Department of Mathematics and Statistics

Approximate Bayesian inference in multivariate Gaussian process regression and applications to species distribution models

Marcelo Hartmann

Academic dissertation

To be presented for public examination with the permission of the Faculty of Science of the university of Helsinki in auditorium PIII, Porthania, City Centre Campus, on 20th of March 2019 at 12 o'clock

University of Helsinki
Finland

Supervisor
Jarno Vanhatalo, University of Helsinki, Finland

Pre-examiners
Theo Damoulas, University of Warwick, United Kingdom
Antti Penttinen, University of Jyväskylä, Finland

Opponent
Mark Girolami, Imperial College London & The Alan Turing institute, United Kingdom

Custos
Samuli Siltanen, University of Helsinki, Finland

Contact information

Department of Mathematics and Statistics
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Email address: domast-info@helsinki.fi
URL: http://mathstat.helsinki.fi/
Telephone: +358 02941 51506

ISBN 978-951-51-4974-9 (paperback)
ISBN 978-951-51-4975-6 (PDF)

Helsinki 2018
Unigrafia Oy

Approximate Bayesian inference in multivariate Gaussian process regression and applications to species distribution models

Marcelo Hartmann

Department of Mathematics and Statistics
P.O. Box 68, FI-00014 University of Helsinki, Finland
marcelo.hartmann@helsinki.fi

PhD Thesis,
Helsinki, February 2019, 56 pages
ISBN 978-951-51-4974-9 (paperback)
ISBN 978-951-51-4975-6 (PDF)

Abstract

Gaussian processes are certainly not a new tool in the field of science. However, alongside the quick increasing of computer power during the last decades, Gaussian processes have proved to be a successful and flexible statistical tool for data analysis. Its practical interpretation as a nonparametric procedure to represent prior beliefs about the underlying data generating mechanism has gained attention among a variety of research fields ranging from ecology, inverse problems and deep learning in artificial intelligence.

The core of this thesis deals with multivariate Gaussian process model as an alternative method to classical methods of regression analysis in Statistics. I develop hierarchical models, where the vector of predictor functions (in the sense of generalized linear models) is assumed to follow a multivariate Gaussian process. Statistical inference over the vector of predictor functions is approached by means of the Bayesian paradigm with analytical approximations.

I developed also new parametrisations for the statistical models in order to improve the performance of the computations related to the inferential task. The methods developed in this thesis are also tightly connected to practical applications. The main applications considered involve multiple species surveys and species distribution modelling in quantitative ecology. This is a field of research which provides a rich variety of applications where statistical methods can be put at test.

Acknowledgements

During almost four years living abroad and among all the experiences I have been through, my family was constantly in my thoughts. I believe that it is mainly due to them and their way of perceiving things that I am able to finalize my PhD studies. This thesis is dedicated to them. To my parents, Ana Hartmann and Luiz Roberto Hartmann for teaching me to have integrity in my choices. To my brother Luiz Hartmann, for dedicate a very tiny amount of his time to talk about Mathematics. I owe you a lot. To my sister Hannah Hartmann for inspiring me. For all the times we spent together whether we were laughing or you were disturbing me. I love you all. Always.

To Carla, Markus, Julia, Henrique, Paulo, Ana Teresa and my grandmother Lydia Sanchez Hartmann (*in memoriam*) who also are a very important part of my life.

To my supervisor Jarno Vanhatalo for giving me the chance to pursue a PhD in Finland and the effort put throughout the years. There was never once I left your office with empty hands.

Without the help of Geoffrey Hosack, Richard Hillary and Lari Veneranta, this thesis would have never reached the shape it is today. I have learned a lot about science working with you.

It is a privilege to have Mark Girolami, Theo Damoulas and Antti Penttinen acting as opponent and pre-examiners respectively. Words could never fulfill how grateful I am. A huge thanks to Samuli Siltanen for acting as custos.

I am also very delighted to have spent three weeks in the Alan Turing institute and in the university of Warwick. During those times I had the opportunity to meet amazing people who would give me strength to keep doing. Thanks to Louis Ellam, Virginia Aglietti, Oliver Hamelijnck, Joe Meagher and Jeremias Knoblauch whom spent a bit of their time showing the city of London and for small chats about their research.

I could not forget to mention my peruvians friends, Juan Pablo Bustamante from whom I have learned valuable lessons in life and Susan Anyosa for being the smartest girl I have ever met.

For the special moments Maiju Männikkö has given me. Whenever we were

able to play together, my thoughts about work would go away, those are particular days I will never forget. Dear Maiju, thank you. To Emmi Vaurola for being constantly happy and transmit all her energy to me. To my friend Dana Helleman for the parties we have been together and the moments shared. To Anna Mötönnen and Oleksandr Zakirov for being great friends from time to time.

For the best times I have had with my brazilian friends Vitória Pacela and Thiago Brito. Thank you for the lunch times and discussions we had, specially for Thiago who had a huge patience to hear me talking about astronomy. Vitória, I will always miss the way you laugh, I was overwhelmed with joy whenever you would do it.

To Ville Tanskanen for spending some of his time with me talking about Gaussian processes. I have this feeling that I was very luck to meet a person like you in my lifetime. Also, thanks for the apple juice supply provided by your parents. I have not forgotten my promise.

To Anna Pivovarova for letting me spend time with her daughter Fenia. I enjoyed each moment I had with her.

My previous supervisors are an important part of this work. To Fernando Moala, who introduced me to Gaussian processes in the Statistics and machine-learning fields and Ricardo Ehlers for giving me total freedom in my master studies.

To my almost forgotten friends back in my hometown, Aroldo Costa, Robson Gimenez, Juliana Cocolo, Marco Pollo and Teodoro Calvo. I have remembered all of you almost every day during these years. I miss you a lot.

Many thanks to Helton Graziadei who was the first one to read the drafts of this thesis. I am in debt with your valuable comments and the mathematical precision of your annotations.

To Arto Klami, Krista Longi, Joseph Sakaya, Aditya Jitta and Tomasz Kusmierczyk for making me feel very welcome before I even start the post-doctoral studies. Thank you all.

A necessary part to complete this thesis is due to the Academy of Finland grants and the research funds of University of Helsinki for which I am sincerely thankful. Special thanks to the staff of the department of Mathematics and Statistics, I had the chance to stay in the best office I could ever have.

Marcelo Hartmann
Helsinki, January 2019

This thesis consists of an overview of the following papers

List of Publications

[I] Marcelo Hartmann, Geoffrey R. Hosack, Richard M. Hillary & Jarno Vanhatalo. (2017). Gaussian processs framework for temporal dependence and discrepancy functions in Ricker-type population growth models. *Annals of Applied Statistics*, 11(3):1375-1402.

[II] Marcelo Hartmann & Jarno Vanhatalo. (2018). Laplace approximation and the natural gradient for Gaussian process regression with heteroscedastic Student-t model. *Statistics and Computing.*

[III] Jarno Vanhatalo, Marcelo Hartmann & Lari Veneranta. Additive multivariate Gaussian process for joint species distribution modelling with heterogeneous data. *Under revision in Bayesian Analysis.*

[IV] Marcelo Hartmann & Jarno Vanhatalo. A new hierarchical Bayesian method for multivariate binary outcomes with Gaussian process priors. *Under revision in Journal of Classification.*

[V] Marcelo Hartmann & Ricardo Ehlers. Bayesian inference for generalized extreme value distributions via Hamiltonian Monte Carlo (2017). *Communications in Statistics - Simulation and Computation 46(7):5285-5302.*

Author's Contributions to the Publications

[I] The research topic and questions were originally conceived by Vanhatalo, Hosack and Hillary. Hartmann contributed significantly to developing the idea further and to theoretical specifications of the models. He also had the main role in writing all the code and running the experiments. The paper was jointly written by all authors.

[II] Hartmann had the main role in all aspects of the work. Vanhatalo contributed to the background consideration, planning the experiments and writing.

[III] The research topic is due to Vanhatalo and the specific research questions were jointly develop by Hartmann and Vanhatalo. Specific model formulation, theoretical derivations, analytic results and inference scheme are due to Hartmann. Vanhatalo conceived the idea to reduce the computational cost via sequential techniques and improved the article by writing and linking the approach with other existing methodologies and with expertise in the study case. Implementation of the computer code is due to Hartmann and running the experiments was done by Hartmann with assistance of Vanhatalo. Veneranta contributed to the analysis by writing case study results

[IV] Hartmann had the main role in all aspects of the work. Vanhatalo contributed to the background consideration, planning the experiments and writing.

[V] Hartmann had main responsibility in the theoretical derivations of the model formulation, computational implementation and running the experiments. Ehlers contributed to background consideration, planning the experiments and writing.

Contents

Chapter 1

Introduction

In this thesis I present a collection of multivariate regression models within the context of the Bayesian approach to statistical inference. Throughout the work, the common methodological goal aims to combine the hierarchical modelling framework with multivariate Gaussian process models (O'Hagan, 1978; Mardia and Goodall, 1993; Rasmussen and Williams, 2006) and to use this as a surrogate to classical linear (nonlinear) methods of regression analysis and generalized linear models (GLM) (Nelder and Wedderburn, 1972; Wild and Seber, 1989; Seber and Lee, 2012). The methodology proposed here involves applications of multivariate Gaussian process regression (MGP) mainly in species distribution models (SDM). However, as this thesis unfolds, those applications will comprise other topics such as population dynamics, quantitative ecology and robust statistical modelling with MGPs. Let us now begin introducing some background information to this work.

In statistical theory, the process of making inferences about some unknown attribute of interest on the basis of measurements (data) is known as statistical inference (Casella and Berger, 2002). The link between what is measured and the attributes is given through a probabilistic model, which encodes the data generating mechanism under the presence of randomness (uncertainty), given that the attributes were supposedly known[1]. In practice, the randomness described via the probabilistic model can appear due to several reasons. For instance, due to the lack of precision in the measurement instrument, or it may reflect the lack of knowledge about the correct values of the attributes. The term attributes comprises broad scenarios. Over very simple cases, attributes may play the role of a unidimensional parameter that represents, for example, the proportion of votes of some candidate in politics. In quantitative ecology,

[1]For real-world phenomena, when trying to make rigorous scientific statements, one may assume our limited knowledge about every aspect of nature. Hence, it appears rather natural to assume that some attributes are never exactly known.

they can represent environmental conditions which filter presence/abundance of species. In medical application, such as X-ray tomography, the inference problem addresses the reconstruction of an object with the goal of seeing its interior without the need of opening it (Kaipio and Somersalo, 2005; Hauptmann, 2017).

In mathematical terms, the probabilistic model for the measurements is usually defined as a function

$$\begin{aligned} &\pi(\cdot|\eta_1,\ldots,\eta_p) : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}_+ \text{ satisfying} \\ &\int_\Omega \pi(y_1,\ldots,y_n|\eta_1,\ldots,\eta_p)\mathrm{d}y_1\ldots\mathrm{d}y_n = 1, \end{aligned} \tag{1.1}$$

wherein for a particular value $\mathbf{y} = (y_1,\ldots,y_n)$ (the measurements), the formula $\pi(\mathbf{y}\,|\eta_1,\ldots,\eta_p)$ expresses its likelihood for any arbitraries values of the attributes $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_p) \in \Xi \subseteq \mathbb{R}^p$. Henceforth we will refer to the attributes as parameters of a probabilistic model[2]. Ξ is referred to as parameter space. Notice that, in theory, these parameters are fixed but they are the commonly unknown for us in real-case scenarios.

Now, given a particular noisy dataset $\mathbf{y}$ and all relevant information about the parameters which possibly comes from other means, we would like to choose the parameters of the probabilistic model as close as possible to its true values based on all the information we have. More important, we seek to quantify the degree of uncertainty attached to any particular decision we make related to parameter values.

From the frequentist point of view (Bain and Engelhardt, 1992; Knight, 1999; Casella and Berger, 2002), statistical inference is performed by choosing a value for $\boldsymbol{\eta}$ based only on the *statistic* (a function of the data only). The task is to find an *estimators* (a function of the statistic) such that it carries good statistical properties. That is to say, unbiasedness, consistence, minimum-variance, etc (Casella and Berger, 2002; Pawitan, 2005). The degree of uncertainty attached to a estimator can be translated into its variance or in its density function since the estimator itself is a function of random variables. The *maximum likelihood estimator* is one example of estimator and it has been widely used over all fields of science. In addition, observe that, in this case no external information about the parameters can be introduced into the inference problem, even if such information was paramount.

Clearly, for numerous realistic situations, there is often relevant external information about the parameters. Put aside prior information of the parameters

[2]In modern probability theory, equation (1.1) may carry a subscript. For example, we could write it as $\pi_Y(y_1,\ldots,y_n|\eta_1,\ldots,\eta_p)$. This is done to underline that the function $\pi_Y(\cdot|\eta_1,\ldots,\eta_p) : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}_+$ is the Radon-Nikodym derivative w.r.t the Lebesgue measure of the probability measure induced by the vector of random variables $Y = (Y_1,\ldots Y_n)$.

in inference purposes might be unwise and lead to misleading results. In this sense, there is a need of a consistent and rigorous probabilistic method for incorporating both sources of information into our inference task (Jaynes, 2003). The Bayesian approach to the inference problem allows us to include prior information via the *prior distribution*, which is a probability model encoding the degrees of uncertainty about the possible values of the parameters. The term Bayesian inference is used whenever the central mechanism of the inference process is given by the means of the Bayes' Theorem (O'Hagan, 2004; Schervish, 2011), which states that

$$\pi_{\text{post}}(\boldsymbol{\eta} \,|\, \mathbf{y}) = \frac{\pi(\mathbf{y} \,|\, \boldsymbol{\eta})\pi_{\text{prior}}(\boldsymbol{\eta})}{\pi_{\text{M}}(\mathbf{y})} \tag{1.2}$$

where $\pi_{\text{prior}}(\boldsymbol{\eta})$ is the prior distribution for the parameters. $\pi_{\text{M}}(\mathbf{y})$ is the marginal likelihood[3] and $\pi_{\text{post}}(\boldsymbol{\eta} \,|\, \mathbf{y})$ is the so-called *posterior distribution* of the parameters for given values of the measurements. The procedure presented through the formula (1.2) represents our updated state of knowledge about the possible value of the parameters.

According to the Bayesian philosophy, the prior distribution must be designed independently from data arising from the experimental set-up; it must reflect our beliefs disregarding the measurements obtained. In practice, the prior information is often of qualitative nature. The parameters of a probabilistic model may possess physical/biological interpretation, and they provide a particular functional form of the probabilistic model for the data in which the analyst would expect the data to be distributed if the parameters were known. The difficult task on the formulation of the prior distribution lies on how to coherently translate the prior information into a quantitative form expressed through the prior distribution (Gosling, 2005; Oakley and O'Hagan, 2007; Akbarov, 2009; Moala and O'Hagan, 2010).

A fundamental difference between the frequentist (classic) and the Bayesian approaches is in the interpretation of probability. From the frequentist point of view, probability is defined as the relative frequency in the limit of infinite number of trials. In practice, this will require large sample sizes to achieve good inferences. Whereas, in the Bayesian approach, the notion of probability has a subjective status. Probability measures degrees of beliefs (De Finetti, 1975; Bernardo and Smith, 1994) and there is no need for a large sample size to perform inference.

Despite the simple appearance of the Bayes' formula, the range of its applications is vast and has been under increasing complexity. Nowadays, regression models are widely used tools in statistical methodology. When the values of the parameters of the probabilistic model (1.1) are expected to vary as a function of

[3] Also referred as normalizing constant or prior predictive distribution.

covariates[4], some functional form is chosen to represent this systematic variation, which in turn depend on unknown parameters which specify the functional form of the regression. Earliest examples of regression methods were present by Carl Friedrich Gauss around 1807. At that time, he proposed the least-squares method which used redundant data to predict the movements of celestial bodies in the celestial vault (Gauss, 1807). In this case, space-time coordinates were taken as covariates (Wild and Seber, 1989; Seber and Lee, 2012). Later on, the least-square method was generalized by Rudolf E. Kalman known as the *Kalman filter*[5] (Kalman, 1960). We refer the paper by Sorenson (1970) for more details.

More recently, the assumption of a fixed functional form in regression models has been relaxed and in the Bayesian approach one can assume the exact functional form of the regression to be unknown and estimate it from the sample data. This is known as *nonparametric Bayesian inference* and there are many ways to approach this problem. See for example the works by Fergusson (1973) to estimate an unknown distribution function, or O'Hagan (1978), Wahba (1990), Neal (1995, 1998), Williams and Barber (1998) and Rasmussen and Williams (2006) to estimate an unknown regression function. For this thesis, we focus on Gaussian process models (GPs) to estimate unknown regression functions (O'Hagan, 1978; Rasmussen and Williams, 2006) and its multivariate extension based on the linear model of coregionalization (LMC) (Mardia and Goodall, 1993; Gelfand et al., 2003). GPs have been increasingly gaining attention as an attractive nonparametric procedure to represent prior beliefs about unknown functions. This is widely recognized due to its flexibility in the sense that linear operations such as, integration, differentiation, linear-filtering and summation with another GP, results also in a GP (Abrahamsen, 1997; Moala, 2006; Oakley and O'Hagan, 2007; Moala and O'Hagan, 2010; Riihimäki and Vehtari, 2010; Wang and Barber, 2014). This way, GPs provide rich methodology to perform statistical inference over functions for large variety of real-case scenarios.

