Department of Computer Science

Modelling non-stationary functions with Gaussian processes

Sami Remes

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 152/2019

Modelling non-stationary functions with Gaussian processes

Sami Remes

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall M1 of the school on 20 September 2019 at 12.

Aalto University
School of Science
Department of Computer Science
Probabilistic Machine Learning

Supervising professor
Prof. Samuel Kaski, Aalto University, Finland

Thesis advisors
Dr. Markus Heinonen, Aalto University, Finland
Dr. Tommi Mononen, University of Helsinki, Finland

Preliminary examiners
Dr. Edwin V. Bonilla, CSIRO's Data61, Australia
Dr. Miguel Lázaro-Gredilla, Vicarious, USA

Opponent
Dr. Seth Flaxman, Imperial College London, UK

Aalto University publication series
DOCTORAL DISSERTATIONS 152/2019

ISBN 978-952-60-8688-0 (printed)
ISBN 978-952-60-8689-7 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-8689-7

Unigrafia Oy
Helsinki 2019

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

A? Aalto University

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Abstract

Author
Sami Remes

Name of the doctoral dissertation
Modelling non-stationary functions with Gaussian processes

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 152/2019

Field of research Machine Learning

Manuscript submitted 24 March 2019 **Date of the defence** 20 September 2019

Permission for public defence granted (date) 14 June 2019 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

Gaussian processes (GP's) are a central piece of non-parametric Bayesian methods, which allow placing priors over functions in settings such as classification and regression. The prior is described using a kernel function that encodes a similarity between any two points in the input space, and thus defines the properties of functions that are modelled by the GP. In applying Gaussian processes the choice of the kernel is crucial, and the commonly used standard kernels often offer unsatisfactory performance due to making the assumption of stationarity. This thesis presents approaches in modelling non-stationarity from two different perspectives in Gaussian processes.

First, this thesis presents a formulation of a non-stationary spectral mixture kernel for univariate outputs, focusing on modelling the non-stationarity in the input space. The construction is based on the spectral mixture (SM) kernel, which has been derived for stationary functions using the Fourier duality implied by Bochner's theorem. The work done in this thesis extends the SM kernel into the non-stationary case. This is achieved by two complementary approaches, based on replacing the constant frequency parameters by input-dependent functions. The first approach is based on modelling the latent functions describing the frequency surface as Gaussian processes. In the second approach the functions are directly modelled as a neural network, parameters of which are optimized with respect to the variational evidence lower bound (ELBO).

Second, this thesis presents a kernel suitable for modelling non-stationary couplings between multiple output variables of interest in the context of multi-task or multi-output GP regression. The construction of the kernel is based on a Hadamard product of two kernels, which model the different aspects of dependencies between the outputs. The part of the kernel modelling the input-dependent couplings is based on a generalized Wishart process, which is a stochastic process on time-varying positive-definite matrices, in this case describing the changing dependencies between the outputs. The proposed Hadamard product kernel is applied in a latent factor model to enrich the latent variable prior distribution, that is, to model correlations within the latent variables explicitly. This results in the latent correlation Gaussian process model (LCGP).

This thesis additionally considers novel, flexible models for classification of multi-view data, specifically one based on a mixture of group factor analyzers (GFA). The model has a close relationship to the LCGP that builds a classifier in the latent variable space, while the classifier in the GFA mixture is based on the mixture assignments. GFA also allows modelling dependencies between groups of variables, which is not done by the LCGP. Applying Gaussian processes and adapting the proposed multi-output kernel would make the multi-view model even more general.

The methods introduced in this thesis now allow modelling non-stationary functions in Gaussian processes in a flexible way. The proposed kernels can be applied very generally, and the approaches introduced to derive them can also be applied to derive other types of non-stationary kernels.

Keywords Gaussian processes, non-stationary kernels, spectral kernels

ISBN (printed) 978-952-60-8688-0 **ISBN (pdf)** 978-952-60-8689-7

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2019

Pages 98 **urn** http://urn.fi/URN:ISBN:978-952-60-8689-7

Aalto-yliopisto

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tiivistelmä

Tekijä
Sami Remes

Väitöskirjan nimi
Epästationääristen funktioiden mallinnus gaussisilla prosesseilla

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietotekniikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 152/2019

Tutkimusala Koneoppiminen

Käsikirjoituksen pvm 24.03.2019 **Väitöspäivä** 20.09.2019

Väittelyluvan myöntämispäivä 14.06.2019 **Kieli** Englanti

☐ **Monografia** ☒ **Artikkeliväitöskirja** ☐ **Esseeväitöskirja**

Tiivistelmä

Gaussiset prosessit (GP) ovat yksi keskeisimpiä epäparametrisiä bayesiläisiä menetelmiä. Niiden avulla on mahdollista asettaa priorijakauma suoraan mallinnettavalle funktiolle esimerkiksi regressiossa tai luokittelussa. Keskeinen osa tätä prioria on niin sanottu kernelifunktio, joka kuvaa kahden datapisteen välistä samankaltaisuutta määräten GP:llä mallinnettavan funktion ominaisuudet. Käytettäessä GP-malleja kernelifunktion oikea valinta on oleellista, ja monet standardikernelit voivat antaa huonoja tuloksia, koska ne olettavat mallinnettavan funktion olevan stationäärinen. Tässä väitöskirjassa esitetään kaksi erilaista lähestymistapaa mallintaa epästationäärisyyttä gaussisissa prosesseissa.

Väitöskirjassa esitetään spektrimikstuurikerneleille epästationäärinen versio, joka mallintaa skalaariarvoisten funktioiden epästationäärisyyttä funktion arvojen käyttäytymisen suhteen. Spektrimikstuuri-kernelifunktio perustuu Bochnerin teoreemaan, jonka mukaan jokainen stationäärinen kerneli on esitettävissä spektritiheyden käänteisenä Fourier-muunnoksena. Mallintamalla kernelin taustalla olevaa spektritiheyttä gaussisena mikstuurina saadaan johdettua hyvin joustava ja monipuolinen kerneli GP-malleille. Tulkitsemalla spektritiheysmallin taajuus-, amplitudi- ja leveys-parametrit ajasta tai sijainnista riippuviksi funktioiksi, saadaan tuloksena yleistetty epästationäärinen spektrimikstuuri-kerneli. Tässä väitöskirjassa parametrifunktioita mallinnetaan kahdella vaihtoehtoisella tavalla: gaussisina prosesseina sekä neuroverkkoina.

Toinen lähestymistapa epästationäärisyyteen on vektori-arvoisten funktioiden tapauksessa, missä on tärkeää mallintaa vektorin eri muuttujien välisiä riippuvuuksia tai korrelaatioita. Tyypillisissä vektori-arvoisissa stationäärisissä gaussisissa prosesseissa muuttujien väliset riippuvuudet oletetaan olevan vakioita riippumatta siitä, missä kohtaa avaruutta funktion arvoja tarkastellaan. Tässä väitöskirjassa esitetään kernelifunktio, joka pystyy mallintamaan ajassa tai avaruudessa muuttuvia riippuvuuksia funktion eri muuttujien välillä. Esitetty kernelifunktio perustuu yleistetyn Wishartin prosessin ja epästationäärisen gaussisen kernelin yhdistelmään Hadamardin tulona.

Tässä väitöskirjassa esitetyt menetelmät mahdollistavat epästationääristen funktioden mallintamisen gaussisilla prosesseilla hyvin joustavilla tavoilla. Esitettyjä kerneleitä voi käyttää hyvin yleisesti eri sovelluksissa, ja kerneleiden johtamisessa käytetyt menetelmät soveltuvat myös uudenlaisten epästationääristen kernelien muodostamisessa.

Avainsanat gaussiset prosessit, epästationääriset kernelit, spektraalikernelit

ISBN (painettu) 978-952-60-8688-0 **ISBN (pdf)** 978-952-60-8689-7

ISSN (painettu) 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2019

Sivumäärä 98 **urn** http://urn.fi/URN:ISBN:978-952-60-8689-7

Preface

This work has been carried out in the Probabilistic Machine Learning (PML) group at the Department of Computer Science in Aalto University School of Science. I have been a part of the Finnish Center of Excellence in Computational Inference Research (COIN) and the Helsinki Institute for Information Technology HIIT. The thesis and the research have been partially funded by the projects ReKnow and ReKnow2 (Tekes), MindSee (EU FP7). The work has also been greatly supported by the computational resources provided by the Aalto Science-IT project.

I would like to thank my supervisor Prof. Samuel Kaski for all of his guidance and his insights as well as introducing me to the world of machine learning. I am also immensely thankful to my co-authors and advisors Dr. Markus Heinonen, with whom I have had the pleasure and privilege to work on three great publications, as well as Dr. Tommi Mononen, with whom I was privileged to work when starting my doctoral studies and writing the first publication that is a part of this thesis.

I wish also to thank the pre-examiners of this thesis, Dr. Edwin Bonilla and Dr. Miguel Lázaro-Gredilla, for their comments that helped to improve this thesis.

Additionally, I would like to thank all the people of the PML group, for fruitful discussions on both life and science, as well as the daily lunches and foosball games (during times when the board wasn't broken and had a place where playing didn't disturb the work of others) many people of the group engaged in. Lastly, but not least, I am grateful for all the support from my friends and family, especially to my wife Janica, who has been there to support me whenever I have needed that.

Helsinki, August 11, 2019,

Sami Remes

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Sami Remes, Tommi Mononen, Samuel Kaski. Classification of weak multi-view signals by sharing factors in a mixture of Bayesian group factor analyzers. In *MLINI 2015 - 5th NIPS Workshop on Machine Learning and Interpretation in Neuroimaging*, arXiv:1512.05610, Montreal, Quebec, Canada, December 2015.

II Sami Remes, Markus Heinonen, Samuel Kaski. A Mutually-Dependent Hadamard Kernel for Modelling Latent Variable Couplings. In *Asian Conference on Machine Learning*, Seoul, Korea, November 2017.

III Sami Remes, Markus Heinonen, Samuel Kaski. Non-Stationary Spectral Kernels. In *Neural Information Processing Systems*, Long Beach, California, USA, December 2017.

IV Sami Remes, Markus Heinonen, Samuel Kaski. Neural Non-Stationary Spectral Kernel. Submitted to *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (Machine Learning Journal Track)*, preprint available as arXiv:1811.10978, 26 November 2018.

