organized as follows. The first chapter describes the preliminary notions of GP and MOGPs along with the previous work. The second chapter describes the spectral representation of multivariate covariance function which yields to the proposed model. In the third chapter experiments of the proposed model are compared against previous approach using synthetic data and real-world data, to end with the conclusions.

Chapter 1

Background

The objective of this chapter is to describe the Gaussian processes methods for univariate regression problems along with the kernel selection problem that they pose and its suggested solution through spectral representations. We also describe the generalization of Gaussian processes to multivariate regression which is known in the Machine Learning field as Multi-Output Gaussian Processes, in addition to its analogous kernel selection problem. We conclude with a review of the state-of-art models for MOGPs.

From now on, the input space, i.e. the set of input variables, will be denoted $\mathcal{X}$ which without loss of generality will be assumed to be $\mathbb{R}^n$, on the other hand, the output space i.e the set of possible output variables, will be denoted $\mathcal{Y}$ and assumed to be $\mathbb{R}^m$, an input-output pair will be denoted (x, y) and a set of N of such pairs, that is $\mathcal{D} = \{(x_i, y_i) : i = 1, \ldots, N\}$ will be called the data set or training set. It is worth mentioning that there are two kinds of training sets, the first kind called *isotopic* where for a given training input x_i, all the coordinates of the training output y_i are known, meanwhile, the second kind called *heterotopic* where for a training input x_i only some coordinates of y_i are known (it can be assumed that only one coordinate its known by considering repetitions of training inputs), under this circumstance the training set it is described as a disjoint union of training sets for each output coordinate, namely, $\mathcal{D} = \cup_{j=1}^{m} \mathcal{D}_j$ where $\mathcal{D}_j = \{(x_i, y_{ij}) : i = 1, \ldots, N_j\}$. For simplicity and understandability the training sets will be assumed to be isotopic, whereas in the experiments, these will be heterotopic.

A regression problem can be thought as the estimation of the output $y \in \mathcal{Y}$ for any given input variable $x \in \mathcal{X}$, for which an estimator $f \colon \mathcal{X} \to \mathcal{Y}$ is built using the data from a training set $\mathcal{D}$ by minimizing, for each pair $(x_i, y_i) \in \mathcal{D}$, some measure of error between the estimation $f(x_i)$ and the observation y_i. From a Bayesian perspective, this problem can be solved by choosing a *prior distribution* for the estimator $f(x)$ which in conjunction with a *likelihood function* for the training set, leads to a *posterior distribution* that can be used for prediction. Within this context, the Gaussian processes methods capitalize on the appealing properties of the Gaussian distribution, in particular its marginalization closure and its explicit and tractable calculations, to produce a full non-parametric Bayesian approach to univariate regression problems. The formalization of this method is the goal of the next section.

1.1 Gaussian Processes

Gaussian processes (GPs [1], also known as *kriging* within Geostatistics [3]) are an univariate non-linear regression method (this is $\mathcal{Y} = \mathbb{R}$) which excel in terms of flexibility, simplicity and tractability. More precisely, GPs are stochastic processes that define probability distributions over continous functions which can be seen as *prior distributions* that whenever combined with observed data through a Gaussian *likelihood function* we obtain a *posterior distribution* which is also a Gaussian process, namely

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

From a classic probabilistic point of view, this definition is equivalent to stating a Gaussian process is a stochastic process such that every finite-dimensional distribution is a Gaussian distribution, then, in virtue of Kolmogorov consistency theorem [9] the resulting stochastic process is well defined and consequently there is no mistake in reasoning Gaussian processes as a probability distributions over continuous functions. A key property of GPs is that they are entirely determined by its mean function $m(x)$ and its covariance function $k(x, x')$, defined as follows

$$\begin{aligned} m(x) &= \mathbb{E}[f(x)], \quad \forall x \in \mathcal{X} \\ k(x, x') &= \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))], \quad \forall x, x' \in \mathcal{X} \end{aligned}$$

without loss of generality, the mean function can be assumed to be zero. The choice of a covariance function determines the sample properties of the prior distribution over functions that a GP embodies by characterizing properties of the process such as: smoothness, periodicity and stationarity among others, but not all two-input function serve as a covariance function, since it must be positive definite and symmetric, specifically

Definition 1.2 *A two-input function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a covariance function (also called kernel) if it is:*

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

(ii) Positive definite, i.e., $\forall N \in \mathbb{N}, c \in \mathbb{R}^N, \{x_p\}_{p=1}^N \subseteq \mathcal{X}$ we have,

$$\sum_{p,q=1}^{N} c_p c_q k(x_p, x_q) \geq 0 \tag{1.1}$$

The terms *covariance function* and *kernel* will be used interchangeably, furthermore, we say that a covariance function is *stationary* if $k(x, x') = k(x - x')$, in this case we denote $\tau = x - x'$. Given two finite sets $X = \{x_i\}_{i=1}^N, X' = \{x'_j\}_{j=1}^{N'} \subseteq \mathcal{X}$ we will denote by $k(X, X')$ the $N \times N'$-matrix where $[k(X, X')]_{ij} = k(x_i, x'_j)$ $\forall i, j \in \{1, \ldots, N\}$ and will be called the *Gram matrix* of the covariance function $k(x, x')$. Usually, covariance functions will have a parametric form and the set of parameters that defines them will be denoted by θ. It is typical to assume that we have *noisy observations*, that is, $y = f(x) + \varepsilon$ for each $(x, y) \in \mathcal{D}$,

where the noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is iid. Training and prediction is performed by *(i)* finding the set of parameters θ that maximize the Gaussian marginal likelihood of the observed data and then *(ii)* conditioning the joint distribution at training inputs and prediction inputs given the observed data which yields to the posterior distribution, namely, given a training set $\mathcal{D} = \{(x_i, y_i) : i = 1, \ldots, N\}$ and let denote $X = [x_1, \ldots, x_N]$ the set of training inputs, $\mathbf{y} = [y_1, \ldots, y_N]$ the set of noisy training outputs, and $\mathbf{f} = f(X) = [f(x_1), \ldots, f(x_N)]$ the set of training latent values, then we have the Gaussian marginal likelihood function

$$p(\mathbf{y}|X, \theta) = \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X, \theta)\mathrm{d}\mathbf{f} = \mathcal{N}\big(\mathbf{y}|\mathbf{0}, k(X, X) + \sigma^2 I_N\big) \tag{1.2}$$

which yields to the log likelihood function

$$\log p(\mathbf{y}|X, \theta) = -\tfrac{1}{2}\mathbf{y}^\top [k(X, X) + \sigma^2 I_N]^{-1}\mathbf{y} - \tfrac{1}{2}\log|k(X, X) + \sigma^2 I_N| - \tfrac{N}{2}\log 2\pi \tag{1.3}$$

and by denoting $\mathbf{y}_* = f(X_*)$, we obtain the posterior distribution by conditioning the joint Gaussian distribution at inputs X, X_* given the noisy observed data $\mathbf{y}$, that is

$$\begin{aligned} p(\mathbf{y}_*|X, \mathbf{y}, X_*) \;=\; & \mathcal{N}\big(\mathbf{y}_*|k(X_*, X)[k(X, X) + \sigma^2 I_N]^{-1}\mathbf{y}, \\ & k(X_*, X_*) - k(X_*, X)[k(X, X) + \sigma^2 I_N]^{-1}k(X, X_*)\big) \end{aligned} \tag{1.4}$$

This posterior distribution its entirely determined by the choice of the covariance function $k(x, x')$. In that regard, flexible families of covariance functions that can automatically learn a proper prior for the data have become of particular interest [2, 10, 11]. An intuitive approach to these automatic kernels are spectral kernels which are the topic of the next section.

1.2 Spectral Representation of Stationary Covariance functions

The spectral representation theorem of stationary complex-valued covariance functions known as Bochner's theorem [12, 6], it is the key result behind of what are called *spectral kernels* which suggest a solution to the kernel selection problem by granting a family of kernels that can approximate any integrable stationary covariance function given enough parameters, to understand this let us remind the definition of Lebesgue-Stieljes measures in $\mathbb{R}^n$ [13].

Let $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n) \in \mathbb{R}^n$, the interval $(a, b] \subset \mathbb{R}^n$ is defined as $\{x = (x_1, \ldots, x_n) \in \mathbb{R}^n \colon a_i < x_i \leq b_i \; for \; all \; i = 1, \ldots, n\}$; the interval $(a, +\infty)$ is defined as $\{x \in \mathbb{R}^n \colon a_i < x_i \; for \; all \; i = 1, \ldots, n\}$, the interval $(-\infty, b]$ as $\{x \in \mathbb{R}^n \colon x_i \leq b_i \; for \; all \; i = 1, \ldots, n\}$. The smallest σ-algebra containing the intervals is denoted by $\mathcal{B}(\mathbb{R}^n)$. Given a real-valued function F in $\mathbb{R}^n$ it is possible to define a measure in $\mathcal{B}(\mathbb{R}^n)$ through this function as long as it fulfill certain conditions, this measure given by F will be called a Lebesgue-Stieljes measure.

Definition 1.3 *Let $F \colon \mathbb{R}^n \to \mathbb{R}$, we say that F is a distribution function if it has the following properties*

- $\lim\limits_{x_i \to -\infty} F(x_1, \ldots, x_n) = 0 \quad \forall i = 1, \ldots, n$

- $F(x_1, x_2, \ldots, x_n)$ *is right-continuous*

These conditions are enough to define *signed* Lebesgue-Stieljes measures in $\mathcal{B}(\mathbb{R}^n)$ through distribution functions, however in order to get a *positive* measure, a monotonic condition its needed, namely

Definition 1.4 *Let* $F\colon \mathbb{R}^n \to \mathbb{R}$*, we define the difference operator* Δ *as*

$$\Delta_{b_i a_i} F(x_1, \ldots, x_n) = F(x_1, \ldots, x_{i-1}, b_i, x_{i+1}, \ldots, x_n) - F(x_1, \ldots, x_{i-1}, a_i, x_{i+1}, \ldots, x_n)$$

Furthermore, for an interval $I = (a, b] \subset \mathbb{R}^n$*, let* $F(I)$ *denote*

$$\Delta_{b_n a_n} \Delta_{b_{n-1} a_{n-1}} \ldots \Delta_{b_1 a_1} F(x_1, \ldots, x_n)$$

We say that a distribution function F *is increasing if* $F(I) \geq 0$ *for any interval* $I \in \mathcal{B}(\mathbb{R}^n)$*.*

Theorem 1.5 *Let* F *an increasing distribution function in* $\mathbb{R}^n$*, for any interval* I*, let define* $\mu_F(I) = F(I)$*, then* μ_F *is a measure over the semi-algebra induced by the intervals and it can be uniquely extended to* $\mathcal{B}(\mathbb{R}^n)$*, this measure will be called a Lebesgue-Stieljes measure [13].*

It is worth mentioning that if we define a Lebesgue-Stieljes measure as in Theorem 1.5 with a bounded distribution function we obtain a *finite measure*, if the distribution function is not increasing we obtain a *signed measure* and if the distribution is complex-valued we obtain a *complex measure*. With this in mind, the following theorem gives a characterization of complex-valued stationary covariance functions in terms of Fourier-Stieljes integrals which can be simplified to simple Fourier transforms through the Lebesgue's decomposition theorem of measures over $\mathcal{B}(\mathbb{R}^n)$.

Theorem 1.6 *(Bochner's theorem [12, 6]) A function* $k : \mathbb{R}^n \to \mathbb{C}$ *is the covariance function of a weakly-stationary mean-square-continuous stochastic process* $f : \mathbb{R}^n \to \mathbb{C}$ *if and only if it admits the following representation*

$$k(\tau) = \int_{\mathbb{R}^n} e^{\iota \omega^\top \tau} \, \mathrm{d}\mu_F(\omega) \tag{1.5}$$

where $F(\omega)$ *is a bounded distribution function on* $\mathbb{R}^n$ *and* ι *denotes the imaginary unit.*

This theorem is the key result behind the Spectral Mixture kernel (SM) proposed by Wilson & Adams [2], although they make an important simplification to this theorem in order to make it suitable to applications, that is, they consider only the absolute continuous part of the finite measure μ_F. Lebesgue's decomposition theorem [14] applied on the finite measure μ_F gives us the representation

$$\mu_F = \mu_F^{(a)} + \mu_F^{(s)} \tag{1.6}$$

where $\mu_F^{(a)}$ is the absolute continuous part of the measure with respect Lebesgue's measure in $\mathcal{B}(\mathbb{R}^n)$ and $\mu_F^{(s)}$ is the singular part of the measure. By neglecting the singular part of the measure and focusing only in the absolute continuous part which has a positive density $S(\omega)\colon \mathbb{R}^n \to \mathbb{R}_+$ with respect Lebesgue's measure, Theorem 1.6 can be simplified to a more convenient version.