Multivariate Gaussian process models (MGP) are the natural extension of univariate GPs. In multivariate settings, we then consider a vector of regression functions whose components are treated as unknown functions with the addition of a possible dependency among them (Gelfand et al., 2003). With this dependency between each component of the vector of function values, we expect that statistical inference over the regression functions is improved as well as the

[4]In statistical literature, covariates are variables that typically are not expected to possess random variation. They can be seen as variables of a function. They are also known as explanatory variables or inputs in machine learning.

[5]In certain cases the method can be seen as a recursive least-squares.

predictive power of the modelling approach[6] (Boyle and Frean, 2004; Teh et al., 2005; Bonilla et al., 2008; Fricker et al., 2013; Vandenberg-Rodes and Shahbaba, 2015). The main challenges with MGPs are in their coherent specification and the computational complexity that unfolds in practical applications (Vanhatalo, 2010).

From a practical point of view, the recent advance of computer power has enabled users to operate complex model with less difficulty in many areas of science. Particularly, the field of environmental sciences and Ecology has provided us with a rich data source in which statistical models can be put in practice. Recently, in Ecology, GPs have been applied in species distribution modelling to tackle one single species and it has shown improved performance compared to classical GLM models (Vanhatalo et al., 2012; Golding and Purse, 2016). However, when databases comprise multiple sources of information from various species surveys, MGPs become particularly interesting as an alternative method to analyse multivariate data. In this thesis, I apply MGPs in joint species distribution modelling, quantitative ecology and robust statistical modelling, and show that MGPs further improves data analysis compared to univariate GPs (papers [**I**], [**II**], [**IV**], [**III**]).

In papers [**I**], [**IV**] and [**III**], I present the multivariate Gaussian process regression methodology applied in joint species distribution models and quantitative ecology. More specifically, paper [**I**] presents a new multivariate Ricker population growth model which is combined with the multivariate GPs regression to improve model's performance. In paper [**IV**], a new probabilistic model for binary outcomes is presented whose construction is based on multivariate GP priors. Paper [**III**] integrate different probabilistic models into one single approach for multivariate data modelling. This is specially important for SDMs, since it will allow us to deal and exploit multi-type measurements which commonly arise in real-case scenarios of multiple species surveys. Paper [**II**] presents robust statistical modelling by putting GPs on the location and scale parameters of the Student-t probabilistic model. The computational inference process is approach by exploiting aspects of parametrisation of the probabilistic models, which are closed related to concepts of differential geometry in Mathematics. Paper [**V**] studies the practical performance of the Hamiltonian Monte Carlo sampler (HMC) (Neal, 2011; Girolami and Calderhead, 2011) in a well-known probabilistic model for extreme value data (Coles, 2004).

This introduction is organized as follows. In Chapter 2, we review the basics and challenges of Gaussian process regression and its multivariate extensions based on the LMC. Chapter 3 reviews the new methodology and models proposed throughout all the papers. Some concepts of parametrisation in sta-

[6] In the sense of smaller posterior variances and measures of predictive power respectively.

tistical models and how it can improve the computational inference process for GP-based models is presented and discussed. We also highlight the importance of species distribution modelling, its standard statistical approach and how we integrate models with field data. In Chapter 4, we discuss the main results from the papers and how they are linked to each other. In Chapter 5, future research possibilities and concluding remarks are presented.

Chapter 2

Multivariate Gaussian process regression

This section gives an overview of Gaussian process models and how they are used as a prior distribution over the function values of the regression model (the predictor function in GLM). We start by introducing the univariate Gaussian process as a surrogate for the regression model in GLM and the multivariate GP extension based on the LMC. We also present how the Bayesian approach plays out in order to update our beliefs about the function values of the regression model and the practical difficulties to perform statistical inference.

2.1 Gaussian process model

Regression models are of central importance in statistical data analysis. When the values of parameters of the probabilistic model are expected to show systematic variation as function of covariates, some functional form is used to represent such variation. Usually, in regression analysis, it is natural to assume that regression functions have a fixed functional form (Wild and Seber, 1989; Seber and Lee, 2012). For example, in GLM, the predictor function is usually chosen to be a polynomial of certain degree, where the coefficients of that polynomial are the parameters over which we want to do statistical inference (Nelder and Wedderburn, 1972). In general, these functions are fully described by only a few parameters and, for an abundance of practical applications, this can severely restrict the type of regression functions which gives rise to the observed data.

GP models have been widely used as flexible alternatives to estimate the regression function. The core idea is to assume that the regression function is distributed according to a Gaussian process, which allows us to treat the regression function values as unknown quantities and estimate it from the sample data. Gaussian processes are particular type of stochastic processes which can be thought as a Gaussian distribution over the space of functions. For more details, see Abrahamsen (1997), Kuo (2005), Rasmussen and Williams (2006)

and Øksendal (2013).

Definition 1 (Stochastic process) *A stochastic process* $f : \Omega \times \mathcal{X} \to \mathbb{R}$ *(this work restricts to* $\mathcal{X} \subseteq \mathbb{R}^d$*)*[1] *is a function of two arguments such that,* $\forall \omega_* \in \Omega$*, the function* $f(\omega_*, \cdot) : \mathcal{X} \to \mathbb{R}$ *is called sample path (deterministic function) and* $\forall\, \mathbf{x}_* \in \mathcal{X}$*, the function* $f(\cdot, \mathbf{x}_*) : \Omega \to \mathbb{R}$ *is a random variable*[2] *.*

We call f a Gaussian process if for any finite collection of index points $\{\mathbf{x}_i\}_{i=1,\ldots,n}$ $= \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the n-dimensional multivariate density function of the random vector $\mathbf{f} = [f(\omega, \mathbf{x}_1), \ldots, f(\omega, \mathbf{x}_n)]^\top$ is multivariate Gaussian (see Kuo, 2005; Rasmussen and Williams, 2006). A Gaussian process is completely specified by its mean function and covariance function. The mean function tells us what is the expected value of f for any $\mathbf{x} \in \mathcal{X}$, and we denote this as $\mathbb{E}[f(\mathbf{x})] = m(\mathbf{x})$. The covariance function expresses the degree of dependency between two different function values as a function of two index points. That is, $\text{Cov}(f(\mathbf{x}), f(\mathbf{x}'))$ $= k(\mathbf{x}, \mathbf{x}')$, where $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, is also known as *kernel* function (see Rasmussen and Williams, 2006, Chapter 4 for more details). In compact notation, this is usually written as

$$f \sim \mathcal{GP}\big(m(\cdot), k(\cdot, \cdot)\big). \tag{2.1}$$

At first, it may seen unwieldy to work in space of functions. However, in the GP regression framework, we usually work with a finite collection of index points so that the computational treatment is reduced to a multivariate Gaussian distribution. The collection of index points $\{\mathbf{x}_i\}_{i=1,\ldots,n} \subseteq \mathcal{X}$, in the previous definition, play the role of covariates in the dataset. The vector of function values whose components are now associated to each of those covariates is then distributed according to a n-dimensional multivariate Gaussian distribution,

$$\begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}_1) \\ \vdots \\ m(\mathbf{x}_n) \end{bmatrix}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \cdots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix} \right). \tag{2.2}$$

In practical settings the mean function is frequently set to zero, as it would be usually hard to specify a function $m(\cdot) : \mathcal{X} \to \mathbb{R}$. However, the form of the covariance function of a GP model plays an important role. It encodes a general assumption about the type of functions over which one wants to do statistical inference and also carries the notion of similarity between the values of the function[3].

[1] Note the set $\mathcal{X}$ can be more general. For example, in the recent literature it has been taken as a manifold. See Niu et al. (2018)

[2] This random variable is defined on some probability space $(\Omega, \mathcal{F}(\Omega), \mathbb{P})$, where Ω is the sample space, $\mathcal{F}(\Omega)$ is a σ-algebra of subsets of Ω and $\mathbb{P}$ is a probability measure on $\mathcal{F}(\Omega)$. See Bain and Engelhardt (1992) for a formal definition and details.

[3] Note that we have omitted ω from the notation in (2.2) and we will do so from now on.

For example, in the one-dimensional case ($d = 1$), a well-known kernel function used to model a variety of real-world phenomena (Stein, 1999) is given by the Laplacian covariance function (*Ornstein-Uhlenbeck* process, paper **[I]**)

$$k_{\text{OU}}(x, x') = \sigma_f^2 \exp\big(-\tfrac{1}{2}|x - x'|/\ell\big), \tag{2.3}$$

where the scalar value ℓ controls how fast the dependency between two function values decay along the distance $|x - x'|$. If the value of ℓ is large, the dependency between two different values of the function decays very slowly. The parameter σ_f^2 controls the level of variation of the function for any x and this kernel gives rise to continuous sample paths which are nowhere differentiable.

Another covariance function that is perhaps the most used in machine-learning and statistical applications is the *squared exponential* covariance function (SE) (papers **[I]**, **[II]**, **[IV]**, **[III]**)

$$k_{\text{SE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\big(-\tfrac{1}{2}\left\|\text{diag}(\boldsymbol{\ell})^{-1}(\mathbf{x} - \mathbf{x}')\right\|_2^2\big). \tag{2.4}$$

The vector of parameters $\boldsymbol{\ell} = [\ell_1 \dots \ell_d]^\top$ controls how fast the dependency between two function values decays in each dimension[4]. Similarly as before, this means that if all components of $\boldsymbol{\ell}$ are large, the dependency between two different values of the function decay very slowly and the function f is expected to vary only a little, resembling almost a constant function. The parameter σ_f^2 controls the level of variation of the function and $\|\cdot\|_p$ denotes the p-norm. In this case, the kernel gives rise to continuous and differentiable sample paths (Stein, 1999).

Both of the aforemetioned covariance functions are stationary, which means that the sample paths will not increase or decrease without bound[5]. Besides, those covariance functions belong to the Mátern class of covariance functions, which possesses an extra parameter controlling the degree of smoothness (differentiability) of the sample paths. In the literature, there exists various other types of covariance functions and it is possible to create new covariance functions from the combination of other ones. For a good review of this topic I refer to Rasmussen and Williams (2006), Chapter 4.

As highlighted in Rasmussen and Williams (2006), the crucial aspects related to this approach lies in the assumption that function values $f(x)$ and $f(x')$ attain similar values when x and x' are close. In predictive tasks, covariates from the dataset that are close to new sets of covariates are informative to the prediction of new regression values. This particular aspect of GPs has been shown to

[4]The notation diag($\boldsymbol{\ell}$) means a $d \times d$ matrix whose main diagonal is composed by the elements of $\boldsymbol{\ell}$ and off-diagonals elements are 0.

[5]Note that, for example, neural-network covariance functions are non-stationary whose sample paths do not increase or decrease without bound, but they are not used in this work.

perform mostly well in interpolation scenarios with scattered data and well designed covariance functions. However, this is not the case in extrapolation tasks when the data does no present clear pattern and the covariance function is not well designed. (Wilson and Adams, 2013; Wilson, 2014). Thus, there is a need to improve predictive accuracy in extrapolation tasks and MGPs are potentially useful for this goal (paper **[III]**).

Besides, GPs have traditionally been used for regression analysis with a single type of response variable. In the present days, databases might comprise distinct type of response variables which somehow are linked to each other[6]. Exploit all information available to us by taking into account distinct types of response variables into one single modelling approach is advantageous. Statistical inference is usually improved when statistical dependency between random variable is taken into account (Nelsen, 2006; Giri et al., 2014). This can be done via the introduction of multivariate GPs to model the regression functions associated to each of the response variables which, consequently, creates the link between the response variables. In the next section, we introduce one type of multivariate GP model which will be used throughout this thesis.

2.2 Multivariate Gaussian Process model

Consider J independent GPs where $g_j : \Omega \times \mathcal{X}_j \to \mathbb{R}$ denotes the j^{th} GP. Let us further consider distinct mean functions $m_j(\cdot)$ and correlation functions $\tilde{k}_j(\cdot,\cdot)$ for each g_j[7]. Now, take a $J \times J$ matrix Σ that is symmetric and positive-definite (PD)[8]. Denote by $\boldsymbol{L} = \text{chol}(\boldsymbol{\Sigma})$ the Cholesky decomposition of $\boldsymbol{\Sigma}$. Recall that this decomposition is unique since $\boldsymbol{\Sigma}$ is PD, see Golub and Van Loan (1996) page 143, Theorem 4.2.5. We construct a multivariate GP model as follows. Define,

$$\boldsymbol{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}_1) \\ \vdots \\ f_J(\mathbf{x}_J) \end{bmatrix} = \begin{bmatrix} L_{1,1} & \cdots & 0 \\ \vdots & \ddots & 0 \\ L_{J,1} & \cdots & L_{J,J} \end{bmatrix} \begin{bmatrix} g_1(\mathbf{x}_1) \\ \vdots \\ g_J(\mathbf{x}_J) \end{bmatrix} \tag{2.5}$$

where $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_J)$. Denote $\boldsymbol{g}(\mathbf{x}) = [g_1(\mathbf{x}_1) \cdots g_J(\mathbf{x}_J)]^\top$. Then, the new multivariate GP $\boldsymbol{f}$ yields the matrix-valued covariance function for two distinct

[6] Binary variables, count variables or continuous variables.

[7] By correlation function we mean a kernel function such that $\tilde{k}(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to (-1, 1)$. Equivalently one can set $\sigma_f^2 = 1$ in the aforementioned covariance functions. Here we also consider $\mathcal{X}_j = \mathcal{X} \subseteq \mathbb{R}^d \ \forall \ j$.

[8] This is the same as a variance-covariance matrix.

vector-valued functions $\boldsymbol{f}(\mathbf{x})$ and $\boldsymbol{f}(\mathbf{x}')$ as

$$\begin{aligned}\operatorname{Cov}(\boldsymbol{f}(\mathbf{x}), \boldsymbol{f}(\mathbf{x}')) &= \boldsymbol{L} \operatorname{Cov}(\boldsymbol{g}(\mathbf{x}), \boldsymbol{g}(\mathbf{x}')) \boldsymbol{L}^\top \\ &= \sum_{j=1}^{J} \tilde{k}_j(\mathbf{x}_j, \mathbf{x}_j') \boldsymbol{L}_j \boldsymbol{L}_j^\top \end{aligned} \tag{2.6}$$

where L_j denotes the j^{th} column of L. In particular, if we look to represent some dependence between two specific processes f_j and $f_{j'}$ at the points $\mathbf{x}_j$ and $\mathbf{x}_{j'}$ respectively, it is not difficult to see that we can write (2.6) concisely. Using the kernel function notation this reads,

$$\operatorname{Cov}(f_j(\mathbf{x}_j), f_{j'}(\mathbf{x}_{j'})) = k(\mathbf{x}_j, \mathbf{x}_{j'}) = \sum_{r=1}^{J} \tilde{k}_r(\mathbf{x}_j, \mathbf{x}_{j'}) u_r(j, j') \tag{2.7}$$

where $u_r(j, j')$ is the entry (j, j') of the matrix $\boldsymbol{U}_r = \boldsymbol{L}_r \boldsymbol{L}_r^\top$. The multivariate GP $\boldsymbol{f}$ with covariance function (2.7) is known as the LMC (Mardia and Goodall, 1993; Grzebyk and Wackernagel, 1994; Gelfand et al., 2003). The multivariate process $\boldsymbol{f}$ is unique and has nice interpretation. If the entry (j, j') $(j \neq j')$ of $\boldsymbol{\Sigma}$ is null, then the processes f_j and $f_{j'}$ are independent. Gelfand et al. (2003) and Álvarez and Lawrence (2011) present alternative ways to construct multivariate GPs which are based on convolution of kernels. However those constructions do not present straightforward interpretation compared to the LMC, for which reason we focus in the multivariate GP that is more attractive in the sense of practical interpretability of the parameter $\boldsymbol{\Sigma}$.

Analogously to the univariate GP regression, consider a collection of index points $\{\mathbf{x}_{j,i_j}\}_{i_j=1,\ldots,n_j}$ for each component f_j of $\boldsymbol{f}$, where n_j is the number of points for the j^{th} process. Then, the vector of function values $\mathbf{f} = [\mathbf{f}_1^\top \cdots \mathbf{f}_J^\top]^\top$ where $\mathbf{f}_j = [f_j(\mathbf{x}_{j,1}) \cdots f_j(\mathbf{x}_{j,n_j})]^\top$, follows the $\sum_j n_j$-dimensional multivariate Gaussian distribution,

$$\begin{bmatrix}\mathbf{f}_1 \\ \vdots \\ \mathbf{f}_J\end{bmatrix} \Big| \boldsymbol{\theta} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{m}_1 \\ \vdots \\ \mathbf{m}_J\end{bmatrix}, \sum_{r=1}^{J} \begin{bmatrix} u_r(1,1)[\tilde{\boldsymbol{K}}_r]_{1,1} & \cdots & u_r(1,J)[\tilde{\boldsymbol{K}}_r]_{1,J} \\ \vdots & \ddots & \vdots \\ u_r(J,1)[\tilde{\boldsymbol{K}}_r]_{J,1} & \cdots & u_r(J,J)[\tilde{\boldsymbol{K}}_r]_{J,J}\end{bmatrix}\right) \tag{2.8}$$

where $[\tilde{\boldsymbol{K}}_r]_{j,j'}$ is a correlation matrix between $\mathbf{f}_j$ and $\mathbf{f}_{j'}$ at their respective collection of points $\{\mathbf{x}_{j,i_j}\}_{i_j=1,\ldots,n_j}$ and $\{\mathbf{x}_{j',i_{j'}}\}_{i_{j'}=1,\ldots,n_{j'}}$, and this matrix is obtained in term of the r^{th} correlation function. The vector $\mathbf{m}_j$ collects the expected function values in their respective collections of points. Observe that, we now have explicitly included conditioning on the vector $\boldsymbol{\theta}$, which embraces the correlation function parameters and the extra variance-covariance parameter $\boldsymbol{\Sigma}$. This is because the vector $\boldsymbol{\theta}$ is a priori unknown to us. In compact notation,

equation (2.8) will be usually written as

$$\boldsymbol{f} \,|\, \boldsymbol{\theta} \sim \mathcal{MGP}\big(\boldsymbol{m}(\cdot), k(\cdot,\cdot)\big). \tag{2.9}$$

Differently from the standard construction of the LMC (Gelfand et al., 2003), we highlight there is no need to assume that the collection of points $\{\mathbf{x}_{j,i_j}\}_{i_j=1,\ldots,n_j}$ are equal for all processes (e.g. spatial locations) and the number of points does not need to be the same. For example, for the process f_1, we might have $n_1 = 10$. For the second process f_2, we may take $n_2 = 1$ and so on. This particular feature is important in multivariate settings. It allow us to naturally tackle missing values in the dataset and introduce statistical dependence across different types of response variables indirectly via the multivariate GP prior[9] (papers [**I**], [**IV**], [**III**]).