Author's Contribution

Publication I: "Classification of weak multi-view signals by sharing factors in a mixture of Bayesian group factor analyzers"

The author had the main responsibility for specifying the details of the model, deriving variational inference updates and implementing the model, as well as designing and running the experiments. Writing of the article was a joint effort.

Publication II: "A Mutually-Dependent Hadamard Kernel for Modelling Latent Variable Couplings"

The author had the main responsibility for specifying the details of the model, deriving variational inference updates and implementing the model as well as designing and running the experiments. Writing of the article was a joint effort.

Publication III: "Non-Stationary Spectral Kernels"

The author came up with the definition of the proposed kernel, and designed and performed the experiments. Writing of the article was a joint effort.

Publication IV: "Neural Non-Stationary Spectral Kernel"

The author together with Dr. Heinonen came up with the idea of using neural networks in the kernel. The author designed and implemented the final formulation of the method, performed the experiments and had the main responsibility in writing the manuscript.

1. Introduction

This introductory chapter first motivates the approach taken in this thesis to model non-stationary functions and summarizes the contributions made in this thesis. Then it continues to briefly discuss the relevant literature related to the presented methods as well as briefly mention other approaches that have been used to model non-stationary functions.

1.1 Modelling Non-Stationarity with Gaussian Processes

This thesis considers modelling of non-stationary functions, which in the context of this thesis is taken to mean functions whose behavior changes according to the point where the function is evaluated. Non-stationarity can appear in modelling in multiple ways. The publications in this thesis approach non-stationary from two different perspectives. First, in vector-valued functions one is often interested in modelling the dependencies between the individual scalar functions. Typically they are assumed to have a covariance structure. This structure is often assumed to be stationary, that is, the covariances between the scalar values do not change according to where the function is evaluated. Methods to estimate covariances that are allowed to depend on the evaluation point of the function are developed as a part of this thesis. Second, even when modelling just one scalar function, the behaviour of the function is often assumed to be the same across the input space. For example, the frequency of a periodic function may realistically change over the years of a long observational period. This thesis presents approaches to model these kinds of changes in the behaviour of the function.

Gaussian processes (GP) provide a flexible way of modelling functions in a Bayesian way, by introducing a way to place a prior directly over functions. The prior information about the behavior of a function of interest in encoded into a *kernel function* that models the covariance structure of the function. The kernel itself is a function of two variables, input points, and determines the similarity between the points in the input space. The definition of the Gaussian process states that if the two points are similar, then the function values should also be similar, to a degree determined by the kernel function.

The choice of the kernel function defines the properties of the function that can modelled by the Gaussian process. Selecting a suitable kernel for the modelling problem at hand can be a difficult task, requiring specific problem domain knowledge. Instead, this thesis considers methods that allow learning an expressive and flexible kernel function directly from data with few assumptions. Many standard choices of kernels assume that the GP is stationary, but as explained above, this can often be a rather strong assumption. This thesis develops several formulations of non-stationary kernels, which can be learned flexibly in a data-driven way.

1.2 Contributions of the Thesis

Publication II constructs a novel multi-output kernel for Gaussian processes, which allows for new kind of analysis of vector-valued functions by modelling input-dependent couplings of the outputs. Specifically the multi-output kernel is a Hadamard product of a Gaussian kernel and an outer product latent Gaussian processes. Publication III and Publication IV propose two alternative constructions of non-stationary spectral mixture kernels that can represent, in addition to non-stationarity, many desirable properties, such as periodicity, or non-monotonicity, in more general way. Concretely, the non-stationary kernels are constructed by replacing the constant parameters of the stationary kernel (Wilson and Adams, 2013) by a function, which is modelled as a GP in Publication III and as a neural network in Publication IV.

Publication I presents a novel multi-view classification model, constructed as a mixture of group factor analyzers (Virtanen et al., 2012; Klami et al., 2015) which is combined with a classification model based on the mixture component assignments, meaning that a single class can be characterized by multiple modes of behavior. While this contribution is not based on Gaussian processes, using GP priors for the factor loadings and scores as well as combining the model with the work done in Publication II could follow in interesting future work.

This thesis considers applications in functional brain imaging (Publication I and Publication II) and in geostatistics for modelling mineral concentrations (Publication II) and land surface temperatures (Publication III). Publication IV furthermore considers time series data of solar activity in terms of the sunspot number and the irradiance as well as motion capture of human locomotion. This shows that the methods that are developed in the thesis are widely applicable in a variety of different fields.

1.3 Related Work

The work done in Publication III and Publication IV draws upon previous works on stationary spectral and sparse spectrum kernels (Wilson and Adams, 2013; Wilson et al., 2014; Lázaro-Gredilla et al., 2010; Kom Samo and Roberts, 2015; Lazaro-Gredilla and Quiñonero Candela, 2007). Using GP priors for latent functions within kernels such as the Gaussian and the Matérn has also been explored previously (Heinonen et al., 2016; Paciorek and Schervish, 2004, 2006). Neural networks have been previously proposed for input transformations in Gaussian processes but not for parameterizing individual parameters or functions (Wilson et al., 2016a). The generalized Wishart process (Wilson and Ghahramani, 2011) as well as Gaussian process regression networks (Wilson et al., 2012; Nguyen and Bonilla, 2013) inspired the work in Publication II.

Variational methods provide scalable approximate inference for complex Bayesian models (Jordan et al., 1999) such as the GP. Particularly for bigger datasets, stochastic variational methods allow handling up to billions of data points even for GP-based models (Titsias, 2009; Hoffman et al., 2013; Hensman et al., 2015), which are traditionally limited by their cubic complexity (Rasmussen and Williams, 2006). Variational approaches for inference are used in Publication I, Publication II and Publication IV of this thesis. For exact inference in GPs, several approaches for speeding up the computations exist, based on e.g. exploiting the Kronecker structure (Flaxman et al., 2015; Wilson et al., 2014) of the kernel matrix as well as other structures (Wilson and Nickisch, 2015). Kronecker structure based approaches are applied in Publication III of this thesis.

Similarly to the notion of non-stationarity that is considered in this thesis, there exists several other approaches, often called e.g. adaptive or local models, that aim to model changes in the behaviour of some function in different parts of the input space.

Change point detection methods find a set of discrete points where the behaviour of e.g. a time series changes suddenly (Aminikhanghahi and Cook, 2017). Change-points have also been proposed for Gaussian process time series models (Saatçi et al., 2010). Change points have also been generalized to change surfaces, for modelling higher-dimensional data (Herlands et al., 2016), and for modelling change-points in spatio-temporal processes (Majumdar et al., 2005; Knoblauch and Damoulas, 2018). The approaches proposed in this thesis assume that the changes in the dynamics are smooth, as opposed to the discrete changes assumed in change-point detection.

Mixtures of simple parametric models have also been adapted to modelling non-stationarity in time series, e.g. by exploiting self-organizing maps to segment the time-series (Ni and Yin, 2008). Multivariate adaptive regression splines (Friedman, 1991) adaptively partition the input space and apply linear models in the discovered segments.

In signal processing, Hilbert-Huang transforms and Hilbert spectral analysis explore input-dependent frequencies, but with deterministic transform functions on the inputs based on splines they do not provide a probabilistic interpretation (Huang, 2008; Huang et al., 1998), in contrast to Gaussian processes, for example.

Deep neural networks (Goodfellow et al., 2016) are also another class of models that do not limit themselves to stationary functions. In the Bayesian framework these models can be formulated either as Bayesian neural networks (Neal, 2012) with priors put on weights of the neural network or as deep Gaussian processes (Salimbeni and Deisenroth, 2017; Cutajar et al., 2017; Damianou and Lawrence, 2013) where each layer is a Gaussian process.

2. Modelling Non-Stationary Functions

This chapter starts by first presenting Gaussian processes as a method for modelling functions and explaining that the choice of the *kernel function* is central in applying Gaussian process models. The chapter continues to show how spectral theory can be applied to design flexible families of kernel functions, including those of Publication III and Publication IV. Last part of this chapter presents several computational approaches to perform inference in Gaussian processes.

2.1 Gaussian Processes as a Prior for Functions

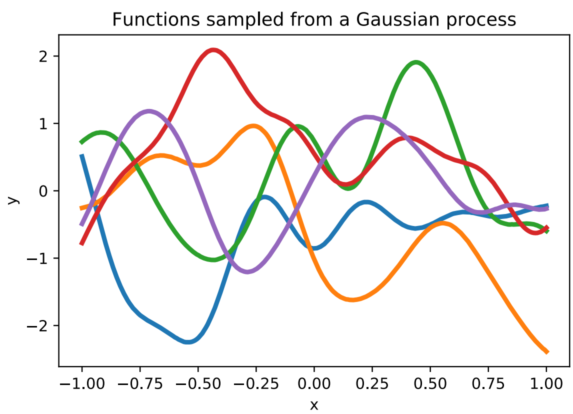

Figure 2.1. Some examples of functions sampled from a Gaussian process with an RBF kernel with lengthscale $\ell = 0.2$ and signal variance $\sigma_f^2 = 1$.

A Gaussian process (Rasmussen and Williams, 2006) defines a distribution over functions $f(x)$, denoted as

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')), \qquad (2.1)$$

Table 2.1. Some popular kernel functions used with Gaussian processes (Rasmussen and Williams, 2006), with $\tau = |x - x'|$ denoting the distance between two inputs.

Kernel	Parameters	$k(x,x')$
RBF	σ_f, ℓ	$\sigma_f^2 \exp\left(-\frac{\tau^2}{2\ell^2}\right)$
Rational Quadratic	σ_f, ℓ, α	$\sigma_f^2 \left(1 + \frac{\tau^2}{2\alpha\ell^2}\right)^{-\alpha}$
Matérn-3/2	σ_f, ℓ	$\sigma_f^2 \left(1 + \frac{\tau\sqrt{3}}{\ell}\right) \exp\left(-\frac{\tau\sqrt{3}}{\ell}\right)$
Cosine	σ_f, ω	$\sigma_f^2 \cos\left(2\pi\omega(x - x')\right)$
Periodic	σ_f, ω	$\sigma_f^2 \exp\left(-\frac{2\sin^2(\pi\omega\|x-x'\|)}{\ell^2}\right)$

where the function $m(x)$ defines the prior mean value $\mathbf{E}[f(x)] = m(x)$ and the kernel function $k(x,x')$ denotes the covariance between function values

$$\mathbf{cov}[f(x), f(x')] = k(x,x') \tag{2.2}$$

at points x and x'. Additionally, for any finite collection of inputs, $x_1, \dots, x_N$, the function values $f_i = f(x_i)$ follow a multivariate Gaussian distribution

$$\begin{pmatrix} f_1, & \dots, & f_N \end{pmatrix}^T \sim \mathcal{N}(\boldsymbol{m}, \mathbf{K}), \tag{2.3}$$

where the elements of $\mathbf{K}$ are $K_{ij} = k(x_i, x_j)$ and of $\boldsymbol{m}$ are $m_i = m(x_i)$.