Theorem 1.7 *(Bochner's theorem simplified [6]) An integrable*[1] *function* $k : \mathbb{R}^n \to \mathbb{C}$ *is the covariance function of a weakly-stationary mean-square-continuous stochastic process* $f : \mathbb{R}^n \to \mathbb{C}$ *if and only if it admits the following representation*

$$k(\tau) = \int_{\mathbb{R}^n} e^{\iota\omega^\top \tau} S(\omega) \mathrm{d}\omega \tag{1.7}$$

where $S(\omega)$ *is a non-negative integrable function on* $\mathbb{R}^n$ *called the spectral density of* $k(\tau)$*.*

This simplified version of Bochner's theorem gives an explicit relationship between the spectral density $S(\omega)$ and the stationary covariance function $k(\tau)$ of the stochastic process $f(x)$, in this sense [2] proposed to model the integrable spectral density $S(\omega)$ as a weighted mixture of Q square-exponential functions, with weights w_q, centres μ_q and diagonal covariance matrices Σ_q, that is,

$$S(\omega) = \sum_{q=1}^{Q} w_q \frac{1}{(2\pi)^{n/2}|\Sigma_q|^{1/2}} \exp\left(-\tfrac{1}{2}(\omega - \mu_q)^\top \Sigma_q^{-1}(\omega - \mu_q)\right). \tag{1.8}$$

This is motivated by the fact that strict positivity is much easier to achieve than positive definiteness, thus, by relying on Theorem 1.6, the stationary covariance function associated to the spectral density $S(\omega)$ in eq. (1.8) is given the spectral mixture kernel defined as follows.

Definition 1.8 *A spectral-mixture (SM) kernel is a positive-definite stationary covariance function given by*

$$k(\tau) = \sum_{q=1}^{Q} w_q \exp\left(-\frac{1}{2}\tau^\top \Sigma_q \tau\right) \cos(\mu_q^\top \tau) \tag{1.9}$$

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

Due to the universal function approximation of the sum-of-Gaussians [15] (considered here in the frequency domain) and the relationship given by Theorem 1.6, the SM kernel is able to approximate continuous stationary kernels at an arbitrary precision given enough spectral components. This concept points in the direction of sidestepping the kernel selection problem in GPs and it will be extended to cater for multivariate regression in the next chapter.

1.3 Multi-Output Gaussian Processes

GPs can be extended to tackle multivariate regression (i.e $\mathcal{Y} = \mathbb{R}^m$, $m > 1$) by considering a ensemble of univariate stochastic processes that are jointly Gaussian, this generalization is known as Multi-Output Gaussian Processes (MOGPs) in the Machine Learning field and is also known as *co-kriging* within Geostatistics [3]. This ensemble yields to a vector-valued stochastic process such that it will also be completely determined by its covariance function,

[1] A function $g(x)$ is said to be integrable if $\int_{\mathbb{R}^n} |g(x)| \mathrm{d}x < +\infty$

but alike the univariate case, the construction of this covariance function will be the principal obstacle to a proper and effective employment of this method in multivariate regression problems since a positive-definite like condition complicates the modeling of dependencies between each component of the output. The formalization of the ideas above is as follows

Definition 1.9 *An m-channel Multi-Output Gaussian process is an m-tuple of stochastic processes $\{(f_1(x), f_2(x), \dots, f_{m-1}(x), f_m(x)) \colon x \in \mathcal{X}\}$, is an m-tuple of stochastic processes $\{f_p(x) \colon x \in \mathcal{X}\}$ $\forall p = 1, \dots, m$, such that for any family $\{X_i\}_{i=1}^m$ of finite subsets of $\mathcal{X}$, the random vector $[f_1(X_1), \dots, f_m(X_m)]$ is a multivariate Gaussian random variable.*

Each component of the vector $\mathbf{f}(x) := (f_1(x), \dots, f_m(x))$ $\forall x \in \mathcal{X}$, will be called a *channel* or an *output*. Recall that the construction of a scalar-valued GP requires choosing a scalar-valued mean function and a scalar-valued covariance function. Conversely, the construction of a m-channel MOGP requires an $\mathbb{R}^m$-valued mean function whose i^{th} element denotes the mean function of the i^{th} channel, and an $\mathbb{R}^m \times \mathbb{R}^m$-valued covariance function whose $(i,j)^{\text{th}}$ element denotes the covariance between the i^{th} and j^{th} channels, this two functions are defined as follows

$$\begin{aligned} \mathbf{m}(x) &= \mathbb{E}[\mathbf{f}(x)], \quad \forall x \in \mathcal{X} \\ \mathcal{K}(x, x') &= \mathbb{E}[(\mathbf{f}(x) - \mathbf{m}(x))(\mathbf{f}(x') - \mathbf{m}(x'))^\top], \quad \forall x, x' \in \mathcal{X} \end{aligned}$$

Analogously to the scalar-valued GPs, the vector-valued mean function can be assumed to be zero, meanwhile, the symmetry and positive-definiteness conditions of the MOGP kernel are defined as follows

Definition 1.10 *A two-input matrix-valued function $\mathcal{K}(x, x') : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m \times m}$ defined element-wise by $\left[\mathcal{K}(x, x')\right]_{ij} = k_{ij}(x, x')$ is a multivariate covariance function (kernel) if it is:*

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

(ii) Positive definite, i.e., $\forall N \in \mathbb{N}, \{c_p\}_{p=1}^N \subseteq \mathbb{R}^m, \{x_p\}_{p=1}^N \subseteq \mathcal{X}$, we have

$$\sum_{i,j=1}^{m} \sum_{p,q=1}^{N} c_{pi} c_{qj} k_{ij}(x_p, x_q) \geq 0 \tag{1.10}$$

Furthermore, we say that a multivariate kernel $\mathcal{K}(x, x')$ is stationary if $\mathcal{K}(x, x') = \mathcal{K}(x - x')$ or equivalently $k_{ij}(x, x') = k_{ij}(x - x')$ $\forall i, j \in \{1, \dots, m\}$, in this case, we denote $\tau = x - x'$. Note that the positive-definite condition (1.10) imposes the diagonal components of a multivariate covariance function $\mathcal{K}(x, x')$ (i.e. the functions $k_{ii}(x, x')$ $\forall i \in \{1, \dots, m\}$), to be positive-definite as in (1.1), that is, they are univariate covariance functions, conversely, the off-diagonal components (i.e the functions $k_{ij}(x, x')$ for $i \neq j$) are not restricted to be positive-definite as in (1.1), hence its design is much more complicated since we cannot rely on representations as Theorem 1.7.

Given two finite sets $X = \{x_i\}_{i=1}^N, X' = \{x'_j\}_{j=1}^{N'} \subseteq \mathcal{X}$ we will denote by $\mathcal{K}(X, X')$ the $mN \times mN'$-block-matrix where the (i,j)-block is $k_{ij}(X, X')$ $\forall i, j \in \{1, \dots, m\}$ and will be called the *Gram matrix* of the multivariate covariance function $\mathcal{K}(x, x')$. Similarly to the univariate case, usually multivariate covariance functions will have a parametric form and the set of parameters them will be denoted by Θ.

The design of multivariate covariance functions involves jointly choosing functions that model the covariance of each channel (diagonal elements in $\mathcal{K}$, called *auto-covariances*) and functions that model the cross-covariance between different channels at different input locations (off-diagonal elements in $\mathcal{K}$, conveniently called *cross-covariances*). Choosing these $m(m+1)/2$ covariance functions is challenging when we want to be as expressive as possible and include, for instance, delays, phase shifts, negative correlations or to enforce specific spectral content while at the same time maintaining positive definiteness of $\mathcal{K}$.

Analogously to the univariate case, training and prediction in MOGPs is done by (i) finding the set of parameters Θ that maximizes the Gaussian marginal likelihood of the observed data and then (ii) conditioning the joint distribution at training inputs and prediction inputs given the observed data which yields to the posterior distribution, namely, given a isotopic training set $\mathcal{D} = \{(x_i, y_i) : i = 1, \ldots, N\}$ and let denote $X = [x_1, \ldots, x_N] \in \mathbb{R}^{nN}$ the concatenated training inputs, $\mathbf{y} = [y_1, \ldots, y_N] \in \mathbb{R}^{mN}$ the concatenation of noisy training outputs, and $\mathbf{F} = [\mathbf{f}(x_1), \ldots, \mathbf{f}(x_N)] \in \mathbb{R}^{mN}$ the contatenated training latent values, then we have the Gaussian marginal likelihood function

$$p(\mathbf{y}|X, \Theta) = \int p(\mathbf{y}|\mathbf{F}, X)p(\mathbf{F}|X, \Theta)\mathrm{d}\mathbf{F} = \mathcal{N}\big(\mathbf{y}|\mathbf{0}, \mathcal{K}(X, X) + \Sigma\big) \tag{1.11}$$

where Σ is the diagonal matrix of output-dependent noises, the above Gaussian marginal likelihood yields to the log likelihood function

$$\log p(\mathbf{y}|X, \Theta) = -\tfrac{1}{2}\mathbf{y}^\top[\mathcal{K}(X, X) + \Sigma]^{-1}\mathbf{y} - \tfrac{1}{2}\log|\mathcal{K}(X, X) + \Sigma| - \tfrac{mN}{2}\log 2\pi \tag{1.12}$$

and by denoting $\mathbf{y}_* = [f_1(X_*), \ldots, f_m(X_*)]$, we obtain the posterior distribution by conditioning the joint Gaussian distribution at inputs X, X_* given the noisy observed data $\mathbf{y}$, that is

$$\begin{aligned} p(\mathbf{y}_*|X, \mathbf{y}, X_*, \Theta) \;=\; & \mathcal{N}\big(\mathbf{y}_*|\mathcal{K}(X_*, X)[\mathcal{K}(X, X) + \Sigma]^{-1}\mathbf{y}, \\ & \mathcal{K}(X_*, X_*) - \mathcal{K}(X_*, X)[\mathcal{K}(X, X) + \Sigma]^{-1}\mathcal{K}(X, X_*)\big) \end{aligned} \tag{1.13}$$

This MOGP posterior distribution allows joint inference between channels through the cross-covariance functions, that is why the proper modelling of these functions is fundamental for an effective use of this method for multivariate regression, unfortunately there is no clear insight on how to build flexible cross-covariances. In the next section, previous work on this issue will be reviewed.

1.4 Previous work

As pointed out in the previous section, MOGPs require a matrix-valued covariance function which has to be symmetric and positive-definite, the design of each component of this multivariate covariance function, while intuitive, it is complicated due to condition (1.10). In this section, previous proposals of multivariate covariance functions will be review and as it will be seen, they all bypass the difficulty on the design by relying on linear operations over latent processes.

1.4.1 Linear Model of Coregionalization (LMC)

The most fundamental and basic idea comes from Geostatistics and it is known as the Linear Model of Coregionalization (LMC [3]). There are two equivalent ways to conceive this proposed model, on one side, it is reasonable to construct m correlated processes by considering each one of them as generated by the same source process $u(x)$ to which an operation has been applied, on the other hand, it is possible to choose an univariate covariance function and *expand* it to a multivariate covariance function through multiplications with positive definite matrices, more formally, let $\{u_q(x)\colon x \in \mathcal{X}, q = 1, \dots, Q\}$ a family of Q Gaussian processes, called *latent processes*, each with covariance function $k_q(x, x')$, the essential idea in the LMC is to consider each channel $(f_1(x), \dots, f_m(x))$ of a m-channel MOGP as a linear combination of these Q latent processes, namely

$$f_j(x) = \sum_{q=1}^{Q} a_{jq} u_q(x) \quad \forall j = 1, \dots, m \tag{1.14}$$

then, the m channels are naturally correlated and the covariance between them can be calculated explicitly which establish a MOGP with the following multivariate covariance function

$$\mathcal{K}(x, x') = \sum_{q=1}^{Q} A_q (A_q)^\top k_q(x, x') \tag{1.15}$$

where $A_q = [a_{1q}, \dots, a_{mq}]^\top \in \mathbb{R}^m$. Note that essentially, this is a multivariate covariance function obtained by multiplying a univariate covariance function with a rank 1 positive definite matrix (usually called *coregionalization matrix*), in order to increment the rank of these matrices, this model proposes to consider *clusters* of latent processes instead, that is, each latent process $u_q(x)$ is replaced by a family of R_q latent processes, all independent between each other with the same covariance function $k_q(x, x')$, see fig 1.1, specifically, each output is now

$$f_j(x) = \sum_{q=1}^{Q} \sum_{c=1}^{R_q} a_{jq}^c u_q^c(x) \quad \forall j = 1, \dots, m \tag{1.16}$$

by defining $A_q^c = [a_{1q}^c, \dots, a_{mq}^c]^\top \in \mathbb{R}^m$, we obtain the multivariate covariance function that the LMC proposes

$$\mathcal{K}(x, x') = \sum_{q=1}^{Q} \left(\sum_{c=1}^{R_q} A_q^c (A_q^c)^\top \right) k_q(x, x') \tag{1.17}$$

This model, while simple, economical and effective at building valid multivariate covariance functions, it is limited, mainly due to the fact that cross-covariances are merely linear combinations of univariate covariance functions which results in symmetric and centered cross-covariances, leaving out an entire spectrum of problems that cannot be modeled properly by this restricted family of multivariate covariance functions, although, as an advantage, it allows the construction of non-stationary multivariate covariance functions by considering the latent processes covariance functions to be non-stationary.