In the recent literature there exists similar models as presented here. However, they differ in the way the multivariate GP is constructed and lead to lack of *identifiability* of the multivariate Gaussian distributions. Note that identifiability concept relates to the probabilistic model rather than to the parameters. The definition of identifiability in a class of probabilistic models is presented below.

Definition 2 (Identifiability) *Let* $\mathbb{A} = \{\pi_Y(\cdot\,|\,\boldsymbol{\eta}) : \Omega \to \mathbb{R}_+ : \boldsymbol{\eta} \in \Xi\}$ *be a class of probabilistic models where* $\Xi \subseteq \mathbb{R}^p$ *and* Ξ *is the parameter space. If for any given distinct values* $\boldsymbol{\eta}$, $\boldsymbol{\eta}' \in \Xi$ *we have* $\pi_Y(\mathbf{y}\,|\,\boldsymbol{\eta}) = \pi_Y(\mathbf{y}\,|\,\boldsymbol{\eta}')\ \forall\, \mathbf{y} \in \Omega$, *then the family* $\mathbb{A}$ *is said to be nonidentifiable. It can also be said that the probabilistic model* $\pi_Y(\cdot\,|\,\boldsymbol{\eta})$ *is nonidentifiable.*

Teh et al. (2005) proposed the *semiparametric latent factor model* which is constructed in a similar manner as in equation (2.5). In their case, the matrix $\boldsymbol{L}$ in (2.5) is replaced with $J \times P$ matrix $\boldsymbol{\Phi}$ of real values where P is the number of indenpendent GPs. In this case the matrix-valued covariance function reads $\mathrm{Cov}(\boldsymbol{f}(\mathbf{x}), \boldsymbol{f}(\mathbf{x}')) = \sum_{p=1}^{P} \boldsymbol{\Phi}_p\, \boldsymbol{\Phi}_p^\top\, \tilde{k}_p(\mathbf{x}_p, \mathbf{x}'_p)$, where $\boldsymbol{\Phi}_p$ is the p^{th} column of $\boldsymbol{\Phi}$. It is clear that the multivariate Gaussian distribution constructed by means of such covariance function is not identifiable. To see this, note that there are many choices of $\boldsymbol{\Phi}_p$ for the same matrix $\boldsymbol{\Phi}_p\, \boldsymbol{\Phi}_p^\top$. Bonilla et al. (2008) restrict $\boldsymbol{\Phi}$ to be $J \times J$ positive-semidefinite (PSD) matrix and assume common correlation function $\tilde{k}(\cdot,\cdot)$ for all independent processes g_j. Hence the covariance function reads $\mathrm{Cov}(\boldsymbol{f}(\mathbf{x}), \boldsymbol{f}(\mathbf{x}')) = \boldsymbol{\Phi}\, \tilde{k}(\mathbf{x}, \mathbf{x}')$. They parametrize $\boldsymbol{\Phi}$ in term of its Cholesky decomposition but since $\boldsymbol{\Phi}$ is PSD, the Cholesky decomposition is not guaranteed to be unique (Golub and Van Loan, 1996; Horn and Johnson, 2012). Álvarez and Lawrence (2011) follow similar approaches as described

[9] This is also known as multi-task GP. The statistical dependence is known as transfer learning or information sharing in machine-learning.

before but they use incomplete Cholesky decomposition (sparse approximation of a Cholesky decomposition). Note also that a real symmetric PSD matrix does not guarantee the existence of its inverse. In Chapter 6, we review some facts and definitions of PD and PSD matrices.

Our starting point in this work avoids such model constructions and there is no lack of identifiability in (2.8) with uniqueness of the Cholesky decomposition for PD matrices. Pinheiro and Bates (1996) pointed out that the lack of uniqueness of the Cholesky decomposition in the optimization of some objective function causes trouble in numerical procedures. However, observe that this is more generally linked to the following line of reasoning in Statistics. A model that is nonidentifiable is not able to "learn" the parameters as $n \to \infty$. As the dataset increases, we would never be able to find the true value in the parameter space which would have had defined a unique probabilistic model generating the observed data (Casella and Berger, 2002; Wechsler et al., 2013). Consequentially, it would be natural to expect troubles in computational inference algorithms. For example, in the optimization of some log-posterior function (or the log-likelihood) when different optimal solutions are close together in the parameter space, or in Monte Carlo Markov chain (MCMC) methods when the parametrisation induces a posterior distribution with distant modes and the MCMC algorithm is not able to explore all the regions of the parameter space.

In the paper by Gelfand et al. (2003), inference on $\mathbf{\Sigma}$ is conducted via MCMC methods directly in the space of covariance matrices. Differently, this work uses the separation strategy of covariance matrices (Barnard et al., 2000) and uses the closed-form mapping between the space of correlation matrices to the space $\mathbb{R}^{\binom{J}{2}}$ (Kurowicka and Cooke, 2003; Lewandowski et al., 2009). Thus, this gives us more flexibility in the sense that advanced MCMC methods such as Hamiltonian Monte Carlo (Neal, 2011) or optimization techniques (Pinheiro and Bates, 1996) can be straighforwardly used in the unconstrained space $\mathbb{R}^{\binom{J}{2}}$. Throughout the next sections, I elaborate in more details how inference on $\mathbf{\Sigma}$ will be conducted in the unconstrained space $\mathbb{R}^{\binom{J}{2}}$.

2.3 Hierarchical Bayesian approach for MGP regression

We briefly outline how the MGP regression plays out in multivariate settings. The basic idea is to approach statistical regression hierarchically and set the dependency in the second level of the hierarchy via the MGP. This alleviates the possibly many choices of multivariate probabilistic models and allow us to combine well-known univariate models into one single multivariate modelling approach.

The model building is done similarly as in Wikle (2003), Cressie and While

(2011) and Banerjee et al. (2015), the levels of the hierarchy are built as follows,

$$\begin{aligned} \boldsymbol{Y} \,|\, \boldsymbol{f}, \boldsymbol{\eta} &\sim \pi_{\boldsymbol{Y}} \\ \boldsymbol{f} \,|\, \boldsymbol{\theta} &\sim \mathcal{MGP} \\ \boldsymbol{\eta}, \boldsymbol{\theta} &\sim \pi_{\text{hyper}}. \end{aligned} \tag{2.10}$$

The first layer in this hierarchy defines the probabilistic model $\pi_{\boldsymbol{Y}}$ for multivariate data $\boldsymbol{Y}$ given the MGP regression values $\boldsymbol{f}$ and another parameters $\boldsymbol{\eta}$ of the probabilistic model. The second layer defines the MGP prior given the processes hyperparameters $\boldsymbol{\theta}$, and the third layer defines the hyperprior distribution π_{hyper} for all unknown parameters and hyperparameters.

Let us start assuming a J-variate random vector $\boldsymbol{Y} = [Y_1 \cdots Y_J]^\top$ such that each of its components is respectively associated with each component of $\boldsymbol{f} = [f_1 \cdots f_J]^\top$. Besides, for each component of $\boldsymbol{f}$, there is a set of associated covariates $\mathbf{x}_j$, $j = 1, \ldots, J$. A common assumption in regression analysis is that samples from the statistical model (2.10) are obtained independently. By further assuming that the components of $\boldsymbol{Y}$ are conditionally independent given $\boldsymbol{f}$ and $\boldsymbol{\eta}$, the *sample distribution*[10] is given by,

$$\pi_{\boldsymbol{Y}}(\mathbf{y} \,|\, \mathbf{f}, \boldsymbol{\eta}) = \prod_{j=1}^{J} \prod_{i_j=1}^{n_j} \pi_{Y_j}(y_{j,i_j} | f_{j,i_j}, \eta_j) \tag{2.11}$$

where y_{j,i_j} is the i_j'th observation related to the j'th process with vector of covariates $\mathbf{x}_{j,i_j}$. The observed vector $\mathbf{y} = [\mathbf{y}_1^\top \cdots \mathbf{y}_J^\top]^\top$ with $\mathbf{y}_j^\top = [y_{j,1} \cdots y_{j,n_j}]$ collects all the observations and $\mathbf{f} = [\mathbf{f}_1^\top \ldots \mathbf{f}_J^\top]^\top$, where $f_j(\mathbf{x}_{j,i_j}) = f_{j,i_j}$ collects the regression function values. From (2.11), we will also assume that the probabilistic models for $Y_j | f_j, \eta_j$, $j = 1, \ldots, J$ have only one possible extra scalar parameter η_j. Applying the Bayes' rule considering the hierarchical structure we obtain,

$$\pi_{\text{post}}(\mathbf{f}, \boldsymbol{\eta}, \boldsymbol{\theta} \,|\, \mathbf{y}) = \frac{\pi_{\boldsymbol{Y}}(\mathbf{y} \,|\, \mathbf{f}, \boldsymbol{\eta}) \pi(\mathbf{f} \,|\, \boldsymbol{\theta}) \pi_{\text{hyper}}(\boldsymbol{\eta}, \boldsymbol{\theta})}{\pi_{\text{M}}(\mathbf{y})} \tag{2.12}$$

where $\pi(\mathbf{f} \,|\, \boldsymbol{\theta}) := \mathcal{N}(\mathbf{f} \,|\, \mathbf{0}, \boldsymbol{K})$ is our MGP prior (2.8) for the multivariate regression. The vector $\mathbf{m}$ collects the expected function values and $\boldsymbol{K}$ is the variance-covariance matrix from the multivariate Gaussian distribution in equation (2.8). $\pi_{\text{M}}(\mathbf{y})$ is the marginal likelihood.

Hyperpriors are chosen accordingly to the background knowledge of the problem and the structure of the model. Given the nonparametric nature of the MGP prior, the choice of the hyperpriors for the hyperparameters combines the weakly informative principle from Gelman (2006) and the penalised model

[10] The joint distribution of the random sample (Knight, 1999).

component complexity priors (PC-priors) (Simpson et al., 2017). The general idea is that the density function for the hyperparameters should give more weight to simple regression functions such as straight lines, planes, etc. That is, the prior should favour small variability of the sample paths in the MGP prior and more strongly correlated function values within the same unknown function, e.g., between $f_j(\mathbf{x}_j)$ and $f_j(\mathbf{x}'_j)$. This has been done in order to avoid overfitting. See Gelman (2006) and Simpson et al. (2017) for more details.

For the variance-covariance matrix parameter $\boldsymbol{\Sigma}$ (paper [**IV**]-[**III**]), careful treatment is necessary. We first rewrite it in terms of variances and correlations, that is, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_J)\ \boldsymbol{\mathcal{P}}\ \text{diag}(\sigma_1, \ldots, \sigma_J)$, where $\boldsymbol{\mathcal{P}}$ is a correlation matrix of dimension $J \times J$. A correlation matrix has the following properties. Each component of its main diagonal is 1, off-diagonal elements $\rho_{j,j'} \in (-1, 1)$ and $\boldsymbol{\mathcal{P}}$ is PD. See, Rousseeuw and Molenberghs (1994) for details and particular illustration of the space of correlation matrices. A particular off-diagonal entry of this matrix measures the statistical dependency between two processes. Values of $\rho_{j',j}$ close to one indicates that processes j' and j have strong positive linear dependency. If $\rho_{j',j}$ is close to minus one, processes j' and j have strong negative linear dependency and if $\rho_{j',j}$ is 0 no dependency exists.

For the correlation matrix $\boldsymbol{\mathcal{P}}$, we assume a prior distribution that induces marginally noninformative priors. That is, since we would expect lack of information about the dependency between the GPs, the marginal distribution for every correlation parameter $\rho_{j,j'}$ is uniform over $(-1, 1)$. This is achieved with the distribution of Barnard et al. (2000) and Tokuda et al. (2012). Note also that, the separation strategy of covariance matrices and the prior choice for the correlation matrix parameter in the MGPs is presented for the first time in here (paper [**IV**]-[**III**]).

2.3.1 Prediction of new outcomes

Consider a new set of covariates $\{\mathbf{x}_{j,i_j,*}\}_{i_j=1,\ldots,n_{j,*}}$, $j = 1, \ldots, J$ that for each of which we want to predict new outcomes $Y_{j,i_j,*}$. Lets denote the vector of regression values at the new set of covariates as $\mathbf{f}_* = [\mathbf{f}_{1,*}^\top \ldots \mathbf{f}_{J,*}^\top]^\top$ where $f_j(\mathbf{x}_{i_j,*}) = f_{j,i_j,*}$. From properties of GPs, the joint distribution of $\mathbf{f}_*$ and $\mathbf{f}$ conditioned on the parameters $\boldsymbol{\theta}$ is multivariate Gaussian

$$\begin{bmatrix} \mathbf{f}_* \\ \mathbf{f} \end{bmatrix} \mid \boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{K}_{*,*} & \boldsymbol{K}_* \\ \boldsymbol{K}_*^\top & \boldsymbol{K} \end{bmatrix}\right) \tag{2.13}$$

where $\mathbf{m}_*$ is the expected values of $\mathbf{f}_*$ and the matrix $\boldsymbol{K}_{*,*}$ is constructed using the new set of covariates. Its construction is done the same way as $\boldsymbol{K}$ (equation

(2.8)). The cross-covariance matrix between $\mathbf{f}_*$ and $\mathbf{f}$ is given by,

$$\boldsymbol{K}_* = \sum_{r=1}^{J} \begin{bmatrix} u_r(1,1)[\tilde{\boldsymbol{K}}_r]_{1,1,*} & \cdots & u_r(1,J)[\tilde{\boldsymbol{K}}_r]_{1,J,*} \\ \vdots & \ddots & \vdots \\ u_r(J,1)[\tilde{\boldsymbol{K}}_r]_{J,1,*} & \cdots & u_r(J,J)[\tilde{\boldsymbol{K}}_r]_{J,J,*} \end{bmatrix} \tag{2.14}$$

where $[\tilde{\boldsymbol{K}}_r]_{j,j',*}$ is the $n_{j,*} \times n_{j'}$ cross-correlation matrix between $\mathbf{f}_{j,*}$ and $\mathbf{f}_j$ at the points $\{\mathbf{x}_{j,i_j,*}\}_{i_j=1,\ldots,n_{j,*}}$ and $\{\mathbf{x}_{j,i_j}\}_{i_j=1,\ldots,n_j}$, obtained with the r^{th} correlation function. From equation (2.14), the conditional distribution of $\mathbf{f}_* \,|\, \mathbf{f}, \boldsymbol{\theta}$ is another multivariate Gaussian given by

$$\mathbf{f}_* \,|\, \mathbf{f}, \boldsymbol{\theta} \sim \mathcal{N}\big(\boldsymbol{K}_* \, \boldsymbol{K}^{-1} \, \mathbf{f}, \boldsymbol{K}_{*,*} - \boldsymbol{K}_* \, \boldsymbol{K}^{-1} \, \boldsymbol{K}_*^\top \big). \tag{2.15}$$

Now, the posterior predictive distribution for $\mathbf{f}_* \,|\, \mathbf{y}$ can be obtained using the hierarchical structure (2.10) as,

$$\pi(\mathbf{f}_* \,|\, \mathbf{y}) = \mathbb{E}_{\mathbf{f},\boldsymbol{\theta} \,|\, \mathbf{y}}\big(\pi(\mathbf{f}_* \,|\, \mathbf{f}, \boldsymbol{\theta})\big) \tag{2.16}$$

where $\pi(\mathbf{f}_* \,|\, \mathbf{f}, \boldsymbol{\theta}) := \mathcal{N}(\mathbf{f}_* \,|\, \boldsymbol{K}_* \, \boldsymbol{K}^{-1} \, \mathbf{f}, \boldsymbol{K}_{*,*} - \boldsymbol{K}_* \, \boldsymbol{K}^{-1} \, \boldsymbol{K}_*^\top)$ and $\pi_{\text{post}}(\mathbf{f}, \boldsymbol{\theta} \,|\, \mathbf{y})$ is the marginal of (2.12) integrating out $\boldsymbol{\eta}$. The posterior predictive distribution for all new outcomes conditioned only on the observed data becomes,

$$\pi_{\text{pred}}(\mathbf{y}_* \,|\, \mathbf{y}) = \mathbb{E}_{\mathbf{f}_*,\boldsymbol{\eta} \,|\, \mathbf{y}}\big(\pi_{\boldsymbol{Y}}(\mathbf{y}_* \,|\, \mathbf{f}_*, \boldsymbol{\eta})\big). \tag{2.17}$$

Finally, point estimates and measures of dispersion, particularly for any $Y_{j,i_j,*} \,|\, \mathbf{y}$, can be obtained for example as

$$\mathbb{E}(Y_{j,i_j,*} \,|\, \mathbf{y}) = \mathbb{E}\big(\mathbb{E}(Y_{j,i_j,*} | f_j(\mathbf{x}_{i_j,*}), \eta_j, \mathbf{y})\big) \tag{2.18}$$

and

$$\mathbb{V}(Y_{j,i_j,*} \,|\, \mathbf{y}) = \mathbb{E}\big(\mathbb{V}(Y_{j,i_j,*} | f_j(\mathbf{x}_{i_j,*}), \eta_j, \mathbf{y})\big) + \mathbb{V}\big(\mathbb{E}(Y_{j,i_j,*} | f_j(\mathbf{x}_{i_j,*}), \eta_j, \mathbf{y})\big).$$

Some facts about full Bayesian inference in GP-based models

Full Bayesian inference for $\mathbf{f}, \boldsymbol{\eta}, \boldsymbol{\theta} \,|\, \mathbf{y}$ may be practically unfeasible due to lack of closed-form expressions and the dimensionality of the posterior (2.12). Markov chain Monte Carlo (MCMC) would provide precise answers in the limit of large number of posterior samples (Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984; Neal, 2003; Robert and Casella, 2004; Neal, 2011; Girolami and Calderhead, 2011; Calderhead, 2012). However, in the case of a GP model, there exists a hard task to engineer good sampling strategies for efficient exploration of the target distribution (2.12). In particular, each evaluation of (2.12) requires inversion of the matrix K that scales to $\mathcal{O}\big((\sum_j n_j)^3\big)$ computational operations.