The key property of Gaussian processes is that they can encode the properties of the modelled functions by correlating function values at input points that are similar according to the kernel $k(x,x')$. A very commonly used kernel is the squared exponential, also called the radial basis function (RBF), which for a univariate input $x \in \mathbb{R}$ is given by

$$k_{\text{RBF}}(x,x') = \sigma_f^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right), \tag{2.4}$$

where the signal variance σ_f^2 corresponds to the range scale of the function values, and the length-scale ℓ encodes how fast the function values can change with respect to the distance $\tau = x - x'$ between datapoints x and x'. The mean function $m(x)$ is often assumed to be a constant, $m(x) \equiv 0$, when the data have been normalized to zero mean. The multivariate case $\mathbf{x} \in \mathbb{R}^D$ can be handled by, e.g,. replacing the difference by $\tau = ||\mathbf{x} - \mathbf{x}'||$. Some other common kernel functions are listed in Table 2.1.

The choice of the kernel function that is used in the Gaussian process constrains the functions that can be modelled by the GP. Some important properties of the kernel function are (non-)stationarity, monotonicity and periodicity. A stationary kernel is only a function of the difference $\tau = x - x'$, that is $k(x,x') = k(x-x')$, and cannot have a different behaviour depending on the actual values of x or x'. Many of the standard kernels, such as the Gaussian, Matérn and rational quadratic, are stationary kernels. A non-stationary kernel, on the other hand, can encode dependencies that are different in different regions of the input space. Monotonicity refers to the property that the covariance between two points decreases as a function of their distance $|x-x'|$. Many of the popular kernels, such as the RBF and rational quadratic, behave this way. The opposite of monotonicity are periodic kernels that give rise to periodic functions, where the covariance is high at regular intervals with respect to the distance $|x-x'|$. Kernels that are neither monotonic nor strictly periodic can be said to be *non-monotonic*.

As the kernel defines the properties of the modelled function, the choice of the kernel is thus the most crucial decision to be made when applying Gaussian processes to any task. A common default choice is using e.g. the Gaussian or Matérn kernels, but they may not be able to model all the properties that the function has. Standard kernels can also be combined by taking sums or products of kernels having desired properties, resulting in more complex kernels, with some restrictions to ensure positive-definiteness. This can be, however, a laborious task and may require substantive domain knowledge. Some automated approaches to construct these compositions of kernels have been explored (Duvenaud et al., 2013; Lloyd et al., 2014). Kernels based on spectral decompositions provide another way of learning very flexible kernels from data, which is the topic of the next section.

2.2 Spectral Mixture Kernels

Spectral theory provides another approach for constructing kernel functions, as opposed to constructing them as compositions of other known base kernels. Bochner's theorem stipulates that a stationary kernel function has a corresponding spectral density, implying that specifying the spectral density at the same time specifies a corresponding kernel function. This idea has been used to constsruct several kernels and kernel approximations, such as random Fourier features (Rahimi and Recht, 2008), sparse spectrum kernels (Lazaro-Gredilla and Quiñonero Candela, 2007; Lázaro-Gredilla et al., 2010) as well as the spectral mixture kernel (Wilson and Adams, 2013). This section presents the background on spectral mixtures and then covers the developments done to introduce the non-stationary spectral mixtures of Publication III and Publication IV.

2.2.1 The Stationary Case

The building block for various (stationary) spectral kernel constructions is the Bochner's theorem for stationary signals, whose covariance can be written, with minor abuse of notation, as $k(\tau) = k(x-x') = k(x,x')$.

Theorem 1 (Bochner's Theorem)**.** *The function $k(\tau)$ is a continuous positive-definite function if and only if there exists a positive finite measure μ_S such that $k(\tau) = \int \exp(2\pi i s \tau) d\mu_S(s)$.*

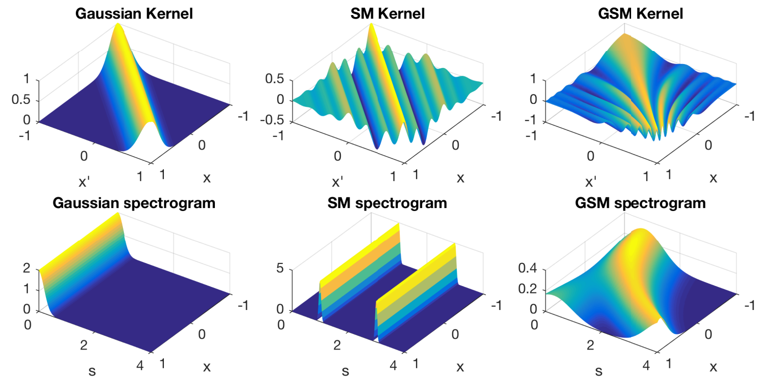

Figure 2.2. Illustration of the Gaussian, SM and GSM kernels on inputs on an interval $[-1,1]$. Gaussian and SM kernels' corresponding frequency representations are constant across input x, while the GSM contains a smoothly changing frequency component.

The theorem implies a Fourier dual (Wilson and Adams, 2013)

$$k(\tau) = \int S(s) e^{2\pi i s \tau} ds \tag{2.5}$$

$$S(s) = \int k(\tau) e^{-2\pi i s \tau} d\tau \,. \tag{2.6}$$

Here $S(s)$ denotes the *spectral density* describing the spectral content of the signal, and s is expressed in some units of frequency (e.g. for time series this could be in Hz). In other words, a kernel function can be obtained as the Fourier transform of its spectral density. The variables x and τ are within the input domain (e.g. time in seconds). The complex exponential function can be decomposed using Euler's formula as

$$e^{ix} = \cos(x) + i\sin(x)\,, \tag{2.7}$$

where the complex parts cancel out when the Fourier transform is applied to positive definite functions.

This dual representation has been exploited to design rich, yet stationary (due to being based on Bochner's theorem) kernel representations (Sinha and Duchi, 2016; Yang et al., 2015) and used for large-scale inference based on random Fourier features (Rahimi and Recht, 2008). Lázaro-Gredilla et al. (2010) proposed to directly learn the spectral density as a mixture of Dirac delta functions

$$S(s) = \sum_i \delta(s - s_i) \tag{2.8}$$

leading to the sparse spectrum (SS) kernel

$$k_{\text{SS}}(\tau) = \frac{1}{Q} \sum_{i=1}^{Q} \cos(2\pi s_i^T \tau)\,. \tag{2.9}$$

Wilson and Adams (2013) derived a stationary spectral mixture (SM) kernel by modelling the univariate spectral density using a mixture of normals

$$S_{\mathrm{SM}}(s) = \sum_i w_i^2 [\mathcal{N}(s|\mu_i, \sigma_i^2) + \mathcal{N}(s| - \mu_i, \sigma_i^2)]/2\,, \tag{2.10}$$

corresponding to the kernel function

$$k_{\mathrm{SM}}(\tau) = \sum_i w_i^2 \exp(-2\pi^2 \sigma_i^2 \tau^2) \cos(2\pi\mu_i \tau)\,. \tag{2.11}$$

The SM kernel was also extended for multidimensional inputs using Kronecker structure for scalability (Wilson et al., 2014; Flaxman et al., 2015; Saatçi, 2011).

Kernels derived from the spectral representation are particularly well suited to encoding long-range, non-monotonic or periodic kernels; however, they have been unable to handle non-stationarity fully before developments done in Publication III. Wilson (2014) in his thesis also presented a partly non-stationary SM kernel that has input-dependent mixture weights $w(x)$, similar to what has been done also by Tolvanen et al. (2014) and Heinonen et al. (2016). Kom Samo and Roberts (2015) in their unpublished work have also derived a general family of non-stationary kernels based on a more generalized version of the Bochner's theorem; their theorem however is missing a necessary condition that the mixture coefficients have to be positive. Random Fourier features for non-stationary kernels were also developed by Ton et al. (2018), The construction can be seen as a direct generalization of the stationary sparse spectrum kernel (Lázaro-Gredilla et al., 2010), using the generalized spectral density discussed in Kom Samo and Roberts (2015) and Publication III.

2.2.2 Non-Stationary Kernel by Extending the Stationary SM

Many non-stationary kernels, on the other hand, have been constructed by non-stationary extensions of Matérn and Gaussian kernels with input-dependent length-scales (Gibbs, 1997; Heinonen et al., 2016; Paciorek and Schervish, 2004, 2006), input space warpings (Sampson and Guttorp, 1992; Snoek et al., 2014), and with local stationarity with products of stationary and non-stationary kernels (Genton, 2001; Silverman, 1957). The simplest non-stationary kernel is arguably the dot product kernel (Rasmussen and Williams, 2006), which has also been used as a way to assign input-dependent signal variances (Tolvanen et al., 2014). These types of non-stationary kernels are a good match for functions with transitions in their dynamics, yet are unsuitable for modelling non-monotonic properties such as periodicity or other long-range structures.

Publication III presents the generalised spectral mixture (GSM) kernel, which can be seen as a generalisation of the spectral mixture (SM) kernel of Wilson and Adams (2013). The GSM kernel essentially parameterises the hyperparameters of the stationary SM kernel by Gaussian processes. That is, the frequencies μ, length-scales ℓ and mixture weights w are given GP priors:

$$\log w_i(x) \sim \mathcal{GP}(0, k_w(x,x')), \tag{2.12}$$

$$\log \ell_i(x) \sim \mathcal{GP}(0, k_\ell(x,x')), \tag{2.13}$$

$$\operatorname{logit} \mu_i(x) \sim \mathcal{GP}(0, k_\mu(x,x')). \tag{2.14}$$

Here the log transform is used to ensure the mixture weights $w(x)$ and length-scales $\ell(x)$ are non-negative, and the logit transform $\operatorname{logit}\mu(x) = \log \frac{\mu}{F_N - \mu}$ limits the learned frequencies between zero and the Nyquist frequency F_N, which is defined as half of the sampling rate of the signal (or for non-equispaced signals as the inverse of the smallest time interval between the samples). The functions $w(x)$, $\ell(x)$ and $\mu(x)$ can be thought intuitively as forming a smooth spectrogram surface

$$S(x,s) = w(x)^2 \exp\left(-\ell(x)^2 (s - \mu(x))^2\right), \tag{2.15}$$

by simply replacing the constant parameters in the spectral density (2.10) of the SM kernel with the functions $w(x)$, $\ell(x)$ and $\mu(x)$.