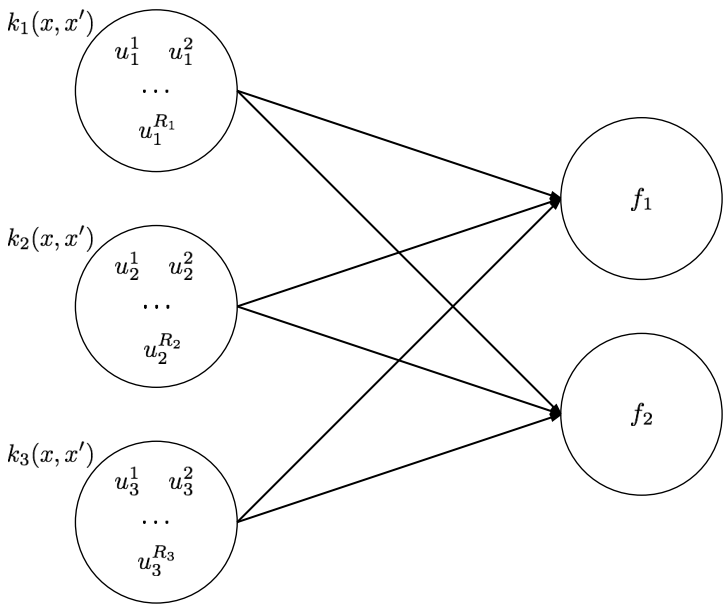

Figure 1.1: Graphical model of the Linear Model of Coregionalization with $Q = 3$ and $m = 2$. Each cluster has R_1, R_2 and R_3 latent processes respectively.

1.4.2 Intrinsic Model of Coregionalization (IMC)

The Intrinsic Model of Coregionalization (IMC [3]) is a particular case of the LMC, where in expression (1.17), the factorization $v_{ij}b_q := \sum_{c=1}^{R_q} a_{iq}^c a_{jq}^c$ its considered, which is equivalent to consider the LMC with $Q = 1$, thus the IMC proposes the following multivariate covariance function

$$[\mathcal{K}(x, x')]_{ij} = \sum_{q=1}^{Q} v_{ij} b_q k_q(x, x') = v_{ij} \sum_{q=1}^{Q} b_q k_q(x, x') = v_{ij} k(x, x')$$

that is, the proposed multivariate covariance function is $\mathcal{K}(x, x') = Bk(x, x')$ where $B \in \mathbb{R}^{m \times m}$ is a positive definite matrix. This kind of multivariate covariance functions it is understood as the detachment of the covariance between inputs x, x' and the covariance across outputs (i, j), which gives them the name of *separable kernels*.

1.4.3 Semi-parametric latent factor model (SLFM)

This model proposed by Teh et. al. [16], as pointed out in [7], results in a particular case of the LMC where $R_q = 1$, that is

$$\mathcal{K}(x, x') = \sum_{q=1}^{Q} A_q (A_q)^\top k_q(x, x') \tag{1.18}$$

The semi-parametric name comes from the non-parametric latent model that is the latent Gaussian processes and the linear parametric combination of these latent processes.

1.4.4 Convolution Model (CONV)

The Convolution model [8, 4, 17] is a generative model for multivariate covariance functions based on latent processes where the operation applied on the latent process is the convolution operation, more formally, let $\{u_q(x)\colon x \in \mathcal{X}, q = 1, \ldots, Q\}$ a family of Q latent processes, independent of each other, each with covariance function $k_q(x, x')$ and let $\{k_{jq}(\tau)\colon j = 1, \ldots, m,\ q = 1, \ldots, Q\}$ a family of stationary kernels, called *smoothing kernels*, the essential idea in the Convolution model is to consider each channel $(f_1(x), \ldots, f_m(x))$ of a m-channel MOGP as the sum of these Q latent processes convolved with its corresponding smoothing kernels, namely

$$f_j(x) = \sum_{q=1}^{Q} \int_{\mathcal{X}} k_{jq}(x - z) u_q(z) \mathrm{d}z \quad \forall j = 1, \ldots, m \tag{1.19}$$

Since the convolution of a Gaussian process with a smoothing kernel is also a Gaus-

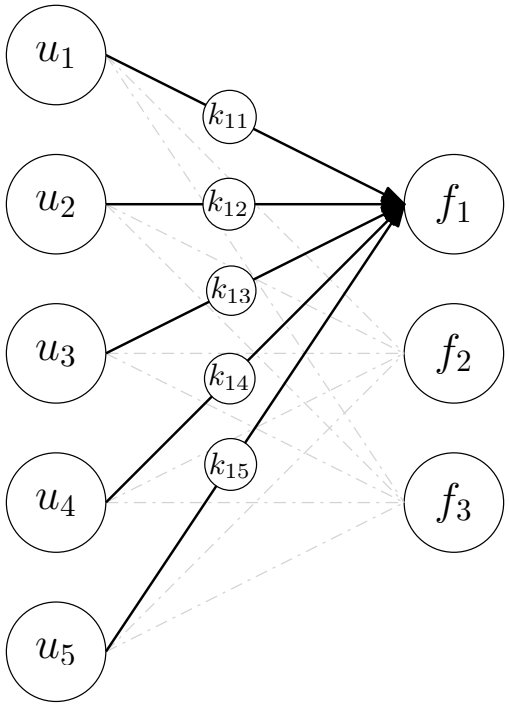

Figure 1.2: Graphical model of the Convolution model with $Q = 5$ and $m = 3$, for this model $Q(m + 1)$ kernels are needed.

sian process, the above expression allows to define a MOGP with the following multivariate covariance function

$$[\mathcal{K}(x, x')]_{ij} = \sum_{q=1}^{Q} \int_{\mathcal{X}} \int_{\mathcal{X}} k_{iq}(x - z) k_{jq}(x' - z') k_q(z, z') \mathrm{d}z \mathrm{d}z' \tag{1.20}$$

The principal obstacle to the use of this multivariate covariance function is the choice of smoothing kernels and covariance functions for the latent processes that allows the integrals in expression (1.23) to have an explicit and tractable form, the proposed approach [4] to

tackle this problem is to consider Gaussian functions as smoothing kernels and the covariance functions of the latent processes due to the fact Gaussian functions are closed under convolution, then

Definition 1.11 *The Gaussian Convolution model (denoted CONV) is given by the following choice of smoothing kernels and latent covariance functions*

$$k_{jq}(\tau) = \frac{w_{jq}|\Sigma_{jq}|^{1/2}}{(2\pi)^{n/2}} \exp\left(\tfrac{1}{2}\tau^\top \Sigma_{jq}\tau\right) \tag{1.21}$$

$$k_q(\tau) = \exp\left(\tfrac{1}{2}\tau^\top \Sigma_q \tau\right) \tag{1.22}$$

by denoting $\Sigma_{ijq} = \Sigma_{iq}^{-1} + \Sigma_{jq}^{-1} + \Sigma_q^{-1}$*, we obtain*

$$[\mathcal{K}(\tau)]_{ij} = k_{ij}(\tau) = \sum_{q=1}^{Q} \frac{w_{iq}w_{jq}|\Sigma_q^{-1}|^{1/2}}{|\Sigma_{ijq}|^{1/2}} \exp\left(\tfrac{1}{2}\tau^\top \Sigma_{ijq}^{-1}\tau\right) \tag{1.23}$$

This multivariate covariance function is a generalization of the LMC and it is not within the separable family, which yields an increase in the complexity of the cross-covariances generated, but they are still limited to linear combinations of univariate covariance functions, but more limiting is the fact that the model in (1.23) only allows covariances that are either positive or negative.

1.4.5 Cross-Spectral Mixture kernel (CSM)

The Cross-Spectral Mixture (CSM) kernel [18] proposed by Ulrich et al. is an innovation to the LMC where the coregionalization matrices have complex coefficients which, when chosen wisely, will produce non-symmetric cross-covariances. The source idea of this model starts by considering the LMC with spectral mixture kernels as the covariance functions of the latent processes, that is

$$\mathcal{K}(\tau) = \sum_{q=1}^{Q} \left(\sum_{c=1}^{R_q} A_q^c (A_q^c)^\top \right) k_q(\tau) \tag{1.24}$$

where $k_q(\tau) = \mathcal{F}^{-1}(S_q(\omega)) = \exp\left(-\frac{1}{2}\tau^\top \Sigma_q \tau\right)\cos(\mu_q^\top \tau)$ is a basic spectral mixture kernel, the fundamental idea of the CSM kernel comes from considering the coefficients of the coregionalization matrices as complex numbers of the form $a_{iq}^c = b_{iq}^c \exp(\iota\psi_{iq}^c)$ for $b_{iq}^c, \psi_{iq}^c \in \mathbb{R}$, which yields to

$$\mathcal{K}(\tau) = \sum_{q=1}^{Q} \left(\sum_{c=1}^{R_q} A_q^c (A_q^c)^H \right) k_q(\tau) \tag{1.25}$$

where $(\cdot)^H$ is the Hermitian (transpose and conjugate) operator and by using the spectral representation of the SM kernels, we obtain

$$[\mathcal{K}(\tau)]_{ij} = \sum_{q=1}^{Q} \sum_{c=1}^{R_q} b_{iq}^c b_{jq}^c \int_{\mathbb{R}^n} e^{\iota(\omega^\top \tau + \psi_{iq}^c - \psi_{jq}^c)} S_q(\omega) \mathrm{d}\omega \tag{1.26}$$

which in conjunction with the following reparameterization $\psi_{iq}^c = \mu_q^\top \phi_{iq}^c$ where μ_q is the frequency of the SM kernel $k_q(\tau)$ and $\phi_{iq}^c \in \mathbb{R}^n$, the CSM kernel is as follows

Definition 1.12 *The Cross-Spectral Mixture Kernel (CSM) is defined as follows*

$$[\mathcal{K}(\tau)]_{ij} = \sum_{q=1}^{Q} \sum_{c=1}^{R_q} b_{iq}^c b_{jq}^c \exp(\tau^\top \Sigma_q \tau) \cos\left((\tau + \phi_{iq}^c - \phi_{jq}^c)^\top \mu_q\right) \tag{1.27}$$

This multivariate covariance function is the first, among the previous approaches, that allows non-symmetric cross-covariance functions given by the phase parameters ϕ_{iq}^c, this hints the importance of having complex cross-spectral densities but since it is constructed from the LMC framework, the auto-covariance functions across channels cannot be to different from each other, which presuppose a strong correlation between channels. As it will be seen, this model is a particular case of the model to be proposed in the next chapter.

Chapter 2

Multi-Output Spectral Mixture Kernel

In the previous chapter it has been described how covariance functions for Gaussian processes can be built in the frequency domain in virtue of Bochner's theorem which yielded the Spectral Mixture kernel, whereas the purpose of this chapter is to give details of the generalization to multiple outputs of the aforementioned theorem. This generalization is known as Cramér's theorem and as we shall see it gives necessary and sufficient conditions for the construction of multivariate covariance functions in the frequency domain, which will be used in order to propose a generalization of the SM kernel to multiple outputs that will be called the Multi-Output Spectral Mixture kernel (MOSM).

2.1 Spectral Representation of Stationary Multivariate Covariance functions

The design of multivariate covariance functions is challenging because we have to deal with a trade off between choosing auto-covariances and cross-covariances flexible enough that model a extensive class of possible dependencies while at the same time these choices have to fulfill condition (1.10).

In that regard, motivated by the univariate spectral kernels, it is reasonable to ask whether auto-covariance and cross-covariance functions can be built in the spectral domain in conjunction, thus, bypassing the troublesome positive definite condition (1.10). The required framework for this is provided by the following theorem which presents a spectral representation for multivariate covariance functions that with similar simplifications to those that had been done in the univariate case, will allow us to build, with relative ease, multivariate covariance functions in the frequency domain as in [2].

Theorem 2.1 *(Cramér's Theorem [5, 6]) A family* $\{k_{ij}(\tau)\}_{i,j=1}^{m}$ *of complex-valued functions are the covariance functions of a weakly-stationary multivariate stochastic process if and only*

if they (i) admit the representation

$$k_{ij}(\tau) = \int_{\mathbb{R}^n} e^{\imath\omega^\top\tau} \, \mathrm{d}\mu_{F_{ij}}(\omega) \qquad \forall i, j \in \{1, \ldots, m\} \tag{2.1}$$

where each F_{ij} is a complex-valued distribution function $F_{ij} : \mathbb{R}^n \to \mathbb{C}$ and (ii) fulfil the positive definiteness condition, for any interval I

$$\sum_{i,j=1}^{m} \overline{z}_i z_j F_{ij}(I) \geq 0 \qquad \forall \{z_1, \ldots, z_m\} \subset \mathbb{C} \tag{2.2}$$

where $\overline{z}$ is the complex conjugate of $z \in \mathbb{C}$. Likewise in the construction of the Spectral Mixture kernel, it is possible to only consider the absolute continuous part of the complex-valued measures $\mu_{F_{ij}}$ which leads to a complex-valued density $S_{ij}(\omega) \colon \mathbb{R}^n \to \mathbb{C}$ with respect Lebesgue's measure and is equivalent to reduce the representation to integrable multivariate covariance functions, this simplifies Theorem 2.1 as follows

Theorem 2.2 *(Cramér's Theorem simplified [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 (i) admit the representation*

$$k_{ij}(\tau) = \int_{\mathbb{R}^n} e^{\imath\omega^\top\tau} \, S_{ij}(\omega) \mathrm{d}\omega \qquad \forall i, j \in \{1, \ldots, m\} \tag{2.3}$$

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

$$\sum_{i,j=1}^{m} \overline{z}_i z_j S_{ij}(\omega) \geq 0 \qquad \forall \{z_1, \ldots, z_m\} \subset \mathbb{C}, \ \omega \in \mathbb{R}^n \tag{2.4}$$

Note that eq.(2.3) states that each covariance function $k_{ij}(\tau)$ is the inverse Fourier transform of a spectral density $S_{ij}(\omega)$, therefore, we will say that these functions are *Fourier pairs* and refer to the set of arguments of the covariance function $\tau \in \mathbb{R}^n$ as *time* or *space domain* depending of the application considered, and to the set of arguments of the spectral densities $\omega \in \mathbb{R}^n$ as *Fourier* or *frequency domain*. Furthermore, a direct consequence of the above theorem and the symmetry property of covariance is that for any element ω in the Fourier domain, the matrix defined by $S(\omega) = [S_{ij}(\omega)]_{i,j=1}^m \in \mathbb{C}^{m \times m}$ is Hermitian, i.e., $S_{ij}(\omega) = \overline{S}_{ji}(\omega) \ \forall i, j, \omega$, which implies that diagonal spectral densities ($i = j$) must be real-valued.