In case of large datasets the posterior will also have large dimension, in which case the number of MCMC iterations will drastically increase in order to obtain a representative random sample of the posterior (2.12)[11]. This will be extremely more challenging when the probabilistic models for the data involved in (2.10) are non-Gaussian and the number of processes in the MGP starts to be relatively large.

2.4 Approximate inference

Analytical approximations such as Laplace approximation (LP) (Tierney and Kadane, 1986; Tierney et al., 1989) or expectation-propagation (EP) (Minka, 2001a; Dehaene and Barthelmé, 2018)[12] have also proved to provide accurate approximate inference for univariate GP models with much lower computational time requirements (Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008; Rue et al., 2009; Riihimäki et al., 2013). The core idea is to approximate the posterior distribution for the regression values with conditioning on the data, parameters and hyperparameters. This can be done via LP or EP method. In this case, the conditional distribution we aim to approximate is

$$\pi_{\text{post}}(\mathbf{f} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y}) = \frac{\pi_{\boldsymbol{Y}}(\mathbf{y} \,|\, \mathbf{f}, \boldsymbol{\eta}) \pi(\mathbf{f} \,|\, \boldsymbol{\theta})}{\pi_{\text{M}}(\mathbf{y} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta})} \tag{2.19}$$

where the $\pi_{\text{M}}(\mathbf{y} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta})$ is the marginal likelihood with conditioning on parameters and hyperparameters. The choice of the hyperparameters to be used in (2.19) is based on the marginal posterior distribution of the parameters and hyperparameters given by,

$$\pi_{\text{post}}(\boldsymbol{\eta}, \boldsymbol{\theta} \,|\, \mathbf{y}) \propto \pi_{\text{M}}(\mathbf{y} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}) \, \pi_{\text{hyper}}(\boldsymbol{\eta}, \boldsymbol{\theta}) \tag{2.20}$$

where $\boldsymbol{\eta}, \boldsymbol{\theta}$ is chosen such that (2.20) attains its maximum value (Gibbs, 1997; Riihimäki, 2013). Usually, it is not possible to find a closed-form for $\pi_{\text{M}}(\mathbf{y} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta})$ but we can also approximate it with LP or EP.

2.4.1 Laplace approximation

The Laplace approximation is based on the second-order Taylor expansion of $\log \pi_{\text{post}}(\mathbf{f} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y})$ around the maximum a posteriori estimate (MAP), that is $\hat{\mathbf{f}} = \arg\max_{\mathbf{f} \in \mathbb{R}^{\sum_j n_j}} \log \pi_{\text{post}}(\mathbf{f} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y})$. The method yields a multivariate

[11] Problems with mixing, unusual dependence structure and exploration of the whole posterior domain are common.

[12] Although EP lacks theoretical guarantees of its convergence, this method has been widely used in GP-based models and other areas. Recently, EP has been proven to converge as the number of data points grows to infinity, see Dehaene and Barthelmé (2018).

Gaussian approximation for the conditional posterior distribution (2.19) given by

$$\tilde{\pi}_{\text{post}}(\mathbf{f} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}\big(\mathbf{f} \,|\, \hat{\mathbf{f}}, (\boldsymbol{K}^{-1} + \widehat{\mathbf{W}})^{-1}\big) \tag{2.21}$$

where $\boldsymbol{K}$ is the covariance matrix of the multivariate Gaussian process prior. $\widehat{\mathbf{W}} = \mathbf{W}\,|_{\mathbf{f}=\hat{\mathbf{f}}}$ where $\mathbf{W} = -\nabla\nabla \log \pi_{\boldsymbol{Y}}(\mathbf{y} \,|\, \mathbf{f}, \boldsymbol{\eta})$ is the diagonal Hessian matrix of the negative log-likelihood function with respect to $\mathbf{f}$. Particularly, each diagonal element of W is given by $\mathbf{W}_{j,i_j} = -\partial^2 / \partial f_{j,i_j}^2 \log \pi_{Y_j}(y_{j,i_j} | f_{j,i_j}, \eta_j)$.

The Newton method is usually used to find the mode $\hat{\mathbf{f}}$. However, when the probabilistic model leads to non log-concave likelihoods, there are difficulties in finding the mode (Rasmussen and Williams, 2006; Vanhatalo et al., 2009). An alternative solution for this difficulties is given by the natural gradient (Amari, 1998). In this case, the Fisher information matrix of the probabilistic model is used instead of its Hessian matrix. To find the maximum of (2.19), we can use the natural gradient with metric proposed by Calderhead (2012) (page 87, Section 4.1.4, equation 4.2)[13] which leads to the iteration

$$\mathbf{f}^{\text{new}} = (\boldsymbol{K}^{-1} + \mathbb{E}_{\boldsymbol{Y} \,|\, \mathbf{f}, \boldsymbol{\eta}}[\mathbf{W}])^{-1} \big(\mathbb{E}_{\boldsymbol{Y} \,|\, \mathbf{f}, \boldsymbol{\eta}}[\mathbf{W}]\, \mathbf{f} + \nabla_{\mathbf{f}} \log \pi_{\boldsymbol{Y}}(\mathbf{y} \,|\, \mathbf{f}, \boldsymbol{\eta})\big) \tag{2.22}$$

where $\mathbb{E}_{\boldsymbol{Y} \,|\, \mathbf{f}, \boldsymbol{\eta}}[\mathbf{W}])$ is the Fisher information matrix of the probabilistic models and $\nabla_{\mathbf{f}} \log \pi_{\boldsymbol{Y}}(\mathbf{y} \,|\, \mathbf{f}, \boldsymbol{\eta})$ is the gradient vector of the log-likelihood function w.r.t. to the latent values $\mathbf{f}$. This approach is used in paper **[II]**.

In general, the LP method provides good approximations. However, if the posterior distribution we aim to approximate has narrow (or very broad) peak, the approximate posterior shape will have lighter tails (or stronger) compared to the true posterior shape. This may lead to smaller (or higher variances). This is because the approximate posterior variance can also be seen as a curvature evaluated at the MAP estimate, where the function's graph is defined by the log-posterior density function and the approximation is Gaussian.

The approximate posterior predictive distribution is obtained similar as (2.16), but we use (2.21) in place of (2.19) with conditioning on the parameters and hyperparameters. The result is

$$\tilde{\pi}(\mathbf{f}_* \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}\big(\mathbf{f}_* \,|\, \boldsymbol{K}_*\, \boldsymbol{K}^{-1}\, \hat{\mathbf{f}}, \boldsymbol{K}_{*,*} - \boldsymbol{K}_* (\boldsymbol{K} + \widehat{\mathbf{W}}^{-1})^{-1}\, \boldsymbol{K}_*\big) \tag{2.23}$$

Since the parameter and hyperparameters are fixed, the approximate posterior predictive distribution for the new outcomes is obtained as in (2.17) and we integrate only w.r.t $\mathbf{f}_*$, whose distribution is given in (2.23). Point estimates

[13] From another viewpoint, if each point in the posterior domain (equation (2.19)) $\mathbf{f} \in \mathbb{R}^{\sum_j n_j}$ is associated with a metric/inner product $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{f}} = \mathbf{u}^\top \boldsymbol{G}(\mathbf{f}) \mathbf{v}$, where $\boldsymbol{G}$ is PD, the pair $(\mathbb{R}^{\sum_j n_j}, \boldsymbol{G})$ is known as a Riemannian manifold (Do Carmo, 2013). Thereof, the natural gradient can be defined.

and measures of dispersion are calculate using equation (2.18). In this case, closed-form expressions might be available for well-known probabilistic models and this speeds up computations considerably.

Approximate marginal likelihood with LP

As a by-product of the Laplace approximation, the approximate marginal likelihood is given by (Rasmussen and Williams, 2006),

$$\tilde{\pi}_{\mathrm{M}}(\mathbf{y} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}) = \pi_{\boldsymbol{Y}}(\mathbf{y} \,|\hat{\mathbf{f}}, \boldsymbol{\eta})| \, \boldsymbol{I} + \widehat{\mathbf{W}} \, \boldsymbol{K} \, |^{-\frac{1}{2}} \exp\big(-\tfrac{1}{2} \hat{\mathbf{f}}^{\top} \, \boldsymbol{K}^{-1} \, \hat{\mathbf{f}} \big) \tag{2.24}$$

where $\boldsymbol{I}$ is the identity matrix with dimensions $\sum_j n_j \times \sum_j n_j$. If all the probabilistic models π_{Y_j} are log-concave, the evaluation of approximate marginal likelihood has a stable computational treatment, see for example discussion Section in Vanhatalo et al. (2009) and Jylänki et al. (2011). However, if some probabilistic models are not log-concave, then evaluation of (2.24) may require a more refined computational treatment. An alternative way to choose hyperparameters is to replace $\widehat{\mathbf{W}}$ with the Fisher information matrix evaluated at $\hat{\mathbf{f}}$ in equation (2.24). Hence, we have

$$\tilde{\pi}_{\mathrm{MF}}(\mathbf{y} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}) = \pi_{\boldsymbol{Y}}(\mathbf{y} \,|\hat{\mathbf{f}}, \boldsymbol{\eta})| \, \boldsymbol{I} + \mathbb{E}_{\boldsymbol{Y} \,|\hat{\mathbf{f}}, \boldsymbol{\theta}}[\mathbf{W}] \, \boldsymbol{K} \, |^{-\frac{1}{2}} \exp\big(-\tfrac{1}{2} \hat{\mathbf{f}}^{\top} \, \boldsymbol{K}^{-1} \, \hat{\mathbf{f}} \big). \tag{2.25}$$

This leads to great stability in computer implementations. If there a difficulty in the evaluation of (2.24) one can also choose the parameters and hyperparameters based on (2.25). In our experience (paper **[II]**), equation (2.25) has provided very good inferences for parameters and hyperparameters in the sense that point estimates are also close to their true values. Full investigation whether this would be always a good alternative for parameter and hyperparameter point estimation is left for future. For other probabilistic models the approach would follow similarly.

2.4.2 Expectation-propagation

Consider the class of multivariate Gaussian distributions $\mathcal{C} = \{q(\cdot) = \mathcal{N}(\cdot \,|\, \boldsymbol{\mu}, \boldsymbol{\Lambda}) : \boldsymbol{\mu} \in \mathbb{R}^N, \ \boldsymbol{\Lambda} \text{ is covariance matrix}\}$. Given the conditional posterior density (2.19), we search for a member $q(\cdot) \in \mathcal{C}$ such that the Kullback-Leibler divergence (Kullback and Leibler, 1951) between $\pi_{\mathrm{post}}(\cdot \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y})$ and $q(\cdot)$ attains the minimum. It can be shown that the choice of $q(\cdot)$ such that $\boldsymbol{\mu} = \mathbb{E}(\mathbf{f} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y})$ and $\boldsymbol{\Lambda} = \mathbb{V}(\mathbf{f} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y})$ is the minimizer (Seeger, 2005; Stephenson and Broderick, 2016). However, direct computation of first and second-order moments of $\mathbf{f} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y}$ may be unfeasible in high dimensions. The EP approach proposed by Minka (2001a,b) calculates those moments via simple sequential iterations,

which alleviates the computational overhead. The marginal EP approximation (Rasmussen and Williams, 2006) now yields the following multivariate Gaussian approximation,

$$\tilde{\pi}_{\text{post}}(\mathbf{f} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y}) = \frac{1}{Z_{\text{EP}}} \pi(\mathbf{f} \,|\, \boldsymbol{\theta}) \prod_{j=1}^{J} \prod_{i_j=1}^{n_j} \tilde{t}_{j,i_j}(f_{j,i_j} | \tilde{Z}_{j,i_j}, \tilde{\mu}_{j,i_j}, \tilde{\sigma}^2_{j,i_j}). \tag{2.26}$$

where $\tilde{t}_{j,i_j}(f_{j,i_j}|\tilde{Z}_{j,i_j}, \tilde{\mu}_{j,i_j}, \tilde{\sigma}^2_{j,i_j}) = \tilde{Z}_{j,i_j}\mathcal{N}(f_{j,i_j}|\tilde{\mu}_{j,i_j}, \tilde{\sigma}^2_{j,i_j})$, $i_j = 1, \ldots, n_j$, $j = 1, \ldots, n_j$, are referred as *approximate local likelihood terms* with *site parameters* $\tilde{Z}_{j,i_j}$, $\tilde{\mu}_{j,i_j}$ and $\tilde{\sigma}^2_{j,i_j}$. Each of which is associated with the j^{th} probabilistic model and the i_j^{th} observation (see Rasmussen and Williams, 2006, Section 3.6, page 52 for more details). Z_{EP} denotes the approximate marginal likelihood. Equation (2.26) can also be rewritten so that

$$\tilde{\pi}_{\text{post}}(\mathbf{f} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}\big(\mathbf{f} \,|\, (\boldsymbol{K}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1})^{-1} \tilde{\boldsymbol{\Sigma}}^{-1} \, \tilde{\boldsymbol{\mu}}, (\boldsymbol{K}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1})^{-1}\big), \tag{2.27}$$

where $\tilde{\boldsymbol{\mu}} = [\tilde{\mu}_{1,1} \cdots \tilde{\mu}_{J,n_J}]^\top$ and $\tilde{\boldsymbol{\Sigma}} = \text{diag}(\tilde{\sigma}^2_{1,1}, \ldots, \tilde{\sigma}^2_{J,n_J})$. We point out that $\mathbb{E}(\mathbf{f} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y}) = (\boldsymbol{K}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1})^{-1} \tilde{\boldsymbol{\Sigma}}^{-1} \, \tilde{\boldsymbol{\mu}}$ and the diagonal elements of $\mathbb{V}(\mathbf{f} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y})$ match those diagonal elements of $(\boldsymbol{K}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1})^{-1}$. This makes the EP approximation a highly desirable approach when compared with gold-standard MCMC methods. Observe that this is due to practical reasons since the end product of the Bayesian analysis is often some functional of the true posterior. For instance, users usually report posterior mean and variance for each regression values for given new sets of covariates[14].

The approximate posterior predictive distribution obtained through the EP framework is given by

$$\tilde{\pi}(\mathbf{f}_* \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}\big(\mathbf{f}_* \,|\, \boldsymbol{K}_*(\boldsymbol{K} + \tilde{\boldsymbol{\Sigma}})^{-1} \tilde{\boldsymbol{\mu}}, \boldsymbol{K}_{*,*} - \boldsymbol{K}_*(\boldsymbol{K} + \tilde{\boldsymbol{\Sigma}})^{-1} \, \boldsymbol{K}_*\big), \tag{2.28}$$

and the procedure to make prediction for new outcomes closely resembles the LA method. The difference in here is that the parameters of the multivariate Gaussian approximation for the regression values $\mathbf{f}_*$ has mean vector and variance-covariance matrix given in (2.28).

Approximate marginal likelihood with EP

The approximate marginal likelihood Z_{EP} is obtained by integrating the numerator of (2.26) w.r.t $\mathbf{f}$. This is done by rewriting the product of the approximate local likelihood terms as a multivariate Gaussian. Besides, since the product of

[14] There exists another ways to approximate posterior distributions such as variational methods. These are similar to EP in that the reverse of the aforementioned KL divergence is minimized.

two Gaussians is another unnormalized multivariate Gaussian, integration w.r.t $\mathbf{f}$ is possible in closed-form. The resulting expression is given by

$$\tilde{\pi}_{\mathrm{M}}(\mathbf{y} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{0} \,|\, \tilde{\boldsymbol{\mu}}, \boldsymbol{K} + \tilde{\boldsymbol{\Sigma}}) \prod_{j=1}^{J} \prod_{i_j=1}^{n_j} \tilde{Z}_{j,i_j}. \tag{2.29}$$

As in the Laplace approximation, with not log-concave probabilistic models, convergence problems can occur in the EP algorithm. If that is the case, these problems can be alleviate with damping or fractional updates (e.g. Seeger, 2005; Jylänki et al., 2011).