The non-stationary generalised spectral mixture (GSM) kernel is then given by:

$$k_{\mathrm{GSM}}(x,x') = \sum_{i=1}^{Q} w_i(x) w_i(x') k_i^{\mathrm{G}}(x,x') \cos(2\pi(\mu_i(x)x - \mu_i(x')x')). \tag{2.16}$$

The kernel is a product of three PSD terms. The GSM kernel encodes the similarity between two data points based on their combined signal variance $w(x)w(x')$, and the frequency surface based on the frequencies $\mu(x), \mu(x')$ and frequency lengthscales $\ell(x), \ell(x')$ associated with both inputs. The input-dependent length-scales are accommodated by replacing the exponential part of (2.11) by the Gibbs kernel

$$k_i^{\mathrm{G}}(x,x') = \sqrt{\frac{2\ell_i(x)\ell_i(x')}{\ell_i(x)^2 + \ell_i(x')^2}} \exp\left(-\frac{(x-x')^2}{\ell_i(x)^2 + \ell_i(x')^2}\right),$$

which is a non-stationary generalisation of the Gaussian kernel (Gibbs, 1997; Heinonen et al., 2016; Paciorek and Schervish, 2004).

The GSM kernel is able to model functions that have complex, non-monotonic and non-stationary covariances. The kernel is also able to model periodic functions where e.g. the amplitude or period changes over time. Further, it also reduces to the stationary SM kernel when the functions $\mu(x)$, $\ell(x)$, and $w(x)$ are constant, which can be obtained by setting their associated length-scales to large values.

2.2.3 Neural Networks in the Spectral Mixture

Publication IV proposes modelling the latent frequency $\mu(x)$, variance $w(x)$ and length-scale $\ell(x)$ functions using neural networks instead, as opposed to giving them a GP prior as in Publication III. This can be seen as similar to variational autoencoders (Kingma and Welling, 2013) proposed for inference in latent variable models. Here the neural network parameterized functions are applied to model the parameters of the kernel, instead of the parameters of a variational distribution, and as such are actually a part of the model (the kernel).

A typical, fully-connected feed-forward neural network (Goodfellow et al., 2016) is defined as a composition of simple affine maps (referred to as layers) followed by element-wise non-linear transformations (activations). This can be written as

$$\mathbf{h}_1 = a(\mathbf{W}_1 x + \mathbf{b}_1) \tag{2.17}$$

$$\mathbf{y} = a(\mathbf{W}_2 \mathbf{h}_1 + \mathbf{b}_2) \tag{2.18}$$

for a network with just one hidden layer, for a function from x to $\mathbf{y}$. Here matrices $\mathbf{W}_l$ are the *weights* and vectors $\mathbf{b}_l$ are the *biases* of the lth layer of the network, and $a(\cdot)$ is the *activation function*. We can also denote this mapping as $\mathbf{y} = \mathrm{NN}(x)$. For the hidden layer activation functions, the scaled exponential linear units (SELU) that have been proposed recently for feed-forward networks are used (Klambauer et al., 2017). At the top layer, the softplus activation

$$\mathrm{softplus}(x) = \log(1 + \exp(x)) \tag{2.19}$$

ensures that the functions are positive-valued. Now the latent functions within the GSM kernel are given as

$$\mathbf{w}(x) = \mathrm{NN}_w(x), \tag{2.20}$$

$$\boldsymbol{\ell}(x) = \mathrm{NN}_\ell(x), \tag{2.21}$$

$$\boldsymbol{\mu}(x) = \mathrm{NN}_\mu(x), \tag{2.22}$$

where all elements of each function are given by the same neural network, that is, they share all weights except on the final layer before the softmax activation. Thus, the neural network is able to find a common representation of the input data that is useful in determining the amplitudes $\mathbf{w}$, the lengthscales $\boldsymbol{\ell}$ and frequencies $\boldsymbol{\mu}$. This is in contrast to the GP-GSM, where the functions within each mixture component were modelled independently.

2.3 Inference in GPs

This section covers the topic of performing inference in Gaussian processes. The first subsection provides a recap of standard GP regression, and the rest of this section provides two approaches to speed up the inference, by exploiting properties of Kronecker products as well as applying stochastic variational inference.

2.3.1 Inference in GP Regression

Gaussian processes in the regression framework assume a Gaussian likelihood over N data points $(\mathbf{x}_j, y_j)_{j=1}^N$ with all outputs collected into a vector $\mathbf{y} \in \mathbb{R}^N$,

$$y_j = f(\mathbf{x}_j) + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}(0, \sigma_n^2) \tag{2.23}$$

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}' \mid \boldsymbol{\theta})), \tag{2.24}$$

where (2.23) implies a Gaussian likelihood

$$p(\mathbf{y} \mid \mathbf{f}) = \mathcal{N}(\mathbf{y} \mid \mathbf{f}, \sigma_n^2 \boldsymbol{I}). \tag{2.25}$$

A standard approach for finding the kernel parameters $\boldsymbol{\theta}$ is to marginalize over the function values $\mathbf{f}$, and optimize the marginal log likelihood

$$\log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log \mathbf{E}_{p(\mathbf{f})}[p(\mathbf{y} \mid \mathbf{f})] \tag{2.26}$$

$$= -\frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K} + \sigma_n^2\mathbf{I}| + \text{const.} \tag{2.27}$$

using standard optimization techniques, such as gradient descent. This requires $\mathcal{O}(N^3)$ computation due to inverting and computing the determinant of the $N \times N$ kernel matrix $\mathbf{K}$, making it unfeasible to apply GPs to large datasets. The same applies for computing the predictive distribution of the function values f^* for new inputs $\mathbf{x}^*$, calculated using standard Gaussian formulas as

$$p(f^* \mid \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(f^* \mid \mu_{f^*}, \sigma_{f^*}^2) \tag{2.28}$$

$$\mu_{f^*} = \boldsymbol{k}^T(\mathbf{K} + \sigma_n^2)^{-1}\mathbf{y} \tag{2.29}$$

$$\sigma_{f^*}^2 = k(\mathbf{x}^*, \mathbf{x}^*) - \boldsymbol{k}^T(\mathbf{K} + \sigma_n^2)^{-1}\boldsymbol{k}, \tag{2.30}$$

where the $\boldsymbol{k}$ collects the kernel values $k(\mathbf{x}^*, \mathbf{x}_i)$ evaluated between the new point $\mathbf{x}^*$ and all training data points $\mathbf{x}_i$. The gradient of the marginal log likelihood (2.27) in general is

$$\frac{\partial \log p(\mathbf{y} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{2}\left(\boldsymbol{\alpha}^T \frac{\partial \mathbf{K}}{\partial \boldsymbol{\theta}} \boldsymbol{\alpha} - \operatorname{tr}\left((\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\frac{\partial \mathbf{K}}{\partial \boldsymbol{\theta}}\right)\right), \tag{2.31}$$

where $\boldsymbol{\alpha} = (\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y}$.

2.3.2 Speeding Up Inference with Kronecker Products

If a kernel function defined over a multivariate input, that is $\mathbf{x} \in \mathbb{R}^D$ with $D \geq 2$, can be written as

$$k(\mathbf{x}, \mathbf{x}') = \prod_{d=1}^{D} k_d(x_d, x_d') \tag{2.32}$$

and the data are located on a multi-dimensional grid, the kernel matrix can be written as a Kronecker product of kernel matrices

$$\mathbf{K} = \bigotimes_{d=1}^{D} \mathbf{K}_d \tag{2.33}$$

where each $\mathbf{K}_d$ is computed only using the kernel function $k_d(\cdot,\cdot)$ using the inputs along the dimension x_d (Saatçi, 2011; Wilson et al., 2014).

Experiments done in Publication III exploit Kronecker inference in modelling land surface temperatures (LST) over the United States and in an in-painting task using texture images. In these cases the inputs are three and two dimensional and on a grid, in the LST case all longitude and latitude coordinates are observed at all time points and in textures the full pixel information is present except for missing some horizontal or vertical blocks that does not break the grid structure. With some additional approximations, Kronecker inference can also be made to work with incomplete grids and non-Gaussian likelihoods (Flaxman et al., 2015; Wilson et al., 2014; Wilson and Nickisch, 2015).

The Kronecker product of two matrices $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ and $\boldsymbol{B} \in \mathbb{R}^{P\times Q}$ is defined as

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1M}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2M}\boldsymbol{B} \\ \vdots & & \ddots & \\ a_{N1}\boldsymbol{B} & & & a_{NM}\boldsymbol{B} \end{bmatrix} \in \mathbb{R}^{NP\times MQ}. \tag{2.34}$$

The most relevant properties of the Kronecker product in applying to GP inference include the inverse, assuming invertible square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1} \tag{2.35}$$

and the following applying to the eigen decomposition

$$\boldsymbol{A} \otimes \boldsymbol{B} = \boldsymbol{Q}\boldsymbol{V}\boldsymbol{Q}^T = (\boldsymbol{Q}_A \otimes \boldsymbol{Q}_B)(\boldsymbol{V}_A \otimes \boldsymbol{V}_B)(\boldsymbol{Q}_A \otimes \boldsymbol{Q}_B)^T, \tag{2.36}$$

where the matrix $\boldsymbol{Q}$ collects the eigenvectors and the matrix $\boldsymbol{V}$ consists of the eigenvalues on the diagonal. Additionally a matrix-vector product between a Kronecker structured matrix and a vector is efficiently expressed as

$$(\boldsymbol{A} \otimes \boldsymbol{B})\,\mathrm{vec}(\mathbf{Y}) = \mathrm{vec}(\boldsymbol{B}^T \mathbf{Y} \boldsymbol{A}), \tag{2.37}$$

where $\mathbf{Y}$ is a matrix with compatible dimensions to $\boldsymbol{A}$ and $\boldsymbol{B}$.