Theorem 2.2 gives the guidelines to construct multivariate covariance functions for MOGPs by designing their corresponding spectral densities instead, i.e., the design is performed in the Fourier rather than the space domain. The simplicity of design in the Fourier domain stems from the positive-definiteness condition of the spectral densities in eq. (2.4) which is much easier to achieve than that of the covariance functions in eq. (1.10). This can be understood through an analogy with the univariate model: in the single-output case the positive-definiteness condition of the covariance function only requires positivity of the

spectral density, whereas in the multi-output case the positive-definiteness condition of the multivariate covariance function only requires that the matrix $S(\omega),\ \forall\omega \in \mathbb{R}^n$, is positive definite which is a way more weak condition plus there are no severe constraints on each function $S_{ij} : \omega \mapsto S_{ij}(\omega)$.

2.2 Spectral Densities

Following the idea behind the Spectral Mixture kernel, the goal now is to propose a family of Hermitian positive-definite complex-valued functions $\{S_{ij}(\cdot)\}_{i,j=1}^m$, thus fulfilling the requirements of Theorem 2.1, eq. (2.2), to use them as cross-spectral densities within multivariate covariance functions. As we shall see the complex-valued requirement is going to be fundamental for the proper modeling of a broad family of cross-covariances which is going to be the main improvement and contribution over the latent processes models, in addition with more autonomy across auto-covariances. The proposed family of functions is designed with the aim of providing physical parametric interpretation and closed-form covariance functions after applying the inverse Fourier transform.

An intuitive procedure for constructing positive-definite complex matrix-valued functions is to construct the Cholesky decomposition instead. Recall that complex-valued positive-definite matrices be decomposed in the form $S(\omega) = R^H(\omega)R(\omega)$, meaning that the $(i,j)^{\text{th}}$ entry of $S(\omega)$ can be expressed as $S_{ij}(\omega) = R_{:i}^H(\omega)R_{:j}(\omega)$; where $R(\omega) \in \mathbb{C}^{Q\times m}$, $R_{:i}(\omega)$ is the i^{th} column of $R(\omega)$, and $(\cdot)^H$ denotes the Hermitian (transpose and conjugate) operator. Clearly, this factor decomposition fulfills eq. (2.4):

$$\sum_{i,j=1}^m \overline{z_i} R_{:i}^H(\omega)R_{:j}(\omega)z_i = \left\|\sum_{i=1}^m z_i R_{:i}(\omega)\right\|^2 = \left\|R(\omega)z\right\|^2 \geq 0 \quad \forall z = [z_1,\dots,z_m]^\top \in \mathbb{C}^m,\ \omega \in \mathbb{R}^n$$

We refer to Q as the rank of the decomposition, since by choosing $Q < m$ the rank of $S(\omega) = R^H(\omega)R(\omega)$ can be at most Q. For ease of notation we choose[1] $Q = 1$, 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 $S_{ij}(\omega) = \overline{R}_i(\omega)R_j(\omega)\ \forall i,j = 1,\dots,m$.

Squared-exponential (SE) functions have many useful properties that suggest them as suitable spectral densities, such as closed-form expressions in Fourier's transform calculations, they are closed under multiplications i.e the product of two SE functions is also a SE function, but also they allow an easy integration with imaginary parts, hence it is proposed

$$R_i(\omega) = w_i \exp\left(-\tfrac{1}{4}(\omega-\mu_i)^\top\Sigma_i^{-1}(\omega-\mu_i)\right)\exp\left(-\iota(\theta_i^\top\omega+\phi_i)\right), \quad i = 1,\dots,m \tag{2.5}$$

where $w_i, \phi_i \in \mathbb{R}$, $\mu_i, \theta_i \in \mathbb{R}^n$ and $\Sigma_i = \operatorname{diag}([\sigma_{i1}^2,\dots,\sigma_{in}^2])^\top \in \mathbb{R}^{n\times n}$. This choice of the functions $\{R_i\}_{i=1}^m$ in conjunction with the following property of SE functions

Proposition 2.3 *The product of SE functions is closed up to a constant, that is*

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

[1]The extension to arbitrary Q will be presented at the end of this section.

where:

$$\begin{aligned}\alpha_{ij} &= e^{\left(-\frac{1}{2}(\mu_i-\mu_j)^\top(\Sigma_i+\Sigma_j)^{-1}(\mu_i-\mu_j)\right)}\\ \mu_{ij} &= (\Sigma_i+\Sigma_j)^{-1}(\Sigma_i\mu_j+\Sigma_j\mu_i)\\ \Sigma_{ij} &= \Sigma_i(\Sigma_i+\Sigma_j)^{-1}\Sigma_j\end{aligned}$$

yields the spectral densities $\{S_{ij}\}_{i,j=1}^m$ which are given by

$$S_{ij}(\omega) = w_{ij}\exp\left(-\tfrac{1}{2}(\omega-\mu_{ij})^\top\Sigma_{ij}^{-1}(\omega-\mu_{ij}) + \iota\left(\theta_{ij}^\top\omega+\phi_{ij}\right)\right), \quad \forall i,j=1,\ldots,m \tag{2.6}$$

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

- covariance: $\Sigma_{ij} = 2\Sigma_i(\Sigma_i+\Sigma_j)^{-1}\Sigma_j$
- mean: $\mu_{ij} = (\Sigma_i+\Sigma_j)^{-1}(\Sigma_i\mu_j+\Sigma_j\mu_i)$
- 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: $\phi_{ij} = \phi_i - \phi_j$

where the so-constructed magnitudes w_{ij} ensure positive definiteness and, in particular, the auto-spectral densities S_{ii} are real-valued SE functions (since $\theta_{ii} = \phi_{ii} = 0$) as in the standard (scalar-valued) spectral mixture approach [2].

The proposed spectral densities in eq. (2.6) yields complex-valued covariance functions and therefore complex-valued GPs. In order to restrict this generative model only to real-valued GPs, there are two equivalent procedures, on the one hand, the real part of a complex-valued covariance function is a real-valued covariance function, thus we can focus only in the real part of the right-side expression (2.3), on the other hand, the proposed family of spectral densities can be symmetrized with respect to ω [19] which ensures that the covariance function is real-valued, we then make $S_{ij}(\omega)$ symmetric simply by reassigning $S_{ij}(\omega) \mapsto \frac{1}{2}(S_{ij}(\omega)+S_{ij}(-\omega))$, this is equivalent to choosing $R_i(\omega)$ to be a vector of two *mirrored* complex SE functions. Therefore, the proposed multivariate covariance function is obtained by calculating the inverse Fourier transform of the spectral densities $S_{ij}(\omega)$, that is, the integral

$$\begin{aligned}k_{ij}(\tau) &= \int_{\mathbb{R}^n} e^{\iota\omega^\top\tau}S_{ij}(\omega)\,\mathrm{d}\omega = w_{ij}\int_{\mathbb{R}^n} e^{\iota\omega^\top\tau}e^{\left(-\frac{1}{2}(\omega-\mu_{ij})^\top\Sigma_{ij}^{-1}(\omega-\mu_{ij})+\iota(\theta_{ij}^\top\omega+\phi_{ij})\right)}\,\mathrm{d}\omega\\ &= \int_{\mathbb{R}^n}\exp\left(-\tfrac{1}{2}\omega^\top\Sigma_{ij}^{-1}\omega + \left(\Sigma_{ij}^{-1}\mu_{ij}+\iota(\tau+\theta_{ij})\right)^\top\omega - \tfrac{1}{2}\mu_{ij}^\top\Sigma_{ij}^{-1}\mu_{ij}+\iota\phi_{ij}\right)\mathrm{d}\omega\end{aligned}$$

the above integral has a closed-form determined by the fact that for any diagonal matrix $\Lambda = \mathrm{diag}(\lambda^{(1)},\ldots,\lambda^{(n)})$ and for any $b\in\mathbb{C}^n, c\in\mathbb{C}$ we have

$$\int_{\mathbb{R}^n}\exp\left(-x^\top\Lambda x - 2b^\top x + c\right)\mathrm{d}x = \frac{\pi^{\frac{n}{2}}}{|\Lambda|^{1/2}}\exp\left(b^\top\Lambda^{-1}b+c\right) \tag{2.7}$$

by making use of the preceding formula with $\Lambda = \frac{1}{2}\Sigma_{ij}^{-1}$, $b = -\frac{1}{2}\left(\Sigma_{ij}^{-1}\mu_{ij} + \iota(\tau + \theta_{ij})\right)$ and $c = -\frac{1}{2}\mu_{ij}^{\top}\Sigma_{ij}^{-1}\mu_{ij} + \iota\phi_{ij}$, we get

$$\begin{aligned} k_{ij}(\tau) &= \alpha_{ij} \exp\left(\frac{1}{2}\left(\Sigma_{ij}^{-1}\mu_{ij} + \iota(\tau + \theta_{ij})\right)^{\top}\Sigma_{ij}\left(\Sigma_{ij}^{-1}\mu_{ij} + \iota(\tau + \theta_{ij})\right) - \frac{1}{2}\mu_{ij}^{\top}\Sigma_{ij}^{-1}\mu_{ij} + \iota\phi_{ij}\right) \\ &= \alpha_{ij} \exp\left(-\frac{1}{2}(\tau + \theta_{ij})^{\top}\Sigma_{ij}(\tau + \theta_{ij}) + \iota\left((\tau + \theta_{ij})^{\top}\mu_{ij} + \phi_{ij}\right)\right) \end{aligned}$$

where $\alpha_{ij} = w_{ij}(2\pi)^{\frac{n}{2}}|\Sigma_{ij}|^{1/2}$. By taking the real part of the above complex-valued covariance function $k_{ij}(\tau)$ or by symmetrizing the spectral densities $S_{ij}(\omega)$, the desired real-valued covariance function is obtained

$$k_{ij}(\tau) = \alpha_{ij} \exp\left(-\frac{1}{2}(\tau + \theta_{ij})^{\top}\Sigma_{ij}(\tau + \theta_{ij})\right) \cos\left((\tau + \theta_{ij})^{\top}\mu_{ij} + \phi_{ij}\right) \tag{2.8}$$

Note that the auto-covariances ($i = j$) are, up to normalization, Spectral Mixture kernels since $\alpha_{ii} \geq 0$ and $\theta_{ii} = \phi_{ii} = 0$. Conversely, the proposed model for the cross-covariance between different channels ($i \neq j$) allows for (i) both negatively- and positively-correlated channels ($\alpha_{ij} \in \mathbb{R}$), (ii) delayed channels through the delay parameter $\theta_{ij} \neq 0$ and (iii) out-of-phase channels where the covariance is not symmetric with respect to the delay for $\phi_{ij} \neq 0$. Fig. 2.1 shows cross-spectral densities and their corresponding covariance function for a choice of different delay and phase parameters. In addition, unlike latent processes models, each auto-covariance function has it own set of parameters which only interact through the cross-covariances functions. Although, this extra complexity comes at a price: a higher number of parameters, the proposed multivariate covariance function (2.8) has $3mn + 2m$ parameters which is significant more compared than those in the latent process models.

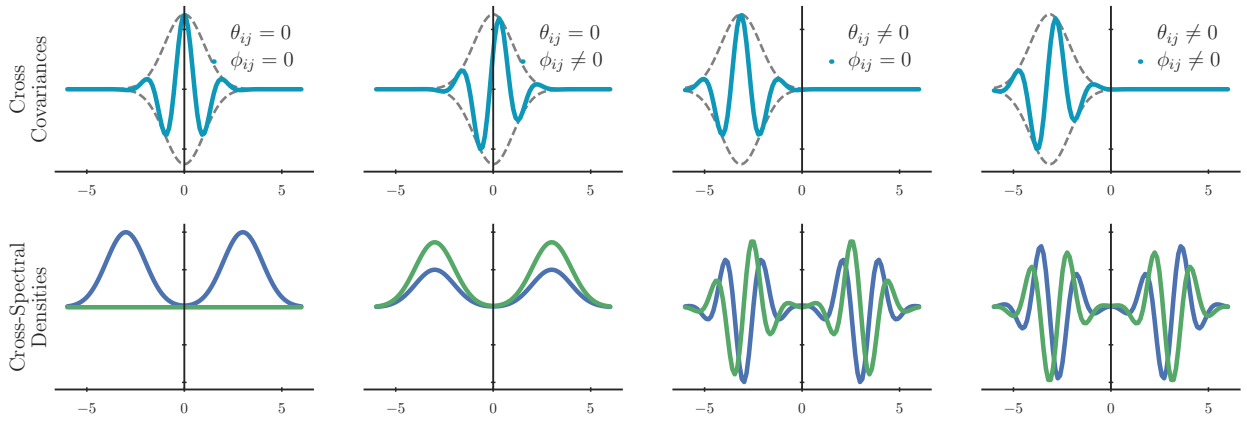

Figure 2.1: Cross-spectral densities (bottom, real part in blue and imaginary part in green) and cross-covariance function (top) generated by the proposed model in eq. (2.8): From left to right: zero delay and zero phase; zero delay and non-zero phase; non-zero delay and zero phase; and non-zero delay and non-zero phase. The dashed lines denote the SE envelopes.