In paper **[IV]**, we have derived the marginal likelihood in closed-form for the multivariate Gaussian process classification which generalized the result for Gaussian process classification (see Rasmussen and Williams, 2006, subsection 3.3, page 39). This formula takes the form

$$\pi(\mathbf{y} \,|\, \boldsymbol{\theta}) = F(\mathbf{0} \,|\, \mathbf{0}, \boldsymbol{I} + \boldsymbol{I}_{\mathbf{y}} \boldsymbol{K} \boldsymbol{I}_{\mathbf{y}}), \tag{2.30}$$

where $F(\cdot \,|\, \mathbf{0}, \boldsymbol{I} + \boldsymbol{I}_{\mathbf{y}} \boldsymbol{K} \boldsymbol{I}_{\mathbf{y}})$ is the multivariate Gaussian cumulative distribution function. The EP algorithm designed by Cunningham et al. (2011) to evaluate multivariate Gaussian probabilities is used to approximate (2.30). This is because it has been shown to have a good degree of approximation in extensive numerical experiments (see Cunningham et al., 2011, for details).

2.4.3 Hyperparameter inference

As mentioned earlier in this section, there is a need to tackle parameter and hyperparameter inference more carefully. Besides, both of the aforementioned posterior approximations assume known parameters and hyperparamaters whose values are obtained from the MAP estimate of marginal posterior distribution (2.20). When non-Gaussian probabilistic models are involved in (2.10), there is frequently lack of analytical treatability of $\pi_{\mathrm{M}}(\mathbf{y} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta})$. Hence, we replace this with (2.24) or (2.29) in expression (2.20) whether we perform approximate inference with LP or EP respectively. We then denote the approximate marginal likelihood as $\tilde{\pi}_{\mathrm{M}}(\mathbf{y} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta})$. Thereof, the approximate marginal posterior distribution of the parameters and hyperparameters reads

$$\tilde{\pi}_{\mathrm{post}}(\boldsymbol{\eta}, \boldsymbol{\theta} \,|\, \mathbf{y}) \propto \tilde{\pi}_{\mathrm{M}}(\mathbf{y} \,|\, \boldsymbol{\eta}, \boldsymbol{\theta}) \pi_{\mathrm{hyper}}(\boldsymbol{\eta}, \boldsymbol{\theta}). \tag{2.31}$$

Our main concern related to the MGP prior is in the parametrisation of the multivariate Gaussian distribution (2.8). This is mainly because of two reasons. Firstly, the parameters quantifying the degree of dependency among processes is given by the correlation matrix $\boldsymbol{\mathcal{P}}$ and, secondly, the space where $\boldsymbol{\mathcal{P}}$ lives is more "complex" than usual. As an example, suppose that we would like to

use gradient descent to optimize an objective function whose argument is a PD matrix. Then, it is natural to see that in each step of the algorithm, the new proposed value will not usually live in the space of PD matrices. Hence, this requires careful treatment in order to improve stability of numerical procedures in computer algorithms. To achieve this goal, we proceed by using a one-to-one mapping between the space of correlation matrices and the real space $\mathbb{R}^{\binom{J}{2}}$. See Kurowicka and Cooke (2003) and Lewandowski et al. (2009) for details in this mapping[15].

In short, to map from $\mathbb{R}^{\binom{J}{2}}$ to the space of correlation matrices, we consider the upper triangular Cholesky decomposition of $\boldsymbol{\mathcal{P}}$ as

$$\boldsymbol{U} = \begin{bmatrix} 1 & z_{1,2} & z_{1,3} & \cdots & z_{1,J} \\ 0 & \prod_{i=1}^{1}(1-z_{i,2}^2)^{\frac{1}{2}} & z_{2,3}\prod_{i=1}^{1}(1-z_{i,3}^2)^{\frac{1}{2}} & \cdots & z_{2,J}\prod_{i=1}^{1}(1-z_{i,J}^2)^{\frac{1}{2}} \\ 0 & 0 & \prod_{i=1}^{2}(1-z_{i,3}^2)^{\frac{1}{2}} & \cdots & z_{3,J}\prod_{i=1}^{2}(1-z_{i,J}^2)^{\frac{1}{2}} \\ 0 & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \prod_{i=1}^{J-1}(1-z_{i,J}^2)^{\frac{1}{2}} \end{bmatrix} \tag{2.32}$$

where each $z_{i,j} \in (-1,1)$ is knowns as *partial correlation* (see Kurowicka and Cooke, 2003, for details). Since each $z_{i,j}$ can freely vary in the interval $(-1,1)$ without violating the positive definiteness property of $\boldsymbol{\mathcal{P}}$ (see Kurowicka and Cooke, 2003; Lewandowski et al., 2009), we then introduce the map $z_{i,j}(\delta_{i,j}) = 2/[1+\exp(-a\delta_{i,j})] - 1$ where $\delta_{i,j} \in \mathbb{R}$. Thus we can now safely work with $\boldsymbol{\mathcal{P}}$ through the space $\mathbb{R}^{\binom{J}{2}}$. The scalar value a is positive and it stretches or squeezes the real axis.

The inverse transform which takes an element on the set of positive-definite correlation matrices and maps to $\mathbb{R}^{\binom{J}{2}}$ can be obtained recursively. Take the upper Cholesky decomposition of $\boldsymbol{\mathcal{P}}$, $U = \text{chol}(\boldsymbol{\mathcal{P}})_{\text{up}}$. Denote each entry of $\boldsymbol{U}$ as $U_{i,j}$, where $U_{i,j} = 0$ whenever $i > j$. Calculate $z_{i,j}$ as,

$$z_{i,j} = \begin{cases} 0, & \text{if } i \geq j, \\ U_{i,j}, & \text{if } i = 1, j = 2, \ldots, J \\ U_{i,j}\prod_{k=1}^{i-1}(1-z_{k,j})^{-2} & \text{if } 2 \leq i < j, \end{cases} \tag{2.33}$$

and obtain $\delta_{i,j} = -\log((1-z_{i,j})/(z_{i,j}+1))$. Observe that, Lewandowski et al. (2009) also provide the determinant of the Jacobian of the previous mapping. This transformation provides us more flexibility in any modelling approach in the sense that, if a prior distribution on the space of correlation matrices is

[15] This map is also a diffeomorphism.

given, then we can obtain a prior distribution on $\mathbb{R}^{\binom{J}{2}}$ and vice-versa. The multivariate Gaussian distributions used in papers [**III**] and [**IV**] follow this parametrisation and this is how we conduct computer implementation of Matlab codes in the software GPstuff (Vanhatalo et al., 2013). We also note the paper of Pinheiro and Bates (1996), which presents many mappings from real space $\mathbb{R}^{\binom{J}{2}}$ to the space of correlation matrices. But in none of those mappings the determinant of the Jacobian transformation and the inverse mapping are presented[16]. The variance and length-scale hyperparameters whose values live in $\mathbb{R}_+$ are transformed to the real-line using the logarithm function.

In papers [**II**], [**III**] and [**IV**], the MAP estimate is obtained with respect to the transformed original parameter space and hyperparameter space. In general notation, we first consider the transformation $(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}}) = H(\boldsymbol{\eta}, \boldsymbol{\theta})$, where $H(\cdot)$ is a one-to-one mapping which takes a value on the original parameter/hyperparameter space and transforms to the real space ($\mathbb{R}$). Using the Jacobian method for the transformation of random variables (Casella and Berger, 2002), the approximate marginal posterior density for the transformed parameters and hyperparameters reads

$$\tilde{\pi}_{\text{post},H}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}} | \, \mathbf{y}) = \tilde{\pi}_{\text{post}}(H^{-1}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}}) | \, \mathbf{y}) |\det J_{H^{-1}}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}})|. \tag{2.34}$$

where $J_{H^{-1}}$ denotes the Jacobian of the mapping H^{-1}.

Optimization is then conducted via conjugate-gradient on the log scale of (2.34). Once the MAP estimate has been found, say $(\underline{\tilde{\boldsymbol{\eta}}}, \underline{\tilde{\boldsymbol{\theta}}})$, we transform these values to their original space and plug them into (2.24) or (2.27). Closed-form gradients for $\log \pi_{\text{post},H}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}} | \, \mathbf{y})$ are obtained following the general formulation presented by Rasmussen and Williams (2006), Section 5.5 page 124, for the LP and EP approximations. See paper [**III**] for details on how to calculate gradients when considering the correlation matrix $\boldsymbol{\mathcal{P}}$. Note that the aforementioned parameter/hyperparameter inference procedure is the same one presented throughout all the book *Gaussian process for machine learning* (Rasmussen and Williams, 2006) but now tends to be more complex due to particular properties of the unknown matrix $\boldsymbol{\mathcal{P}}$.

[16] This could be another line of investigation.

Chapter 3

New models and methods

In this chapter a more detailed background knownledge about the models, methods and general motivation presented throughout the papers [**I**], [**II**] and [**III**]-[**IV**] are presented. This is done to foster applications of MGPs in species distribution models for quantitative ecology and to complement the different ideas presented in this thesis.

We start introducing two distinct views of mathematical modelling in ecological studies. I briefly discuss the main statistical approach for species distribution models and the mathematical models which are based on mechanistic assumptions of the underlying ecological processes giving rise to field data.

The rest of this chapter focuses in strengthen the link between theoretical methodology and practical applications. For paper [**I**], the discrete-time population Ricker model is extended to a multivariate version and we show how to derive it. Other aspects of statistical modelling for this model is discussed. For paper [**II**], notions of differential geometry are discussed. Those notions are closely related to aspects of reparametrisation of statistical models which can improve the inference process. We end the chapter discussing the multivariate modelling and the link between papers [**III**] and [**IV**].

3.1 General overview of species distribution models

Mathematical modelling is usually based on the background knowledge of the underlying context and the question one looks to address. Species distribution models follow the same principles. Recent literature describes them as any type of models which aim at explaining and predicting variation of species observations (abundance or presence/absence) as a function of environmental variables (covariates) (Austin, 2007; Elith and Leathwich, 2009; Seber and Lee, 2012). The formalism of GLMs in statistics is dominating in practical applications with ecological data (Nelder and Wedderburn, 1972; Guisan et al., 2002;

Bergström et al., 2013; Clark et al., 2014; Angelieri et al., 2016). The core idea is to assume a data generating mechanism (probability model usually from the exponential family) for species observations and link such mechanism with environmental covariates. Once the data have been collected, statistical inference is performed over the parameters of the model. In turn, the parameter estimates quantify the amount of variation (uncertainty) in the data explained by each of the possible environmental variables in the model (Guisan et al., 2013; Ovaskainen and Soinnien, 2011; Ovaskainen et al., 2016a,c, 2017).

It is also widely recognized that ecological processes are complex by nature. They are influenced by many small and large-scale dynamics, such as intrinsic species behaviours, species-to-species associations, species-environment interactions, spatial distribution of habitat characteristic conditions or even weather conditions. To account for all those processes with parametric model may not be the ideal approach. Gaussian process regression have been recently introduced to the SDM literature and they have shown to improve models performance compared to GLM-based models (Vanhatalo et al., 2012; Golding and Purse, 2016). In this sense, GPs are flexible tools that can accommodate such complex phenomena since the regression function can be estimated without imposition of restrictive assumption about its form (O'Hagan, 1978; Rasmussen and Williams, 2006; Golding and Purse, 2016).

Generally, as a final product of the statistical analysis, end-users are frequently interested in building thematic maps. These maps have the goal to highlight and provide a big picture of particular features of species populations across the study region. This is important to influence future policy decision and management actions (Sinclair et al., 2010).

From another point of view, mathematical models often aim to understanding species observations according to first principles. That is to say, the model building is developed from mechanistic underpinnings of ecological processes that govern species behaviour (Murray, 2004). In these cases, species observations are usually described by a discrete-time population dynamic model, via partial differential equations or stochastic partial differential equations. See the works by Murray (2004) or Chapter 1 and page 7 of Ovaskainen et al. (2016b) for more details. Once the models are established, the goal is to study the quantitative and qualitative behaviour of model's predictions by varying those parameters which describe the functional form of the equations. In practice, one has to decide proper values of the parameters and confront the model with measured data. If there is a good match between model prediction and data, one may conclude the validity of the proposed models. Yet, this does not mean that the model is the best one, since different models can provide similar predictions. The point here in case is that, usually no formal method is chosen to select model's parameters. Moreover, there will always be uncertainty related to

which model is the true one. In this sense, statistical inference becomes a useful tool to formally account for uncertainty in parameter estimation and model selection. Recently, this has been an active area of research. See for instance the works by Berliner (1991), Kaipio and Somersalo (2007), Ovaskainen et al. (2008), Calderhead and Girolami (2011), Campbell and Chkrebtii (2013) and Chkrebtii et al. (2016).

3.2 A new multivariate Ricker population model

Discrete-time population dynamic models for iteroparity single species are often modelled as difference equations (Murray, 2004; Brännstron and Sumpter, 2005). In other words, the population size at time $t+1$ is a function of the population size at previous time t. For multispecies population dynamics, more than one species is considered in the model building. In this case, we have a population vector whose components are the population sizes for each species. The population vector at time $t+1$ is then taken as a function of the population vector at time t.

Let us consider J species and denote the difference equation as $\boldsymbol{N}_{t+1} = F(\mathbf{N}_t)$, where $\boldsymbol{N}_t = [N_{1,t} \dots N_{J,t}]^\top$ is the population vector at time t. We follow the reference Brännstron and Sumpter (2005) and similarly assume that each individual's reproductive success in species j is described as function of other members of the population within an region D centred on that individual. We denote this as

$$r_{j,D}(Z_{1,t}, \dots, Z_{J,t}) = \prod_{s=1}^{J} c_{j,s}^{Z_{s,t}} \tag{3.1}$$

where $Z_{j,t}$ is the number of members of population j at time t within the region D. The parameters $c_{j,s} = c_{s,j} \in \mathbb{R}_+$ for $s \neq j$. If $s = j$ then we restrict $c_{j,j} < 1$. This can have many interpretations. For example, if $c_{j,r} > 1$ for some $r \neq j$, interspecific cooperation is plausible. Otherwise, interspecific competition for basic resources (e.g. space or food), cannibalism or potential spread of disease becomes present.

If each population j of size $N_{j,t}$ is assumed to be randomly spread over a large environment region A at any fixed time t, then the number of individuals of species j within a smaller area D, can be assumed to be distributed according to the Poisson distribution as

$$Z_{j,t} \overset{\text{i.i.d}}{\sim} \text{Poisson}\left(\frac{D}{A} N_{j,t}\right) \tag{3.2}$$

for $j = 1, \dots, J$. Moreover, if all individuals of the population j are assumed to be equal, the expected growth ratio of the j^{th} population is proportional to

the expected reproductive success rate of any individual in species j, given the population size $N_{j,t}$,

$$\mathbb{E}\left[\frac{N_{j,t+1}}{N_{j,t}}\middle| N_{j,t}\right] = d_j\ \mathbb{E}\big[r_{j,D}(Z_{j,t},\ldots,Z_{j,t})\big] = d_j \prod_{s=1}^{J} \mathbb{E}\big[c_{j,s}^{Z_{s,t}}\big]. \tag{3.3}$$

where d_j is the proportionality constant. It is possible to obtain that $\mathbb{E}\big[c_{s,j}^{Z_{s,t}}\big] = e^{-(1-c_{j,s})(D/A)N_{j,t}}$. Therefore, the multivariate discrete-time population dynamic model is given by,

$$\mathbb{E}[N_{j,t+1}|N_{j,t}] = N_{j,t}\ d_j\ e^{-\sum_{s=1}^{J}(1-c_{j,s})\frac{D}{A}N_{s,t}}. \tag{3.4}$$

Denote $\alpha_j = d_j$ and $\beta_{j,s} = (1-c_{j,s})\frac{D}{A}$. Changing the notation and using the terminology from fisheries sciences, $R_{j,t} = \mathbb{E}[N_{j,t+1}]$ for the number of elements that will be added into the total population of species j (recruit population for species j) and $S_{j,t} = N_{j,t}$ for the spawning stock size in the species j, the multivariate stock-recruitment Ricker model reads,

$$R_{j,t}(S_{1,t},\ldots,S_{J,t}) = \alpha_j S_{j,t} e^{-\sum_{s=1}^{J}\beta_{j,s}S_{s,t}} \tag{3.5}$$

for $j = 1,\ldots,J$. The parameters $\beta_{j,j}$ stand for the density-dependency of species j and $\beta_{j,s}$ is interspecific density-dependent parameter between species j and s. The parameter α_j is referred as the maximum reproductive rate and it is widely recognized as a measure of renewal potential for the fish populations. Hence it is the most fundamental parameter in the fisheries sciences (Hilborn and Walters, 1992; Myers et al., 1997; Quinn and Deriso, 1999; Myers, 2001; Rose et al., 2001).

One can now study the dynamics of the population via mathematical properties of the model. For example, by studying the values of the parameters such that they lead to populations in stable equilibrium or possible exctintion of some species. With respect to the statistical modelling framework, equation (3.5) is transformed with the logarithm function and we assume the log reproductive rates to be time-varying according to the MGP prior. That is, $(\log\alpha_1(t),\ldots,\log\alpha_J(t)) \sim \mathcal{MGP}(\mathbf{0},k)$. This is the approach taken in paper **[I]** and summarized in Section 4.