Computing the two terms in the marginal log likelihood (2.27) now becomes

$$(\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1}\mathbf{y} = \boldsymbol{Q}(\boldsymbol{V} + \sigma_n^2\mathbf{I})^{-1}\boldsymbol{Q}^T\mathbf{y} = \left(\bigotimes_d \boldsymbol{Q}_d\right)\left((\boldsymbol{V} + \sigma_n^2\mathbf{I})^{-1}\left(\left(\bigotimes_d \boldsymbol{Q}_d^T\right)\mathbf{y}\right)\right) \tag{2.38}$$

$$\log|\mathbf{K} + \sigma_n^2\mathbf{I}| = \sum_i \log(\boldsymbol{V}_{ii} + \sigma_n^2). \tag{2.39}$$

The inversion of $\boldsymbol{V} + \sigma_n^2 \boldsymbol{I}$ is trivial as it is a diagonal matrix. The matrix-vector products such as $\left(\otimes_d \boldsymbol{Q}_d^T\right)\mathbf{y}$ can also be computed efficiently using formula (2.37) successively.

For optimizing the marginal log likelihood, the gradient of the Kronecker product kernel can be computed as

$$\frac{\partial \mathbf{K}}{\partial \theta_d} = \mathbf{K}_1 \otimes \ldots \otimes \frac{\partial \mathbf{K}_d}{\partial \theta_d} \otimes \ldots \otimes \mathbf{K}_D \tag{2.40}$$

with respect to dimension-specific hyperparameters θ_d. The trace term in the gradient (2.31) can be computed using the Kronecker eigenvalue decomposition (2.36) and the cyclic property of the trace operator.

2.3.3 Stochastic Variational Inference in Sparse GPs

Variational Inference

Variational inference (VI) is one of the most popular approximate inference methods for Bayesian inference (Jordan et al., 1999; Wainwright et al., 2008; Hoffman et al., 2013). For this more general treatment, the following assumes a probabilistic model for a data point x with a local latent variable z (size of which depends on the size of the data) and a global parameter β (fixed size specified by the specific model), factorizing as $p(x,z,\beta) = p(x \mid z,\beta)p(z \mid \beta)p(\beta)$. In addition to the global parameter β, the model may also have some hyperparameters θ that are not treated fully probabilistically within the VI framework, which can be denoted by having all distributions conditioned on θ as $p(x,z,\beta \mid \theta)$.

The variational inference approach is based on lower bounding the model log evidence

$$\log p(x \mid \theta) = \log \int p(x,z,\beta \mid \theta) dz d\beta. \tag{2.41}$$

The approximation to the posterior $p(z,\beta \mid x,\theta)$ is denoted by the variational distribution $q(z,\beta)$. Plugging in q into the evidence,

$$\log p(x \mid \theta) = \log \int p(x,z,\beta \mid \theta) \frac{q(z,\beta)}{q(z,\beta)} dz d\beta \tag{2.42}$$

$$= \log \mathbf{E}_{q(z,\beta)} \left[\frac{p(x,z,\beta \mid \theta)}{q(z,\beta)} \right] \tag{2.43}$$

$$\geq \mathbf{E}_{q(z,\beta)}[\log p(x,z,\beta \mid \theta)] - \mathbf{E}_{q(z,\beta)}[\log q(z,\beta)] \equiv \mathscr{L}(q), \tag{2.44}$$

where exchanging the order of log and $\mathbf{E}_q$ in (2.44) is due to the Jensen's inequality. The evidence lower bound (ELBO) $\mathscr{L}$ acts as the objective function that is optimized in variational inference. The additional hyperparameters θ may also be optimized with respect to the ELBO to determine suitable values.

Mean-field Approximations

A typical form of approximation q is the so-called mean-field factorization

$$q(z,\beta) = q(\beta \mid \lambda)\prod_n q(z_n \mid \phi_n), \tag{2.45}$$

where the product is taken over the data points n. The variables λ and ϕ_n are referred to as the global and local variational parameters, respectively. For example, if $q(\beta \mid \lambda)$ were a Gaussian distribution, then λ would be its mean and variance parameters or the corresponding natural parameters in the exponential family formulation. Distributions in the exponential family can be written as

$$p(\beta \mid \theta) = h(\beta)\exp\left(\eta(\theta)^T T(\beta) - A(\theta)\right), \tag{2.46}$$

where η is the natural parameter, T the sufficient statistic, h an arbitrary function and A the log-normalizer ensuring that the distribution integrates to one.

When maximizing the ELBO (2.44) with respect to a single term (e.g. $q(\beta)$) in the factorization (2.45), the optimum is of the form

$$\log q(\beta) = \mathbf{E}_{q(z)}\left[\log p(x,z,\beta \mid \theta)\right] + \text{const.}. \tag{2.47}$$

In the exponential family this corresponds (Hoffman et al., 2013) to setting the parameter to the value of the expected natural parameter η

$$\lambda = \mathbf{E}_{q(z)}\left[\eta(x,z,\theta)\right], \tag{2.48}$$

where $\eta(x,z,\theta)$ is the natural parameter of the conditional $p(\beta \mid x,z,\theta)$.

Stochastic Variational Inference

Stochastic variational inference (SVI) employs stochastic optimization to the variational objective (Hoffman et al., 2013). The approach is based on an observation that in the exponential family the mean-field updates correspond to natural gradients of the ELBO. In SVI, only a partial update is made using a subset of the data points as a weighted average

$$\lambda^{(t)} = (1-\rho_t)\lambda^{(t-1)} + \rho_t\hat{\lambda}^{(t)}, \tag{2.49}$$

where $\hat{\lambda}^{(t)}$ is computed using just one data point replicated N times using the standard parameter update rule (2.48), and ρ_t is the step size.

So-called black-box variational inference (Ranganath et al., 2014; Rezende et al., 2014; Kingma and Welling, 2013) methods enable, together with automatic differentiation, variational inference without assuming exponential family distributions or solving the approximate model update rules (2.47) by hand. The methods are based on applying standard stochastic gradient descent or similar stochastic optimization algorithms directly to the variational ELBO using Monte Carlo samples from the variational distribution to compute a noisy gradient, if the analytical form of the ELBO and its gradient are not available.

Sparse Gaussian Processes

In sparse Gaussian processes approaches (Titsias, 2009; Hensman et al., 2015; Snelson and Ghahramani, 2006; Lawrence et al., 2003; Smola and Bartlett, 2001; Seeger et al., 2003), the GP is approximated by working on a smaller set of $M \ll N$ pseudo- or inducing inputs. Approximations based on the variational framework are the most common presently. In the variational approaches (Titsias, 2009; Hensman et al., 2015), the function values are decomposed by adding new extra variables $\mathbf{u}$ giving the joint distribution as

$$p(\mathbf{y},\mathbf{f},\mathbf{u} \mid \boldsymbol{\theta}) = p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f} \mid \mathbf{u},\boldsymbol{\theta})p(\mathbf{u} \mid \boldsymbol{\theta}) \tag{2.50}$$

Here $\mathbf{u}$ denote the function values at *inducing points* $\mathbf{z}$, which according to the definition of Gaussian process, have the prior

$$p(\mathbf{u} \mid \boldsymbol{\theta}) = \mathcal{N}(\mathbf{u} \mid 0, \mathbf{K}_{\mathbf{zz}}), \tag{2.51}$$

where $\mathbf{K}_{\mathbf{zz}}$ denotes the kernel matrix computed at the inducing points. The kernel computation depends on the kernel hyperparameter values $\boldsymbol{\theta}$. The posterior $p(\mathbf{u} \mid \mathbf{y})$ is approximated by $q(\mathbf{u}) = \mathcal{N}(\mathbf{u} \mid \boldsymbol{m}, \boldsymbol{S})$, where $\boldsymbol{m}$ and $\boldsymbol{S}$ are the variational parameters. Furthermore, still under the assumption that $\mathbf{f}$ and $\mathbf{u}$ are drawn from the same GP, the conditional distribution $f \mid \mathbf{u}$ is given by the standard formula

$$p(\mathbf{f} \mid \mathbf{u},\boldsymbol{\theta}) = \mathcal{N}(\mathbf{f} \mid \boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f) \tag{2.52}$$

$$\boldsymbol{\mu}_f = \mathbf{K}_{\mathbf{xz}}\mathbf{K}_{\mathbf{zz}}^{-1}\mathbf{u} \tag{2.53}$$

$$\boldsymbol{\Sigma}_f = \mathbf{K}_{\mathbf{zz}} - \mathbf{K}_{\mathbf{xz}}^T\mathbf{K}_{\mathbf{xx}}^{-1}\mathbf{K}_{\mathbf{xz}}. \tag{2.54}$$

The following inequality applies for the joint distribution (2.50), due to Jensen's inequality,

$$\log p(\mathbf{y} \mid \mathbf{u}) = \log \int p(\mathbf{y} \mid \mathbf{f})p(\mathbf{f} \mid \mathbf{u})d\mathbf{f} \tag{2.55}$$

$$= \log\left(\mathbf{E}_{p(\mathbf{f}|\mathbf{u})}[p(\mathbf{y} \mid \mathbf{f})]\right) \tag{2.56}$$

$$\geq \mathbf{E}_{p(\mathbf{f}|\mathbf{u})}[\log p(\mathbf{y} \mid \mathbf{f})]. \tag{2.57}$$

Recall also the standard variational bound (2.44),

$$\log p(\mathbf{y} \mid \boldsymbol{\theta}) \geq \mathbf{E}_{q(\mathbf{u})}[\log p(\mathbf{y} \mid \mathbf{u},\boldsymbol{\theta})] - \mathrm{KL}(q(\mathbf{u})||p(\mathbf{u} \mid \boldsymbol{\theta})), \tag{2.58}$$

which is now written using the approximate posterior $q(\mathbf{u})$. The evidence lower bound (ELBO) for the sparse variational Gaussian process is constructed by applying the two separate lower bounds to the marginal likelihood.

$$\log p(\mathbf{y} \mid \boldsymbol{\theta}) \geq \mathbf{E}_{q(\mathbf{u})}[\log p(\mathbf{y} \mid \mathbf{u},\boldsymbol{\theta})] - \mathrm{KL}(q(\mathbf{u})||p(\mathbf{u} \mid \boldsymbol{\theta})) \tag{2.59}$$

$$\geq \mathbf{E}_{q(\mathbf{u})}[\mathbf{E}_{p(\mathbf{f}|\mathbf{u},\boldsymbol{\theta})}[\log p(\mathbf{y} \mid \mathbf{f})]] - \mathrm{KL}(q(\mathbf{u})||p(\mathbf{u} \mid \boldsymbol{\theta})) \tag{2.60}$$

The ELBO is then jointly maximized with respect to the variational parameters $\boldsymbol{m}$ and $\boldsymbol{S}$, inducing points $\mathbf{z}$ as well as any hyperparameters $\boldsymbol{\theta}$ belonging to the kernel functions and possible hyperparameters of the likelihood (e.g. the noise variance in case of regression). The covariance $\boldsymbol{S}$ is decomposed into $\boldsymbol{S} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is constrained to a lower triangular matrix whose elements can be freely optimized.

