Department of Biomedical Engineering and Computational Science

Advances in Approximate Bayesian Inference for Gaussian Process Models

Jaakko Riihimäki

Aalto University

DOCTORAL
DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 144/2013

Advances in Approximate Bayesian Inference for Gaussian Process Models

Jaakko Riihimäki

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the auditorium F239a of the school on 11 October 2013 at 12.

Aalto University
School of Science
Department of Biomedical Engineering and Computational Science

Supervising professor
Prof. Jouko Lampinen

Thesis advisor
Dr. Aki Vehtari

Preliminary examiners
Prof. Mark Girolami, University College London, UK
Prof. Håvard Rue, Norwegian University of Science and Technology, Norway

Opponent
Associate Prof. Ole Winther, Technical University of Denmark, Denmark

Aalto University publication series
DOCTORAL DISSERTATIONS 144/2013

ISBN 978-952-60-5332-5 (printed)
ISBN 978-952-60-5333-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5333-2

Unigrafia Oy
Helsinki 2013

Finland

NORDIC ECOLABEL
441 697
Printed matter

Aalto University

Abstract

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

Author
Jaakko Riihimäki

Name of the doctoral dissertation
Advances in Approximate Bayesian Inference for Gaussian Process Models

Publisher School of Science

Unit Department of Biomedical Engineering and Computational Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 144/2013

Field of research Computational Engineering

Manuscript submitted 11 June 2013 **Date of the defence** 11 October 2013

Permission to publish granted (date) 16 August 2013 **Language** English

□ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

Gaussian processes (GPs) provide a flexible approach to construct probabilistic models for Bayesian data analysis. In the Bayesian approach, GPs are used to specify prior assumptions on the latent function values that describe the underlying relationships between the explanatory variables and the associated target variables. These prior assumptions are combined with information from the observations using Bayes' rule. The obtained result is the posterior distribution that represents the uncertainty about the latent function values of interest, conditioned on the observations and the model assumptions. A challenge with the Bayesian approach is that exact inference is analytically intractable to calculate for most GP models used in practice. Therefore, approximate methods are needed in order to evaluate the posterior distribution and to make predictions for new observations.

This thesis develops methods for approximate Bayesian inference in various modelling problems involving GP models. The focus is on efficient ways to form Gaussian posterior approximations based on Laplace's method or expectation propagation (EP). The inference for the studied GP models is challenging in two aspects. Firstly, observation models are generalized in the way that the probability distribution for each observation can depend on multiple values of the latent function instead of only one value, or on the derivative values of the latent function. Secondly, instead of one prior process, the models can have multiple uncorrelated prior processes that are coupled through the observation model.

This thesis presents improvements to approximate Bayesian inference for GP models in density estimation, survival analysis, and multiclass classification. We describe Laplace's method for a logistic GP model and for a Cox-type survival model constructed from GP priors to speed up the inference. We develop a novel nested EP algorithm for multinomial probit GP classification that does not require numerical quadratures and scales linearly in the number of classes. In addition, we extend the existing methodology proposed for regression and binary classification by introducing monotonicity information into a GP model with EP. We demonstrate the practical accuracy of the described methods with several experiments and apply them to real-life modelling problems.

Keywords Bayesian modelling, approximate inference, Gaussian processes, Laplace's method, expectation propagation

ISBN (printed) 978-952-60-5332-5 **ISBN (pdf)** 978-952-60-5333-2

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

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2013

Pages 173 **urn** http://urn.fi/URN:ISBN:978-952-60-5333-2

A? Aalto-yliopisto

Tiivistelmä

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

Tekijä
Jaakko Riihimäki

Väitöskirjan nimi
Menetelmiä likimääräiseen bayesilaiseen päättelyyn gaussisia prosesseja käyttäville malleille

Julkaisija Perustieteiden korkeakoulu

Yksikkö Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 144/2013

Tutkimusala Laskennallinen tekniikka

Käsikirjoituksen pvm 11.06.2013 **Väitöspäivä** 11.10.2013

Julkaisuluvan myöntämispäivä 16.08.2013 **Kieli** Englanti

□ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Gaussiset prosessit (GP) mahdollistavat joustavan lähestymistavan todennäköisyysmallien muodostamiseen bayesilaisessa tilastotieteessä. Bayesilaisessa päättelyssä gaussisilla prosesseilla voidaan määritellä priorioletuksia latentista funktiosta, jolla mallinnetaan selittävien ja ennustettavien muuttujien välistä tuntematonta yhteyttä. Näitä priorioletuksia päivitetään havainnoista saatavalla tiedolla Bayesin kaavaa käyttäen. Tuloksena saadaan posteriorijakauma, joka esittää tarkasteltavien latentin funktion arvojen epävarmuutta ehdollistettuna havainnoille ja mallioletuksille. Bayesilaisen mallinnuksen haasteena on, että täsmällinen päättely on useimmille GP-malleille laskennallisesti vaikeaa. Siksi posteriorijakauman ja uusien havaintojen ennusteiden laskemisessa joudutaan usein käyttämään likimääräisiä menetelmiä.

Tässä väitöskirjassa kehitetään menetelmiä, jotka mahdollistavat likimääräisen bayesilaisen päättelyn GP-malleille erilaisissa mallinnusongelmissa. Väitöstyö keskittyy tehokkaisiin tapoihin muodostaa gaussisia posteriorijakauma-approksimaatioita käyttäen Laplacen menetelmää tai expectation propagation (EP) -algoritmia. Päättelyyn liittyvä laskenta tutkittavilla GP-malleilla on haastavaa kahdessa suhteessa. Ensinnäkin havaintomalleja on yleistetty siten, että jokaiseen havaintoon liittyvä todennäköisyysjakauma voi riippua latentin funktion derivaatan arvosta tai useammasta kuin yhdestä latentin funktion arvosta. Toiseksi tutkittavat GP-mallit voivat rakentua yhden prioriprosessin sijaan useammasta riippumattomasta prioriprosessista, jotka kytkeytyvät toisiinsa havaintomallin kautta.

Tässä väitöstyössä esitetään menetelmiä likimääräiseen bayesilaiseen päättelyyn GP-malleille tiheysjakauman estimoinnissa, elinaika-analyysissa sekä usean luokan luokitteluongelmassa. Työssä kuvataan Laplacen menetelmä logistiselle GP-mallille sekä GP-prioreista rakennetulle Coxin elinaikamallille laskennan nopeuttamiseksi. Työssä kehitetään multinomi-probit-havaintomallille uudenlainen sisäkkäinen EP-algoritmi, joka ei tarvitse numeerisia integrointimenetelmiä toimiakseen, ja jonka laskennallinen rasite kasvaa enintään lineaarisesti luokkien lukumäärän suhteen. Lisäksi työssä esitetään regressio- ja luokittelumalleille menetelmä, joka mahdollistaa monotonisuustiedon lisäämisen GP-prioreihin EP-algoritmin avulla. Kuvattujen menetelmien tarkkuutta tutkitaan työssä monenlaisien kokeiden avulla ja menetelmiä sovelletaan käytännön mallinnusongelmiin.

Avainsanat bayesilainen mallintaminen, likimääräinen päättely, gaussiset prosessit, Laplacen menetelmä, expectation propagation -algoritmi

ISBN (painettu) 978-952-60-5332-5 **ISBN (pdf)** 978-952-60-5333-2

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

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2013

Sivumäärä 173 **urn** http://urn.fi/URN:ISBN:978-952-60-5333-2

Preface

This thesis is the result of my research work that has been carried out at the Laboratory of Computational Engineering at Helsinki University of Technology and later at the Department of Biomedical Engineering and Computational Science at Aalto University. I gratefully acknowledge the financial support provided by Academy of Finland and by Finnish Funding Agency for Technology and Innovation.

First and foremost, I would like to express my gratitude to my instructor Dr. Aki Vehtari for his inspiring guidance and encouragement throughout these years. I appreciate how you always found time to share your knowledge of Bayesian statistics that has been invaluable to this work. I am also grateful to Prof. Jouko Lampinen for supervising this thesis and for providing the excellent research facilities.

I am thankful to Prof. Mark Girolami and Prof. Håvard Rue for acting as the preliminary examiners of this thesis and for providing encouraging comments on the manuscript.

I wish to express my gratitude to all my collaborators during my doctoral studies. Especially, I would like to thank Prof. Heikki Joensuu, Dr. Reijo Sund, and Dr. Aki Havulinna for productive research collaboration.

My warmest thanks to all my colleagues and friends, including Pasi Jylänki, Dr. Jouni Hartikainen, Juho Kettunen, Janne Ojanen, Dr. Jarno Vanhatalo, Dr. Tommi Mononen, Dr. Eli Parviainen, Dr. Simo Särkkä, Dr. Mari Myllymäki, Tomi Peltola, Arno Solin, Ville Tolvanen, Juho Kokkala, Dr. Pekka Marttinen, Dr. Jussi Kumpula, Dr. Riitta Toivonen, Dr. Riku Linna, Dr. Margareta Segerståhl, among others, who created a pleasant and inspiring working atmosphere. I am grateful for your collaboration and for all the interesting coffee-break discussions we have had during these years. I would also like to thank Eeva Lampinen, Laura Pyysalo, Katri Kaunismaa, Katja Korpinurmi, and Marita Stenman for helping with work facilities and practicalities, and the IT support personnel for helping with computational resources.

Beyond the academic world, I want to thank all my friends, with special thanks to Tommi, Jari, Matti, and Elina for encouragement. Finally, I express my gratitude to my family for supporting me during these years and for always being there for me.

Helsinki, September 11, 2013,

Jaakko Riihimäki

Contents

List of Publications

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

I Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. In *Journal of Machine Learning Research: Workshop and Conference Proceedings*, Volume 9: AISTATS 2010, pages 645–652, Chia Laguna Resort, Sardinia, Italy, May 2010.

II Jaakko Riihimäki and Aki Vehtari. Laplace approximation for logistic Gaussian process density estimation. *Submitted to Bayesian Analysis, 20 pages, preprint: arXiv:1211.0174v1*, 2012.

III Jaakko Riihimäki, Pasi Jylänki and Aki Vehtari. Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. *Journal of Machine Learning Research*, 14(Jan):75–109, 2013.

IV Heikki Joensuu, Aki Vehtari, Jaakko Riihimäki, Toshirou Nishida, Sonja E Steigen, Peter Brabec, Lukas Plank, Bengt Nilsson, Claudia Cirilli, Chiara Braconi, Andrea Bordoni, Magnus K Magnusson, Zdenek Linke, Jozef Sufliarsky, Federico Massimo, Jon G Jonasson, Angelo Paolo Dei Tos and Piotr Rutkowski. Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts. *The Lancet Oncology*, 13(3):265–274, 2012.

V Jaakko Riihimäki, Reijo Sund and Aki Vehtari. Analysing the length of care episode after hip fracture: a nonparametric and a parametric Bayesian approach. *Health Care Management Science*, 13(2):170–181, 2010.

VI Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen and Aki Vehtari. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research*, 14(Apr):1175–1179, 2013.

Author's Contribution

Publication I: "Gaussian processes with monotonicity information"

Riihimäki had the main responsibility in writing the article, implementing the described models and methods, and running all the experiments. Vehtari contributed to the initiation of the research problem and in background considerations.

Publication II: "Laplace approximation for logistic Gaussian process density estimation"

Riihimäki had the main responsibility in writing the article and implementing the described models and methods. The original idea was proposed by Vehtari who also participated in writing the paper and running the experiments.

Publication III: "Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood"

Riihimäki and Jylänki developed the methodology and wrote the article jointly. Jylänki contributed more in the theoretical derivations and Riihimäki had the main responsibility for implementing and running the experiments. Vehtari contributed to the methodological developments through discussions and commented on the manuscript.

Publication IV: "Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts"

Joensuu designed the study, wrote the study protocol, did the literature search, and compiled the study database. TN, SES, PB, LP, BN, CC, CB, AB, MKM, ZL, JS, FM, JGJ, APDT, and PR collected and provided the original data. TN, SES, PB, LP, BN, CC, CB, AB, MKM, FM, JGJ, and PR reviewed the transmitted data. Joensuu, Vehtari, and Riihimäki did the statistical analyses. Vehtari and Riihimäki designed the nonlinear model. Joensuu drafted the manuscript, and Vehtari and Riihimäki participated in writing. All authors reviewed and commented on the manuscript.