3.3 Some aspects of parametrisation in statistical models

Usually, in statistical modelling, the choice of parametrisation of a probabilistic model is mostly left aside by practictioners (MacKay, 1998). This is not without any apparent reason. In most of the models used in practice, the original

parametrisation usually has direct interpretation with the observed data, hence there would be no reason to change model parametrisation.

From a different viewpoint, different parametrisations of the probability model are important to achieve better inferences in approximation techniques[1] and to improve efficiency of estimation procedures in computer algorithms. See for example works by Cox and Reid (1987), Kass (1989), Achcar and Smith (1990), Achcar (1994), Kass and Slate (1994), MacKay (1998) and Calderhead (2012). However, by doing so, one might think that the probabilistic model changes with the choice of the parametrisation. This is not true. parametrisation of the model is closely related to notions of differential geometry which perceives the probabilistic model as a set of points which can be expressed by many different ways without changing the data generated by the model. In other words, the model can be expressed with different set of parameters.

The key point is the notion of *smooth manifold*. A smooth manifold can be thought as a set $\mathcal{M}$ and a family $\mathcal{A}$ of injective mappings $\boldsymbol{\xi}_r : U_r \subseteq \mathbb{R}^D \to \mathcal{M}$ such that they satisfy two properties,

(i) $\bigcup_r \boldsymbol{\xi}_r(U_r) = \mathcal{M}$

(ii) $\forall\, \boldsymbol{\xi}_r, \boldsymbol{\xi}_k \in \mathcal{A}$ the mapping $\boldsymbol{\xi}_r^{-1} \circ \boldsymbol{\xi}_k$ is differentiable $(r \neq k)$[2].

Each pair $(\boldsymbol{\xi}_r, U_r)$ is called system of coordinates of $\mathcal{M}$ and the set $\{(\boldsymbol{\xi}_r, U_r)\}$ is called differentiable structure on $\mathcal{M}$. In Statistics, the sets U_r play the role of the parameter space and $\boldsymbol{\xi}_r$ is the parametrisation of the probabilistic model. The set $\mathcal{M}$ can be taken as any given family of probabilistic models. For instance, consider the Weibull family of probabilistic models (Lawless, 2002) in two different parametrisations (this model is widely used in survival/lifetime analysis). For the first and most common used parametrisation, the Weibull class of probabilistic models can be represented as $\mathcal{M} = \{\pi_{Y|\,\boldsymbol{\alpha}}(\cdot|\,\boldsymbol{\alpha}) : \boldsymbol{\alpha} = (\alpha_1, \alpha_2) \in \mathbb{R}^2_+\}$ where $\pi_{Y|\,\boldsymbol{\alpha}}(y|\alpha_1, \alpha_2) = \alpha_1\alpha_2(\alpha_2 y)^{\alpha_1 - 1}\exp(-(\alpha_2 y)^{\alpha_1})$. Another parametrisation (which may be seen as an uncommon way of representing the Weibull class) of this model is presented in paper **[II]**. In that case the set $\mathcal{M} = \{\pi_{Y|\,\boldsymbol{\eta}}(\cdot|\,\boldsymbol{\eta}) : \boldsymbol{\eta} = (\eta_1, \eta_2) \in \mathbb{R}^2\}$ with $\pi_{Y|\,\boldsymbol{\eta}}(y|\,\boldsymbol{\eta}) := \pi_{Y|\,\boldsymbol{\alpha}}(y|\,\boldsymbol{\alpha}(\boldsymbol{\eta}))$ is the same, hence the set $\mathcal{M}$ is only expressed through a different parametrisation. To see this, note that the transformation $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\boldsymbol{\eta}) = (\exp(\eta_1), \exp(C\exp(-\eta_1) + C\,\eta_2))$ is one-to-one, which means that whether the data Y is generated through those different parametrisations is irrevelant. Given any value $\boldsymbol{\alpha} \in \mathbb{R}^2_+$ we can always find its

[1] Laplace, expectation-propagation, variational methods or MCMC approximations.

[2] We can also say that $\mathcal{M}$ is D-dimensional manifold. In statistical applications we will just restrict $\boldsymbol{\xi}_r : U_r \subseteq \mathbb{R}^D \to \mathcal{M}$ to be diffeomorphism, a differentiable bijective mapping with differentiable inverse mapping.

correspondent value $\boldsymbol{\eta} \in \mathbb{R}^2$ and vice-versa[3].

The concept of *tangent space* and *Riemannian manifold* are also brielfy introduced. These concepts are focused on this work to extend the notion of distance, rate of change and gradient of a function on spaces which are more general than Euclidean. Like the set $\mathcal{M}$, whose elements are density functions.

Tangent Space and Riemannian manifold

Observe that, the modulus of a vector interpreted as the length of a straight line connecting two points in the set $\mathcal{M}$ may not make sense. Two points in the parameter space can be close together but they might still produce great disparity between their correspondent density functions. In this sense, vectors on $\mathcal{M}$ are *tangent vectors* at each $p \in \mathcal{M}$ and the *tangent space* formed by the set of all those vectors is a local approximation of the manifold (Calderhead, 2012).

We say that a function on the manifold $f : \mathcal{M} \to \mathbb{R}$ is differentiable on $\mathcal{M}$ if for any given parametrisation $\boldsymbol{\xi} : U \subseteq \mathbb{R}^D \to \mathcal{M}$, the composite function $f \circ \boldsymbol{\xi} : U \subseteq \mathbb{R}^D \to \mathbb{R}$ is differentiable at $\boldsymbol{\xi}^{-1}(p)$, $\forall\, p \in \mathcal{M}$.

Let $\gamma : (-\epsilon, \epsilon) \to \mathcal{M}$ be a differentiable curve in $\mathcal{M}$ for which $\gamma(0) = p$. Denote by $\mathbb{D}$ the set of functions which are differentiable on $\mathcal{M}$. The tangent vector to the curve $\gamma(t)$ at $t = 0$ is a function $\gamma'(0)(\cdot) : \mathbb{D} \to \mathbb{R}$ given by

$$\gamma'(0)(f) := \frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)\big|_{t=0}.$$

The tangent vector at p is the tangent vector of the curve $\gamma(t)$ at $t = 0$. The set of all tangent vectors to $\mathcal{M}$ at p is indicated as $T_p\mathcal{M}$. If we choose a parametrisation $\boldsymbol{\xi} \in \mathcal{A}$, where $\boldsymbol{\xi}^{-1}(p) = (\xi_1, \ldots, \xi_D) \in U \subseteq \mathbb{R}^D$ and $p = \boldsymbol{\xi}(\mathbf{0})$, we can express both of the functions f and γ in $\boldsymbol{\xi}^{-1}$ and obtain

$$\gamma'(0)(f) = \Big(\sum_{i=1}^{D} \xi_i'(0) \frac{\partial}{\partial \xi_i}\Big|_{t=0} \Big)(f).$$

From the above expression we remark that the set of differential operators $\frac{\partial}{\partial \xi_1}|_{t=0}$, $\ldots$, $\frac{\partial}{\partial \xi_D}|_{t=0}$ are interpreted as linearly independent tangent vectors at $p \in \mathcal{M}$. Thus, the choice of parametrisation determines the associated basis $\{\frac{\partial}{\partial \xi_1}|_{t=0}, \ldots, \frac{\partial}{\partial \xi_D}|_{t=0}\}$ in $T_p\mathcal{M}$ and, consequently, $T_p\mathcal{M}$ is a vector space of dimension D. Besides, any reparametrisation is also invariant with respect to the tangent space, since any choice of parametrisation changes only the associated basis in $T_p\mathcal{M}$ (Do Carmo, 2013; Pressley, 2001).

At each point in p we can now associate an inner product of vectors in $T_p\mathcal{M}$ and this will allow us to measure distances (or angles) on the manifold $\mathcal{M}$ (see

[3]This transformation represents the property *(ii)*, in the above concept of a smooth manifold.

Pressley, 2001, Chapter 6, page 121). The inner product is also usually termed as *metric tensor* and is in general defined as a real-valued function acting on the vectors of the tangent space

$$g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}.$$

(Pressley, 2001; Do Carmo, 2013). Since any vector in $T_p\mathcal{M}$ can be decomposed as a linear combination of the associate basis, the function g takes a quadratic form. To see this, let $\mathbf{u} = \sum_{i=1}^{D} a_i \partial/\partial\xi_i$, $\mathbf{v} = \sum_{i=1}^{D} b_i \partial/\partial\xi_i \in T_p\mathcal{M}$, then

$$\begin{aligned} g_p(\mathbf{u}, \mathbf{v}) &= \langle \mathbf{u}, \mathbf{v} \rangle \\ &= \sum_{i=1}^{D}\sum_{j=1}^{D} a_i b_i \langle \tfrac{\partial}{\partial\xi_i}, \tfrac{\partial}{\partial\xi_j} \rangle_{g_p} \\ &= \mathbf{a}^\top \boldsymbol{G}_p(\xi_1, \ldots, \xi_D)\mathbf{b} \end{aligned}$$

where $\boldsymbol{G}_p(\xi_1, \ldots, \xi_D)$ a is symmetric matrix and each of its entries are functions which we denoted as $G_{i,j}(\xi_1, \ldots, \xi_D)_p = \langle \partial/\partial\xi_i, \partial/\partial\xi_j \rangle_{g_p}$. Notice that, if the space is "flat" and the coordinate system (parametrisation) is Cartesian, the set of basis vectors is $\{e_1, \ldots, e_D\}$, the entries $G_{i,j}(\xi_1, \ldots, \xi_D)_p = \delta_{i,j}$ and g reduces to the common inner product, $g(\mathbf{u}, \mathbf{v}) = \mathbf{a}^\top \mathbf{b}$.

A *Riemannian manifold* is a smooth manifold $\mathcal{M}$ together with a choice of a metric tensor g in $T\mathcal{M}$ that is positive $\forall p \in \mathcal{M}$ (except for null vectors) and varies smoothly with p. This means that for a given parametrisation $\boldsymbol{\xi}$, the matrix entries $G_{i,j}(\xi_1, \ldots, \xi_D)_p$ are differentiable at $\boldsymbol{\xi}^{-1}(p)$ and the matrix $\boldsymbol{G}(\xi_1, \ldots, \xi_D)_p$ is positive-definite (hence invertible). In this case g is called *Riemannian metric* and $\boldsymbol{G}$ is the matrix which collects the coefficients of the metric.

In practical settings, the evident challenge lies in the choice of $\boldsymbol{G}$, which defines the manifold $(\mathcal{M}, g)$ from within (intrinsically). Usually, this choice requires extensive knownledge of the structure of the problem in question (Amari and Douglas, 1998; Nakahara, 2003). However, in the field of Statistics, Rao (1945) showed that the well-known Fisher information matrix (Schervish, 2011, Section 2.3) satisfies the properties of a metric tensor (see Calderhead, 2012, Section 3.2.3). This way, the pair $(\mathcal{M}, g)$, with $\boldsymbol{G}$ given by the Fisher information matrix, specifies a Riemannian manifold. Note that, in the majority of the probabilistic models used in practice the Fisher information matrix is known in closed-form (Johnson et al., 1995, 2005).

Now, consider a real-valued function $h : \mathcal{M} \to \mathbb{R}$ defined on a Riemmanian manifold $(\mathcal{M}, g)$ and a parametrisation $\boldsymbol{\xi} : U \subseteq \mathbb{R}^D \to \mathcal{M}$ of the manifold. Lets denote $\boldsymbol{\xi}^{-1}(p) = (\xi_1, \ldots, \xi_D) \in U$ where $p \in \mathcal{M}$. Petersen (2000), Section 2, pointed out that the *rate of change* of the function h in the direction of a

tangent vector $\mathbf{v} \in T_p\mathcal{M}$ (directional derivative) can be defined as

$$\begin{aligned} \mathrm{d}h_p(\mathbf{v}) &= \langle \mathbf{v}, \boldsymbol{G}_p^{-1}(\xi_1, \ldots, \xi_D)\nabla h \rangle \\ &= \sum_{i=1}^{D}\sum_{j=1}^{D} g_p^{i,j}\ \partial_{\xi_i} h\ \partial/\partial\xi_j. \end{aligned}$$

where $g_p^{i,j}$ are the elements of the inverse matrix $\boldsymbol{G}_p(\xi_1, \ldots, \xi_D)^{-1}$. Then, $\mathrm{d}h_p(\mathbf{v})$ is invariante under reparametrization. The symbol ∇ is the gradient of h with respect to the parametrisation $\boldsymbol{\xi}$ and the vector $\boldsymbol{G}(\xi_1, \ldots, \xi_D)^{-1}\nabla h$ is known as *natural gradient* due to Amari (1998)[4]. Finally, Amari (1998) showed that the choice of $\mathbf{v}$ in the direction of the natural gradient provides the stepest ascend direction of the function h (the highest rate of change) within a small neighbourhood of p.

In paper **[II]**, a closed-form expression of the Fisher information matrix for the Student-t model was derived by Fonseca et al. (2008). Moreover, we noticed that the model has a special type of parametrisation. The location and scale parameters are *orthogonal* in the sense of Jeffreys (1998)[5]. This enabled us to easily expand the models presented in Vanhatalo et al. (2009) and Jylänki et al. (2011) to heteroscedastic settings more easily. By exploiting this particular property and using the *natural gradient* (Amari, 1998), we were able to efficiently implement numerical optimization and consenquetly perform approximate inference with the Laplace's method with high stability of computer codes. This is in contrast with the tuning of computer algorithms presented by Vanhatalo et al. (2009) and Jylänki et al. (2011) for a less complex GP-model with the homocedastic Student-t probabilistic model.

3.4 Dealing with multiple-type observations

Multivariate statistical modelling often requires a joint probabilistic model for multivariate data (one can use the term simultaneous data). In this settings, the choice of the joint model imposes the type of dependency the data can assimilate. For example, in data types whose values are continuous, the multivariate Gaussian model is frequently used due to the easiness of interpretation of correlation parameters (also known as Pearson correlation). The practical interpretation is that they directly measure the strength of the linear dependency between two random variables.

However, in general, joint probabilistic models are usually difficult to formulate from basic principles for any type of data, and the dependency structure

[4] This has been long known in statistics as Fisher score, see Longford (1987)

[5] When off-diagonals entries of the Fisher information matrix are null, then the pair of parameters associated to those entries are called orthogonal parameters

might not have straightforward interpretation. There exists several types of dependency structure. This can be studied in the theory of copula functions, which is a good starting point for building multivariate probabilistic models and the study of statistical dependence. See for example the work by Nelsen (2006).

In many practical applications, experiments may provide us with a rich variety of databases of which are fraught with different data types. An easy way to account for all sources of information and introduce dependency among different data types is via the assumptions of hierachical model formulation (2.10). This idea is conducted in papers [**III**] and [**IV**]. In paper [**IV**], we exploit the hierarchical construction (2.10). The probabilistic model for each observable variable is Bernoulli, where the inverse link function is given by the one-dimensional Gaussian distribution function. The dependency is introduced via the MGP and we were able to show that the marginal distribution for the data have the form of (2.29). This distribution inherits Gaussian properties which is attractive and shows much more flexibility when compared to that of Ashford and Sowden (1970), Chib and Greenberg (1998) and Dai et al. (2013).

In paper [**III**], distinct probabilistic models for different types of data are combined into one single approach where the dependency is introduced via the MGP. By doing so, the predictive power of the model is increased and the estimates are more reliable. This is seen at least in the sense of smaller variances in the predictive posterior variance for the regression values when compared with independent GPs.

Chapter 4

Publication's summary

4.1 Article [**I**]

As the maximum reproductive rate represent the species renewal ability at low populational size levels, it has often been seen as an important factor to be measured in fisheries sciences. This measure is usually treated as time-invariant and usually species are treated separately. In real-case scenarios, interspecific-cooperation or interspecific-competition may be present and there is a clear interplay between species and its environment.

In this work, we allow the maximum reproductive rate to be time-varying and extend the Ricker stock-recruitment model to multispecies settings. We frame the statistical modelling under the Bayesian approach with hierarchical structure (2.10) and confront different models with real-data. We also investigate the performance of semiparametric discrepancy functions which have gained lot of interest in ecology more recently. The performance of all models is checked in terms of their posterior probabilities and leave-one-out cross-validation prediction task.

The data strongly support two models. The time-varying maximum reproductive rate with temporal cross-dependency between species and, in addition, the same model with inclusion of interspecific density-dependency. However, data, historical facts of changing ecosystem and expert knowledge reveal that the former model is more plausible. These findings have an important impact in practical policies. Usually, the maximum sustainable yield (MSY), which is a measure of sustainable harvesting of species population, is set considering species in isolation. This work shows strong evidence of temporal dependence between species which indicates that management decision must take the relationship between species into account.

4.2 Article [**II**]

The Student-t probabilistic model is commonly used in order to robustify data analysis in the presence of outliers. In the context of approximate inference with the LP method and GP priors, there has been difficulties in inference procedures due to lack of concavity of the log-likelihood function.

This paper extends previous models presented in Vanhatalo et al. (2009) and Jylänki et al. (2011) in two ways. First, we assume the regression noise is distributed according to the Student-t distribution and allow the noise level to vary as a function of covariates by putting a GP prior on the scale parameter of the model[1]. Second, in order to improve inference procedure, we exploited the orthogonal parametrisation which the Student-t model naturally possesses (or a parametrisation in which the Fisher information is diagonal, Section 3.3).