GPflow (Matthews et al., 2017) is a Python package based on Tensorflow (Abadi et al., 2016) implementing various GP methods, including the variational bound (2.60) for several non-conjugate likelihoods and link functions exactly (including Gaussian as well as Bernoulli with a probit link), and falls back to numerical integration for the ones that not tractable. GPyTorch (Gardner et al., 2018) is another software package providing variational inference, exploiting many other techniques to speed-up GPs.

3. Non-Stationary Dependencies in Vector-Valued Functions

This chapter starts by considering Gaussian processes for vector-valued functions in a general case by constructing a kernel that is defined over pairs of the output vector elements and input variables. Many standard formulations of multi-output kernels are stationary. Publication II proposes a flexible non-stationary kernel that considers input-dependent dependencies between the outputs based on an extension of the generalized Wishart process. The chapter concludes with another type of multi-output model that constructs a classifier for single trials in a brain imaging study using magnetoencephalography (MEG) with a mixture of group factor analyzers.

3.1 Multi-Output GPs

Gaussian processes for functions that are vector-valued, i.e. $\mathbf{f}(x) \in \mathbb{R}^M$, are constructed considering pairs (x,p) that index the single elements, or outputs, f_p of the vector function evaluated at x. The kernel then is denoted as

$$k((x,p),(x',p')) = \mathbf{E}[f_p(x) f_{p'}(x')], \tag{3.1}$$

assuming a zero mean function. The simplest way of dealing with the issue of defining a multi-output kernel is modelling the dependencies with a positive-definite matrix $\mathbf{A}$ whose elements $a_{pp'}$ specify a similarity between the outputs p and p' (Bonilla et al., 2007). With this matrix, the kernel is expressed as

$$k((x,p),(x',p')) = a_{pp'} k(x,x'). \tag{3.2}$$

The kernel can also be written as a Kronecker product $\mathbf{K} = \mathbf{A} \otimes \mathbf{K}_{xx}$. Stegle et al. (2011); Rakitsch et al. (2013) presented an efficient inference scheme for kernels expressed as (3.2) by exploiting the Kronecker structure of the kernel matrix $\mathbf{K}$.

More expressive multi-output kernels have been proposed using convolutional processes (Álvarez and Lawrence, 2011; Alvarez et al., 2010; Alvarez and Lawrence, 2008). The semi-parametric latent factor model (Teh and Seeger, 2005) and the Gaussian process regression network (Nguyen and Bonilla, 2013; Wilson et al., 2012) construct multi-output GP by modelling the outputs as sums

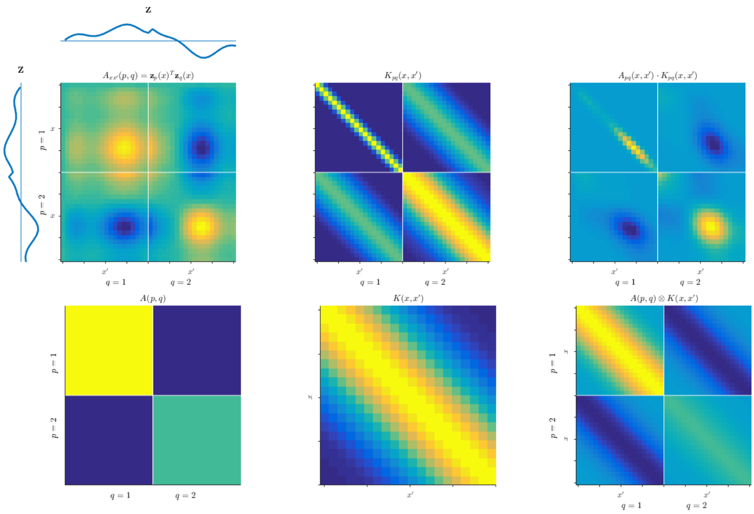

Figure 3.1. Illustration of the Wishart-Gibbs kernel. Top: The input-dependent signal kernel $A_{xx'}(p,p')$ and signal-dependent input kernel $K_{pp'}(x,x')$ are mixed into a rich and flexible pairwise combined kernel. Bottom: A standard Kronecker kernel between signals and inputs can not reproduce dependent correlations.

of latent Gaussian processes. More recently, multi-output kernels based on the spectral mixture kernel (Wilson and Adams, 2013) have also been proposed (Ulrich et al., 2015; Gallagher et al., 2017; Parra and Tobar, 2017; Chen et al., 2018), modelling the time, power and phase correlations and delays between the outputs, while still being in the class of stationary kernels.

3.2 Non-Stationary Hadamard Wishart-Gibbs Kernel

Publication II proposes a multi-output kernel where the dependencies between two elements (p and p') of the vector-valued function $\mathbf{f}(x)$ are also a function of the inputs x and x'. The resulting Gaussian process is written as

$$f_p(x) \sim \mathcal{GP}\left(0, A_{xx'}(p,p')k_{pp'}(x,x')\right), \tag{3.3}$$

such that the joint covariance $\mathbf{cov}[f_p(x), f_{p'}(x')] = A_{xx'}(p,p')k_{pp'}(x,x')$ is a product of signal and input similarities. Both similarities depend on each other to produce a *non-stationary* joint covariance. This is in contrast to the common Kronecker product kernel, where the dependency between the elements is a constant $a_{pp'}$.

The pairwise, mutually dependent Hadamard product kernel

$$k((x,p),(x',p')) = A_{xx'}(p,p')k_{pp'}(x,x') \tag{3.4}$$

encodes a rich similarity between input x of latent signal p and x' of signal p' as the product of the two conditional kernels. The kernel $A_{xx'}(p,q)$ encodes the *signal similarity* at inputs x and x', while the kernel $K_{pp'}(x,x')$ specifies the *input similarity* at latent signals p and p'. Since the two kernels depend on each other, a simple model such as Kronecker product kernel is not suitable. Both kernels can be interpreted as cross-covariances. The full kernel matrix can be written[1] also as

$$\mathbf{K} = \mathbf{A} \circ \mathbf{K}, \tag{3.5}$$

where $\circ$ denotes the Hadamard product.

The construction of the mutually dependent covariance structure $A_{xx'}(p,q)$ was based on a form of Wishart processes. The Wishart distribution is a matrix-variate probability distribution. Specifically a sum of outer products of Gaussian distributed random vectors $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{V})$ follows

$$\mathbf{A} = \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^T \sim \mathcal{W}(\mathbf{V}, n). \tag{3.6}$$

The generalized Wishart process (Wilson and Ghahramani, 2011) defines a prior on *input-dependent* covariance matrices $\mathbf{A}(x)$. Compared to the standard Wishart distribution, the Gaussian random vectors $\mathbf{z}$ are replaced by Gaussian processes $\mathbf{z}(x)$, giving the construction

$$\mathbf{A}(x) = \sum_{i=1}^{n} \mathbf{L}\mathbf{z}_i(x)\mathbf{z}_i(x)^T \mathbf{L}^T \sim \mathcal{GWP}(\mathbf{V}, n, k_z),$$

where the prior mean covariance $\boldsymbol{V} = \mathbf{L}\mathbf{L}^T$ is Cholesky decomposed, and all random vectors $z_{ij}(x) \sim \mathcal{GP}(0, k_z(x,x'))$ are independent Gaussian processes. The kernel k_z determines the smoothness of $\mathbf{A}(x)$ in the input space. The joint kernel $A_{xx'}(p,p')$ is defined such that the GWP marginal for $\mathbf{A}(x)$ is preserved by extending the GWP into *cross-covariances* of two variables as

$$A_{xx'}(p,p') = \mathbf{z}_p(x)^T \mathbf{z}_{p'}(x'), \tag{3.7}$$

where each element of $\mathbf{z}_p(x) \in \mathbb{R}^n$ has a GP prior. With this choice the prior expectation of the covariance is the identity matrix, corresponding to a prior belief that the signals are independent (non-correlated).

Within the input domain x the non-stationary version of the Gaussian kernel (Gibbs, 1997)

$$k_{pp'}(x,x') = \sqrt{\frac{2\ell_p \ell_{p'}}{\ell_p^2 + \ell_{p'}^2}} \exp\left(-\frac{(x-x')^2}{\ell_p^2 + \ell_{p'}^2}\right) \tag{3.8}$$

correctly handles the specific length-scales ℓ_p for each signal. The kernel within a single signal $k_{pp}(x,x')$ reduces into the standard Gaussian kernel, while the cross-covariance similarity $k_{pp'}(x,x')$ measures similarity of two inputs with

[1]Bold-face symbols refer to variables that are matrices or vectors. For example, $\mathbf{A}(x)$ is an $M \times M$ matrix while $A_{pp'}(x,x')$ is scalar value of the kernel function evaluated at (x,p) and (x',p').

different associated length-scales, and retains the positive-definiteness for the similarity of pairs (x,p) and (x',p').

The kernel is applied in Publication II to enrich the latent variable prior distribution in the Gaussian process regression network (Wilson et al., 2012; Nguyen and Bonilla, 2013) as well as in a direct multi-output regression problem.

3.3 Mixture of Group Factor Analyzers for Classification

3.3.1 Group Factor Analysis

Group factor analysis (Virtanen et al., 2012; Klami et al., 2015; Zhao et al., 2016; Leppäaho et al., 2017) can be seen from two distinct perspectives. Perhaps most intuitively the model can be seen as extending factor analysis to model the connections between multiple groups of variables.