The multivariate covariance function in eq. (2.8) resulted from a low rank choice for the PSD matrix S_{ij}, therefore, increasing the rank in the proposed model for S_{ij} is equivalent to

consider several spectral components. Arbitrarily choosing Q of these components yields to the expression for the proposed multivariate covariance function:

Definition 2.4 *The Multi-Output Spectral Mixture Kernel (MOSM) has the form:*

$$k_{ij}(\tau) = \sum_{q=1}^{Q} \alpha_{ij}^{(q)} \exp\left(-\frac{1}{2}(\tau + \theta_{ij}^{(q)})^\top \Sigma_{ij}^{(q)} (\tau + \theta_{ij}^{(q)})\right) \cos\left((\tau + \theta_{ij}^{(q)})^\top \mu_{ij}^{(q)} + \phi_{ij}^{(q)}\right) \tag{2.9}$$

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

This multivariate covariance function has spectral-mixture positive-definite kernels as auto-covariances, while the cross-covariances are non-symmetric, delayed, and not necessarily positive-definite, spectral mixture kernels with different parameters for different pairs of outputs. Therefore, the MOSM kernel is a multi-output generalization of the spectral mixture approach [2] where the positive definiteness is guaranteed by the factor decomposition of S_{ij} as shown in eq. (2.2).

Note that, in virtue of the Fourier convolution theorem, the factor decomposition of the spectral densities $S_{ij}(\omega) = \overline{R_i(\omega)} R_j(\omega)$, is equivalent to convolutions of complex-valued squared exponential functions, that is

$$\begin{aligned} k_{ij}(\tau) &= \mathcal{F}^{-1}(S_{ij}(\omega))(\tau) \\ &= (\mathcal{F}^{-1}(\overline{R_i}) * \mathcal{F}^{-1}(R_j))(\tau) \\ &= (\overline{k_i}(-\tau) * k_j(\tau))(\tau) \quad \forall i, j \in \{1, \ldots, m\} \end{aligned}$$

where $k_i(\tau)$ is a complex-valued squared exponential function given by

$$\begin{aligned} k_i(\tau) &= \mathcal{F}^{-1}(R_i(\omega))(\tau) \\ &= (4\pi)^{n/2} |\Sigma_i|^{1/2} w_i \exp\left(-(\tau - \theta_i)^\top \Sigma_i (\tau - \theta_i) + \iota \left((\tau - \theta_i)^\top \mu_i - \phi_i\right)\right) \end{aligned}$$

This hints the possibility of building non-parametric cross-covariances as in [11] given the convolution formalism.

2.3 Relationship with other models

Generalizing the scalar spectral mixture kernel to multivariate covariance functions can be achieved from the LMC framework by considering scalar spectral mixture kernels as the covariance functions of the latent processes (this approach is denoted SM-LMC), as pointed out in [18], this formulation its equivalent to consider only real-valued cross-spectral densities, that is why the authors propose a multivariate covariance function by including a complex component, within the LMC framework, to the cross spectral densities to cater for phase differences across channels. The CSM model can be seen as a particular case of the proposed MOSM model with $\mu_i = \mu_j$, $\Sigma_i = \Sigma_j$, $\theta_i = \theta_j \ \ \forall i, j \in \{1, \ldots, m\}$ and $\phi_i = \mu_i^\top \psi_i$ for $\psi_i \in \mathbb{R}^n$,

therefor, the SM-LMC is a particular case of the MOSM model where the parameters $\mu_i, \Sigma_i, \theta_i$ are restricted to be same for all channels and therefore only phase shifts and no delays are allowed unlike the example in Fig. 2.1.

2.4 t-Student Spectral Densities

As we saw in the previous sections, squared exponential spectral densities are useful and handy, since they allow an easy integration of imaginary parts and they have explicit Fourier calculations, on the downside, these spectral densities yield infinite differentiable covariance functions, thus, inference brings forth infinite differentiable estimators. Such an excessive smoothness can be unrealistic on several physical problems where the smoothness of the data is known. In this section we explore the possibility of using a similar approach to the MOSM model but with t-Student spectral densities instead, as in [20, 21], since they allow parameters that restricts the smoothness of the resultant covariance functions, for this purpose, let us remember the Matérn covariance functions for one-dimensional input spaces (that is $n = 1$):

Definition 2.5 *A t-Student distribution with 2ν degrees of freedom and scale parameter α has the form:*

$$S(\omega) = w^2 \frac{\Gamma(\nu + 1/2)}{(\sqrt{2\nu}\alpha)\pi^{1/2}\Gamma(\nu)} \left(1 + \frac{\omega^2}{2\nu\alpha^2}\right)^{-(\nu+1/2)} \tag{2.10}$$

$$= w^2 \frac{\Gamma(\nu + 1/2)(2\nu\alpha^2)^\nu}{\pi^{1/2}\Gamma(\nu)} \frac{1}{(2\nu\alpha^2 + \omega^2)^{\nu+1/2}} \tag{2.11}$$

where $\Gamma(z)$ is the Gamma function and $\nu, \alpha \in \mathbb{R}_+$, $w \in \mathbb{R}$

Bochner's theorem can be used with this distribution as spectral density which yields the renowned Matérn covariance function, defined as follows

Definition 2.6 *The Matérn covariance function has the form:*

$$k(\tau) = w^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu}|\tau|)^\nu K_\nu(\sqrt{2\nu}|\tau|) \tag{2.12}$$

where K_ν is the modified Bessel function of the second kind, defined by

$$K_\nu(c\tau) = \frac{\Gamma(\nu + 1/2)(2c)^\nu}{\pi^{1/2}\tau^\nu} \int_0^\infty \frac{\cos(\omega\tau)}{(c^2 + \omega^2)^{\nu+1/2}} \mathrm{d}\omega \quad \forall c, \tau > 0 \tag{2.13}$$

For this class of covariance functions we have the following property: a Gaussian process $f(x)$ with a Matérn covariance function is mean-squared k-times differentiable if and only if $\nu > k$, thus by choosing an specific value of ν we can limit the differentiability of the estimators. A regular choice for this parameters is $\nu = p + 1/2$ for $p \in \mathbb{N}$, mainly due the fact that the Bessel function can be simplified into a product of a polynomial and a exponential function, for example:

- $p = 0, \nu = 1/2, \quad k(\tau) = w^2 \exp(-\alpha|\tau|)$ (also known as *Ornstein-Uhlenbeck* kernel)
- $p = 1, \nu = 3/2, \quad k(\tau) = w^2 \left(1 + \sqrt{3}\alpha|\tau|\right) \exp(-\sqrt{3}\alpha|\tau|)$
- $p = 2, \nu = 5/2, \quad k(\tau) = w^2 \left(1 + \sqrt{3}\alpha|\tau| + \sqrt{5}\alpha^2|\tau|^2\right) \exp(-\sqrt{5}\alpha|\tau|)$

Following the approach of [21], it is possible to propose spectral densities of the form

$$S_{ij}(\omega) = \int_0^\infty R_i(\omega, \xi) R_j(\omega, \xi) \mathrm{d}\xi \quad \forall i, j \in \{1, \ldots, m\} \tag{2.14}$$

where each $R_i(\omega, \xi)$ is defined as:

$$R_i(\omega, \xi) = w_i \frac{2^{\nu_i/2 - 1/2}}{\pi^{1/4}} \xi^{\left(\frac{\nu_i}{2} + \frac{1}{4} - \frac{1}{2}\right)} e^{-\left(\frac{\alpha_i^2}{2} + \frac{\omega^2}{2}\right)\xi} \quad \forall i \in \{1, \ldots, m\} \tag{2.15}$$

thus, by using the fact that $\Gamma(b)/a^b = \int_0^\infty \xi^{b-1} e^{-a\xi} \mathrm{d}\xi$, the spectral densities S_{ij} are given by

$$S_{ij}(\omega) = w_i w_j \frac{2^{\nu_{ij} - 1} \Gamma(\nu_{ij} + \frac{1}{2}) \alpha_{ij}^{2\nu_{ij}}}{\pi^{1/2} \alpha_{ij}^{2\nu_{ij}}} \frac{1}{(\alpha_{ij}^2 + \omega^2)^{\nu_{ij} + 1/2}} \quad \forall i, j \in \{1, \ldots, m\} \tag{2.16}$$

where $\alpha_{ij}^2 = \frac{\alpha_i^2 + \alpha_j^2}{2}$ and $\nu_{ij} = \frac{\nu_i + \nu_j}{2}$. These spectral densities fullfil condition (2.4) due to the decomposition in (2.14), therefore, in virtue of Cramér's theorem, the covariance functions are given by

$$k_{ij}(\tau) = \frac{w_i w_j}{\alpha_{ij}^{2\nu_{ij}}} (\alpha_{ij}|\tau|)^{\nu_{ij}} K_{\nu_{ij}}(\alpha_{ij}|\tau|) \quad \forall i, j \in \{1, \ldots, m\} \tag{2.17}$$

This multivariate covariance function is a simplification of the work in [21] and it allows different length-scales and different differentiability across outputs, but it doesn't allows the modeling of phases and delays due to the fact that the spectral densities proposed are real valued and centered around zero, provided that, it is possible to extend these spectral densities by considering a non-centered t-Student distribution and by adding an imaginary component, as in the MOSM model, that is, to propose

$$S_{ij}(\omega) = w_i w_j \frac{2^{\nu_{ij} - 1} \Gamma(\nu_{ij} + \frac{1}{2}) \alpha_{ij}^{2\nu_{ij}}}{\pi^{1/2} \alpha_{ij}^{2\nu_{ij}}} \frac{e^{\iota\left(\theta_{ij}\omega + \phi_{ij}\right)}}{(\alpha_{ij}^2 + (\omega - \mu)^2)^{\nu_{ij} + 1/2}} \quad \forall i, j \in \{1, \ldots, m\} \tag{2.18}$$

where θ_{ij}, ϕ_{ij} are defined alike the MOSM model and μ is the shift of the t-Student spectral density of the origin, by denoting $w_{ij} = \frac{w_i w_j}{\alpha_{ij}^{2\nu_{ij}}}$, we obtain the following multivariate covariance function:

$$k_{ij}(\tau) = w_{ij} (\alpha_{ij}|\tau + \theta_{ij}|)^{\nu_{ij}} K_{\nu_{ij}}(\alpha_{ij}|\tau + \theta_{ij}|) \cos\left(\mu(\tau + \theta_{ij}) + \phi_{ij}\right) \tag{2.19}$$

Analogously to the MOSM model, we can consider a mixture of these spectral densities in order to increase the rank of the PSD matrix S_{ij}, which, in virtue of Cramér's theorem, yield the following multivariate covariance function:

Definition 2.7 *The extended multivariate Matérn covariance function has the form:*

$$k_{ij}(\tau) = \sum_{q=1}^{Q} w_{ij}^{(q)} (\alpha_{ij}^{(q)}|\tau + \theta_{ij}^{(q)}|)^{\nu_{ij}^{(q)}} K_{\nu_{ij}^{(q)}}(\alpha_{ij}^{(q)}|\tau + \theta_{ij}^{(q)}|) \cos\left(\mu^{(q)}(\tau + \theta_{ij}^{(q)}) + \phi_{ij}^{(q)}\right) \tag{2.20}$$

where $w_{ij}^{(q)} = w_i^{(q)} w_j^{(q)} / (\alpha_{ij}^{(q)})^{2\nu_{ij}^{(q)}}$ *and the superindex* $(\cdot)^{(q)}$ *denotes the parameter of the* q^{th} *component of the spectral mixture.*

This multivariate covariance function allows for different length-scales, smoothness, delays and phases across channels. On the downside, it doesn't allow different for frequencies μ across channels as in the MOSM model, also, is limited to one-dimensional inputs only, in this regard, this model is incomplete. This poses the question whether a more general family of t-Student spectral densities can be formulated, a family of the form:

$$S_{ij}(\omega) \propto \frac{e^{\iota(\theta_{ij}^{\top}\omega + \phi_{ij})}}{\left(1 + \frac{1}{2\nu_{ij}}(\omega - \mu_{ij})^{\top}\Sigma_{ij}^{-1}(\omega - \mu_{ij})\right)^{\nu_{ij}+1/2}} \tag{2.21}$$

is it possible to choose the parameters of this family as in the MOSM model such that fullfil condition (2.4) and when the degree of freedom parameters goes to infinity ($\nu_{ij} \to \infty$) this model converges to the MOSM model? (due to the convergence of t-Student distributions to Gaussian distributions). For now this question will remain unanswered.