The heteroscedastic Student-t model shows better performance in all datasets analysed when compared to many other classical models previously used for those data in the literature. We also show that alternative Laplace approximation based on the Fisher information matrix and traditional Laplace approximation, give almost the same results in their respective approximate posterior distributions. Inference concerning on hyperparameters is also improved noting that there is stability of computer evaluation of the approximate marginal likelihood and similarity in results of experiments. Moreover, we have solved the instability problem in the Newton's method to find the mode of (2.19) with non log-concave likelihoods by using the natural gradient.

The approach presented in this paper can be used with any other probabilistic model whether the likelihood function for that model is concave or not. Aspects of parametrisation also deserves more careful attention since there is freedom of choice in the parametrisation of the model. Approximative methods and numerical procedures may have different performance with different parametrisation.

4.3 Articles [**III**]-[**IV**]

Most of traditional species distribution models do not take species interactions into account, hence the joint modelling of species distribution has been in increasing focus of attention (Elith and Leathwich, 2009; Ovaskainen et al., 2016c). Nowadays, with the fast advancement of computers power and technology, multispecies surveys provide us with great variety of information in large databases (e.g. global positioning system, remote sensing, photogrametry, geographic information system). Besides, traditional approaches to model species distribution can not be used to integrate different sources of data.

[1] This is also referred as heteroscedasticity in the noise.

In both of these articles multivariate data modelling is tackled using hierarchical MGPs where the dependency for the data is introduced in the second layer of the model building as presented in (2.10). This approach allow us to accommodate scenarios with missing observations and unequal amount of observations for different species. For species distribution models this is particularly important. For example, when dealing with rare or endangered species, data might be complicated to obtain due to difficulty of inaccessible regions or because there is a lack of knowledge of its presence, this makes data sparse, patchy or totally missing.

In particular, paper [**IV**] generalizes the probit Gaussian process model to multivariate binary settings and acommodate scenarios with missing observations in the entry of the binary random vector. As a by-product, we have shown a new multivariate Bernoulli model which is closed under marginalization and uncorrelatedness implies statistical independence. This is new for the literature of probabilistic models in statistics community and machine-learning. The EP algorithm develop by Cunningham et al. (2011) is central to evaluate the multivariate probability mass function with good precision.

Paper [**III**] focuses on the same hierarchical structure for multivariate data, but instead we consider different probabilistic models for different types of data. In particular, we have shown how this strategy for model building plays out in the case where we assume the Binomial and/or Negative-Binomial probabilistic model for the data. In the real case study, where we consider data with seven distinct species from the coastal region of the Gulf of Bothnia, the model which considers dependency clearly improves prediction task in extrapolation.

Both of the proposed models extend two recently well known species distribution models in the ecology literature. See the works by Pollock et al. (2014) and Ovaskainen et al. (2017).

4.4 Article [**V**]

This paper is a simulated study with the aim of showing the empirical performance of Hamiltonian Monte Carlo (HMC) and Riemannian manifold HMC (RMHMC) (with fixed metric) for Bayesian inference in the parameters of the extreme value models (Neal, 2011; Calderhead, 2012; Girolami and Calderhead, 2011; Coles and Powell, 1996; Coles, 2004). We also study the performance of those methods when modelling time dependence with a new autoregressive models where the error is distributed according to the extreme value probabilistic model. It is noted that first and second order geometric information involved in HMC and RMHMC are compensated by faster exploration of the parameter space when compared to standard Metropolis-Hastings algorithms.

This study shows that parameter estimation is relatively robust to the choice

of algorithm. HMC and RMHMC are much faster to reach stationary distribution by noting the smaller number of iterations in the simulations. Note also that both HMC and RMHMC requires only two parameters while Metropolis-Hastings algorithms requires the specification of a whole distribution such that this distribution is similar to the target distribution (Hastings, 1970; Chib and Greenberg, 1995).

Chapter 5

Future outlook and concluding remarks

This chapter summarizes the main contribution of this work. We also highlight and discuss future research directions in which the ideas presented throughout the thesis will be potentially extended.

5.1 Future outlook

There exists many extension for the type of modelling approach presented in this thesis. From the author's viewpoint, the inclusion of monotonicity constrains over the regression functions $f_1, \ldots, f_J$ in regions where we would have prior information about their increasing/decreasing value is attractive. See for example works by Riihimäki and Vehtari (2010) and Wang and Yang (2016). In the aforementioned multivariate settings, this can be now achieved straightforwardly by taking the derivates of the covariance function (2.7) w.r.t to some argument of that function (Abrahamsen, 1997). Recall the covariance function (2.7), denote distinct points for distinct processes as $\mathbf{x}_{j,i}$ and $\mathbf{x}_{j',i'}$. The covariance function expressing the dependency between the rate of change of $f_j(\mathbf{x}_{j,i})$ in the direction of the variable x_d with the value of any process $f_{j'}(\mathbf{x}_{j',i'})$ $(j' \neq j)$ reads

$$\operatorname{Cov}\left(\frac{\partial}{\partial_{x_d}} f_j(\mathbf{x}_{j,i}), f_{j'}(\mathbf{x}_{j',i'})\right) = \sum_{r=1}^{J} \frac{\partial}{\partial_{x_d}} \tilde{k}_r(\mathbf{x}_{j,i}, \mathbf{x}_{j',i'}) u_r(j, j') \tag{5.1}$$

and the covariance function expressing the dependency between the rate of change of $f_j(\mathbf{x}_{j,i})$ in the direction of the index variable x_d and the rate of change of $f_{j'}(\mathbf{x}_{j',i'})$ $(j' \neq j)$ in the direction of the variable $x_{d'}$, is given by

$$\operatorname{Cov}\left(\frac{\partial}{\partial_{x_d}} f_j(\mathbf{x}_{j,i}), \frac{\partial}{\partial_{x_{d'}}} f_{j'}(\mathbf{x}_{j',i'})\right) = \sum_{r=1}^{J} \frac{\partial^2}{\partial_{x_d} \partial_{x_{d'}}} \tilde{k}_r(\mathbf{x}_{j,i}, \mathbf{x}_{j',i'}) u_r(j, j') \tag{5.2}$$

Notice that, equations (5.1) and (5.2) take into account the dependency between all processes if they are all correlated. This way, we can exploit prior information about the monotonicity of different processes in different regions and the correlation between processes would "share" that monotonicity information among the processes. This is a promising approach in multivariate modelling for species distribution models, animal movement in multivariate settings (Hooten et al., 2017), inverse problems and other variety of applications. This has been in implementation phase (in Julia language and GPStuff) with minor applications in animal movement and SDMs.

In paper **[II]**, we have presented an alternative solution to the Laplace approximation of the posterior distribution when considering the heteroscedastic Student-t model and Gaussian process priors. In that case, we exploited the natural orthogonal parametrisation of the probabilistic model and showed similar performance between the classical Laplace approximation and the alternative Laplace approximation based on the Fisher information matrix. Therefore, the natural gradient approach and the alternative Laplace approximation presented in **[II]** deserves more attention. Nickisch and Rasmussen (2008) and Kuss and Rasmussen (2005) did an extension analysis for the quality of the approximation with LP and EP method. They concluded that the EP method performs better than LP in the GP classification case. However different parametrisation of the probabilitisc model may give different degrees of accuracy in the analytical approximation (Cox and Reid, 1987; Achcar and Smith, 1990; Achcar, 1994; Kass and Slate, 1994; MacKay, 1998).

In addition to the result presented in paper **[II]**, an example of reparametrisation for the Weibull model was presented. This is achieved by setting the off-diagonal terms from equality (Huzurbazar, 1956; Cox and Reid, 1987),

$$G_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = J^{\top} G_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}(\boldsymbol{\eta}))J \tag{5.3}$$

to 0 and we were able to find a pair of orthogonal parameters. As an extension of this approach, it would be interesting if this new parametrisation could be improved by choosing another reparametrisation such that the Fisher information matrix would be diagonal with constant diagonal terms[1].

Those previous aspects of reparametrisation are also important to advanced MCMC methods such as RMHMC or Metropolis-adjusted Langevin algorithms (MALA) (Calderhead, 2012). In this case, if the chosen parametrisation of the probabilistic model induces a diagonal Fisher information matrix G, then it is straightforward to see that computer eficiency would be improved as the implementation of the Hamiltonian dynamics is greatly simplified. See the work by Girolami and Calderhead (2011), Section 6, page 132.

[1] This would mean that the space is Euclidean.

In [**IV**] and [**III**] papers, the hierarchical construction (2.10) to deal with multi-type observation showed improved predictive power and a new multivariate Bernoulli model has been proposed (paper [**III**]). This is novel in statistics and machine-learning literature and might be interesting for practical applications at least. Those ideas foster future research for multi-type observations with distinct probabilistic models and even other types of multivariate Gaussian processes.

Still, another possibility would be to use the bijective mapping between the space of correlation matrices and $\mathbb{R}^{\binom{J}{2}}$ to introduce Gaussian process covariance regression. Following the hierarchical model building presented in (2.10), we could particularly write that

$$\begin{aligned} \boldsymbol{Y} \mid \boldsymbol{f} &\sim \mathcal{N}(0, \Sigma(\boldsymbol{f})) \\ \boldsymbol{f} \mid \boldsymbol{\theta} &\sim \mathcal{MGP} \\ \boldsymbol{\theta} &\sim \pi_{\text{hyper}}. \end{aligned} \tag{5.4}$$

In this case, $\boldsymbol{Y}$ has dimension J and the MGP would have dimension $\binom{J}{2} + J$. The notation $\Sigma(\cdot)$ is the mapping which transforms $\mathbb{R}^{\binom{J}{2}+J}$(taking the variance parameters into account) to the space of covariance matrices. The challenge with this model clearly resides on the computational complexity and implementation, which for large data-sets, would require sparse approximation for the full covariance matrices. Similar modelling approaches using GPs to model covariance matrix is presented by Fox and Dunson (2015), where GPs are introduced in the elements of a factor loading matrix via a latent factor model viewpoint.

All the aforementioned approaches can be put together to possibly investigate, for example, the performance of multivariate log-Gaussian Cox processes (Diggle et al., 2013), to better correct the observer bias with the inclusion of monotonicty constrains in multivariate GPs for presence-only data in species distribution models (Warton et al., 2013), and to improve GP methods in multi-objective Bayesian optimization (Swersky et al., 2013; Hernandez-Lobato et al., 2016).

5.2 Conclusions

In its own right, paper [**I**] introduces a new multivariate Ricker population model. Moreover, the paper shows that maximum reproductive rate provides us with great insight that sustainable harvesting must consider not only the relationship species-environment, but also species-to-species associations.

Paper [**II**] is of particular importance for all the other papers presented in this thesis. The use of natural gradient with notions of Riemannian geometry, naturally improves the inference process in multivariate GP-based models. This

comes without any additional difficulties in computational implementation and possibly simplifies the inference process.

The important contribution of papers [**III**]-[**IV**] lies in the alternative way to deal with multi-type observation in regression analysis under the GP formalism. Besides, we highlight how one can introduce statistical dependency in the second layer of the Bayesian hierarchical model and discuss the notion of dependency in statistical modelling. This is fundamental if one wants to enhance the capabilities of a probabilistic model used to accommodate real data behaviour.

All in all, this dissertation presents a building block on how to carefully construct Bayesian hierarchical models based on multivariate Gaussian processes. Although there exists many approaches in the literature (Gelfand et al., 2003; Boyle and Frean, 2004; Teh et al., 2005; Bonilla et al., 2008; Álvarez and Lawrence, 2011), the way in which the models are built in this dissertation have strong foundations and the methods presented here were not tackled before in GP-based modelling. This opens up an avenue for new models and foster new ideas.

Chapter 6

Positive-definite and positive-semidefinite matrices

The goal of this section is to make a clear meaning of what is a PD and PSD matrices throughout the thesis. For this we review some facts and definitions. Henceforth we will denote $\boldsymbol{M}$ as a real and symmetric matrix of dimensions $J \times J$ and its entries as $M_{j,j'}$ for $j, j' = 1, \ldots, J$.

Definition 3 (Positive-semidefinite matrix) *The matrix $\boldsymbol{M}$ is said to be positive-semidefinite if $\mathbf{a}^\top \boldsymbol{M}\, \mathbf{a} \geq 0 \;\forall\; \mathbf{a} \in \mathbb{R}^J$. (Note that it can happen $\mathbf{a} \neq \mathbf{0}$ and $\mathbf{a}^\top \boldsymbol{M}\, \mathbf{a} = 0$).*

Definition 4 (Positive-definite matrix) *A real and symmetric $J \times J$ matrix M is said to be positive-definite if $\mathbf{a}^\top \boldsymbol{M}\, \mathbf{a} > 0 \;\forall\; \mathbf{a} \in \mathbb{R}^J \setminus \{\mathbf{0}\}$.*

Theorem 6.1 *If M is positive-semidefinite, its diagonal elements are nonnegative. If M is positive-definite its diagonal elements are positive.*

Proof. Take $\mathbf{a} = (0 \ldots 0\ 1\ 0 \ldots 0)^\top$. Then $\mathbf{a}^\top \boldsymbol{M}\, \mathbf{a} = M_{j,j}$. The conclusion from the above definition is that $M_{j,j} \geq 0$ if $\boldsymbol{M}$ is PSD and $M_{j,j} > 0$ if $\boldsymbol{M}$ is PD. □

Lemma 1 *Let $\boldsymbol{M}$ be PSD. Then $\mathbf{a}^\top \boldsymbol{M}\, \mathbf{a} = 0$ if and only if $\boldsymbol{M}\, \mathbf{a} = \mathbf{0}$.*

Proof. If $\boldsymbol{M}\, \mathbf{a} = \mathbf{0}$ then $\mathbf{a}^\top \boldsymbol{M}\, \mathbf{a} = 0$. For the converse proof, consider the quadratic polynomial $p(\lambda)$ as

$$\begin{aligned} p(\lambda) &= (\mathbf{a} + \lambda \mathbf{b})^\top \boldsymbol{M} (\mathbf{a} + \lambda \mathbf{b}) \\ &= \mathbf{a}^\top \boldsymbol{M}\, \mathbf{a} + 2\lambda \mathbf{b}^\top \boldsymbol{M}\, \mathbf{a} + \lambda^2 \mathbf{b}^\top \boldsymbol{M}\, \mathbf{b} \end{aligned}$$

where $\mathbf{a}$ and $\mathbf{b}$ are of appropriate dimensions and λ is scalar. Then for all $\mathbf{a}$, $\mathbf{b}$ and λ we have

$$p(\lambda) \geq 0.$$

Then from the Bhaskara formula we get

$$\Delta = 4\big[(\mathbf{b}^\top \boldsymbol{M}\,\mathbf{a})^2 - (\mathbf{b}^\top \boldsymbol{M}\,\mathbf{b})(\mathbf{a}^\top \boldsymbol{M}\,\mathbf{a})\big] \leq 0$$

which must be nonpositive. This expression shows that, if $\mathbf{a}^\top \boldsymbol{M}\,\mathbf{a} = 0$ then $\Delta = 0$ only if $\mathbf{b}^\top \boldsymbol{M}\,\mathbf{a} = 0$. But this would hold $\forall\ \mathbf{b}$. Hence $\boldsymbol{M}\,\mathbf{a} = 0$. □

Theorem 6.2 *M is nonsingular if and only if it is PD.*

Proof. If M is PSD then $\mathbf{a}^\top \boldsymbol{M}\,\mathbf{a} \geq 0$. By the previous Lemma $\mathbf{a}^\top \boldsymbol{M}\,\mathbf{a} = 0$ if and only if $M\mathbf{a} = 0$. Suppose that M is PD. Then, $\mathbf{a}^\top \boldsymbol{M}\,\mathbf{a} = 0$ if and only if $\mathbf{a} = 0$. Therefore, $\boldsymbol{M}\,\mathbf{a} = 0$ only if $\mathbf{a} = \mathbf{0}$. Then, $\boldsymbol{M}$ is nonsingular (remember the null space of a matrix). Conversely, if $\boldsymbol{M}$ is nonsingular, $\boldsymbol{M}\,\mathbf{a} = 0$ only if $\mathbf{a} = \mathbf{0}$. Then, $\mathbf{a}^\top \boldsymbol{M}\,\mathbf{a} = 0$. Therefore M is PD. □

Corollary 1 *If $\boldsymbol{M}$ is PD then its inverse M^{-1} is also PD.*

Proof. $\boldsymbol{M}$ is PD, then $\mathbf{a}^\top \boldsymbol{M}\,\mathbf{a} = \mathbf{a}^\top \boldsymbol{M}\,\boldsymbol{M}^{-1}\,\boldsymbol{M}\,\mathbf{a} = \mathbf{c}^\top \boldsymbol{M}^{-1}\,\mathbf{c} > 0$, where $\mathbf{c} = \boldsymbol{M}\,\mathbf{a}$. We have $\mathbf{c}^\top \boldsymbol{M}^{-1}\,\mathbf{c} = 0$ if and only if $\mathbf{c} = \mathbf{0}$. □

Corollary 2 *If $\boldsymbol{M}$ is PSD but not PD then it is singular.*

Proof. From the previous Theorem 6.2, the matrix $\boldsymbol{M}$ is PD if and only if $\boldsymbol{M}$ is nonsingular. Therefore, if it is not positive, it is singular. □

References

Abrahamsen, P. (1997). A review of Gaussian random fields and correlation functions. Technical report, Norwegian Computing Center.

Achcar, J. A. (1994). Some aspects of reparametrization in statistical models. *Pakistan Journal of Statistics*, 10(3):597–616.