Origins of factor analysis are in the field of psychometrics (Spearman, 1904; Pearson, 1901; Thurstone, 1931). There the goal was in discovering underlying traits that would generate the scores of individual tests, for example finding a trait interpreted as intelligence that would affect the exam scores in different school subjects; factor analysis models are, however, generally unidentifiable and hard to interpret. Mathematically the model is expressed as

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\varepsilon} \tag{3.9}$$

where $\mathbf{x} \in \mathbb{R}^D$ are the observed features and $\mathbf{z} \in \mathbb{R}^K$ the latent factor scores that are mapped through the loading matrix $\mathbf{W} \in \mathbb{R}^{D\times K}$. The noise term $\boldsymbol{\varepsilon}$ is assumed to be a Gaussian with diagonal covariance. Thus the factor analysis likelihood is written as

$$p(\mathbf{x} \mid \mathbf{z}, \mathbf{W}, \boldsymbol{\sigma}^2) = \mathcal{N}\left(\mathbf{x} \mid \mathbf{W}\mathbf{z}, \mathrm{diag}(\boldsymbol{\sigma}^2)\right). \tag{3.10}$$

If the diagonal covariance were isotropic, then the model would correspond to the probabilistic (Bayesian) principal component analysis (Tipping and Bishop, 1999b; Bishop, 1999).

The GFA model considers a group of feature vectors $\{\mathbf{x}^{(m)}\}_{i=1}^M$. The groups could correspond to different feature sets computed from distinct sources, for example in encephalography (EEG) or MEG to features from the sensors in different locations over the scalp (see the experiment in Publication I).

Another view of GFA is looking at it as a generalization of canonical correlation analysis (CCA). CCA (Hotelling, 1936; Hardoon et al., 2004) seeks to find linear projections for two groups of variables such that the correlation between the projections is maximized. In a probabilistic interpretation of CCA (Bach and Jordan, 2005), CCA is cast as a latent variable model, similar to PCA or factor analysis. A Bayesian formulation of CCA (Virtanen et al., 2011; Klami et al., 2013) is then exactly a special case of GFA with $M = 2$ groups.

In GFA, each group of variables is modelled as being generated from the same latent variable $\mathbf{z}$ but with their own loading matrix $\mathbf{W}^{(m)}$, giving rise to the isotropic Gaussian likelihood

$$p(\mathbf{x}^{(m)} \mid \mathbf{W}^{(m)}, \mathbf{z}, \sigma_m^2) = \mathcal{N}\left(\mathbf{x}^{(m)} \mid \mathbf{W}^{(m)}\mathbf{z}, \sigma_m^2 \mathbf{I}\right). \tag{3.11}$$

The columns of the loading matrix $\mathbf{W}^{(m)}$ determine which of the K latent factors are relevant in modelling the group m. The columns are given a sparsifying automatic relevance determination (ARD) prior

$$p(\mathbf{w}_k^{(m)} \mid \alpha_k^{(m)}) = \mathcal{N}(\mathbf{w}_k^{(m)} \mid \mathbf{0}, \alpha_k^{(m)}\mathbf{I}), \tag{3.12}$$

where the value of the variance $\alpha_k^{(m)}$ determines whether the factor k will be relevant in explaining the variation in $\mathbf{x}^{(m)}$, either by zeroing out the corresponding column $\mathbf{w}_k^{(m)}$ or letting it be free.

3.3.2 Mixture of GFA's

Publication I presents a model based on a mixture of group factor analyzers for classification of weak multi-view signals. *Weak* in this context refers to that most variance within the signals is independent of the class but weaker class-specific factors exist that explain the variation between the classes. The model incorporates factors that are present in all mixture components as well as factors that are present in one mixture component. The model further assumes that each observation is also assigned to a class. In essence, the model provides a generative process for both the real-valued data $\mathbf{x}$ as well as the related class label that is denoted by r. The mixture model allows the modelling of different modes present in the data, by allowing the factor model to be different depending on the value of $\mathbf{x}$, demonstrating a form of non-stationarity as well as freeing the linearity assumption of the base GFA model. Mixture models of probabilistic PCA and other related models have been also proposed by other authors (Tipping and Bishop, 1999a; Archambeau et al., 2008; Ghahramani and Beal, 2000).

Mixture model. The likelihood for a group of features $\mathbf{x}^{(m)}$ originating from a mixture component, or cluster, c is denoted by $p(\mathbf{x}^{(m)} \mid c)$. Assuming that all factors are independent, the likelihood can therefore be written using cluster-specific factor loadings $\mathbf{W}_c$ as well as shared factors $\boldsymbol{V}$ with their factor scores $\mathbf{z}$ and $\boldsymbol{h}$, respectively, as

$$p(\mathbf{x}^{(m)} \mid c) = \mathcal{N}(\mathbf{x}^{(m)} \mid \mathbf{W}_c^{(m)}\mathbf{z} + \boldsymbol{V}^{(m)}\mathbf{h}, \sigma^2\mathbf{I}). \tag{3.13}$$

The cluster-specific factor loadings have a similar prior to GFA, with sparsity inducing prior per variable group m,

$$p([\mathbf{W}_c^{(m)}]_k \mid \alpha_{c,k}^{(m)}) = \mathcal{N}([\mathbf{W}_c^{(m)}]_k \mid \mathbf{0}, \alpha_{c,k}^{(m)}\mathbf{I}), \tag{3.14}$$

while the shared factors have loadings that are active for all variable groups,

$$p(\boldsymbol{V}_k^{(m)} \mid \beta_k) = \mathcal{N}\left(\boldsymbol{V}_k^{(m)} \mid \mathbf{0}, \beta_k\mathbf{I}\right). \tag{3.15}$$

Furthermore, the cluster assignments c are given a standard prior

$$p(c \mid \boldsymbol{\pi}) = \text{Multinomial}(c \mid \boldsymbol{\pi}) \tag{3.16}$$

with a conjugate Dirichlet prior for the cluster prior probabilities $\boldsymbol{\pi}$.

Classification. The GFA mixture model further extends to include a part that ties the mixture components also to the class labels $r \in \{0,1\}$. The likelihood for the class r is then a Bernoulli distribution

$$p(r \mid c, \boldsymbol{\gamma}) = \text{Bernoulli}(r \mid \gamma_c) \tag{3.17}$$

where the variable $\boldsymbol{\gamma}$ denotes the class distribution related to each of the mixture components, denoting the prior probability $p(r=1 \mid c)$. It is given a non-informative conjugate prior

$$p(\gamma_c) = \text{Beta}\left(\gamma_c \mid \frac{1}{2}, \frac{1}{2}\right). \tag{3.18}$$

Publication I applies the GFA mixture classifier in a brain imaging study using MEG. Each MEG data channel is considered to be its own variable group in the GFA models and the classifier is applied to discriminate between a listening and a speaking condition.

4. Discussion and Conclusions

In summary, the publications that are a part of this thesis contribute several novel approaches to modelling non-stationary data. Publication I presents a novel mixture model of group factor analyzers which is combined with a classification model based on the mixture component assignments, meaning that a single class can be characterized by multiple modes of behaviour. Publication II constructs a novel multi-output kernel for Gaussian processes, which allows for a new kind of analysis of vector-valued functions by modelling input-dependent couplings of the outputs. Publication III and Publication IV propose two alternative constructions of non-stationary spectral mixture kernels that can represent, in addition to non-stationarity, many desirable properties, such as periodicity, or non-monotonicity in more general way. This is in contrast to standard kernels that focus on modelling only one such feature.

As a result of this thesis, there now exists evidence that including non-stationarity in spectral mixture kernels can make the performance of these already very flexible kernels even better. The proposed kernels are well applicable in various problem domains, making the potential impact of the developed methods high, as many real-world processes exhibit non-stationarity. Publication II also shows that non-stationarity can be important in modelling vector-valued functions.

The methods that were introduced in making the spectral mixture kernel non-stationary would be interesting to apply also to other approaches in constructing stationary kernels function such as the Lévy process spectral kernel (Jang et al., 2017). Furthermore, modelling non-stationarity in the multi-output spectral mixtures (Parra and Tobar, 2017; Chen et al., 2018) could open new possibilities for multi-output Gaussian processes, by being more flexible and scalable than the non-stationary vector-valued kernel developed in Publication II. Overall, combining deep neural networks with Gaussian processes in different ways is a very promising line of research, including Publication IV as well as work such as deep kernel learning (Wilson et al., 2016a,b) that constructs kernels by transforming input variables through a deep network, also extending to a semi-supervised setup (Jean et al., 2018).

Interpretability of machine learning models is increasingly a relevant topic.

Typical neural network methods are often deemed as black boxes or as non-interpretable[1], and do not provide probabilistic predictions which would make it possible to assess the confidence of the model. Gaussian processes as Bayesian models do provide the confidence, and interpretation can be done via the kernel parameters (which often have a natural meaning, describing e.g. the period or smoothness of the function) that are learned to fit to the data, as opposed to neural network weights that in themselves are not easily interpretable. In spectral kernels, the parameters are interpretable as describing the spectral composition of the modelled functions or signals, a property which is important in many applications. This point of view has not been greatly explored thus far in the publications in this thesis or significantly in other literature on stationary spectral mixtures.

The Wishart-Gibbs kernel of Publication II, as presented there, is not scalable to situations where the product of the number of outputs M and the number of samples N is $NM \gg 1000$, as in that regime the size of the kernel matrix becomes too large to fit in memory and operations such as inversions are not feasible, making the method impractical. Thus, it would be interesting study the method in the sparse Gaussian process framework (Titsias, 2009; Hensman et al., 2015) or finding other avenues including e.g. the grid-structured eigenfunctions (Evans and Nair, 2018), where the complexity of the method could be handled. Another interesting approach would be to apply the neural network parameterization done in Publication IV to the generalized Wishart process (Wilson and Ghahramani, 2011), in order to replace the latent Gaussian processes.

In the work done in this thesis, it remains unclear whether the spectral mixture kernels can be applied for datasets whose dimensionality is $D \gg 10$. The highest dimension within the datasets used in Publication III and Publication IV was limited to $D = 9$ in the Protein dataset. It also remains a question how the methods compare with truly bigger datasets, in the regime of e.g. $N \gg 1$ million, as the datasets explored in this thesis are limited only to hundreds of thousands in Publication III.

The extrapolation performance of the non-stationary kernels (as well as their stationary versions) remains also a question mark, as this has not been studied to a significant extent in the experiments performed in this thesis, except for the initial steps taken for the land surface temperature forecasting in Publication III. The Neural-GSM variant would likely perform better than the GP-parameterized GSM, as the neural parameterization can potentially learn a more global support for the parameter functions than GP-GSM that uses RBF kernels for the GP's. Extrapolation performance in general is difficult to objectively measure and definitely deserves more research.