Publication V: "Analysing the length of care episode after hip fracture: a nonparametric and a parametric Bayesian approach"

Riihimäki had the main responsibility in writing the article and running the experiments. The expertise for the application was provided by Sund who also participated in writing the article and in analysing the results. Sund and Vehtari contributed to the initiation of the research problem and in background considerations.

Publication VI: "GPstuff: Bayesian modeling with Gaussian processes"

This publication is a link to a computer software package in which Riihimäki implemented methodology described in Publications II–IV and contributed also to other parts in developing the software.

1. Introduction

Gaussian processes (GPs) are popular methods for analysing data in a wide range of applications, such as regression and classification (O'Hagan, 1978; Williams and Rasmussen, 1996; MacKay, 1998; Neal, 1998; Williams and Barber, 1998), density estimation (Leonard, 1978; Tokdar, 2007), dimension reduction (Lawrence, 2005), and spatial statistics (Cressie, 1993; Best et al., 2005). GPs provide a flexible approach to construct models for making probability-based Bayesian inference from data and computing predictions for new observations. As an example, nonlinear effects and interactions between explanatory variables can be modelled with GPs without explicitly specifying parametric forms for relationships among the variables.

The Bayesian approach provides a unified framework to express uncertainties as probabilities and to combine information from different sources (e.g. Bernardo and Smith, 2000; Gelman et al., 2003). In the Bayesian approach, GPs can be used to specify prior assumptions on the latent function values that describe the underlying relationships between the explanatory variables and the associated target variables (e.g. Rasmussen and Williams, 2006). These prior assumptions are combined with information from the observations using Bayes' rule. The obtained result is the posterior distribution which represents the uncertainty about the latent function values of interest, conditioned on the observations and model assumptions. A challenge with the Bayesian approach is that exact inference is analytically intractable to calculate for most GP models of practical interest. Therefore, approximate methods are needed in order to evaluate the posterior distribution and to make predictions for new observations.

A general solution to approximate the posterior distribution for GP models is to use sampling-based stochastic methods (e.g. Neal, 1998). However, these methods can be slow in practice because the computational time (in most cases) scales cubically in the number of observations, which complicates the analysis of large-scale data sets with GPs. An active area of research in the machine learning community is approximate Bayesian inference using analytical Gaussian approximations that facilitate the posterior computations considerably. Especially deterministic methods, such as Laplace's approximation (LA, Williams and Barber, 1998), expectation propagation (EP, Minka, 2001b), and variational bounding (Gibbs and Mackay, 2000) and factorized variational (Csató et al., 2000; Girolami and Rogers, 2006) methods have been considered for GP models (see, e.g., Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008).

This work develops methods for approximate Bayesian inference in various modelling problems involving GP models. The focus is on efficient ways to form Gaussian posterior approximations based on LA or EP. The main objective of this thesis is to develop methodological improvements to approximate Bayesian inference for GP models in nonlinear regression, density estimation, survival analysis, and in binary and multiclass classification. The inference for the studied GP models is challenging in two aspects. Firstly, observation models are generalized in the way that the probability distribution for each observation can depend on multiple values of the latent function instead of only one value, or on the derivative values of the latent function. Secondly, instead of one prior process, the models can have multiple uncorrelated prior processes that are coupled through the observation model.

This thesis consists of Publications I–VI with the following more detailed research aims. In Publication I our objective is to improve the existing methodology proposed for nonlinear GP regression and binary classification (e.g. Rasmussen and Williams, 2006) by introducing additional monotonicity information into a GP model. We develop a method based on the EP algorithm to approximate a monotonic GP. Publication II aims to facilitate practical Bayesian inference for a logistic GP density estimation model (Leonard, 1978) by designing an approximation based on LA. In Publication III we aim to improve inference for multiclass GP classification by developing further the EP approximations for the multinomial probit model (Seeger et al., 2006; Girolami and Zhong, 2007). Publications IV–V are application studies where our objective is on accurate modelling of time-to-event survival data by considering nonlinear effects and interdependencies between explanatory variables in order to obtain a high predictive accuracy. In Publication IV our aim is also to facilitate the posterior computations of a Cox-type survival model (Cox, 1972) based on GP priors by constructing an approximation with the LA method. Finally, in Publication VI our objective is to develop a unifying software package to improve practical approximate Bayesian inference with various GP models.

The rest of this overview part is structured as follows. In Chapter 2, we review probability models constructed from Gaussian process priors and provide an overview of models considered in this work. Chapter 3 discusses approximate inference methods and computational approaches used in this work. These two chapters give the essential background theory for Publications I–VI included in this thesis. Chapter 4 provides brief summaries of Publications I–VI and Chapter 5 concludes the work.

2. Gaussian Process Models

This section gives an overview of models constructed from Gaussian process priors. We begin in Section 2.1 by discussing GPs from a machine learning point of view. Section 2.2 considers probability models constructed from GP priors and reviews the Bayesian approach for GP regression and binary classification. Section 2.3 focuses on GP models where multiple latent values are associated with each observation and on models based on multiple prior processes.

2.1 Gaussian Processes

Gaussian processes are flexible nonparametric models to define distributions directly over functions of one or more input variables (see, for example, O'Hagan, 1978; MacKay, 1998; Neal, 1998; Rasmussen and Williams, 2006, for an overview of GPs in the context of machine learning). Formally, a Gaussian process is defined to be "a collection of random variables, any finite number of which have a joint Gaussian distribution" (Rasmussen and Williams, 2006). A Gaussian process over the latent function can be written as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}')),$$

where x is an arbitrary input vector. The process is specified completely by the second order statistics, that is, by the mean function $m(\mathbf{x}) = \mathrm{E}[f(\mathbf{x})]$ and the covariance function $\kappa(\mathbf{x}, \mathbf{x}') = \mathrm{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$. Often, the prior mean function is assumed to be zero (also in Publications I, III and IV), although non-zero prior mean functions (Publication II) can be specified (e.g. Rasmussen and Williams, 2006). The covariance function defines the smoothness and scale properties of the GP and it is required to be positive semi-definite. We consider here two particular covariance functions from literature: a squared exponential covariance and a neural network covariance.

The squared exponential covariance function is widely applied in machine learning and also mainly used in Publications I–III. It can be written as

$$\kappa_{\mathrm{se}}(\mathbf{x}, \mathbf{x}') = \sigma_{\mathrm{m}}^2 \exp\left(-\frac{1}{2}\sum_{i=1}^{d} \rho_i^{-2}(x_i - x_i')^2\right),$$

where the lengthscale parameters $\rho_1, \ldots, \rho_d$ define the correlation lengths with respect to d input dimensions, and the magnitude parameter σ_{m} controls the magnitude of the function. From now on, the covariance function parameters are denoted with θ and they are called hyperparameters of the model. The squared exponential function is infinitely differentiable, and thus it produces smooth sample functions. It is also stationary (invariant to translations in the input space).

As shown by Neal (1996), in the limit of infinite hidden units a Bayesian neural network model converges to a Gaussian process. Using this connection, Williams (1998) derived a neural network covariance which corresponds to a neural network with an infinite number of hidden units with specific transfer functions and weight priors. The neural network covariance function is an example of a non-stationary covariance function, and it is given by

$$\kappa_{\mathrm{nn}}(\mathbf{x}, \mathbf{x}') = \frac{2}{\pi}\sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^{\mathrm{T}}\Sigma_{\mathrm{nn}}\tilde{\mathbf{x}}'}{(1+2\tilde{\mathbf{x}}^{\mathrm{T}}\Sigma_{\mathrm{nn}}\tilde{\mathbf{x}})(1+2\tilde{\mathbf{x}}'^{\mathrm{T}}\Sigma_{\mathrm{nn}}\tilde{\mathbf{x}}')}\right),$$

where $\tilde{\mathbf{x}} = [1, x_1, \ldots, x_d]^{\mathrm{T}}$ is an input vector augmented with 1. The diagonal matrix $\Sigma_{\mathrm{nn}} = \mathrm{diag}([\sigma_0^2, \sigma_1^2, \ldots, \sigma_d^2]^{\mathrm{T}})$ is the weight prior, where σ_0^2 is the variance for the bias parameter controlling the functions offset from the origin and $\sigma_1^2, \ldots, \sigma_d^2$ are the variances for the weight parameters.[1] The neural network covariance function can be used for modelling saturating effects, because of the sigmoidal shapes of the network transfer functions. We assume a GP model with the neural network covariance in Publication IV and a multilayer perceptron (MLP) neural network model (e.g. Neal, 1996) in Publication V. The MLP model has a hierarchical prior structure for the network weights and biases, but only a finite number of hidden units. Therefore, the MLP model can be thought to be an approximation for a GP model with the neural network covariance function and an infinite number of hidden units.

[1] We use the following notation: $\mathrm{diag}(\mathbf{a})$ with a vector argument is a square matrix with a on the main diagonal, and $\mathrm{diag}(A)$ with a matrix argument is a column vector containing the diagonal elements of matrix A.

The squared exponential and neural network covariance functions are only two examples of many possible covariance functions. For example, Rasmussen and Williams (2006, Chapter 4) present a thorough list of alternative covariances, their properties and possible combinations for creating even a wider class of covariance functions for modelling purposes.

2.2 Bayesian Inference for Gaussian Process Models

Gaussian processes are convenient for defining prior distributions over functions in a Bayesian framework. We begin by considering probability models constructed from GP priors with the following hierarchical structure:

$$\begin{aligned} \text{Observation model:} &\qquad \mathbf{y}|\mathbf{f} \sim \prod_{i=1}^{n} p(y_i|f_i) \\ \text{GP prior:} &\qquad f(\mathbf{x})|\theta \sim \mathcal{GP}\left(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'|\theta)\right) \\ \text{hyperprior:} &\qquad \theta \sim p(\theta), \end{aligned}$$

where the n observations $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ associated with inputs (or, explanatory variables or covariates) $X = \{\mathbf{x}_i = [x_{i,1}, \ldots, x_{i,d}]^{\mathrm{T}}\}_{i=1}^{n}$ are assumed to be conditionally independent given a latent function $f(\mathbf{x})$ in a way that the likelihood $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i)$, where $f_i = f(\mathbf{x}_i)$ and $\mathbf{f} = [f_1, f_2, \ldots, f_n]^{\mathrm{T}}$, factorizes over cases. A prior distribution $p(\theta)$ is set for the hyperparameters of the covariance function, but for now, we assume that the hyperparameters are given and condition the inference on θ. We return to estimating θ in Section 3.3.

By the definition of a Gaussian process, the GP prior results in a multivariate Gaussian distribution for the latent function values evaluated at X as

$$p(\mathbf{f}|X, \theta) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K_{\mathbf{f},\mathbf{f}}),$$

where $K_{\mathbf{f},\mathbf{f}} = K(X, X)$ is the $n \times n$ covariance matrix whose entries depend on inputs X according to the covariance function. In Bayesian inference, the posterior distribution for $\mathbf{f}$ is obtained by combining the prior distribution and the likelihood using Bayes' rule:

$$p(\mathbf{f}|X, \mathbf{y}, \theta) = \frac{p(\mathbf{f}|X, \theta)p(\mathbf{y}|\mathbf{f})}{p(\mathbf{y}|X, \theta)} = \frac{\mathcal{N}(\mathbf{f}|\mathbf{0}, K_{\mathbf{f},\mathbf{f}})}{p(\mathbf{y}|X, \theta)} \prod_{i=1}^{n} p(y_i|f_i). \tag{2.1}$$

The posterior distribution (2.1) represents the uncertainty about $\mathbf{f}$, conditioned on the prior assumptions and the observations $\mathbf{y}$. The normalization term, $p(\mathbf{y}|X, \theta) = \int p(\mathbf{f}|X, \theta)p(\mathbf{y}|\mathbf{f})d\mathbf{f}$, in Equation (2.1) is known as the marginal likelihood, and it is useful for hyperparameter inference as discussed in Section 3.3.