Achcar, J. A. and Smith, A. F. (1990). Aspects of reparametrization in approximate Bayesian inference. *Bayesian and Likelihood Methods in Statistics and Econometrics*, 4(2):439–452.

Akbarov, A. (2009). *Probability elicitation: Predictive approach*. PhD thesis, University of Salford.

Álvarez, M. A. and Lawrence, N. (2011). Computationally Efficient Convolved Multiple Output Gaussian Processes. *Journal of Machine Learning*, (12):1425–1466.

Amari, S. (1998). Natural Gradient Works Efficiently in Learning. *Neural Computation (communicated by Steven Nowlan and Erkki Oja)*, 10:251–276.

Amari, S. and Douglas, S. C. (1998). Why natural gradient ? In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1998*, volume 2, pages 1213–1216.

Angelieri, C. C. S., Christine, A.-H., de Barros, F. K. M. P. M., de Souza Marcelo Pereira, and Alexander, M. C. (2016). Using Species Distribution Models to Predict Potential Landscape Restoration Effects on Puma Conservation. *PLOS ONE*, 11(1):1–18.

Ashford, J. R. and Sowden, R. R. (1970). Multivariate probit analysis. *Biometrics*, 26:535–46.

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, 200(1-2):1–19.

Bain, L. and Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics*. Brooks/Cole.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2015). *Hierarchical Modelling and Analysis for Spatial Data*. Chapman Hall/CRC, second edition.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modelling covariance matrices in terms of standart deviations and correlations, with applications to shrinkage. *Statistical Sinica*, 4(10):1281–1311.

Bergström, U., Sundblad, G., Downie, A.-L., Snickars, M., Boström, C., and Lindegarth, M. (2013). Evaluating eutrophication management scenarios in the Baltic Sea using species distribution modelling. *Journal of Applied Ecology*.

Berliner, L. M. (1991). Likelihood and Bayesian prediction of chaotic systems. *Journal of the American Statistical Association*, 86(416):938–952.

Bernardo, J.-M. and Smith, A. F. M. (1994). *Bayesian Theory*, volume 90. John Wiley and Sons.

Bonilla, E. V., Chai, K. M., and Williams, C. (2008). Multi-task Gaussian Process Prediction. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 153–160.

Boyle, P. and Frean, M. (2004). Dependent Gaussian Processes. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'04, pages 217–224. MIT Press.

Brännstron, Å. and Sumpter, D. J. T. (2005). The role of competition and clustering in population dynamics. *Proceedings of Royal Society B*, 272:2065–2072.

Calderhead, B. (2012). *Differential geometric MCMC methods and applications*. PhD thesis, University of Glasgow.

Calderhead, B. and Girolami, M. (2011). Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *Interface Focus*.

Campbell, D. and Chkrebtii, O. (2013). Maximum profile likelihood estimation of differential equation parameters through model based smoothing state estimates. *Mathematical Biosciences*, 246(2):283–292.

Casella, G. and Berger, R. (2002). *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–361.

Chkrebtii, O. A., Campbell, D. A., Calderhead, B., and Girolami, M. A. (2016). Bayesian solution uncertainty quantification for differential equations. *Bayesian Analysis*, 11(4):1239–1267.

Clark, J. S., Gelfand, A., Woodall, C. W., and Zhu, K. (2014). More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, 24(5):990–999.

Coles, S. (2004). *An Introduction to Statistical Modelling of Extreme Values*. Springer Series in Statistics.

Coles, S. G. and Powell, E. A. (1996). Bayesian Methods in Extreme Value Modelling: A Review and New Developments. *International Statistical Review*, 64(1):119–136.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–39.

Cressie, N. and While, C. K. (2011). *Statistics for Spatial-Temporal Data*. Wiley Series in Probability and Statistics.

Cunningham, J. P., Hennig, P., and Lacoste-Julien, S. (2011). Gaussian Probabilities and Expectation Propagation. *ArXiv e-prints*.

Dai, B., Ding, S., and Wahba, G. (2013). Multivariate Bernoulli distribution. *Bernoulli*, 19(4):1465–1483.

De Finetti, B. (1975). *Theory of Probability*, volume 1-2. Wiley, New York.

Dehaene, G. and Barthelmé, S. (2018). Expectation propagation in the large data limit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):199–217.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-gaussian cox processes. extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563.

Do Carmo, M. (2013). *Riemannian Geometry*. Mathematics: Theory & Applications. Birkhäuser Boston.

Elith, J. and Leathwich, J. R. (2009). Species Distributions Models: Ecological Explanation and Predictions Across Space and Time. *The Annual Review of Ecology, Evolution and Systematics*, 40(677-697).

Fergusson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistic*, 1(2):209–230.

Fonseca, T. C. O., Ferreira, M. A. R., and Migon, H. S. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika*, 95(2):325.

Fox, E. B. and Dunson, D. B. (2015). Bayesian nonparametric covariance regression. *Journal of Machine Learning research*, 16(1):2501–2542.

Fricker, T. E., Oakley, J. E., and Urban, N. M. (2013). Multivariate Gaussian process emulators with nonseparable covariance structures. *Technometrics*, 55(1):47–56.

Gauss, C. F. (1807). *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections: A Translation of Gauss's Theoria Motus with an Appendix (by Charles Henry Davis, T)*. Wentworth Press.

Gelfand, A., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial Modelling with Spatially Varying Coefficient Processes. *Journal of American Statistical Association*, 98(462):387–396.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Learning*, 6(6):721–741.

Gibbs, M. N. (1997). *Bayesian Gaussian Processes for Regression and Classification.* PhD thesis, Department of Physics, University of Cambridge.

Giri, N., Birnbaum, Z., and Lukacs, E. (2014). *Multivariate Statistical Inference*. Probability and mathematical statistics. Elsevier Science.

Girolami, M. and Calderhead, B. (2011). Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods. *Journal of the Statistical Royal Society B*, 73(2):123–214.

Golding, N. and Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, 7:598–608.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press.

Gosling, J. (2005). *Elicitation: A nonparametric view*. PhD thesis, University of Sheffield.

Grzebyk, M. and Wackernagel, H. (1994). Multivariate analysis and spatial/temporal scales: Real and complex models. In *XVII-th International Biometric Conference*, pages 19–33.

Guisan, A., Edwards, T. C., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, 157(2-3):89–100.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M. R., Possingham, H. P., and Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, 16(12):1424–1435.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and their Applications. *Biometrika*, 57(1):97–109.

Hauptmann, A. (2017). *Advances in D-Bar methods for partial boundary data electrical impedance tomographY - From continuum to electrode models and back*. PhD thesis, University of Helsinki.

Hernandez-Lobato, D., Hernandez-Lobato, J., Shah, A., and Adams, R. (2016). Predictive entropy search for multi-objective bayesian optimization. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1492–1501, New York, New York, USA.

Hilborn, R. and Walters, C. J. (1992). *Quantitative fisheries stock assessment: choice, dynamics and uncertainty*. Springer US, 1 edition.

Hooten, M., Johnson, D., McClintock, B., and Morales, J. (2017). *Animal Movement: Statistical Models for Telemetry Data*. CRC Press.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, 2nd edition.

Huzurbazar, V. (1956). Sufficient statistics and orthogonal parameters. *The Indian journal of Statistics*, 17(3).

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffreys, H. (1998). *The Theory of Probability*. Oxford Classic Texts in the Physical Sciences. OUP Oxford, 3rd edition.

Johnson, N., Kemp, A., and Kotz, S. (2005). *Univariate Discrete Distributions*. Wiley Series in Probability and Statistics. Wiley.

Johnson, N., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley & Sons, 2nd edition.

Jylänki, P., Vanhatalo, J., and Vehtari, A. (2011). Robust Gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research*, 12(12):3227–3257.

Kaipio, J. and Somersalo, E. (2005). *Statistical and Computational Inverse Problems*. Springer.

Kaipio, J. and Somersalo, E. (2007). Statistical inverse problems: Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics*, 198(2):493–504. Special Issue: Applied Computational Inverse Problems.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, (82 (Series D)):35–45.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statistical Science*, 4(3):188–219.

Kass, R. E. and Slate, E. H. (1994). Some diagnostics of maximum likelihood and posterior nonnormality. *The Annals of Statistics*, 22(2):668–695.

Knight, K. (1999). *Mathematical Statistics*. Chapman and Hall/CRC.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *Annals of Mathematical Statististics*, 22(1):79–86.

Kuo, H. (2005). *Introduction to Stochastic Integration*. Springer New York.

Kurowicka, D. and Cooke, R. (2003). A parameterization of positive definite matrices in terms of partial correlation vines. *Linear Algebra and its Applications*, 372(Supplement C):225–251.

Kuss, M. and Rasmussen, C. E. (2005). Assessing approximate inference for binary Gaussian process classification. *Journal of machine learning research*, 6:1679–1704.

Lawless, J. F. (2002). *Statistical Models and Methods for Lifetime Data*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd edition.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74(4):817–827.

MacKay, D. J. (1998). Choice of basis for Laplace approximation. *Machine Learning*, 33(1):77–86.

Mardia, K. V. and Goodall, C. R. (1993). Spatio-Temporal analysis of Multivariate Environmental Monitoring Data. *Multivariate Environmental Statistics*, pages 347–386.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. A., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Minka, T. (2001a). *A family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Tecnology.

Minka, T. P. (2001b). Expectation Propagation for Approximate Bayesian Inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 362–369.

Moala, F. (2006). *Elicitation of multivariate prior distribution*. PhD thesis, University of Sheffield.

Moala, F. and O'Hagan, A. (2010). Elicitation of multivariate prior distributions: A nonparametric Bayesian approach. *Journal of Statistical Planning and Inference*, 140:1635–1655.

Murray, J. D. (2004). *Mathematical Biology I: An Introduction*. Springer. Interdisciplinary applied mathematics.

Myers, R. A. (2001). Stock and recruitment: generalizations about maximum reproductive rate, density dependence, and variability using meta-analytic approaches. *Journal Of Marine Sciences*, 58:937–951.

Myers, R. A., Mertz, G., and Bridson, J. (1997). Spatial scales of interannual recruitment variations of marine, anadromous, and freshwater fish. *Canadian Journal of Fisheries and Aquatic Sciences*, 54(6):1400–1407.

Nakahara, M. (2003). *Geometry, Topology and Physics, Second Edition*. Graduate student series in physics. Taylor & Francis.

Neal, R. (2003). Slice Sampling. *The Annals of Statistics*, 31(3):705–767.

Neal, R. M. (1995). *Bayesian Learning for Neural Networks*. PhD thesis.

Neal, R. M. (1998). Regression and Classification using Gaussian Process Priors. *Bayesian Statistics*, 6.

Neal, R. M. (2011). *Handbook of Markov Chain Monte Carlo*, chapter 5. Chapman and Hall CRC Press.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Association*, 135(3):370–384.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer Series in Statistics.

Nickisch, H. and Rasmussen, C. E. (2008). Approximations for binary Gaussian process classification. *Journal of Machine learning*, 9:2035–2078.

Niu, M., Cheung, P., Lin, L., Dai, Z., Lawrence, N., and Dunson, D. (2018). Intrinsic Gaussian processes on complex constrained domains. *ArXiv e-prints*.

Oakley, J. E. and O'Hagan, A. (2007). Uncertainty in prior elicitations: A nonparametric approach. *Biometrika*, 94.

O'Hagan, A. (1978). Curve Fitting and Optimal Design for Prediction. *Journal of Royal Statistical Society B*, 40(1):1–42.

O'Hagan, A. (2004). *Kendall's Advanced Theory of Statistics: Bayesian Inference*. Oxford University Press.

Øksendal, B. (2013). *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer Berlin Heidelberg.

Ovaskainen, O., Abrego, N., Halme, P., and Dunson, D. (2016a). Using latent variable models to identify species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*, 7(5):549–555.

Ovaskainen, O., Knegt, H. J., and Delgado, M. d. M. (2016b). *Quantitative Ecology and Evolutionary Biology. Integrating models with data*. Oxford University Press.

Ovaskainen, O., Rekola, H., Meyke, E., and Arjas, E. (2008). Bayesian methods for analysing movements in heterogeneous landscapes from mark recapture data. *Ecology*, 89(2):542–554.

Ovaskainen, O., Roy, D. B., Fox, R., Fox, R., and Anderson, B. J. (2016c). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution*, 7:428–436.

Ovaskainen, O. and Soinnien, J. (2011). Making more out of sparse data: Hierarchical modelling of species communities. *Ecology*, 92(2):289–295.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., and Abrego, N. (2017). How to make more out of community data? a conceptual framework and its implementation as models and software. *Ecology Letters*, 20(5):561–576.

Pawitan, Y. (2005). *In all likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Petersen, P. (2000). *Riemannian Geometry*. Graduate Texts in Mathematics. Springer International Publishing.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., Hara, R. B. O., Parris, K. M., Vesk, P. A., and McCarthy, M. A. (2014). Understanding co-occurence by modelling species simultaneously with joint species distribution model (JSDM). *Methods in Ecology and Evolution*, 5(5):397–406.

Pressley, A. (2001). *Elementary Differential Geometry*. Springer undergraduate mathematics series. Springer.

Quinn, T. and Deriso, R. (1999). *Quantitative fish dynamics*. Biological Resource Management. Oxford University Press.

Rao, R. C. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta mathematical society*, 37:81–91.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*, volume 11. The MIT Press.

Riihimäki, J. (2013). *Advances in Approximate Bayesian Inference for Gaussian Process Models*. PhD thesis.

Riihimäki, J., Jylänki, P., and Vehtari, A. (2013). Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. *Journal of Machine Learning*, 14:75–109.

Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. In *Journal of Machine Learning Research: Workshop and Conference Proceedings, 9:645-652, (AISTATS 2010 Proceedings)*.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Text in Statistics.

Rose, K. A., Junior, J. A. C., Winemiller, K. O., Myers, R. A., and Hilborn, R. (2001). Compensatory density dependence in fish populations: importance, controversy, understanding and prognosis. *Fish and Fisheries*, 2:293–327.

Rousseeuw, P. J. and Molenberghs, G. (1994). The shape of correlation matrices. *The American Statistician*, (48):276–9.

Rue, H., Marino, S., and Chopin, N. (2009). Approximate Bayesian Inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society*, 71(2):319–392.

Schervish, M. J. (2011). *Theory of Statistics*. Springer Series in Statistics.

Seber, G. A. F. and Lee, A. J. (2012). *Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley.

Seeger, M. (2005). Expectation propagation for exponential families. Technical report.

Simpson, D. P., Rue, H., Martins, T. G., Riebler, A., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28.

Sinclair, S. J., White, M. D., and Newell, G. R. (2010). How useful are species distribution models for managing biodiversity under future climates. *Ecology and Society. Synthesis*, 15(1).

Sorenson, H. W. (1970). Least-squares estimation: from Gauss to Kalman. *IEEE Spectrum*, 7(7):63–68.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging (Springer Series in Statistics)*. Springer, 1 edition.

Stephenson, W. and Broderick, T. (2016). Understanding covariance estimates in expectation-propagation. In *Advances in approximate Bayesian inference, Neural information processing (NIPS)*.

Swersky, K., Snoek, J., and Adams, R. P. (2013). Multi-task Bayesian optimization. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 2004–2012. Curran Associates, Inc.

Teh, Y. W., Seeger, M., and Jordan, M. I. (2005). Semiparametric latent factor models. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS 2005, Bridgetown, Barbados, January 6-8*.

Tierney, L. and Kadane, J. B. (1986). Accurate Approximation for Posterior Moments and Marginal Densities. *Journal of American Statistical Association*, 81(393):82–86.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions. *Journal of the American Statistical Association*, 84(407):710–716.

Tokuda, T., Goodrich, B., Mechelen, I. V., and Gelman, A. (2012). Visualizing Distributions of Covariance Matrices.

Vandenberg-Rodes, A. and Shahbaba, B. (2015). Dependent Matérn Processes for Multivariate Time Series. *ArXiv*.

Vanhatalo, J. (2010). *Speeding up the inference in Gaussian process models*. PhD thesis, University of Helsinki.

Vanhatalo, J., Jylänki, P., and Vehtari, A. (2009). Gaussian process regression with a Student-t likelihood. *Advances in Neural Information Processing Systems*.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2013). GPstuff : Bayesian Modeling with Gaussian Processes. *Journal of Machine Learning Research*, 14(1):1175–1179.

Vanhatalo, J., Venerante, L., and Hudd, R. (2012). Species distribution modelling with Gaussian processes: A case study with youngest stages of sea spawning whitefish (Coregonus lavatus L. s.l.) larvae. *Ecological Modelling*, (228):49–58.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

Wang, M. and Yang, M. (2016). Posterior property of Student-t linear regression model using objective priors. *Statistics and Probability Letters*, 113:23–29.

Wang, Y. and Barber, D. (2014). Gaussian processes for Bayesian estimation in ordinary differential equations. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1485–II–1493.

Warton, D. I., Renner, I. W., and Ramp, D. (2013). Model-Based control of observer Bias for the Analysis of Presence-only data in ecology. *PLoS One*, 8(11):1–9.

Wechsler, S., Izbicki, R., and Esteves, L. G. (2013). A Bayesian look at non-identifiability: A simple example. *The American Statistician*, 67:90–93.

Wikle, C. K. (2003). Hierarchical Models in Environmental Science. *International Statistical Review*, 71(2):181–199.

Wild, C, J. and Seber, G, A. F. (1989). *Nonlinear regression*. New York: Wiley.

Williams, C. K. I. and Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:1342–1351.

Wilson, A. and Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1067–1075.

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