Chapter 3

Experiments

With the purpose of testing the proposed MOSM model and to compare against other models, we first validated the ability of the model in the identification of known auto- and cross-covariances of synthetic data and compare it against the spectral-mixture linear model of coregionalization (SM-LMC, [3, 2, 18], the Gaussian convolution model (CONV, [4]), and the cross-spectral mixture model (CSM, [18]), to then test the model in the estimation of missing real-world data in two different distributed settings: climate signals and metal concentrations. All models were implemented in Tensorflow [22] using the package GPflow [23]. The performance of all the models in the experiments was measured by the mean absolute error given by

$$\text{MAE} : \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.1}$$

where y_i denotes the true value and $\hat{y}_i$ the MOGP estimate.

3.1 Synthetic example: Learning derivatives and delayed signals

A very important property of Gaussian processes, it is that they are closed under linear operators, in particular, since the derivative operator is linear, then, the *derivative* of a Gaussian process is also a Gaussian process, that is

Proposition 3.1 *Let $f(x)$ be a Gaussian process with stationary covariance function $k(\tau)$, then the derivative stochastic process $f'(x)$ is also a Gaussian process and its stationary covariance function is $-k''(\tau)$, furthermore, $\big(f(x), f'(x)\big)$ form a two-channel Multi-Output Gaussian process ([24], [1], sec. 9.4) with the following multivariate covariance function*

$$\mathcal{K}(\tau) = \begin{pmatrix} k(\tau) & -k'(\tau) \\ k'(\tau) & -k''(\tau) \end{pmatrix} \tag{3.2}$$

With this in mind, all four models were implemented to recover the auto- and cross-covariances of a three-channel MOGP with the following components: (i) a reference function sampled from a GP $f(x) \sim \mathcal{GP}(0, K_{SM})$ with spectral mixture covariance kernel K_{SM} and zero mean, (ii) its derivative $f'(x)$, and (iii) a noiseless delayed version $f_\delta(x) = f(x - \delta)$. The motivation for this illustrative example is that, since the auto-covariances and cross-covariances of the aforementioned processes are known explicitly, we can therefore expect the models to learn approximates of these theoretical covariances and use them for extrapolation.

The reference function was sampled on the domain [-20, 20] and we chose $N_1 = 500$ noisy samples of the reference function for training, the derivative was computed numerically (first order through finite differences) in the same range and $N_2 = 400$ randomly selected noisy samples in the interval [-20, 0] were chosen, the same goes for the delayed signal. Table 3.1 presents the MAE for all models for the estimation of the reference signal and the extrapolation of the derivative and delayed signals over the interval [0, 20] over ten realisations of the experiment.

Fig. 3.1 shows the ground truth and MOSM estimates for all three synthetic signals and the covariances (normalized), where the proposed model successfully learnt all cross-covariances $cov(f(x), f'(x))$ and $cov(f(x), f(x - \delta))$, and auto-covariances without prior information about the delay or the derivative relationship between the two channels and, furthermore, the proposed model is able to extrapolate the derivative and delayed signals in the interval [0,20] where there is no data of these signals.

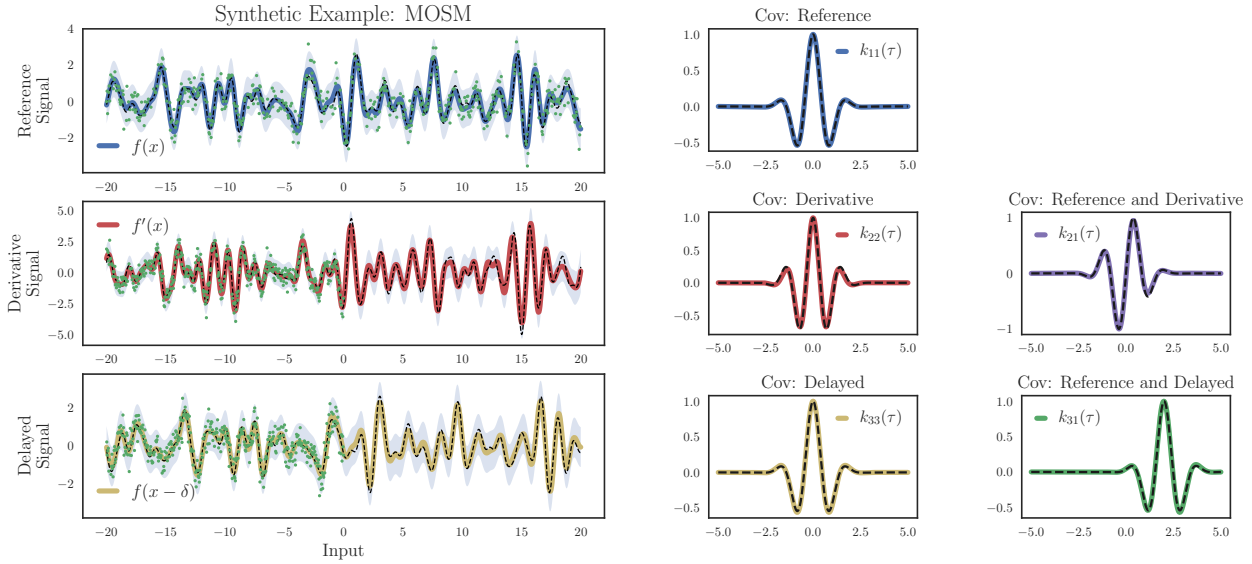

Figure 3.1: MOSM learning of the covariance functions of a synthetic reference signal, its derivative and a delayed version. **Left:** synthetic signals, **middle**: autocovariances, **right**: cross-covariances. The dashed line represents the ground truth and the solid colour lines the MOSM estimates with the respective variances, the training data is shown in green.

For this experiment, the MOSM model was used with only one spectral component ($Q = 1$), meanwhile, the other models were used with five latent processes ($Q = 5$). The

experiment shows the earnings of allowing different parameters across auto-covariances since it yields distinct behaviors across the channels, in contrast with the latent processes models where auto-covariances partake its behavior, on the other hand, learnt cross-covariances shows that the complex parameters of the cross-spectral densities play a fundamental role in this experiment by identifying correctly the nature of the data, allowing non trivial cross-covariances, which implies that the MOSM model can be used for *system identification*.

Table 3.1: Mean absolute error for all four models with one-standard-deviation error bars over five realizations.

Model	Reference	Derivative	Delayed
CONV	0.211 ± 0.085	0.759 ± 0.075	0.524 ± 0.097
SM-LMC	0.166 ± 0.009	0.747 ± 0.101	0.398 ± 0.042
CSM	0.148 ± 0.010	0.262 ± 0.032	0.368 ± 0.089
MOSM	0.127 ± 0.011	0.223 ± 0.015	0.146 ± 0.017

On the other hand, Fig. 3.2 shows the estimates for the CSM model and the learnt auto-covariance and cross-covariances. Note how despite the fact that this model is able to extrapolate the derivative signal with relative low error, it is unable to extrapolate the delayed signal, this is due the fact that in order to model the correlation between a signal and delayed or shifted version of it, a shifted cross-covariance function is needed and this model only allows centered cross-covariances. Additionally, Fig. 3.3 shows the estimates for the LMC model and the learnt auto-covariances and cross-covariances. This model fails in the extrapolation of both derivative and delayed signals as a result of the symmetric cross-covariances product of linear combinations of univariate covariance functions.

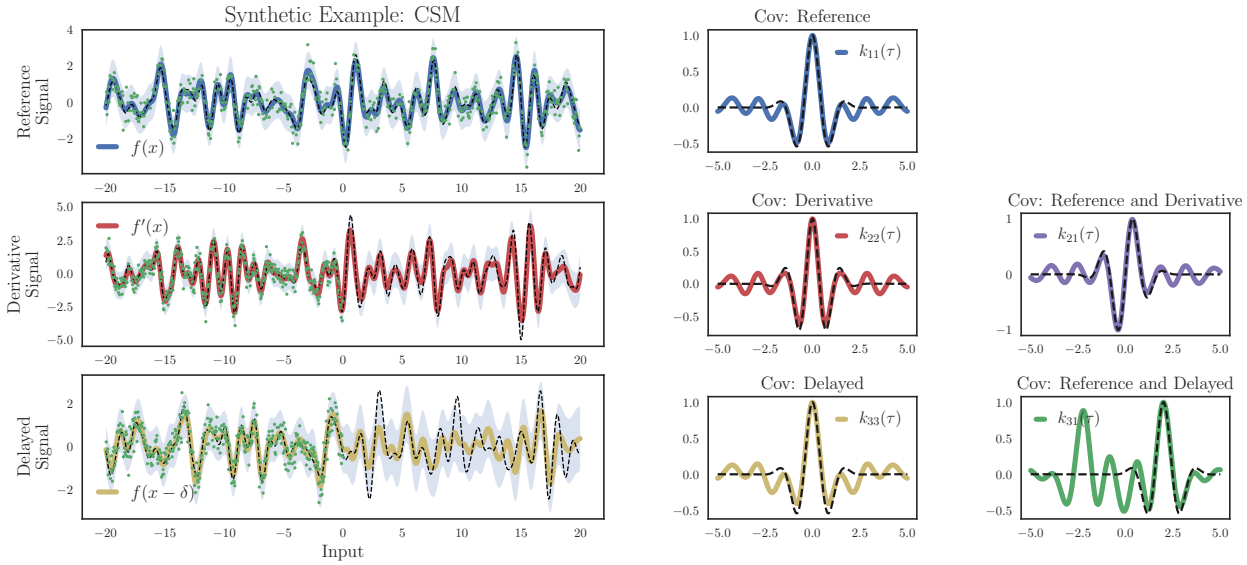

Figure 3.2: CSM learning of the covariance functions of a synthetic reference signal, its derivative and a delayed version. **Left:** synthetic signals, **middle**: autocovariances, **right**: cross-covariances. The dashed line represents the ground truth and the solid colour lines the CSM estimates.

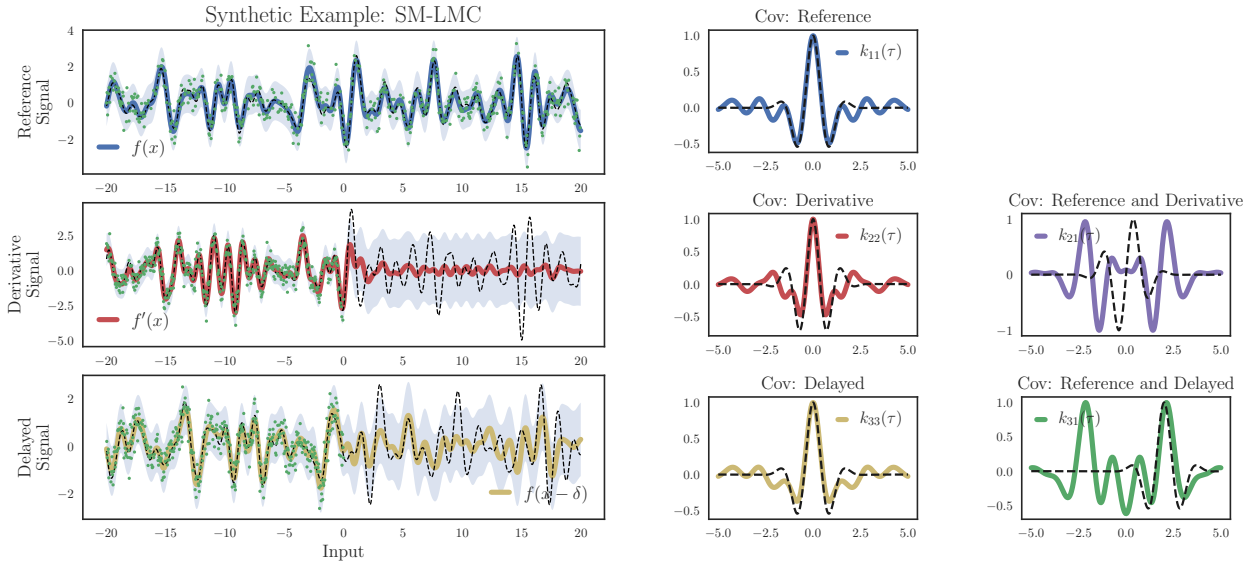

Figure 3.3: SM-LMC learning of the covariance functions of a synthetic reference signal, its derivative and a delayed version. **Left:** synthetic signals, **middle**: autocovariances, **right**: cross-covariances. The dashed line represents the ground truth and the solid colour lines the SM-LMC estimates.

Fig. 3.4 shows the estimates for the CONV model and the learnt auto-covariance and cross-covariances. The expression (1.23) shows that this model is only capable of constructing auto-covariances and cross-covariances that are either positive or negative, which is confirmed in this experiment, resulting in the failure of extrapolating the derivative and delayed signals.

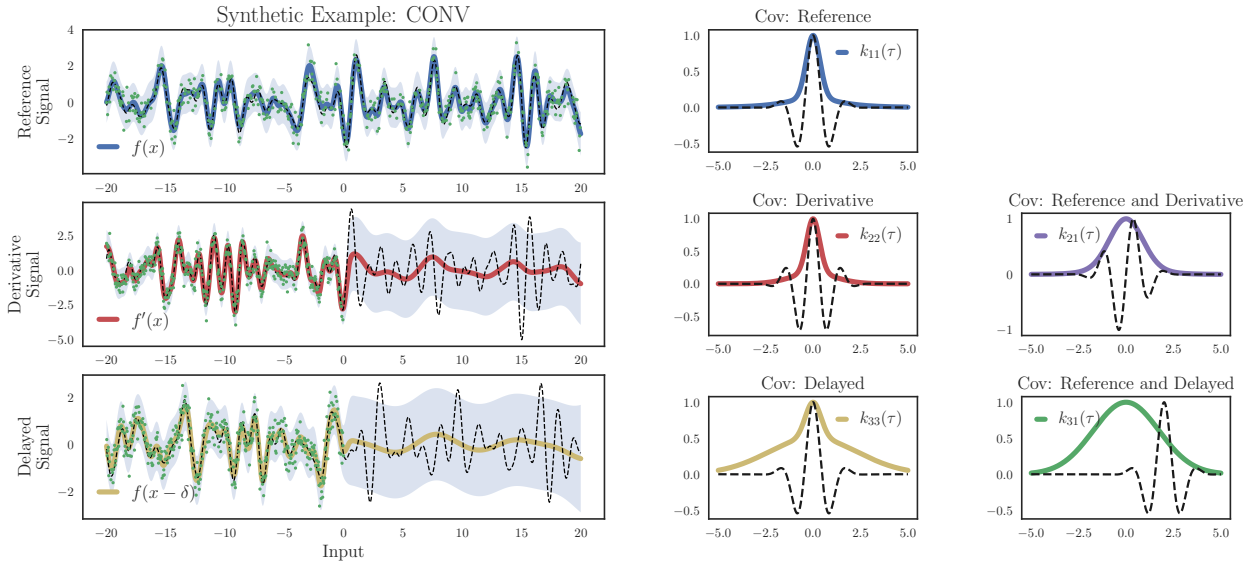

Figure 3.4: CONV learning of the covariance functions of a synthetic reference signal, its derivative and a delayed version. **Left:** synthetic signals, **middle**: autocovariances, **right**: cross-covariances. The dashed line represents the ground truth and the solid colour lines the CONV estimates.

3.2 Climate data

In order to compare the MOSM model against the others models in a real-data setting, a one-dimensional in the input real-world dataset was used, this dataset[1] contains measurements from a sensor network of four climate stations in the south on England, called: Cambermet, Chimet, Sotonmet and Bramblemet. Each sensor measures signals of Air Temperature, Wind Gust, Wind Speed, Tidal Height, among others at 5-minute intervals. For this experiment the normalized Air Temperature signal was considered from March 12, 2017 to March 16, 2017, that is 5692 samples, from where $N = 1000$ were randomly chosen for training.

Following [4] a sensor failure was simulated by removing the second half of the measurements for one sensor and leaving the remaining three sensors operating correctly, all this with the objective of testing the capability of the models to extrapolate the signals of the faulty sensor given the highly correlated signals of the three remaining sensors. This same setup was reproduced across all four sensors thus having four independent experiments. All models considered had five latent processes/spectral components ($Q = 5$ for all models).

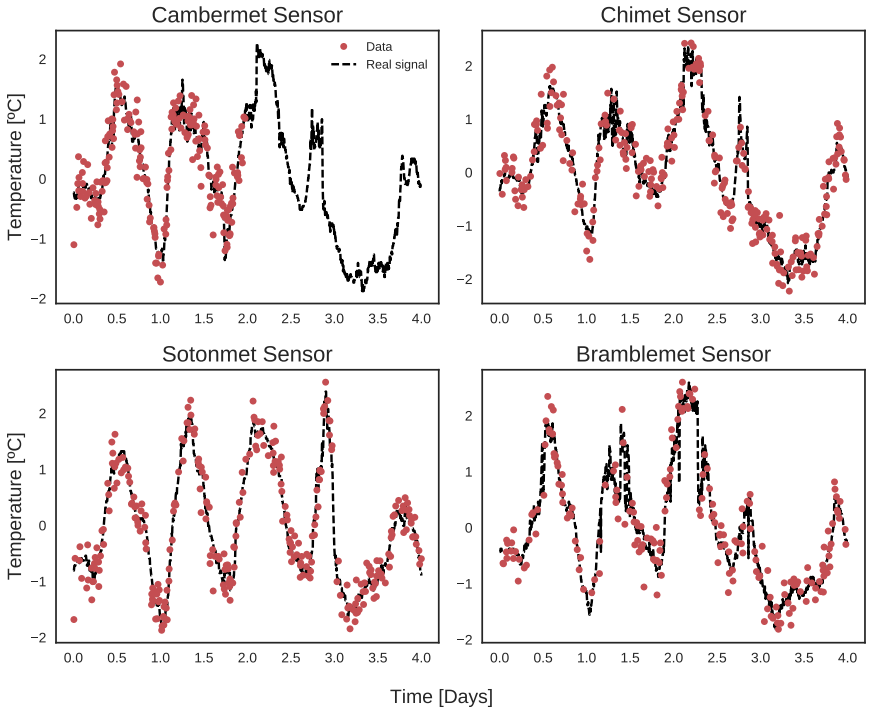

Figure 3.5: Climate data: Signals of all four sensor and the noisy samples, the faulty sensor in this case is the Cambermet sensor.

For all four models considered, Fig. 3.6 shows the estimates of missing data for the case where Cambermet was the simulated faulty sensor. Note how all models were able to capture the behavior of the signal in the missing range, this is because the considered climate signals are very similar to one another. This shows that the MOSM can also collapse to models that share parameters across pairs of outputs when required.

[1]Data can be obtained from `www.cambermet.co.uk.` and the sites therein.

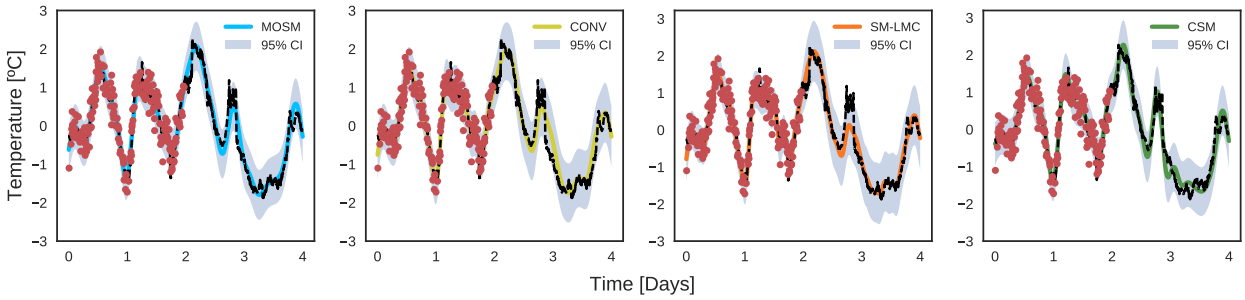

Figure 3.6: Climate data: Prediction of the Cambermet sensor in a simulated missing-data scenario. The red points denote the observations, the dashed black line the true signal, and the solid colour lines the considered models: From left to right: MOSM, CONV, SM-LMC and CSM.

Table 3.2: Mean absolute error for all four experiments with one-standard-deviation error bars over ten realizations.

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

Table 3.2 shows the mean absolute error with one standard deviation of the four experiments over ten realizations. These results do not show a significant contrast between the proposed model and the latent processes based models, in order to test for statistical significance, the Kolmogorov-Smirnov test was used with a significance level $\alpha = 0.05$, concluding that for the Sotonmet sensor we can assure that the MOSM model yields the best results. Conversely, for the Cambermet, Chimet and Bramblemet sensors, MOSM and CSM provided similar results, however, we cannot confirm their difference is statistically significant. Although, given the high correlation of these signals and the relationship between the MOSM model and the CSM model, the similar performance between these two under this dataset is not unexpected.

3.3 Heavy metal concentrations

With the intention of testing the proposed model in a multidimensional setting, the well-known Jura dataset [3] from Geostatistics was used. This dataset, as shown in Fig. 3.7, is a two-dimensional dataset ($n = 2$) that contains, in addition to other geological data, the concentration of seven heavy metals in a region of 14.5 km^2 of the Swiss Jura at different locations.

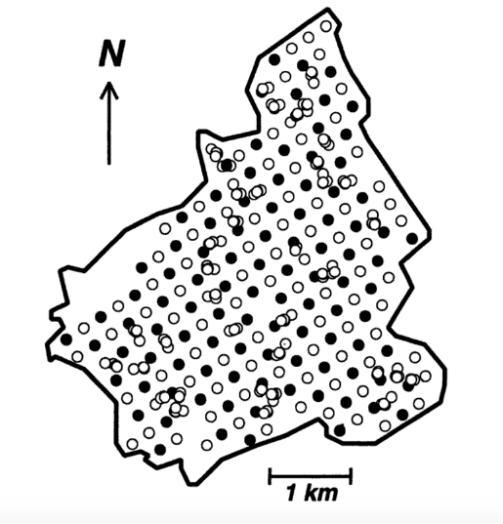

Figure 3.7: Map showing the split of the 359 data locations into a validation set (black circles) and a training set (white circles)

For each heavy metal concentrations data in this dataset, the data is divided into a training set (259 locations) and a validation set (100 locations). A procedure identical to [3, 4] was reproduced where the motivation is to aid the prediction of a variable that is expensive to measure by using abundant measurements of correlated variables which are less expensive to acquire. Specifically, Cadmium and Copper was estimated at the validation locations using measurements of related variables (Nickel and Zinc for Cadmium, and Lead, Nickel and Zinc for Copper) at the training and validation locations.

Table 3.3: Mean absolute error for the estimation of Cadmium and Copper concentrations with one-standard-deviation error bars over ten repetitions of the experiment, the results for the CONV model were obtained from [4]

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

Fig. 3.8 and Fig. 3.9 shows the estimation of the Cadmium and Copper concentrations respectively for the MOSM model, the MAE—see eq. (3.1)—is shown in Table 3.3, where the results for the CONV model were obtained from [4] and all models considered five latent signals/spectral components ($Q = 5$) except the independent Gaussian process (denoted IGP) which had a SM kernel as covariance function.

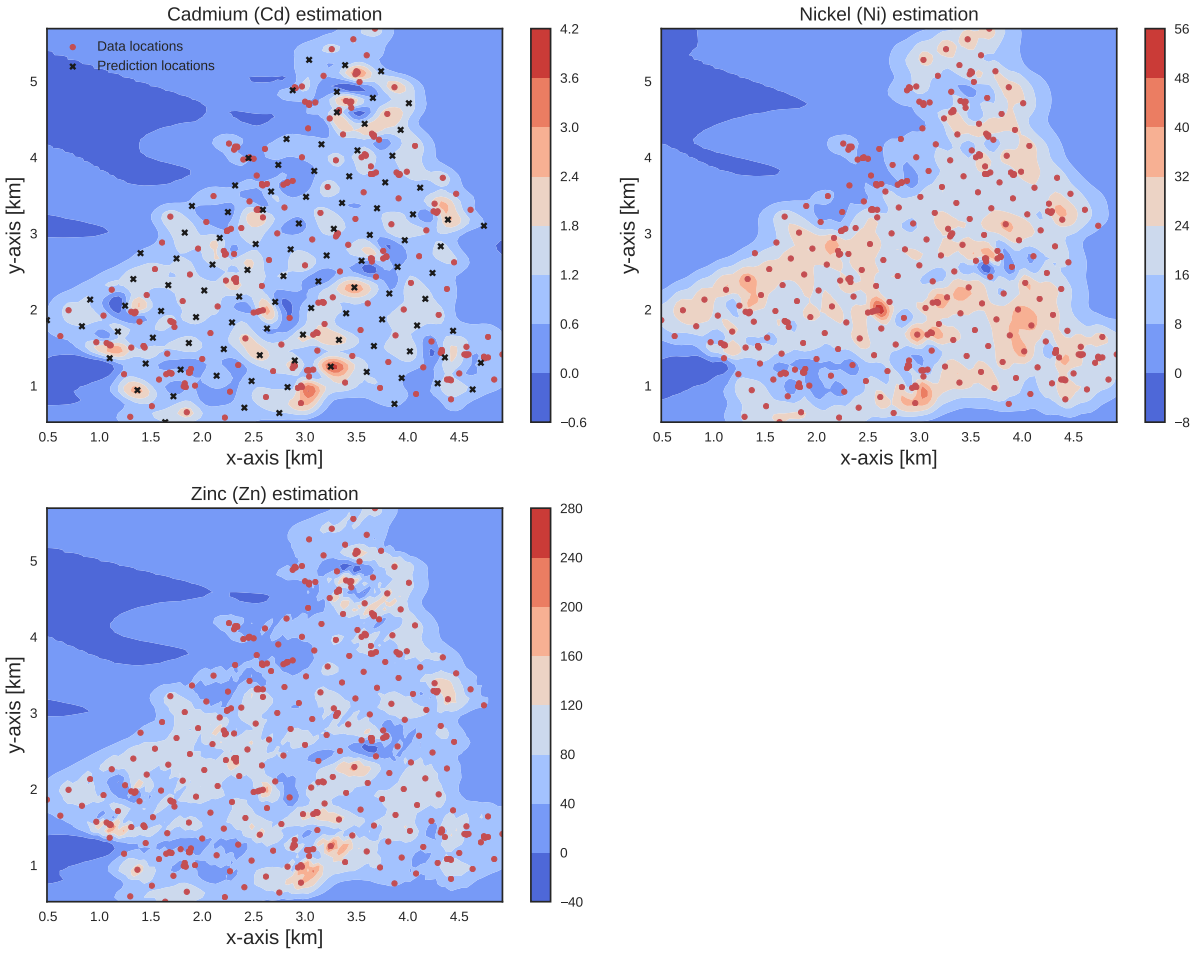

Figure 3.8: Jura dataset: **Upper-left**: estimation of the Cadmium concentration over the region, **Upper-right**: estimation of the Nickel concentration over the region, **Bottom-left**: estimation of the Zinc concentration over the region, the data points are shown in red and the validation (or prediction) locations are shown in black.

Note the proposed MOSM model outperforms all other models over the Cadmium data, which is statistical significant with a significance level $\alpha = 0.05$, while in the Copper case, we cannot guarantee statistical-significance difference between the CSM model and the MOSM. In either cases, testing for statistical significance against the CONV model was not possible since those results were obtained from [4]. On the other hand, the higher variability and non-Gaussianity of the data, as shown in Fig. 3.10 may be the reason of why the simplest MOGP model (SM-LMC) achieves the best results.

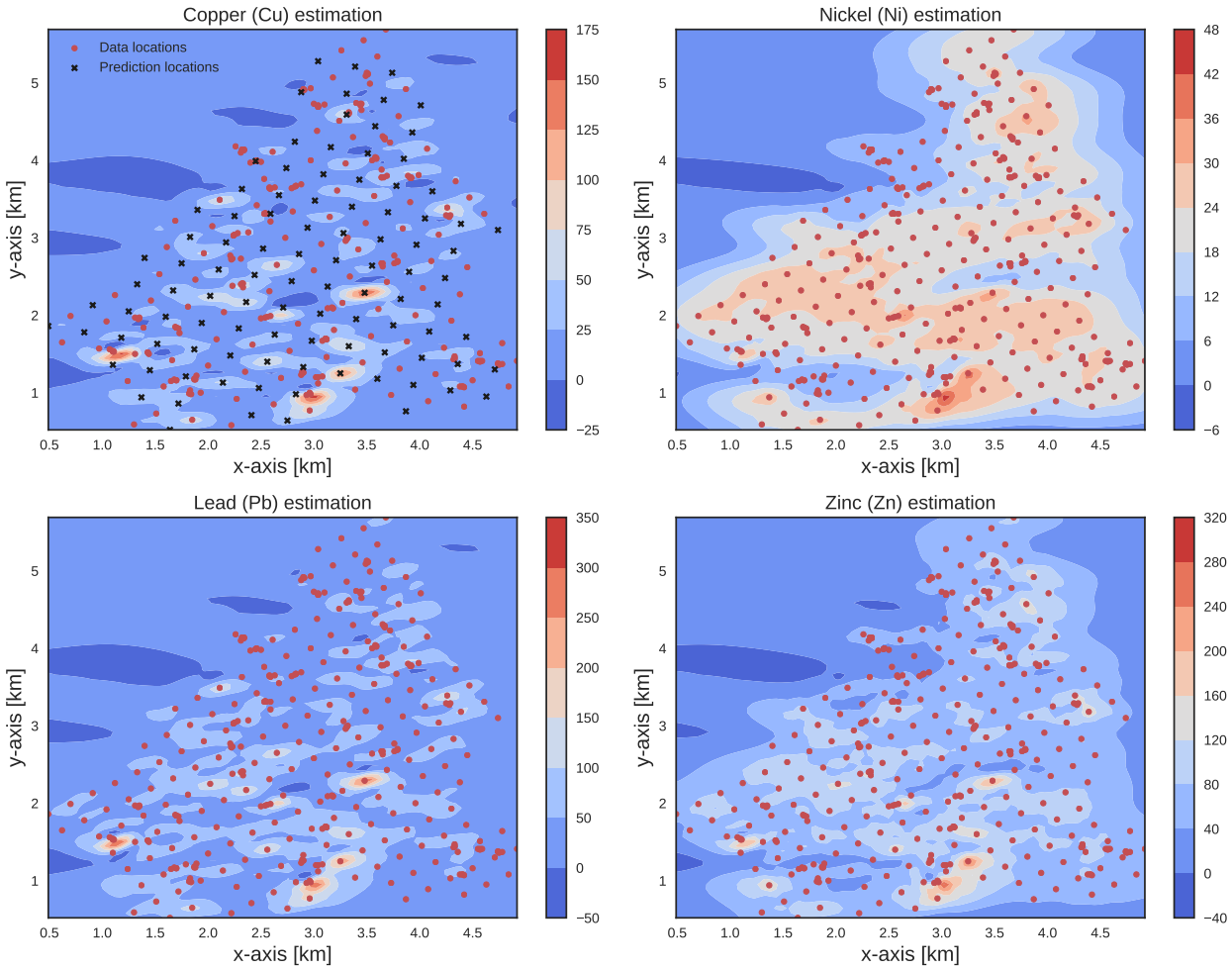

Figure 3.9: Jura dataset: **Upper-left**: estimation of the Copper concentration, **Upper-right**: estimation of the Nickel concentration over the region, **Bottom-left**: estimation of the Lead concentration over the region, **Bottom-right**: estimation of the Zinc concentration over the region, the data points are shown in red and the validation (or prediction) locations are shown in black.

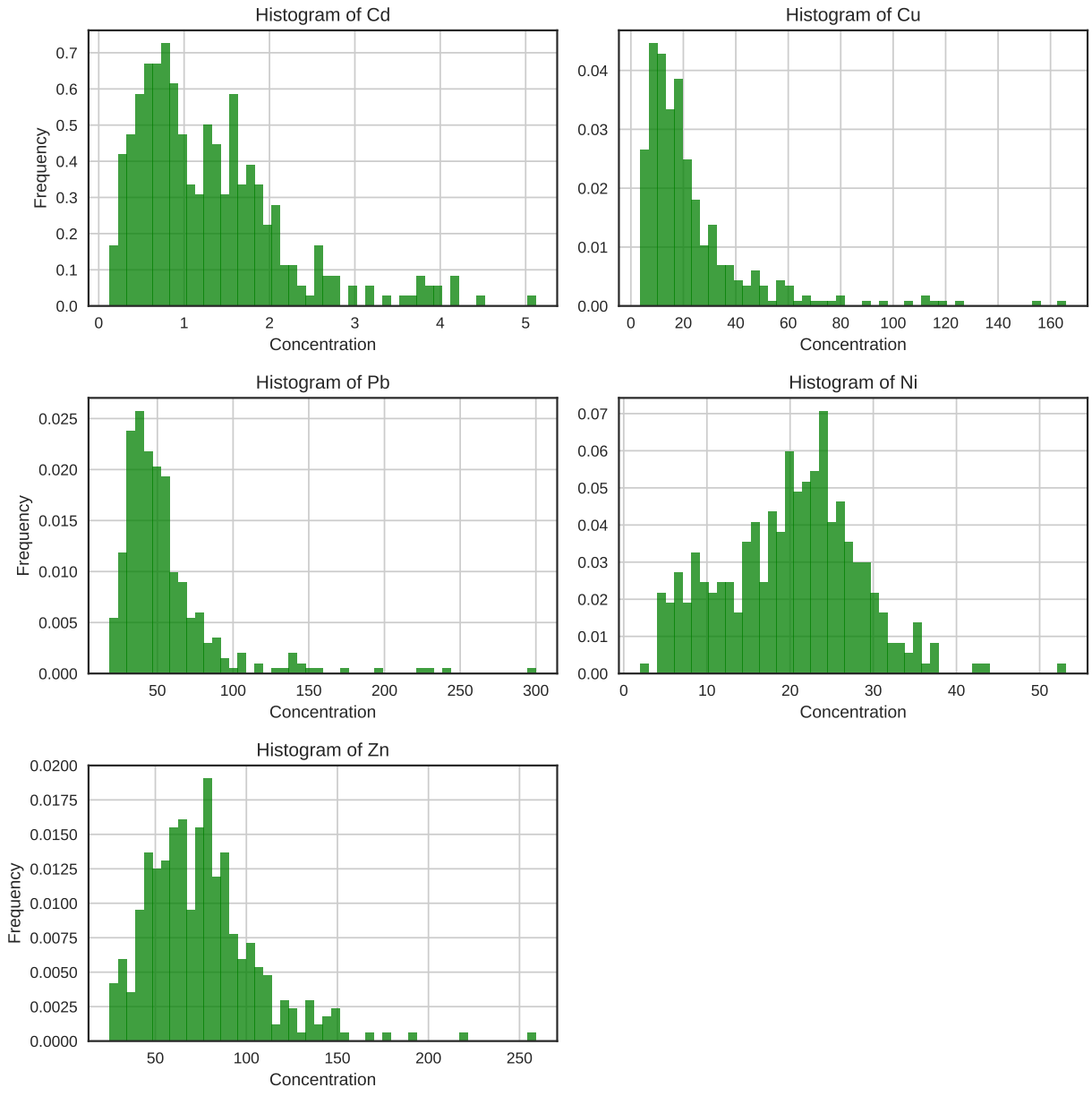

Figure 3.10: Jura dataset: Histogram of the different heavy metal concentrations in the Jura dataset, the non-Gaussian behavior is visible.

Conclusions

In this work, we have proposed the Multi-Output Spectral Mixture model (MOSM), which is a generalization to multi-outputs of the well-known Spectral Mixture kernel [2], this novel multivariate covariance function allows for modelling of rich relationships across multiple outputs within Gaussian processes regression models. This has been achieved by constructing a positive-definite matrix of complex-valued spectral densities and then transforming them by using the inverse Fourier transform according to Cramér's Theorem. The resulting multivariate covariance function provides clear interpretation from a spectral viewpoint, where each of its parameters can be identified with frequency, magnitude, phase and delay for a pair of outputs. A key feature that is unique to the proposed multivariate covariance function is the ability joint model delays and phase differences, this has been possible due to the complex-valued model for the cross-spectral density considered and validated experimentally using a synthetic example—see fig. 3.1. The MOSM kernel has also been compared against existing MOGP models on two real-world datasets, where the proposed model performed competitively in terms of the the mean absolute error. Additionally, we studied possible extensions of the multivariate Matérn model of [20, 21] by using complex-valued t-Student spectral densities instead of complex-valued squared exponential spectral densities.

One aspect left out, that most of the previous models have, is a sparse implementation [4, 25, 17] of the proposed model which is necessary due to the high computational cost that GPs have, also, an sparse representation of the model could allow the design of inducing variables that exploit the spectral content of the processes as in[26, 11]. On the other hand, further research should point towards the development of a general multivariate spectral kernel method such as in [27] where the spectral densities of covariance function given by Bochner's theorem are not modeled as a sum of Gaussian functions, but as a sum of *arbitrary integrable* functions, this allows to prescind of the infinite differentiability of sampled functions given by SM kernels, property which is unfortunately inherited by the MOSM model.

In summary, the proposed model shows the importance of complex components in the cross-spectral densities and hints how Cramér's theorem can be used to build stationary multivariate covariance functions by modelling the Cholesky decomposition instead, which open the possibility of new spectral multivariate models.

Bibliography

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

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

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

[4] M. A. Álvarez and N. D. Lawrence. Sparse convolved Gaussian processes for multi-output regression. In *Advances in Neural Information Processing Systems 21*, pages 57–64, 2008.

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

[6] A. M. Yaglom. *Correlation Theory of Stationary and Related Random Functions*. Number v. 1 in Correlation Theory of Stationary and Related Random Functions. Springer, 1987.

[7] M. A. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Found. Trends Mach. Learn.*, 4(3):195–266, March 2012.

[8] P. Boyle and M. Frean. Dependent gaussian processes. In *Advances in Neural Information Processing Systems 17*, pages 217–224, Cambridge, MA, USA, 2004. MIT Press.

[9] T. Tao. *An Introduction to Measure Theory*. Graduate studies in mathematics. American Mathematical Soc.

[10] D. Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

[11] F. Tobar, T. Bui, and R. Turner. Learning stationary time series using Gaussian processes with nonparametric kernels. In *Advances in Neural Information Processing Systems 28*, pages 3483–3491, 2015.

[12] S. Bochner, H. Pollard, and M. Tenenbaum. *Lectures on Fourier Integrals... With an Au-*

thor's Supplement on Monotonic Functions, Stieltjes Integrals, and Harmonic Analysis. Translated... by Morris Tenenbaum and Harry Pollard. Princeton, 1959.

[13] R.B. Ash and C. Doléans-Dade. *Probability and Measure Theory.* Harcourt/Academic Press, 2000.

[14] P. R. Halmos. *Measure Theory.* Graduate Texts in Mathematics. Springer New York, 1976.

[15] K. Plataniotis and D. Hatzinakos. *Gaussian Mixtures and Their Applications to Signal Processing.* CRC Press, 2017/07/10 2000.

[16] Y. W. Teh and M. Seeger. Semiparametric latent factor models. *Workshop on AISTATS 10*, 10, 2005.

[17] M. Álvarez, D. Luengo, M.K. Titsias, and N. D. Lawrence. Efficient multioutput Gaussian processes through variational inducing kernels. In *AISTATS*, volume 9, pages 25–32, 2010.

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

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

[20] T. Gneiting, W. Kleiber, and M. Schlather. Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491):1167–1177, 2010.

[21] T. V. Apanasovich, M. G. Genton, and Y. Sun. A valid matérn class of cross-covariance functions for multivariate random fields with any number of components. *Journal of the American Statistical Association*, 107(497):180–193, 2012.

[22] M. Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

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

[24] R.J. Adler. *The Geometry of Random Fields.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1981.

[25] M. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574. PMLR, 16–18 Apr 2009.

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

[27] Y. Kom Samo and S. Roberts. Generalized spectral kernels. *arXiv preprint arXiv:1506.02236*, 2015.