In most cases, our objective is to compute predictions for latent function values $\mathbf{f}_*$ (or for new observations $\mathbf{y}_*$) at test points X_*. To proceed with the Bayesian framework, the prior distribution can be written for training latent values $\mathbf{f}$ and for test latent values $\mathbf{f}_*$ jointly as

$$p(\mathbf{f}, \mathbf{f}_* | X, X_*, \theta) = \mathcal{N}\left(\left[\begin{array}{c} \mathbf{f} \\ \mathbf{f}_* \end{array}\right] \middle| \mathbf{0}, \left[\begin{array}{cc} K_{\mathbf{f},\mathbf{f}} & K_{\mathbf{f},*} \\ K_{*,\mathbf{f}} & K_{*,*} \end{array}\right]\right), \tag{2.2}$$

where $K_{\mathbf{f},*}$ defines the covariances between the latent values at training and test points, and $K_{*,*}$ between the latent values at test points. Note also that a Gaussian process can be thought to be an infinite dimensional multivariate Gaussian distribution (Rasmussen and Williams, 2006). However, in practice computations with GP can be done by focusing only on finite index sets (namely training points and arbitrary test points). By using the conditioning properties of a multivariate Gaussian distribution, we can write the conditional distribution for $\mathbf{f}_*$ given $\mathbf{f}$ from Equation (2.2) as

$$p(\mathbf{f}_* | \mathbf{f}, X, X_*, \theta) = \mathcal{N}\left(\mathbf{f}_* | K_{*,\mathbf{f}} K_{\mathbf{f},\mathbf{f}}^{-1} \mathbf{f}, K_{*,*} - K_{*,\mathbf{f}} K_{\mathbf{f},\mathbf{f}}^{-1} K_{\mathbf{f},*}\right). \tag{2.3}$$

By multiplying this conditional distribution with the posterior distribution $p(\mathbf{f}|X, \mathbf{y}, \theta)$ from Equation (2.1), we obtain the joint posterior (predictive) distribution for $\mathbf{f}$ and $\mathbf{f}_*$. According to the Bayesian approach, we need to integrate over the uncertainty related to the (unobserved) latent values $\mathbf{f}$, and by marginalizing we obtain the posterior predictive distribution for $\mathbf{f}_*$:

$$p(\mathbf{f}_* | X, \mathbf{y}, X_*, \theta) = \int p(\mathbf{f}_* | \mathbf{f}, X, X_*, \theta) p(\mathbf{f} | X, \mathbf{y}, \theta) d\mathbf{f}.$$

If we are interested in the predictive distribution for new observations $\mathbf{y}_*$, we can integrate over the uncertainty of $\mathbf{f}_*$ as

$$p(\mathbf{y}_* | X, \mathbf{y}, X_*, \theta) = \int p(\mathbf{y}_* | \mathbf{f}_*) p(\mathbf{f}_* | X, \mathbf{y}, X_*, \theta) d\mathbf{f}_*,$$

to obtain the posterior predictive distribution for $\mathbf{y}_*$.

2.2.1 Regression and Binary Classification

If the observation model $p(\mathbf{y}|\mathbf{f})$ is Gaussian, the integrals over latent values $\mathbf{f}$ and $\mathbf{f}_*$ required for Bayesian inference can be computed analytically. An example of such a case is GP regression with Gaussian noise. The objective in GP regression is to estimate an unknown function $f : \mathbb{R}^d \rightarrow \mathbb{R}$

on an arbitrary test input point $\mathbf{x}_*$, given a training data set $\mathcal{D} = \{X, \mathbf{y}\}$. An additive Gaussian noise model is assumed between the latent function and noisy observations as $y_i = f(\mathbf{x}_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with the noise variance σ^2. To generalize the relation between the input and output variables outside the finite training data points, we need to make assumptions about the underlying function. Thus, by assuming a Gaussian process over the latent functions, the prior assumptions about f can be encoded through the mean and covariance functions of the GP.

For fixed $\mathbf{y}$, we obtain the likelihood

$$p(\mathbf{y}|\mathbf{f}, \sigma) = \prod_{i=1}^{n} \mathcal{N}(y_i|f_i, \sigma^2). \tag{2.4}$$

Because the likelihood (2.4) and the prior (2.2) are both Gaussian, the predictive distribution for $\mathbf{f}_*$ is also Gaussian and it can be evaluated analytically. By integrating over the uncertainty of latent values $\mathbf{f}$, we obtain the posterior predictive distribution for $\mathbf{f}_*$ with the mean and covariance

$$\begin{aligned}
\mathrm{E}[\mathbf{f}_*|X_*, \mathcal{D}, \theta] &= K_{*,\mathbf{f}}(K_{\mathbf{f},\mathbf{f}} + \sigma^2 I_n)^{-1}\mathbf{y} \\
\mathrm{Cov}[\mathbf{f}_*|X_*, \mathcal{D}, \theta] &= K_{*,*} - K_{*,\mathbf{f}}(K_{\mathbf{f},\mathbf{f}} + \sigma^2 I_n)^{-1}K_{\mathbf{f},*},
\end{aligned}$$

where I_n is an identity matrix of size n and θ includes also the σ parameter. The predictive distribution for new observations $\mathbf{y}_*$ can also be computed analytically, and it is Gaussian with the mean $\mathrm{E}[\mathbf{f}_*|X_*, \mathcal{D}, \theta]$ and the covariance $\mathrm{Cov}[\mathbf{f}_*|X_*, \mathcal{D}, \theta] + \sigma^2 I_{n_t}$, where n_t is the number of test points.

In most modelling cases, however, the observation model is non-Gaussian and we cannot obtain a closed-form expression for $p(\mathbf{f}|\mathcal{D}, \theta)$. One widely-studied example of such a generalized case is binary GP classification (e.g. Williams and Barber, 1998; Rasmussen and Williams, 2006, Chapter 3). In binary classification problems, the output variable (that is, the class label) y associated with an input $\mathbf{x}$ is discrete, for example $y \in \{-1, 1\}$, and our objective is to predict the correct class labels for test inputs X_*, given $\mathcal{D}$.[2] In the probabilistic discriminative approach for GP binary classification (e.g. Rasmussen and Williams, 2006), we have a GP prior for the latent function which is squashed through the sigmoid function to construct a model for the class probabilities $p(y = 1|f(\mathbf{x})) = \sigma(f(\mathbf{x}))$ and $p(y = -1|f(\mathbf{x})) = 1 - p(y = 1|f(\mathbf{x}))$. As an example, the sigmoid function can be a symmetric cumulative Gaussian $\sigma(u) = \Phi(u)$ (probit regression),

[2] Note that y represents output variables in general, and whether it can have continuous, binary, or some other values depends on the context.

which results in the likelihood

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} \Phi(f_i y_i), \tag{2.5}$$

where $\Phi(u) = \int_{-\infty}^{u} \mathcal{N}(z|0,1)dz$. Another often used sigmoid function is the logistic response function $\sigma(u) = (1+\exp(-u))^{-1}$ (logistic regression) that can be more robust against outliers than the cumulative Gaussian function $\Phi(u)$ in probit regression (Nickisch and Rasmussen, 2008). By applying Bayes' theorem, we can derive the conditional posterior distribution

$$p(\mathbf{f}|\mathcal{D},\theta) = \frac{1}{Z}\mathcal{N}(\mathbf{f}|\mathbf{0}, K_{\mathbf{f},\mathbf{f}}) \prod_{i=1}^{n} \Phi(f_i y_i), \tag{2.6}$$

where $Z = p(\mathbf{y}|X,\theta) = \int \mathcal{N}(\mathbf{f}|\mathbf{0}, K_{\mathbf{f},\mathbf{f}}) \prod_{i=1}^{n} \Phi(f_i y_i) d\mathbf{f}$. Because the likelihood (2.5) is non-Gaussian, we cannot obtain a closed-form expression for the posterior distribution $p(\mathbf{f}|\mathcal{D},\theta)$. Therefore, to integrate over the uncertainty of latent values f to obtain predictions for $\mathbf{f}_*$ (and for new observations $\mathbf{y}_*$), we need to somehow approximate the integration over $\mathbf{f}$. Approximate methods for inference are discussed in Chapter 3.

In Publication I, we extend GP regression and binary classification to a more general case by introducing additional monotonicity information into a GP model. Due to this additional information, the posterior distribution is no longer Gaussian even in the regression case with Gaussian noise, which is why we must resort to approximate inference (see Section 4.1 and Publication I).

2.3 Models of Multiple Latent Values

In the previous section, the GP models had a fully-factorizing likelihood structure, that is, the distribution for each observation y_i depended only on a single latent value f_i (single-latent models), and not on $f_{j\neq i}$. In this section, we focus on extended GP models where the distribution for each observation y_i depends on multiple latent values $\mathbf{f}_i$ or on all latent values f (multi-latent models). Notice that the output variable can be univariate in these multi-latent models (see, e.g., the model described by Goldberg et al., 1998). This categorization to single-latent or multi-latent GP models is also considered in Publication VI. In addition, instead of having a single GP prior, there can be multiple uncorrelated prior processes that are coupled through a likelihood function.

There are several examples of modelling problems involving multi-latent GP models with either single or multiple output variables. One research question of increasing interest is how to extend GP regression with a single output variable to a scenario of multiple output variables (e.g. Álvarez and Lawrence, 2011), which is known as co-kriging in the geostatistics literature (Cressie, 1993, Section 3.2.3). In regression with multiple output variables, it is assumed that the output variables are correlated and they need to be modelled simultaneously because information can be lost by modelling them separately with multiple single-output GPs. The fixed correlations between output variables can be induced by using uncorrelated GP priors with a correlated noise process or by constructing different correlated prior structures (e.g. Boyle and Frean, 2005; Teh et al., 2005; Bonilla et al., 2008; Álvarez and Lawrence, 2011; Rasmussen and Williams, 2006). Note that regression with multiple output variables can be seen as a special instance of multi-task learning where there are multiple related prediction problems. Often, a multivariate Gaussian model with a global noise level is assumed, but there are many other GP-based models with more general observation models (leading to analytically intractable inference) that can be categorized as multi-latent models. Examples of such models are heteroscedastic noise models (Goldberg et al., 1998; Kersting et al., 2007; Lázaro-Gredilla and Titsias, 2011; Mũnoz-González et al., 2011) and robust regression with a two-component Gaussian mixture model (Naish-Guzman and Holden, 2008), all constructed from two GP priors that are coupled through an observation model. Heteroscedastic regression problems have also been solved by using finite mixtures (Tresp, 2001) or infinite mixtures (Rasmussen and Ghahramani, 2002) of GP priors. Also, more complex network structures based on GP priors have been proposed to model dependencies between multiple output variables (Wilson et al., 2012; Damianou and Lawrence, 2013). In addition to regression, multi-latent GP models are useful for solving other modelling problems, such as, multiclass classification, where latent values from multiple processes are associated with each observation (Neal, 1998; Williams and Barber, 1998). Other multi-latent GP models include, for example, a multinomial model (e.g. Juntunen et al., 2012) or a zero-inflated negative-binomial model (e.g. Vanhatalo et al., 2013). Publication VI lists more examples of multi-latent models constructed from GP priors (see also Vanhatalo et al., 2013).

It is also possible to categorize GP models based on their prior covariance structure. For example, Seeger et al. (2006) considered models that are constructed either from a single GP prior (single-process), or from multiple uncorrelated prior processes that are coupled through the likelihood function (multi-process models), and we adopt this same categorization. It should be noted here that uncorrelated prior processes can be expressed as a single GP prior with a specific hierarchical covariance function. However, we interpret GP models as multi-process models if they have uncorrelated non-additive prior processes, because this specific structure can lead to computational savings (although the posterior processes are correlated) as discussed in Chapter 3.

In the rest of this section, we consider in more detail the three examples of multi-latent GP models from Publications II–IV. We begin in Section 2.3.1 by giving a brief overview of a logistic Gaussian process (Leonard, 1978), where the prior distribution is specified over normalized functions, which makes the prior suitable for density estimation (as also discussed in Publication II). The density model is an example of a single-process multi-latent GP model where each likelihood term depends on all latent values. In Section 2.3.2, we consider multiclass GP classification with multiple prior processes that are coupled through the likelihood function (see also Publication III). In Section 2.3.3, we discuss another multi-process multi-latent model from the field of survival analysis. The survival model is constructed from two a priori uncorrelated processes to model time-to-event data in Publication IV.

2.3.1 Logistic Gaussian Processes

In density estimation the objective is to find an estimate for the unknown density function $p(\mathrm{x})$, based only on the observations X. In Publication II, we consider the logistic density transform for the underlying GP prior to construct prior distributions over densities. By assuming a GP prior over $f(\mathrm{x})$, the logistic Gaussian process (LGP) in a finite region $\mathcal{V}$ of $\mathbb{R}^d$ can be derived as

$$p(\mathbf{x}) = \frac{\exp(f(\mathbf{x}))}{\int_{\mathcal{V}} \exp(f(\boldsymbol{s}))d\boldsymbol{s}}$$

(Leonard, 1978). The LGP prior provides a convenient way to specify prior assumptions about the smoothness properties of density estimates through the covariance structure without restricting to any specific parameterized form. The logistic density transform constrains the density $p(\mathrm{x})$ to non-negative and its integral over the bounded space $\mathcal{V}$ to one.

A challenge in inference with LGP is how to solve this integral required to ensure normalization. In Publication II, we apply a finite-dimensional approximation by evaluating the integral and the GP prior in a grid as described by Tokdar (2007). A finite element approach for density estimation is also presented by Griebel and Hegland (2010), and theoretical studies about the posterior consistency of the LGP density estimation on a closed bounded interval are examined by Tokdar and Ghosh (2007). Furthermore, Tokdar (2007) illustrates that by making the grid finer, the Kullback–Leibler (KL) divergence from the exact posterior to the finite-dimensional approximation converges to zero.

In Publication II, we discretise the finite space $\mathcal{V}$ into n_{grid} subregions, and collect the coordinates of the subregions into an $n_{\mathrm{grid}} \times d$ matrix X, where the i'th row denotes the center point of the i'th subregion. We denote the number of observations that fall within the i'th subregion with y_i, and all the count observations with an $n_{\mathrm{grid}} \times 1$ vector $\mathbf{y}$. By assuming a regular grid, the overall log-likelihood contribution of the n observations after this discretisation can be written as

$$\log p(\mathbf{y}|\mathbf{f}) = \mathbf{y}^{\mathrm{T}}\mathbf{f} - n \log \left(\sum_{j=1}^{n_{\mathrm{grid}}} \exp(f_j) \right), \tag{2.7}$$

where $\mathbf{f}$ is a column vector of n_{grid} latent values associated with each subregion. This likelihood does not factorize over single latent values f_i as each term in the likelihood depends on $\mathbf{f}$ due to the normalization constraint. We discuss approximate inference methods for this non-factorizing likelihood in Publication II.

2.3.2 Multiclass Classification

In multiclass classification the challenge is that the target variables have more than two possible class labels, that is, $y \in \{1, \ldots, c\}$, where $c > 2$ is the number of classes. The objective is to predict the class label for a test input given all training class labels $\mathbf{y}$ (a vector of size n) and the corresponding training inputs. Note that although multiclass classification is possible to implement by using several successive one-vs-rest binary classifiers, these approaches have troubles, for example, in how to combine separate binary classification results and how to perform hyperparameter inference (see, e.g., Seeger and Jordan, 2004).

In the literature for multiclass GP classification, the usual assumption is to use c independent prior processes that are associated with c classes

(Williams and Barber, 1998; Seeger and Jordan, 2004; Rasmussen and Williams, 2006; Girolami and Zhong, 2007). Compared to binary classification in Section 2.2.1, multiclass classification is more challenging because each target class increases the number of unknown latent values by n (the number of observations). By assuming uncorrelated zero-mean GP priors for latent functions associated with different classes, we obtain a zero-mean Gaussian prior for $\mathbf{f} = \left[f_1^1, \ldots, f_n^1, f_1^2, \ldots, f_n^2, \ldots, f_1^c, \ldots, f_n^c\right]^{\mathrm{T}}$ as

$$p(\mathbf{f}|X) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K_{\mathbf{f},\mathbf{f}}),$$

where $K_{\mathbf{f},\mathbf{f}}$ is a $cn \times cn$ block-diagonal covariance matrix with matrices $K^1, K^2, \ldots, K^c$ (each of size $n \times n$) on its diagonal (K^j determines the covariances between the latent values $f_1^j, \ldots, f_n^j$).

Two common observation models for probabilistic multiclass GP classification are the softmax (or multinomial logit) model and the multinomial probit model. The softmax and multinomial probit models are multiclass generalizations of the logistic and probit models respectively. The softmax model is given by

$$p(y_i|\mathbf{f}_i) = \frac{\exp(f_i^{y_i})}{\sum_{j=1}^{c} \exp(f_i^j)} \tag{2.8}$$

(e.g. Neal, 1998; Williams and Barber, 1998). We employ the sofmax model in Publication III with a GP prior and in Publication V with an MLP neural network prior. Note that the softmax likelihood is similar to the likelihood used for LGP density estimation in a grid (cf. Section 2.3.1). The multinomial probit model can be written as

$$p(y_i|\mathbf{f}_i) = \mathrm{E}_{p(u_i)} \left\{ \prod_{j=1, j \neq y_i}^{c} \Phi(u_i + f_i^{y_i} - f_i^j) \right\}, \tag{2.9}$$

where the auxiliary variable u_i is distributed as $p(u_i) = \mathcal{N}(u_i|0,1)$, and $\Phi(u)$ denotes the cumulative density function of the standard normal distribution (e.g. Girolami and Zhong, 2007). The multinomial probit model is applied in Publication III.

Both likelihoods (2.8) and (2.9) leads to an analytically intractable posterior distribution and are examples of a case, where each likelihood term depends on a vector $\mathbf{f}_i$ consisting of c-latent values. However, because of the structure of these likelihood functions, computational savings are possible to obtain if uncorrelated prior processes are assumed over latent functions (Williams and Barber, 1998, see also Publication III).

2.3.3 Models for Survival Analysis

In Publications IV and V, the focus is on survival analysis where the objective is to model time-to-event data (see, e.g., Ibrahim et al., 2001, for a Bayesian approach to survival analysis). In Publication IV, we have time-to-event observations that are possibly right censored. We denote a survival time with y_i and a censoring indicator with δ_i, where $\delta_i = 0$ if the i'th observation is uncensored and $\delta_i = 1$ if the observation is right censored. Recall that the hazard function $h_i(t)$ gives the instantaneous rate of failure at time t for individual i. The traditional way to analyse continuous time-to-event data is to assume the Cox proportional hazard function

$$h_i(t) = h_0(t) \exp(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}),$$

where the baseline hazard rate $h_0(t)$ is unspecified (Cox, 1972). A common approach is to use linear predictor $\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ denotes a vector of regression coefficients and $\mathbf{x}_i$ contains the observed covariates for the i'th individual. However, by extending the linear predictor to more general forms, we can model for example additive and non-linear effects of covariates (Kneib, 2006; Martino et al., 2011).

In Publication IV, we consider a similar Cox-type modelling approach as described recently by Martino et al. (2011). We assume a GP prior over $\eta(\mathbf{x})$, and use an extended proportional hazards model

$$h_i(t) = \exp(\log(h_0(t)) + \eta_i(\mathbf{x}_i)),$$

where the linear predictor is replaced with the latent predictor η_i. In the survival analysis literature, there are many parametric alternatives, for example exponential, Weibull, log-normal, or Gamma distributions, to model the baseline hazard function $h_0(t)$ (Ibrahim et al., 2001). A more flexible alternative is obtained by modelling the hazard function as a piecewise log-constant baseline hazard (e.g. Ibrahim et al., 2001), which is also assumed in Publication IV. We partition the time axis into T intervals with equal lengths: $0 = s_0 < s_1 < s_2 < \ldots < s_T$, where $s_T > y_i$ for all $i = 1, \ldots, n$. In the interval k (where $k = 1, \ldots, T$), we assume a constant baseline hazard

$$h_0(t) = \lambda_k \quad \text{for} \quad t \in (s_{k-1}, s_k].$$

For the i'th individual the hazard rate in the k'th time interval can be written as

$$h_i(t) = \exp(f_k + \eta_i(\mathbf{x}_i)), \quad t \in (s_{k-1}, s_k],$$

where $f_k = \log(\lambda_k)$. To smooth the hazard rate function, we assume another GP prior over $f(t)$. We denote the mean locations of T time intervals with a vector $\mathbf{t} = [t_1, \ldots, t_T]^{\mathrm{T}}$. Then, the GP prior results in the Gaussian distribution

$$p(\mathbf{f}|\mathbf{t}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K_{\mathbf{f},\mathbf{f}}),$$

where $\mathbf{f} = [f_1, \ldots, f_T]^{\mathrm{T}}$. The matrix $K_{\mathbf{f},\mathbf{f}} = K(\mathbf{t}, \mathbf{t})$ is of size $T \times T$ and it determines the covariance structure between the latent values associated with the time points. Thus, the joint prior covariance matrix is a block-diagonal matrix consisting of the matrices $K_{\mathbf{f},\mathbf{f}}$ and $K_{\boldsymbol{\eta},\boldsymbol{\eta}}$ (the covariances between the latent values $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_n]^{\mathrm{T}}$ associated with the n individuals).

The likelihood contribution for the i'th observation (y_i, δ_i) is assumed to be

$$l_i = h_i(y_i)^{(1-\delta_i)} \exp\left(-\int_0^{y_i} h_i(t)dt\right),$$

and with the piecewise log-constant assumption for the hazard rate function, the likelihood contribution leads to

$$l_i = [\lambda_k \exp(\eta_i)]^{(1-\delta_i)} \exp\left(-[(y_i - s_{k-1})\lambda_k + \sum_{g=1}^{k-1}(s_g - s_{g-1})\lambda_g] \exp(\eta_i)\right), \tag{2.10}$$

where $y_i \in (s_{k-1}, s_k]$ (Ibrahim et al., 2001; Martino et al., 2011). The likelihood function couples the two Gaussian processes, and the likelihood contribution for the i'th observation depends on f and η_i. In Section 4.4, we discuss approximate inference issues regarding the Cox proportional hazards model with the GP priors. The Cox proportional hazards model from Publication IV can be thought to be a Gaussian process extension of a similar model considered by Martino et al. (2011) who described the model for latent Gaussian models and implemented the model with Gaussian Markov random field priors. Also, a similar flexible approach for modelling survival times using penalized splines is shown by Kneib and Fahrmeir (2007).

The likelihood contribution of a Cox proportional model with the piecewise log-constant baseline hazard assumption is possible to express using the Poisson likelihood (e.g. Laird and Oliver, 1981). This connection was recently applied by Martino et al. (2011) who expressed the log-likelihood contribution of a Cox model using the Poisson likelihood with an extended data set. In this extended Poisson-distributed data representation, $k-1$ observations are zero with the means $(s_g - s_{g-1})\lambda_g$ and one observation is either zero (if the survival time is censored) or one (observed survival time) with the mean $(y_i - s_{k-1})\lambda_k$. Thus, the likelihood contribution of Equation (2.10) can be written as a fully-factorizing single-latent Poisson model. However, this extended representation increases the number of latent values, which can be challenging with GP priors due to the cubic computational complexity (see Section 4.4).

In Publication V, we have a special case of time-to-event data where there are no censored observations. Therefore, we treat the modelling of time-to-event data in Publication V as a multiclass classification problem and use the softmax model (2.8). For the latent functions, we assume an MLP neural network prior with a finite number of hidden units. The MLP model approximates an infinite neural network GP model, although the MLP model can have a better generalization ability due to its finite complexity (Winther, 2001).

3. Approximate Bayesian Inference

In Bayesian analysis, the posterior distribution expresses the information about unknown quantities, given the observed data and model assumptions. For Gaussian process models, the evaluation of the posterior distribution or the posterior predictive distribution requires integration over a high-dimensional space of unknown latent values. Unfortunately, this integration is in practice analytically intractable if the likelihood function is non-Gaussian. In order to solve the integration problem, we need to resort to approximate methods. This chapter presents an overview of the approximations used in Publications I–VI for Bayesian inference. We begin in Section 3.1 by reviewing general algorithms to approximate the posterior distribution of the latent values. Then, in Section 3.2 we discuss how these algorithms can be tailored to approximate the posterior distribution in various single-latent and multi-latent cases from Chapter 2. Finally, in Section 3.3 we discuss briefly marginal likelihood approximations for model selection.

3.1 Approximate Methods

Numerical sampling is a generic approach to approximate a non-Gaussian posterior distribution without limiting to deterministic approximations of simpler forms (e.g. Gelman et al., 2003). Because direct sampling from a high-dimensional posterior distribution is challenging, stochastic Markov chain Monte Carlo (MCMC) methods are often used to obtain samples from the posterior distribution (e.g. Robert and Casella, 2004). In summary, the idea behind MCMC sampling is to create a Markov chain whose stationary distribution is the posterior distribution $p(\mathbf{f}|\mathcal{D}, \theta)$, and simulate values from such a Markov process long enough, so that the distribution of the simulated values is close enough to the posterior distribution (Gelman et al., 2003). In addition to latent values $\mathbf{f}$, we can also obtain samples from the posterior distribution of the hyperparameters θ to approximate the integration over the uncertainty in θ. By using a finite set of simulated samples representing the posterior distribution, we can compute approximate posterior statistics, such as the mean and the covariance. There are various MCMC algorithms for generating samples from the posterior distribution. For example, in Publication II we approximate the posterior distribution by sampling alternatively from the conditional posterior of the latent values $p(\mathbf{f}|\mathcal{D},\theta)$ by using scaled Metropolis–Hastings sampling (Neal, 1998), and from the conditional posterior of the covariance function parameters $p(\theta|\mathbf{f},\mathcal{D})$ by using hybrid (or Hamiltonian) Monte Carlo (HMC, Duane et al., 1987; Neal, 1996). In Publication III, two different sampling techniques are used depending on the likelihood: scaled Metropolis–Hastings sampling for the softmax function (Neal, 1998), and Gibbs sampling for the multinomial probit function (Girolami and Rogers, 2006). In Publication V, we approximate the integration over the posterior distribution of weight and bias parameters in an MLP model with the HMC algorithm and hyperparameters with Gibbs sampling, as described by Neal (1996). Because sampling-based estimates become exact in the limit of an infinite sample size, sampling techniques are often used as a gold standard for measuring the performance of other approximations (as also in Publication III).

A problem with MCMC methods is that they are computationally demanding and they can be very slow in a practical sense. Posterior computations for a GP model require the evaluation of the inverse of the covariance matrix, which has time complexity $\mathcal{O}(n^3)$, where n is the number of latent values (e.g. Neal, 1998). Therefore, one MCMC iteration scales as $\mathcal{O}(n^3)$, which becomes computationally expensive for large n because thousands of posterior draws may be required to obtain uncorrelated posterior samples, and strong dependency between the hyperparameters and latent values can cause slow mixing of the chains. Also, convergence diagnostics can be challenging as it is difficult to assess whether the sampling mechanism really generates samples from the desired posterior distribution.

To speed up the inference, the non-Gaussian posterior distribution of the latent function values can be approximated with a tractable Gaussian distribution. Although the computational complexity is $\mathcal{O}(n^3)$ also with Gaussian approximations, they require less $\mathcal{O}(n^3)$ operations than

MCMC. By approximating $p(\mathbf{f}|\mathcal{D},\theta)$ with a Gaussian distribution, the integration over f can be done analytically, similarly to the regression case with the Gaussian likelihood discussed in Section 2.2.1. In addition to tractable posterior computations, the Gaussian approximation can be motivated by the asymptotic normality of the posterior distribution (e.g. Gelman et al., 2003). Also, the Gaussian approximation is convenient if $p(\mathbf{f}|\mathcal{D},\theta)$ can be shown to be unimodal. There are different approaches proposed for constructing the Gaussian approximation for $p(\mathbf{f}|\mathcal{D},\theta)$, including Laplace's approximation (Williams and Barber, 1998), expectation propagation (Minka, 2001a), and variational bounding (Gibbs and Mackay, 2000) and factorized variational (Csató et al., 2000; Girolami and Rogers, 2006) methods. For an overview of Gaussian approximations for GP binary classification, see the comprehensive study by Nickisch and Rasmussen (2008). In this work, the focus is on Laplace's approximation (Section 3.1.1) and on the expectation propagation approximation (Section 3.1.2) due to their speed and accuracy. Both approximations also facilitate efficient gradient-based estimation of the covariance function hyperparameters, which can be used to approximate the integration over the uncertainty in θ, as discussed in Section 3.3.

3.1.1 Laplace's Approximation (LA)

Laplace's approximation is based on a second-order Taylor expansion for $\log p(\mathbf{f}|\mathcal{D},\theta)$ around the posterior mode (e.g. Gelman et al., 2003; Rasmussen and Williams, 2006). The mode $\hat{\mathbf{f}}$ can be determined, for example, by Newton's method as described by Williams and Barber (1998) and Rasmussen and Williams (2006). The obtained Gaussian approximation is given by

$$p(\mathbf{f}|\mathcal{D},\theta) \approx q_{\mathrm{LA}}(\mathbf{f}|\mathcal{D},\theta) = \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}},\Sigma), \tag{3.1}$$

where $\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{f}|\mathcal{D},\theta)$ and $\Sigma^{-1} = -\nabla^2_{\mathbf{f}} \log p(\mathbf{f}|\mathcal{D},\theta)|_{\mathbf{f}=\hat{\mathbf{f}}}$ is the Hessian of the negative log posterior at $\hat{\mathbf{f}}$. The posterior covariance is given by $\Sigma = (K_{\mathbf{f},\mathbf{f}}^{-1} + W)^{-1}$, where $W = -\nabla^2_{\mathbf{f}} \log p(\mathbf{y}|\mathbf{f})|_{\mathbf{f}=\hat{\mathbf{f}}}$. To obtain the approximate posterior predictive distribution for $\mathbf{f}_*$, we can combine the conditional distribution $p(\mathbf{f}_*|\mathbf{f},X,X_*,\theta)$ of Equation (2.3) with the obtained Gaussian approximation (3.1), and integrate over the uncertainty of f, similarly as was done in Section 2.2. The approximate posterior predictive distribution for $\mathbf{y}_*$ can be obtained, for example, with sampling methods (see, e.g., Rasmussen and Williams, 2006). The LA method is used in

Publications II–IV.

3.1.2 Expectation Propagation (EP)

Expectation propagation is an algorithm that updates marginal moments iteratively to approximate integrals over functions that factor into simpler terms (Minka, 2001a). For GP models, the posterior distribution of Equation (2.6) can be approximated with EP as

$$q_{\mathrm{EP}}(\mathbf{f}|\mathcal{D},\theta) = \frac{1}{Z_{\mathrm{EP}}} p(\mathbf{f}|X,\theta) \prod_{i=1}^{n} \tilde{t}_i(f_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2), \tag{3.2}$$

where $\tilde{t}_i(f_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = \tilde{Z}_i \mathcal{N}(f_i|\tilde{\mu}_i, \tilde{\sigma}_i^2)$ are local likelihood term approximations parameterized with scalar normalization terms $\tilde{Z}_i$, site locations $\tilde{\mu}_i$, and site variances $\tilde{\sigma}_i^2$. The normalization Z_{EP} is the approximation for the marginal likelihood and can be used in model selection as discussed in Section 3.3. The EP algorithm starts with initialized site approximations. Then, site terms are updated iteratively. In the update step of the EP algorithm, we first remove the i'th site term from the approximate marginal posterior to obtain the cavity distribution

$$q_{-i}(f_i) = \mathcal{N}(f_i|\mu_{-i}, \sigma_{-i}) \propto q(f_i|\mathcal{D},\theta)\tilde{t}(f_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2)^{-1}.$$

Then, this cavity distribution is combined with the exact i'th likelihood term $p(y_i|f_i)$ to obtain the non-Gaussian distribution

$$\hat{p}(f_i) = \hat{Z}_i^{-1} q_{-i}(f_i) p(y_i|f_i), \tag{3.3}$$

which is known as the tilted distribution. If the approximating family is chosen to be Gaussian, we determine a Gaussian approximation $\hat{q}(f_i)$ for $\hat{p}(f_i)$ by minimizing the KL divergence $\mathrm{KL}(\hat{p}(f_i)||\hat{q}(f_i))$. For a Gaussian distribution $\hat{q}(f_i)$, this minimization of the KL divergence is equivalent to matching the first and second moments of $\hat{q}(f_i)$ with the corresponding moments of $\hat{p}(f_i)$. After matching the moments, we can update the i'th site term in a way that the mean and covariance of $q(f_i)$ are consistent with $\hat{q}(f_i)$. After this site update, the posterior distribution (3.2) can be updated with a rank-1 update (sequential EP). Alternatively, the posterior distribution can be refreshed once after all the site approximations have been updated. This is known as parallel EP (see, for example, Van Gerven et al., 2009) and in practice it can result in a computational speed-up. The update steps of EP are repeated until convergence where all the marginal distributions $q(f_i)$ are consistent with $\hat{p}(f_i)$. The Gaussian posterior approximation with EP is obtained by

$$p(\mathbf{f}|\mathcal{D},\theta) \approx \mathcal{N}(\mathbf{f}|(K_{\mathbf{f},\mathbf{f}}^{-1} + \tilde{\Sigma}^{-1})^{-1}\tilde{\Sigma}^{-1}\tilde{\boldsymbol{\mu}}, (K_{\mathbf{f},\mathbf{f}}^{-1} + \tilde{\Sigma}^{-1})^{-1}),$$

where $\tilde{\boldsymbol{\mu}} = [\tilde{\mu}_1, \ldots, \tilde{\mu}_n]^{\mathrm{T}}$ and $\tilde{\Sigma} = \mathrm{diag}([\tilde{\sigma}_1^2, \ldots, \tilde{\sigma}_n^2])$. The posterior predictive distributions for $\mathbf{f}_*$ (or $\mathbf{y}_*$) can be computed similarly to LA. The EP approximation is applied in Publications I and III.

3.2 Computational Strategies

For many models with GP priors the likelihood function can be written in a form that factorizes over cases as $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i)$, that is, each likelihood term depends only on a single latent value f_i. Examples of non-Gaussian likelihood functions with this fully-factorizing structure are the logit and probit likelihoods for binary classification, the Poisson, binomial, and negative-binomial likelihoods for modelling count data, and the Student-t likelihood for robust regression (see, e.g., Publication VI and Vanhatalo et al., 2013, for more examples and references of these single-latent GP models).

The conditional posterior distribution $p(\mathbf{f}|\mathcal{D},\theta)$ can be determined by Bayes' rule (2.1) as discussed in Section 2.2. Note that although the likelihood function factorizes, the posterior distribution $p(\mathbf{f}|\mathcal{D},\theta)$ and the marginal likelihood $p(\mathbf{y}|X,\theta)$ cannot be factorized due to the dependencies induced by the prior covariance structure. By approximating the posterior distribution with the LA method when the likelihood function factorizes, the matrix W of Equation (3.1) is diagonal (W models the precision of the effective likelihood, see e.g. Nickisch and Rasmussen, 2008). If the likelihood function is log-concave, the obtained matrix W has non-negative diagonal elements, and because the prior covariance $K_{\mathbf{f},\mathbf{f}}$ is positive definite by construction, the posterior $p(\mathbf{f}|\mathcal{D},\theta)$ is also log-concave and has a unique maximum. In such a case, the LA method can be implemented similarly as shown by Rasmussen and Williams (2006) for binary classification with the logit or probit likelihood function. For non-log-concave likelihood functions, such as the Student-t with small degrees of freedom, the implementation requires special care as the posterior distribution can be multimodal and W can have negative values (Vanhatalo et al., 2009).

In the EP approximation for the posterior distribution (2.1), each likelihood term is approximated with a univariate unnormalized Gaussian function. At each EP iteration, we need to determine the moments of the tilted distribution $\hat{p}(f_i)$ for all i, which requires solving one-dimensional integrals. In general, these univariate integrals can be computed efficiently using numerical quadrature (Zoeter and Heskes, 2005). For specific likelihood functions (e.g. for the probit function), the integrals are analytically tractable, which facilitates the evaluations of the tilted moments required in binary classification (Rasmussen and Williams, 2006) or in introducing monotonicity information into a GP model (see Section 4.1 and Publication I). Otherwise, the EP algorithm for single-latent likelihood functions with GP priors can be implemented as presented, for example, by Rasmussen and Williams (2006). With non-log-concave likelihoods, convergence problems can occur in the EP algorithm, although these can be alleviated with a more robust EP implementation based on damping, fractional updates, and double-loop algorithms (e.g. Heskes and Zoeter, 2002; Minka, 2004; Seeger, 2005; Jylänki et al., 2011). Note that in Gaussian approximations for Equation (2.1), the approximate likelihood contribution can also have a non-factorizing structure, as discussed by Nickisch and Rasmussen (2008). However, with EP it is assumed that the site terms factorize in the same way as the true likelihood function does (that is, diagonal $\tilde{\Sigma}^{-1}$), and the fully-factorizing approximation for the effective likelihood (diagonal W) arises inherently with LA.

In Publications III–IV, we have a model where each likelihood term depends on multiple latent values $\mathbf{f}_i$ (or on all latent values $\mathbf{f}$, as in Publication II) as $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|\mathbf{f}_i)$. The likelihood factorizes over observations, but we cannot factorize it into terms depending on a single latent value f_i. Publication VI gives examples of GP models with this multi-latent dependency. To approximate the posterior distribution with LA in this multi-latent case, we compute the second derivatives of the log-posterior with respect to the latent values, which gives rise to the structure of W in the posterior covariance of Equation (3.1). The challenge is that now W can be non-diagonal. Depending on the likelihood function, W can have, for example, a full structure (as with the LGP likelihood of Equation 2.7) or a block-matrix structure consisting of diagonal matrices (as with the softmax likelihood of Equation 2.8). It depends on the structures of W and the prior covariance $K_{\mathbf{f},\mathbf{f}}$, whether computational savings can be obtained in the evaluation of the posterior covariance Σ. For example, in a multiclass classification with the softmax likelihood, the structure of W can be exploited when $K_{\mathbf{f},\mathbf{f}}$ is a block-diagonal matrix, to enable efficient posterior computations that scale linearly (instead of cubically) in the number of target classes (Williams and Barber, 1998). This specific multiclass classification case is also discussed in more detail in Publication III. Otherwise, the LA algorithm for multi-latent likelihood functions can be implemented in a similar way as presented by Williams and Barber (1998) and Rasmussen and Williams (2006) for the softmax likelihood (see also Publication II and Section 4.4).

A challenge with EP in the multi-latent setting is how to efficiently evaluate the moments of the tilted distributions (3.3), where now $\hat{p}(\mathbf{f}_i) = \hat{Z}_i^{-1} q_{-i}(\mathbf{f}_i) p(y_i|\mathbf{f}_i)$ is a multivariate distribution. The evaluation of the tilted moments requires determining multi-dimensional integrals over $\mathbf{f}_i$. The moments can be estimated directly using multi-dimensional quadratures, as done for example by Seeger and Jordan (2004), but this can become computationally demanding when the dimensionality of $\mathbf{f}_i$ increases. An alternative approach for approximating the tilted moments is to use the LA method (Ypma and Heskes, 2005; Girolami and Zhong, 2007), which results in an algorithm called Laplace propagation (Smola et al., 2004). However, a problem with this LA approach can be that the mean is replaced with the mode of the distribution and the covariance with the inverse Hessian of the log density at the mode. Because the likelihood function can cause skewness to the tilted distribution, the LA method can lead to inaccurate mean and covariance estimates, in which case the resulting posterior approximation does not correspond to the full EP solution.

The computations with EP can be facilitated by assuming a factorizing approximate posterior distribution that is commonly used in variational approximations (Seeger et al., 2006; Girolami and Zhong, 2007). Explicit likelihood-couplings are omitted, but this fully-factorizing simplification (diagonal $\tilde{\Sigma}^{-1}$) can help estimating the tilted moments. As an example, with the independence assumption, each site update for the multinomial probit likelihood requires only one- and two-dimensional numerical quadratures, instead of c-dimensional (where c is the number of output classes), due to the product form of the likelihood function in Equation (2.9) (Seeger et al., 2006; Girolami and Zhong, 2007). Note also that in multinomial probit GP classification the fully-factorizing structure leads to efficient posterior computations scaling linearly in c. However, this is a special case and for example the softmax likelihood (Equation 2.8) cannot be factorized in a similar way due to the sum terms in the likelihood function. Also, EP with a fully-factorizing likelihood approximations can underestimate the uncertainty on the latent values and in practice it may require more iterations than full EP for convergence especially if the hyperparameter setting results in strong posterior couplings (see Publication III).

One deterministic solution to approximate the tilted moments is to use a secondary (or inner) EP loop that has been considered by Kim and Ghahramani (2006) and Naish-Guzman (2007). As an example, Naish-Guzman (2007) proposed to use inner EP that resembles the EP algorithm for binary classification, to evaluate efficiently tilted moments in a mixture model designed for robust regression. In Publication III, we present a similar nested EP algorithm to approximate the moments of a multivariate distribution for multiclass GP classification with the multinomial probit likelihood.

In addition to the difficulties with the evaluation of the tilted moments, another challenge with EP in the multi-latent and multi-process setting is how to derive a representation for the site precision matrix $\tilde{\Sigma}^{-1}$ that preserves explicit dependencies between the latent values, and that can (possibly) be exploited to obtain efficient posterior computations (if the prior covariance structure $K_{\mathbf{f},\mathbf{f}}$ is assumed to be block-diagonal, as for example in multiclass GP classification). For example, if quadrature rules are used to compute the tilted moments $\hat{p}(\mathbf{f}_i)$ for softmax GP classification, a constrained structure for site approximations (that facilitates the posterior representation scaling linearly in c) can require an additional optimization step (Seeger and Jordan, 2004). In Publication III, we show how this additional optimization step can be avoided in a specific case with the multinomial probit likelihood by computing the tilted moments using an inner EP algorithm that automatically results in an efficient structure for $\tilde{\Sigma}^{-1}$. Also, note that with the LA method for softmax GP classification, a similar constrained structure for W leading to efficient posterior computations is obtained directly by computing $W = -\nabla^2_{\mathbf{f}} \log p(\mathbf{y}|\mathbf{f})|_{\mathbf{f}=\hat{\mathbf{f}}}$ (Williams and Barber, 1998).

3.2.1 Summary of Gaussian Approximations

In this work, we have focused on the LA and EP methods to approximate the conditional posterior distribution $p(\mathbf{f}|\mathcal{D},\theta)$. Overall, the EP approximation for many GP models has been found very accurate with a reasonable computational cost compared to MCMC (e.g. Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008; Girolami and Zhong, 2007; Jylänki et al., 2011). On the other hand, the LA method is fast, although the problem with LA can be that the mean is replaced with the mode of the distribution and the covariance with the inverse Hessian of the log density at the mode. For example, in binary classification problems, where the posterior distribution of latent values can be skewed due to the shape of the sigmoid function, the LA method can lead to inaccurate mean and covariance estimates, whereas EP obtains good practical accuracy (Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008). We use the EP approximation in Publications I and III. In the single-latent case of Publication I, the EP approximation resembles the binary classification case with the probit likelihood function, where EP has been shown to provide accurate results (Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008). In Publication III, the EP approximation is more involved due to the multi-latent and multi-process model, but it is shown that EP obtains accurate results compared to MCMC with slightly longer computational time than LA. In Publications II and IV, we use the LA method for its speed. In both cases, the EP approximation would be challenging because the tilted moment evaluations are computationally demanding due to the multiple latent values associated with each observation. Note that with the LGP model in Publication II, an increase in the number of observations does not increase the number of latent values, which is why LA can be more accurate for LGP than for softmax GP classification. On the other hand, with a finer grid, fewer observations fall into each interval that can emphasise the Gaussian prior contribution to the posterior distribution (the comparisons show that LA for LGP is close to MCMC, see Publication II). Therefore, we believe that EP in this case would improve only slightly the performance of LA, but the computation time would increase considerably due to the multi-dimensional moment matching. However, one approach to solve the moment matching problem of EP could be implementing a similar quadrature-free nested EP approach as done for multinomial probit GP classification, where the normalization is over different output classes (see Publication III). However, based on preliminary testing, the moment matching step can be slow compared to LA for example when $n_{\text{grid}} = 400$. Similarly, an EP-based approximation for the Cox-type model of Publication IV would be difficult due to multivariate moment matching, unless the model is implemented with the Poisson-distributed data representation. However, LA for the Poisson likelihood with GP priors has been observed to be close to MCMC sampling (see comparisons by Vanhatalo et al., 2010).

In addition to LA and EP, variational bounding and factorized variational approximations have been proposed for GP models in the machine learning literature (e.g. Gibbs and Mackay, 2000; Csató et al., 2000; Girolami and Rogers, 2006). For GP models with different likelihood functions, these variational-type approximations have been observed to be less accurate than EP, but close to LA in speed, although they can have troubles in hyperparameter estimation (e.g. Nickisch and Rasmussen, 2008; Jylänki et al., 2011, see also Publication III).

3.3 Hyperparameter Inference

In the previous sections, we have focused on approximating $p(\mathbf{f}|\mathcal{D},\theta)$ conditioned to fixed hyperparameters θ. However, according to the Bayesian approach, we should also integrate over the uncertainty relating to θ as

$$p(\mathbf{f}|\mathcal{D}) = \int p(\mathbf{f}|\mathcal{D},\theta)p(\theta|\mathcal{D})d\theta,$$

where $p(\theta|\mathcal{D})$ is the posterior distribution for θ. In practice, the integration over θ can be approximated by a maximum a posteriori (MAP) point estimate of the hyperparameter values (type-II MAP estimation). In this work, we assume a prior distribution $p(\theta)$ for the hyperparameters to improve the identifiability of the ratio of the covariance function magnitude and lengthscale parameters. As an example, in Publication II we assume a weakly informative half Student-t distribution for θ, as recommended for hierarchical models by Gelman (2006). The MAP estimate for the hyperparameters can be determined by optimizing the marginal posterior distribution $p(\theta|\mathcal{D}) \propto p(\mathbf{y}|X,\theta)p(\theta)$, where $p(\mathbf{y}|X,\theta) = \int p(\mathbf{f}|X,\theta)p(\mathbf{y}|\mathbf{f})d\mathbf{f}$ is the marginal likelihood that normalizes the posterior distribution $p(\mathbf{f}|\mathcal{D},\theta)$. We cannot evaluate $p(\mathbf{y}|X,\theta)$ exactly if the likelihood function is non-Gaussian, but we can approximate it with LA or EP (see, e.g., Rasmussen and Williams, 2006).

Kuss and Rasmussen (2005) and Nickisch and Rasmussen (2008) studied the suitability of the marginal likelihood approximations for selecting hyperparameters in binary classification by comparing the calibration of the predictive performance and the marginal likelihood estimates on a grid of hyperparameter values (see also Publication III, where we consider similar comparisons for multiclass GP classification). The comparisons show that in GP classification, there is a reasonable agreement with the marginal likelihood approximations and classification accuracies with

LA and EP (Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008, Publication III). Also, for example in Publication II the experiments show that LA with type-II MAP estimation achieves an accuracy close to full MCMC.

Instead of using the MAP estimate, it is possible to marginalize over the latent values f with LA or EP, and then apply MCMC methods to integrate over the uncertainty of θ. Or instead of using MCMC, the integration over θ can be approximated by faster grid integration or central composite design (CCD) methods (see Rue et al., 2009). See also comparisons by Vanhatalo et al. (2010) and discussion by Vanhatalo et al. (2013). Although being less accurate, an advantage of the MAP estimate is that it is faster to evaluate compared to grid, CCD, or MCMC sampling, especially when the number of hyperparameters is large.

4. Summary of Studies

This thesis consists of six articles. Their contents and main results are summarised in this chapter.

4.1 Gaussian Processes with Monotonicity Information (I)

The predictive performance of a flexible nonparametric model can be improved by conditioning the inference to imprecise derivative information concerning the function to be learned, in addition to measured observations (Sill and Abu-Mostafa, 1997). As an example, instead of having measurements on derivatives, the output function can be known to be monotonic with respect to an input variable. For univariate and multivariate additive functions, the monotonicity can be forced by construction (e.g. Shively et al., 2009), but for a multivariate flexible GP model, the elicitation of this imprecise expertise belief is more difficult. A generic approach for multivariate models was proposed by Sill and Abu-Mostafa (1997), who introduced monotonicity information to MLP neural networks using hints that are virtual observations placed appropriately in the input space (see also Lampinen and Selonen, 1997).

In Publication I, we propose a method for introducing background information about monotonicity into a GP model. Because the derivative of a Gaussian process remains a Gaussian process (e.g. Rasmussen, 2003; Solak et al., 2003), we can extend the GP model for derivative observations that are formed with EP. To enforce monotonicity information into the GP model, we form the derivative observations, that is, means and corresponding uncertainties by using a (step-like) probit likelihood function for the derivative values of the latent function evaluated at a finite number of locations. The obtained virtual derivative observations are then used in the GP model in addition to real observations. We discuss how the locations of the derivative observations can be iterated to make a monotonic solution more likely. The behaviour of the proposed model is demonstrated, and the model's performance is compared to a standard GP model without monotonicity information, in simulated regression experiments. We also illustrate the behaviour of a GP model with monotonicity information in a binary classification problem, where the challenge is to predict the risk of institutionalisation of elderly persons by using data from health care registers. The simulated experiments and the analysis of the real-life register data set show that the method favours solutions that are more stable and less prone to overfitting, especially in the areas where there are only few or no observations.

4.2 LA for Logistic Gaussian Processes (II)

The flexibility of a Gaussian process makes it an attractive prior for density estimates whose smoothness properties can be controlled through the prior covariance structure (Leonard, 1978). In Publication II, we present approximate inference for LGP density estimation and density regression in a grid using LA to integrate over the non-Gaussian posterior distribution of latent values. The presented LA approach complements the earlier studies of approximate inference for LGP (Leonard, 1978; Thorburn, 1986; Lenk, 1991, 2003), and for the point process intensity estimation with GPs (Cunningham et al., 2008). We propose to use second-order polynomials as explicit basis functions in the GP model to construct a prior that can favour density estimates whose tails go eventually towards zero in regions with only few or no observations. We show how LA can be computed in a numerically stable way. The computational complexity of the proposed LA approach scales cubically in the number of grid points and to speed up the inference for dense grids, we use the fast Fourier transform, and we exploit Kronecker product computations to obtain a reduced-rank approximation of the exact prior covariance structure. To approximate the Bayesian inference for hyperparameters, we determine type-II MAP estimates for the covariance function parameters. The proposed LGP approach with LA is compared to advanced Bayesian kernel methods (Griffin, 2010), because LGP has been shown to outperform simple kernel methods (Tokdar, 2007; Adams, 2009). The results show that LA is useful for practical interactive visualisation of one- and two-dimensional densities. Our experiments with simulated and real one-dimensional data show that the estimation accuracy with LA is close to a logistic Gaussian process model estimated using MCMC and state-of-the-art hierarchical infinite Gaussian mixture models. We also demonstrate the suitability of the LA method for estimating conditional densities with one predictor variable.

4.3 Nested EP for the Multinomial Probit Likelihood (III)

Two challenges with multiclass GP classification are the integration over the non-Gaussian posterior distribution, and the increase of the number of unknown latent values as the number of target classes grows. EP has proven to be a very accurate method for approximate inference but the existing EP approaches for the multinomial probit GP classification uses numerical quadratures, or independence assumptions between the latent values associated with different classes, to facilitate the computations. In Publication III, we complement the earlier work of Seeger and Jordan (2004), Seeger et al. (2006), and Girolami and Zhong (2007) by developing a novel quadrature-free nested EP algorithm for the multinomial probit likelihood with GP priors. The proposed algorithm maintains all between-class posterior dependencies and scales linearly in the number of classes, similar to softmax GP classification (Williams and Barber, 1998). In the moment matching step, we use inner EP to approximate the tilted moments. We show how the tilted distribution can be expressed in a similar functional form as the posterior distribution resulting from a linear binary classifier with a multivariate Gaussian prior on the weights and a probit likelihood function. With this representation, the moments of the tilted distribution can be approximated with EP similarly as in linear classification (Minka, 2001b; Qi et al., 2004). We develop an efficiently scaling implementation where these inner EP approximations can be updated incrementally between the outer EP loops, and derive low-rank site approximations that results in linear computational scaling with respect to the number of target classes.

We test the accuracy of the proposed algorithm with several experiments. We compare nested EP to quadrature-based EP methods with respect to the approximate marginal distributions of the latent values and class probabilities using fixed hyperparameter values, and show that nested EP achieves similar accuracy compared to quadrature in a computationally efficient way. Using nested EP, we study visually the utility of the full EP approximation over an EP approach that assumes the latent values from different classes a posteriori independent (IEP), and compare their convergence properties. Our experiments show that nested IEP can converge more slowly and require more damping than full nested EP. We compare nested EP to the LA (Williams and Barber, 1998) and factorized variational (Girolami and Rogers, 2006) methods, visualise the accuracy of the approximate marginal distributions with respect to MCMC (Girolami and Rogers, 2006), illustrate the suitability of the respective marginal likelihood approximations for type-II MAP estimation of the covariance function hyperparameters, and discuss computational complexities of the methods. We also compare the predictive accuracy of the EP, LA, factorized variational, and MCMC methods with estimation of the hyperparameters using several real-world data sets. The results show that nested EP is the most consistent method compared to MCMC sampling, but in terms of classification accuracy the differences between all the methods are small from a practical point of view. In addition, we show that the predictive probability estimates of LA can be improved using Laplace's method as described by Tierney and Kadane (1986) but the computational cost becomes increasingly demanding if a larger number of predictions are needed.

4.4 LA for the Cox Proportional Hazards Model (IV)

Estimating the risk of recurrence of gastrointestinal stromal tumour (GIST) after surgery is important when considering adjuvant systemic therapy. Adjuvant imatinib therapy increases the time of GIST recurrence (DeMatteo et al., 2009), whereas some patients can be cured by surgery alone. In Publication IV, we create a database by pooling population-based cohorts of patients diagnosed with GIST that were identified from the literature. We assess prognostic factors of the patients, to compare conventional risk-stratification schemes and to develop a nonlinear GP method for estimating the risk of GIST recurrence. The GP method is constructed by replacing the log-linear predictor in a Cox proportional hazard model with a logarithmic GP prior and by smoothing a piecewise log-constant baseline hazard with another GP prior. We approximate the posterior distribution with LA. Instead of using the Poisson-distributed data representation, we approximate the posterior covariance matrix by maintaining a non-diagonal form for the precision matrix of the effective likelihood. With this representation, the posterior computations scale as $\mathcal{O}((n+T)^3)$, whereas the computations with the Poisson-distributed data representation scale as $\mathcal{O}((n\bar{T})^3)$, where $\bar{T}$ denotes the average number of baseline hazard intervals needed to cover observed survival times (see Section 2.3.3). We determine the hyperparameters of neural network covariance functions with type-II MAP estimation. The risk-stratification schemes are compared to the nonlinear GP approach by calculating receiver operating characteristics (ROC) curves and the corresponding areas under the curve (AUC) using ten-fold cross-validation. The generalization of the GP model is also validated with an independent data set. The results show that the nonlinear GP approach produces accurate estimates for the risk of GIST recurrence, although the risk-stratification criteria identify also well low-risk and high-risk patients. To facilitate the estimation of individualised outcomes with the GP approach, we also provide novel prognostic contour maps and heat maps to illustrate the continuous effects of the key prognostic factors for the risk of GIST recurrence.

4.5 Modelling Length-of-Stay in a Care Episode (V)

The accurate modelling of patient length-of-stay (LOS) in a care episode can provide information for health care providers to improve the effective planning of limited resources (e.g. Fisher and Altaffer, 1992). In Publication V, we consider a case study where the objective is to model LOS in a care episode after a fractured hip (see Sund, 2008). We discuss the challenges related with the register-based data of hip fractures, and present a Bayesian nonparametric approach to model LOS as a multiclass classification problem. In the modelling, we apply the softmax likelihood whose latent values are given an MLP neural network prior with a hierarchical prior structure for weight and bias parameters of the network. In order to evaluate the performance of the nonparametric MLP approach, we model LOS also with an alternative parametric approach based on a finite mixture of Weibull distributions (e.g. Ibrahim et al., 2001). In both approaches, inference is done with MCMC. We compare the predictive performances of both models, and identify patient explanatory variables by their predictive relevances. The results show that the flexible modelling approach produces more accurate predictions. Our experiments also demonstrate advantages of the nonparametric approach over the parametric approach by visualising nonlinear effects and interdependencies between explanatory variables found in the data set.

4.6 Software for Gaussian Process Models (VI)

Publication VI is a brief manual for the GPstuff software package that is a versatile collection of many Gaussian process models and tools for approximate Bayesian inference and model assessment. The software is fully compatible with Matlab[1] (version r2009b or later) and most features are compatible with Octave[2] (tested with 3.6.4). In Publication VI, we illustrate how to construct and use a GP model, and describe the modularity of the model construction. Publication VI presents the key features of GPstuff, including available covariance functions, mean functions, single-latent observation models, multi-latent observation models, priors for hyperparameters, inference methods, and model assessment tools. The modularity of the software package makes it useful for many modelling applications and facilitates the implementation of new features. A more specific description of the GPstuff toolbox is given by Vanhatalo et al. (2013).

[1] http://www.mathworks.com/

[2] http://www.gnu.org/software/octave/

5. Discussion

The main aim of this thesis was to develop the methodology for accurate and efficient approximate Bayesian inference to solve various modelling problems involving probability-based models constructed from Gaussian process priors. Through application motivated case studies, we have shown how Laplace's approximation and the expectation propagation algorithm can be tailored to various multi-latent GP models to facilitate the multi-dimensional integration over the posterior distribution of the latent values. In Publication III it was shown how the EP approaches for multinomial probit GP classification (Seeger et al., 2006; Girolami and Zhong, 2007) can be developed further to obtain an EP approximation that maintains all between-class posterior dependencies and scales linearly in the number of classes, without relying on quadratures. Earlier studies (e.g. Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008) point out that EP approximations can provide accuracy similar to MCMC in a computationally efficient manner for binary GP classification, and our various predictive comparisons in Publication III extend these studies by showing that EP for multiclass classification gives similar results. Also, the findings of our experiments in a case study dealing with logistic GP models (II) show that the Gaussian approximation based on LA with type-II MAP estimation obtains practical accuracy close to MCMC with considerably faster posterior computations.

In general, one important question in Bayesian modelling is how accurately the chosen probability model captures the characteristics of the phenomenon being modelled. In modelling with GPs, one advantage is that we do not need to make strong explicit assumptions, for example, about how different explanatory variables can interact with other variables. The usefulness of a flexible Bayesian modelling approach was supported by the results of predictive comparisons obtained from the real-life applications (IV–V), where nonlinear effects and interactions were observed between explanatory variables. Although the flexibility of GP models can enable a high predictive accuracy (that was one of the objectives in the case studies of this thesis), the flexible modelling approach can also cause troubles in situations where, for example, data is scarce. In Publication I, we developed an approach for introducing background monotonicity information into GP models with EP, and in the light of the experiments, this additional information can constrain the flexibility of GP models and improve the predictions in regression and binary classification in the cases where the target function is monotonic. The main drawback with GP models is the infamous cubic computational complexity in the number of observations, arising from the prior covariance structure. This scaling complicates the applicability of the GP models studied in this thesis for large-scale data sets, although sparse alternatives can be considered to speed up the inference (e.g. Quiñonero-Candela and Rasmussen, 2005; Rue and Held, 2005).

The approximate methods considered in this thesis have been known for long, but the conversion of these general methods into efficient algorithms to approximate inference for various multi-latent and multi-process GP models requires tailored solutions. One obvious line of future work is to study the suitability of EP-based methods for models with multi-latent dependencies especially if sampling-based methods are slow for practical inference and the LA method does not achieve accuracy close to MCMC. As an example, the results from Publication III imply that the nested EP approach could be applicable also for other similar multi-latent models that involve integral representations consisting of simple factorized functions each depending on linear transformations of the latent values.

Another possible extension of this work is correcting the Gaussian approximation for the marginal posterior distribution of multiple latent values, in a similar manner as done for single latent values by Rue et al. (2009) and Cseke and Heskes (2011). Simple corrections based on Gaussian copula have already been suggested (Rue et al., 2009), but improving the multivariate marginal posterior distribution can be challenging if multi-dimensional integrals are required. Also, due to the empirical accuracy of EP in several experiments, it seems that the predictive distribution for the observations can be approximated fairly precisely without the need to average over the actual posterior distribution, as long as the approximate distribution shares lower order statistics with the actual posterior (Paquet et al., 2009). Although EP has proven to be a very accurate method for approximate Bayesian inference (as also supported by the results of Publication III), quadrature or inner-EP methods in a moment matching step of the EP algorithm can be computationally infeasible for models in which each likelihood term related to an observation depends on multiple latent values. In the future, it would be interesting to see whether in such multi-latent cases, the LA method could be improved by considering corrections for multivariate marginal distributions with efficient nested approximations.

Bibliography

Adams, R. P. (2009). *Kernel Methods for Nonparametric Bayesian Inference of Probabilities and Point Processes*. PhD thesis, University of Cambridge.

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

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. John Wiley & Sons, Ltd.

Best, N., Richardson, S., and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14:35–59.

Bonilla, E., Chai, K. M. A., and Williams, C. K. I. (2008). Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems 20*, pages 153–160. MIT Press.

Boyle, P. and Frean, M. (2005). Dependent Gaussian processes. In *Advances in Neural Information Processing Systems 17*, pages 217–224. MIT Press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, 34:187–220.

Cressie, N. A. C. (1993). *Statistics for Spatial Data (Revised Edition)*. John Wiley & Sons, Inc.

Csató, L., Fokoué, E., Opper, M., Schottky, B., and Winther, O. (2000). Efficient approaches to Gaussian process classification. In *Advances in Neural Information Processing Systems 12*, pages 251–257. MIT Press.

Cseke, B. and Heskes, T. (2011). Approximate marginals in latent Gaussian models. *Journal of Machine Learning Research*, 12:417–454.

Cunningham, J. P., Shenoy, K. V., and Sahani, M. (2008). Fast Gaussian process methods for point process intensity estimation. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 192–199. Omnipress.

Damianou, A. and Lawrence, N. (2013). Deep Gaussian processes. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS), Journal of Machine Learning Research: Workshop and Conference Proceedings 31*, pages 207–215.

DeMatteo, R. P., Ballman, K. V., Antonescu, C. R., Maki, R. G., Pisters, P. W., Demetri, G. D., Blackstein, M. E., Blanke, C. D., von Mehren, M., Brennan, M. F., Patel, S., McCarter, M. D., Polikoff, J. A., Tan, B. R., and Owzar, K. (2009). Adjuvant imatinib mesylate after resection of localised, primary gastrointestinal stromal tumour: a randomised, double-blind, placebo-controlled trial. *The Lancet*, 373(9669):1097–1104.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222.

Fisher, W. H. and Altaffer, F. B. (1992). Inpatient length of stay measures: statistical and conceptual issues. *Administration and Policy in Mental Health*, 19(5):311–320.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis (Second Edition)*. Chapman & Hall/CRC.

Gibbs, M. N. and Mackay, D. J. C. (2000). Variational Gaussian process classifiers. *IEEE Transactions on Neural Networks*, 11(6):1458–1464.

Girolami, M. and Rogers, S. (2006). Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 8:1790–1817.

Girolami, M. and Zhong, M. (2007). Data integration for classification problems employing Gaussian process priors. In *Advances in Neural Information Processing Systems 19*, pages 465–472. MIT Press.

Goldberg, P. W., Williams, C. K. I., and Bishop, C. M. (1998). Regression with input-dependent noise: A Gaussian process treatment. In *Advances in Neural Information Processing Systems 10*, pages 493–499. MIT Press.

Griebel, M. and Hegland, M. (2010). A finite element method for density estimation with Gaussian process priors. *SIAM Journal of Numerical Analysis*, 47(6):4759–4792.

Griffin, J. (2010). Default priors for density estimation with mixture models. *Bayesian Analysis*, 5(1):45–64.

Heskes, T. and Zoeter, O. (2002). Expectation propagation for approximate inference in dynamic Bayesian networks. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 216–223.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer.

Juntunen, T., Vanhatalo, J., Peltonen, H., and Mäntyniemi, S. (2012). Bayesian spatial multispecies modelling to assess pelagic fish stocks from acoustic- and trawl-survey data. *ICES Journal of Marine Science*, 69(1):95–104.

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

Kersting, K., Plagemann, C., Pfaff, P., and Burgard, W. (2007). Most likely heteroscedastic Gaussian process regression. In *Proceedings of the 24th international conference on Machine learning (ICML)*, pages 393–400. ACM.

Kim, H.-C. and Ghahramani, Z. (2006). Bayesian Gaussian process classification with the EM-EP algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1948–1959.

Kneib, T. (2006). Mixed model-based inference in geoadditive hazard regression for interval-censored survival times. *Computational Statistics & Data Analysis*, 51:777–792.

Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, 34:207–228.

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

Laird, N. and Oliver, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374):231–240.

Lampinen, J. and Selonen, A. (1997). Using background knowledge in multilayer perceptron learning. In *Proceedings of the 10th Scandinavian Conference on Image Analysis, volume 2*, pages 545–549.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816.

Lázaro-Gredilla, M. and Titsias, M. K. (2011). Variational heteroscedastic Gaussian process regression. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 841–848. ACM.

Lenk, P. J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, 78(3):531–543.

Lenk, P. J. (2003). Bayesian semiparametric density estimation and model verification using a logistic-Gaussian process. *Journal of Computational and Graphical Statistics*, 12(3):548–565.

Leonard, T. (1978). Density estimation, stochastic processes, and prior information. *Journal of the Royal Statistical Society. Series B*, 40(2):113–146.

MacKay, D. J. C. (1998). Introduction to Gaussian processes. In *Neural Networks and Machine Learning*, pages 133–166. Springer-Verlag.

Martino, S., Akerkar, R., and Rue, H. (2011). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics*, 38:514–528.

Minka, T. P. (2001a). Expectation Propagation for approximative Bayesian inference. In *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 362–369.

Minka, T. P. (2001b). *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology.

Minka, T. P. (2004). Power EP. Technical report, Microsoft Research, Cambridge.

Mũnoz-González, L., Lázaro-Gredilla, M., and Figueiras-Vidal, A. R. (2011). Heteroscedastic Gaussian process regression using expectation propagation. In *Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Naish-Guzman, A. (2007). *Sparse and Robust Kernel Methods*. PhD thesis, University of Cambridge.

Naish-Guzman, A. and Holden, S. (2008). Robust regression with twinned Gaussian processes. In *Advances in Neural Information Processing Systems 20*, pages 1065–1072. MIT Press.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer.

Neal, R. M. (1998). Regression and classification using Gaussian process priors. In *Bayesian Statistics 6*, pages 475–501. Oxford University Press.

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

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B*, 40(1):1–42.

Paquet, U., Winther, O., and Opper, M. (2009). Perturbation corrections in approximative inference: mixture modelling applications. *Journal of Machine Learning Research*, 10:1263–1304.

Qi, Y., Minka, T. P., Picard, R. W., and Ghahramani, Z. (2004). Predictive automatic relevance determination by expectation propagation. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 671–678.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959.

Rasmussen, C. E. (2003). Gaussian processes to speed up Hybrid Monte Carlo for expensive Bayesian integrals. In *Bayesian Statistics*, volume 7, pages 651–659. Oxford University Press.

Rasmussen, C. E. and Ghahramani, Z. (2002). Infinite mixtures of Gaussian process experts. In *Advances in Neural Information Processing Systems 14*, pages 881–888. MIT Press.

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

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods (Second Edition)*. Springer.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields Theory and Applications*. Chapman & Hall/CRC.

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

Seeger, M. (2005). Expectation propagation for exponential families. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Seeger, M. and Jordan, M. I. (2004). Sparse Gaussian process classification with multiple classes. Technical report, University of California, Berkeley, CA.

Seeger, M., Lawrence, N., and Herbrich, R. (2006). Efficient nonparametric Bayesian modelling with sparse Gaussian process approximations. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

Shively, T. S., Sager, T. W., and Walker, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society. Series B*, 71(1):159–175.

Sill, J. and Abu-Mostafa, Y. S. (1997). Monotonicity hints. In *Advances in Neural Information Processing Systems 9*, pages 634–640. MIT Press.

Smola, A., Vishwanathan, V., and Eskin, E. (2004). Laplace propagation. In *Advances in Neural Information Processing Systems 16*. MIT Press.

Solak, E., Murray-Smith, R., Leithead, W. E., Leith, D. J., and Rasmussen, C. E. (2003). Derivative observations in Gaussian process models of dynamic systems. In *Advances in Neural Information Processing Systems 15*, pages 1033–1040. MIT Press.

Sund, R. (2008). *Methodological Perspectives for Register-Based Health System Performance Assessment: Developing a Hip Fracture Monitoring System in Finland*. PhD thesis, National Research and Development Centre for Welfare and Health.

Teh, Y. W., Seeger, M., and Jordan, M. I. (2005). Semiparametric latent factor models. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, pages 333–340. Society for Artificial Intelligence and Statistics.

Thorburn, D. (1986). A Bayesian approach to density estimation. *Biometrika*, 73(1):65–75.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Tokdar, S. T. (2007). Towards a faster implementation of density estimation with logistic Gaussian process priors. *Journal of Computational and Graphical Statistics*, 16(3):633–655.

Tokdar, S. T. and Ghosh, J. K. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137(1):34–42.

Tresp, V. (2001). Mixtures of Gaussian processes. In *Advances in Neural Information Processing Systems 13*, pages 654–660. MIT Press.

Van Gerven, M., Cseke, B., Oostenveld, R., and Heskes, T. (2009). Bayesian source localization with the multivariate Laplace prior. In *Advances in Neural Information Processing Systems 22*, pages 1901–1909.

Vanhatalo, J., Jylänki, P., and Vehtari, A. (2009). Gaussian process regression with Student-t likelihood. In *Advances in Neural Information Processing Systems 22*, pages 1910–1918.

Vanhatalo, J., Pietiläinen, V., and Vehtari, A. (2010). Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, 29(15):1580–1607.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., and Vehtari, A. (2013). Bayesian modeling with Gaussian processes using the GPstuff toolbox. *ArXiv:1206.5754*.

Williams, C. K. I. (1998). Computation with infinite neural networks. *Neural Computation*, 10(5):1203–1216.

Williams, C. K. I. and Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351.

Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. In *Advances in Neural Information Processing Systems 8*, pages 514–520. MIT press.

Wilson, A. G., Knowles, D. A., and Ghahramani, Z. (2012). Gaussian process regression networks. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*. Omnipress.

Winther, O. (2001). Computing with finite and infinite networks. In *Advances in Neural Information Processing Systems 13*, pages 336–342. MIT Press.

Ypma, A. and Heskes, T. (2005). Novel approximations for inference in nonlinear dynamical systems using expectation propagation. *Neurocomputing*, 69:85–99.

Zoeter, O. and Heskes, T. (2005). Gaussian quadrature based expectation propagation. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, pages 445–452. Society for Artificial Intelligence and Statistics.

Aalto-DD 144/2013

ISBN 978-952-60-5332-5
ISBN 978-952-60-5333-2 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Biomedical Engineering and Computational Science