[1]Frameworks such as SHAP (Lundberg and Lee, 2017) do exist for generating explanations of model predictions even in a model-agnostic way, however.

References

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

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

M. A. Álvarez and N. D. Lawrence. Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research*, 12:1459–1500, 2011.

M. A. Alvarez, D. Luengo, M. K. Titsias, and N. D. Lawrence. Efficient multioutput Gaussian processes through variational inducing kernels. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9, pages 25–32, 2010.

S. Aminikhanghahi and D. J. Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

C. Archambeau, N. Delannay, and M. Verleysen. Mixtures of robust probabilistic principal component analyzers. *Neurocomputing*, 71(7-9):1274–1282, 2008.

F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, 2005.

C. M. Bishop. Bayesian PCA. In *Advances in Neural Information Processing Systems*, pages 382–388, 1999.

E. Bonilla, K. Chai, and C. Williams. Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems 20*, pages 153–160, 2007.

K. Chen, P. Groot, J. Chen, and E. Marchiori. Generalized spectral mixture kernels for multi-task Gaussian processes. *arXiv preprint arXiv:1808.01132*, 2018.

K. Cutajar, E. V. Bonilla, P. Michiardi, and M. Filippone. Random feature expansions for deep Gaussian processes. In *Proceedings of the 34th International Conference on Machine Learning*, pages 884–893. JMLR. org, 2017.

A. Damianou and N. Lawrence. Deep Gaussian processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 207–215, 2013.

D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1166–1174, 2013.

T. Evans and P. Nair. Scalable Gaussian processes with grid-structured eigenfunctions (GP-GRIEF). In *Proceedings of the 35th International Conference on Machine Learning*, pages 1417–1426, 2018.

S. Flaxman, A. G. Wilson, D. Neill, H. Nickisch, and A. Smola. Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. In *Proceedings of the 32th International Conference on Machine Learning*, pages 607–616, 2015.

J. H. Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991.

N. Gallagher, K. R. Ulrich, A. Talbot, K. Dzirasa, L. Carin, and D. E. Carlson. Cross-spectral factor analysis. In *Advances in Neural Information Processing Systems 30*, pages 520–534, 2017.

J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Advances in Neural Information Processing Systems 31*, 2018.

M. Genton. Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2001.

Z. Ghahramani and M. J. Beal. Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems 13*, pages 449–455, 2000.

M. Gibbs. *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, University of Cambridge, 1997.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, 2016.

D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

M. Heinonen, H. Mannerström, J. Rousu, S. Kaski, and H. Lähdesmäki. Non-stationary Gaussian process regression with Hamiltonian Monte Carlo. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 732–740, 2016.

J. Hensman, A. Matthews, and Z. Ghahramani. Scalable Variational Gaussian Process Classification. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38, pages 351–360, 2015.

W. Herlands, A. Wilson, H. Nickisch, S. Flaxman, D. Neill, W. Van Panhuis, and E. Xing. Scalable Gaussian processes for characterizing multidimensional change surfaces. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1013–1021, 2016.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

N. Huang. A review on Hilbert-Huang transform: Method and its applications to geophysical studies. *Reviews of Geophysics*, 46, 2008.

N. Huang, S. Zheng, S. Long, M. Wu, H. Shih, Q. Zheng, N.-Q. Yen, C. Tung, and H. Liu. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *In Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 454:903–995, 1998.

P. A. Jang, A. Loeb, M. Davidow, and A. G. Wilson. Scalable Lévy process priors for spectral kernel learning. In *Advances in Neural Information Processing Systems 30*, pages 3940–3949, 2017.

N. Jean, S. M. Xie, and S. Ermon. Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance. In *Advances in Neural Information Processing Systems 31*, 2018.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems 30*, pages 971–980, 2017.

A. Klami, S. Virtanen, and S. Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(Apr):965–1003, 2013.

A. Klami, S. Virtanen, E. Leppäaho, and S. Kaski. Group factor analysis. *IEEE transactions on neural networks and learning systems*, 26(9):2136–2147, 2015.

J. Knoblauch and T. Damoulas. Spatio-temporal Bayesian on-line changepoint detection with model selection. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2718–2727, 2018.

Y.-L. Kom Samo and S. Roberts. Generalized spectral kernels. Technical report, University of Oxford, 2015. arXiv:1506.02236.

N. D. Lawrence, M. Seeger, and R. Herbrich. Fast sparse Gaussian process methods: The informative vector machine. In *Advances in Neural Information Processing Systems 16*, pages 625–632, 2003.

M. Lazaro-Gredilla and J. Quiñonero Candela. Sparse spectral sampling gaussian processes. Technical Report MSR-TR-2007-152, November 2007. URL `https://www.microsoft.com/en-us/research/publication/sparse-spectral-sampling-gaussian-processes/`.

M. Lázaro-Gredilla, J. Quiñonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11:1865–1881, 2010.

E. Leppäaho, M. Ammad-ud din, and S. Kaski. GFA: exploratory analysis of multiple data sources with group factor analysis. *Journal of Machine Learning Research*, 18 (1):1294–1298, 2017.

J. R. Lloyd, D. K. Duvenaud, R. B. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Automatic construction and natural-language description of nonparametric regression models. In *AAAI*, pages 1242–1250, 2014.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774, 2017.

A. Majumdar, A. E. Gelfand, and S. Banerjee. Spatio-temporal change-point modeling. *Journal of Statistical Planning and Inference*, 130(1-2):149–166, 2005.

A. G. d. G. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, and J. Hensman. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18(40):1–6, 2017.

R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

T. Nguyen and E. Bonilla. Efficient variational inference for Gaussian process regression networks. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, pages 472–480, 2013.

H. Ni and H. Yin. Self-organising mixture autoregressive model for non-stationary time series modelling. *International journal of neural systems*, 18(06):469–480, 2008.

C. Paciorek and M. Schervish. Nonstationary covariance functions for Gaussian process regression. In *Advances in Neural Information Processing Systems 17*, 2004.

C. Paciorek and M. Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506, 2006.

G. Parra and F. Tobar. Spectral mixture kernels for multi-output Gaussian processes. In *Advances in Neural Information Processing Systems 30*, pages 6681–6690, 2017.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 21*, pages 1177–1184, 2008.

B. Rakitsch, C. Lippert, K. Borgwardt, and O. Stegle. It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals. In *Advances in Neural Information Processing Systems 26*, pages 1466–1474, 2013.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 814–822, 2014.

C. E. Rasmussen and C. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning*, pages 1278–1286, 2014.

Y. Saatçi. *Scalable Inference for Structured Gaussian Process Models*. PhD thesis, University of Cambridge, 2011.

Y. Saatçi, R. D. Turner, and C. E. Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning*, pages 927–934, 2010.

H. Salimbeni and M. Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems 30*, pages 4588–4599, 2017.

P. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87, 1992.

M. Seeger, C. Williams, and N. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In *Proceedings of the 9th International Conference on Artificial Intelligence and Statistics*, 2003.

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

A. Sinha and J. Duchi. Learning kernels with random features. In *Advances in Neural Information Processing Systems 29*, pages 1298–1306, 2016.

A. J. Smola and P. L. Bartlett. Sparse greedy gaussian process regression. In *Advances in Neural Information Processing Systems 14*, pages 619–625, 2001.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 19*, pages 1257–1264, 2006.

J. Snoek, K. Swersky, R. Zemel, and R. Adams. Input warping for Bayesian optimization of non-stationary functions. In *Proceedings of the 31th International Conference on Machine Learning*, pages 1674–1682, 2014.

C. Spearman. "general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

O. Stegle, C. Lippert, J. Mooij, N. D. Lawrence, and K. Borgwardt. Efficient inference in matrix-variate Gaussian models with iid observation noise. In *Advances in Neural Information Processing Systems 24*, pages 630–638, 2011.

Y. Teh and M. Seeger. Semiparametric latent factor models. In *Proceedings of the 10th International Conference on Artificial Intelligence and Statistics*, pages 333–340, 2005.

L. L. Thurstone. Multiple factor analysis. *Psychological Review*, 38(5):406, 1931.

M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999a.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999b.

M. Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pages 567–574, 2009.

V. Tolvanen, P. Jylänki, and A. Vehtari. Expectation propagation for nonstationary heteroscedastic Gaussian process regression. In *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*, pages 1–6. IEEE, 2014.

J.-F. Ton, S. Flaxman, D. Sejdinovic, and S. Bhatt. Spatial mapping with Gaussian processes and nonstationary Fourier features. *Spatial Statistics*, 2018.

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.

S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In *Proceedings of the 28th International Conference on Machine Learning*, pages 457–464, 2011.

S. Virtanen, A. Klami, S. Khan, and S. Kaski. Bayesian group factor analysis. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pages 1269–1277, 2012.

M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

A. Wilson and H. Nickisch. Kernel interpolation for scalable structured gaussian processes (KISS-GP). In *Proceedings of the 32th International Conference on Machine Learning*, pages 1775–1784, 2015.

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

A. G. Wilson and R. Adams. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1067–1075, 2013.

A. G. Wilson and Z. Ghahramani. Generalised Wishart processes. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 736–744, 2011.

A. G. Wilson, D. Knowles, and Z. Ghahramani. Gaussian process regression networks. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1139–1146, 2012.

A. G. Wilson, E. Gilboa, A. Nehorai, and J. P. Cunningham. Fast kernel learning for multidimensional pattern extrapolation. In *Advances in Neural Information Processing Systems 27*, pages 3626–3634, 2014.

A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 370–378, 2016a.

A. G. Wilson, Z. Hu, R. R. Salakhutdinov, and E. P. Xing. Stochastic variational deep kernel learning. In *Advances in Neural Information Processing Systems 29*, pages 2586–2594, 2016b.

Z. Yang, A. Smola, L. Song, and A. Wilson. A la carte: Learning fast kernels. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 1098–1106, 2015.

S. Zhao, C. Gao, S. Mukherjee, and B. E. Engelhardt. Bayesian group factor analysis with structured sparsity. *Journal of Machine Learning Research*, 17(1):6868–6914, 2016.

Aalto-DD 152/2019

9 789526 086880

ISBN 978-952-60-8688-0 (printed)
ISBN 978-952-60-8689-7 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS